WWC Intervention Report

A summary of findings from a systematic review of the evidence

ies INSTITUTE OF EDUCATION SCIENCES

Charter Schools January 2018

Knowledge is Power Program (KIPP)

Intervention Description[1]

The *Knowledge Is Power Program (KIPP)* is a nonprofit network of more than 200 public charter schools educating early childhood, elementary, middle, and high school students. Every *KIPP* school obtains approval to operate from a charter school authorizer. Students, parents, and teachers must sign a commitment to abide by a set of responsibilities, including high behavioral and disciplinary expectations. *KIPP* also has an active alumni network and set of partnerships with scholarship organizations to help guide former students through college. *KIPP* schools have an extended school day and an extended school year compared with traditional public schools. When demand for enrollment exceeds enrollment capacity at a *KIPP* school, student admission is based upon a lottery. Funding for *KIPP* schools comes primarily through public federal, state, and local finances, along with supplemental funding through charitable donations from foundations and individuals.

Research[2]

The What Works Clearinghouse (WWC) identified four studies of *KIPP* that fall within the scope of the Charter Schools topic area and meet WWC group design standards. One study meets WWC group design standards without reservations, and three studies meet WWC group design standards with reservations. Together, these studies included approximately 21,000 students in middle and high schools across 16 states and the District of Columbia.[3]

According to the WWC review, the extent of evidence for *KIPP* on the academic achievement of middle and high school students was medium to large for four outcome domains—mathematics achievement, English language arts achievement, science achievement, and social studies achievement, and was small for one outcome domain—student progression. No studies meet WWC group design standards in the five other domains, so this intervention report does not report on the effectiveness of *KIPP* for those domains.[4] (See the Effectiveness Summary on p. 4 for more details of effectiveness by domain.)

Effectiveness

KIPP had positive effects on mathematics achievement and English language arts achievement, and potentially positive effects on science achievement and social studies achievement for middle and high school students, and no discernible effects on student progression for high school students.

Report Contents

Overview	p. 1
Intervention Information	p. 3
Research Summary	p. 4
Effectiveness Summary	p. 5
References	p. 9
Research Details for Each Study	p. 12
Outcome Measures for Each Domain	p. 19
Findings Included in the Rating for Each Outcome Domain	p. 20
Supplemental Findings for Each Outcome Domain	p. 25
Endnotes	p. 30
Rating Criteria	p. 32
Glossary of Terms	p. 33

This intervention report presents findings from a systematic review of the *Knowledge is Power Program (KIPP)* conducted using the WWC Procedures and Standards Handbook (version 3.0) and the Charter Schools review protocol (version 3.0).

Table 1. Summary of findings[5]

Outcome domain	Rating of effectiveness	Improvement index *(percentile points)*		Number of studies	Number of students	Extent of evidence
		Average	Range			
Mathematics achievement	Positive effects	+12	+7 to +20	4	19,542	Medium to large
English language arts achievement	Positive effects	+8	+6 to +13	4	20,804	Medium to large
Science achievement	Potentially positive effects	+11	+10 to +13	2	18,712	Medium to large
Social studies achievement	Potentially positive effects	+5	+1 to +9	2	10,363	Medium to large
Student progression	No discernible effects	+5	na	1	852	Small

na = not applicable

Intervention Information

Background

Launched in 1994 with two schools in Houston, Texas, *KIPP* is now a nationwide network of more than 200 public charter schools. Address: *KIPP* Foundation, 520 8th Avenue, Suite 2005, New York, NY 10018. Web: http://www.kipp.org. Telephone: (212) 233-5477.

Intervention details

KIPP is a nonprofit network of more than 200 public charter schools serving students in prekindergarten through high school. Student admission is based upon a lottery when there is excess demand for enrollment. For some *KIPP* schools, students living within specified ZIP codes receive preference in the lottery. Funding for *KIPP* schools comes primarily through public federal, state, and local finances, along with supplemental funding through charitable donations from foundations and individuals.

KIPP schools, like all charter schools, are publicly funded schools that operate autonomously, outside the direct control of the local school district. Every *KIPP* school obtains approval to operate from a charter school authorizer. The charter schools are exempt from certain state or local rules and regulations. In return for flexibility and autonomy, the charter school must meet the accountability standards outlined in its charter. The group or jurisdiction that granted the charter reviews it periodically (typically every 3 to 5 years) and can revoke the charter if the school does not follow guidelines on curriculum and management or does not meet the standards.

KIPP schools have an extended day and extended school year compared with traditional public schools. Students, parents, and teachers sign a pledge called the Commitment to Excellence that describes the roles and expectations for each group in forming a partnership that puts learning first. These include attendance, homework, and behavior for students; assistance and support for parents; and preparation and availability for teachers.

KIPP regional organizations and schools have significant autonomy in setting leadership practices, hiring and dismissing principals, and training teachers and future school leaders. *KIPP* principals have the ability to hire and fire staff and teachers based on performance, as well as authority to allocate school resources based on student needs.

Cost

KIPP receives funding from a combination of charitable donations, as well as from local districts. Per-pupil expenditure in *KIPP* schools appears to be comparable or slightly higher than that of local school districts serving similar students. One study found that a Massachusetts *KIPP* school spent approximately $13,500 per pupil in fiscal year 2008, including rental and capital costs, in comparison to approximately $13,000 in local district schools in the same year (Angrist, 2012). Another study found that nationally, in the 2007–08 school year, *KIPP* schools received a combined revenue of $12,731 per student, in comparison to the average of $11,960 in local school districts (Miron, Urschel, & Saxton, 2011).

Research Summary

The WWC identified six eligible studies that investigated the effects of *KIPP* on the academic achievement of middle and high school students. An additional 15 studies were identified but do not meet WWC eligibility criteria (see the Glossary of Terms in this document for a definition of this term and other commonly used research terms) for review in this topic area. Citations for all 21 studies are in the References section, which begins on p. 9.[6]

Table 2. Scope of reviewed research

Grade	PK–12
Delivery method	Whole school
Program type	School level

The WWC reviewed six eligible studies against group design standards. One study is a randomized controlled trial that meets WWC group design standards without reservations, and three studies are randomized controlled trials or use quasi-experimental designs that meet WWC group design standards with reservations. This report summarizes those four studies. The remaining two studies do not meet WWC group design standards.

Summary of study meeting WWC group design standards without reservations

Tuttle et al. (2015, Middle School, RCT)[7] conducted a randomized controlled trial in eight states that randomly assigned students the opportunity to attend a *KIPP* school through an admissions lottery.[8] This analysis compared fifth- or sixth-grade students who received an admission offer to one of 16 *KIPP* middle schools during the 2011–12 school year with applicants who did not receive an admission offer to a *KIPP* school. Outcomes were measured in the first, second, and third follow-up years after random assignment. The WWC based its effectiveness ratings on the analysis from the third follow-up year of 455 students for mathematics and 458 students for English language arts.[9]

Summary of studies meeting WWC group design standards with reservations

Tuttle et al. (2015, Middle School, QED) conducted a quasi-experimental design that included 37 *KIPP* middle schools in at least 10 states, comparing students who entered a *KIPP* middle school in grades 5 or 6 from 2001–02 through 2013–14 with similar students who never enrolled in *KIPP*. The WWC based its effectiveness ratings on analyses focused on the *KIPP* middle schools that opened in fall 2011 or later, which included a sample of 13,624 students for mathematics, 14,551 students for English language arts, 17,413 students for science, and 9,762 students for social studies.

Tuttle et al. (2015, High School) conducted a quasi-experimental design that included schools in nine states, comparing *KIPP* students with similar non-*KIPP* students. Two analyses in this study contributed to the effectiveness rating.[10] The first of these compared students who entered a *KIPP* high school for the first time in grade 9 with students who never enrolled in *KIPP*. The sample included 14 *KIPP* high schools. The second analysis compared students attending *KIPP* middle schools in grade 8 who had the option to attend *KIPP* high schools in grade 9 with a comparison group of *KIPP* students in grade 8 from different middle schools in regions with no *KIPP* high school open at the time. The sample included eight *KIPP* high schools. The WWC based its effectiveness ratings on findings from analysis of a sample of 1,928 students for mathematics, 2,260 students for English language arts, 1,299 students for science, 601 students for social studies, and 852 students for student progression.

Woodworth et al. (2008) conducted a quasi-experimental design in two districts in the San Francisco Bay area, comparing *KIPP* students with similar students in other schools, matched using propensity scores based on baseline reading and math test scores, gender, race, special education classification, limited English proficiency, and free or reduced-price lunch status. The intervention group included students who entered one of three *KIPP* middle schools as fifth-grade students in the 2003–04 and 2004–05 school years, or students who entered *KIPP* as sixth-grade students in the 2004–05 school year. The comparison group included similar students in the same districts, grades, and cohorts who never enrolled in *KIPP*.[11] The WWC based its effectiveness rating on findings from analysis of a sample of 3,535 students for mathematics and 3,535 students for English language arts.

Effectiveness Summary

The WWC review of *KIPP* for the Charter Schools topic area includes student outcomes in 10 domains: mathematics achievement, English language arts achievement, science achievement, social studies achievement, general achievement, social-emotional competence, disciplinary experiences, student attendance, student progression, and earnings in adulthood. The four studies of *KIPP* that met WWC group design standards reported findings in five of the 10 domains: mathematics achievement, English language arts achievement, science achievement, social studies achievement, and student progression. The following findings present the authors' estimates and WWC-calculated estimates of the size and statistical significance of the effects of *KIPP* on middle and high school students. Additional comparisons are available 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. 32.

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 the four studies that reported findings, the estimated impact of the intervention on outcomes in the *mathematics achievement* domain was positive and statistically significant, and one of these studies meets WWC group design standards without reservations. No studies show statistically significant or substantively important negative effects.
Extent of evidence	**Criteria met**
Medium to large	Four studies that included 19,542 students reported evidence of effectiveness in the *mathematics achievement* domain.

Four studies that met WWC group design standards with or without reservations reported findings in the mathematics achievement domain.

Tuttle et al. (2015, Middle School, RCT) examined one outcome in the mathematics achievement domain: a state standardized mathematics test score from each state included in the sample. The authors reported, and the WWC confirmed, a positive and statistically significant difference between *KIPP* students and the comparison students for the middle school lottery sample measured in the third follow-up year after *KIPP* entry.[12] The WWC characterizes this study finding as a statistically significant positive effect.

Tuttle et al. (2015, Middle School, QED) examined one outcome in the mathematics achievement domain: a state standardized mathematics test score from each state included in the sample. The authors reported, and the WWC confirmed, a positive and statistically significant difference between incoming fifth- or sixth-grade students in *KIPP* schools and the comparison students in the fourth follow-up year after *KIPP* entry. The WWC characterizes this study finding as a statistically significant positive effect.

Tuttle et al. (2015, High School) examined two outcomes in the mathematics achievement domain: a state standardized mathematics test score from each state included in the sample, and a TerraNova mathematics test score. The authors reported, and the WWC confirmed, a positive and statistically significant difference on the third year follow-up state standardized mathematics test score between students who entered *KIPP* high schools for the first time in grade 9 and the comparison students. The authors found, and the WWC confirmed, no statistically significant effects of *KIPP* on the third year follow-up TerraNova mathematics test score between students who entered *KIPP* high schools for the first time in grade 9 and the comparison students. Taken together, the WWC characterizes this study finding as a statistically significant positive effect.

Woodworth et al. (2008) examined one outcome in the mathematics achievement domain: California's State Standardized Assessment in Mathematics. The authors reported findings on the first follow-up year after *KIPP* entry for two cohorts of fifth-grade students and one cohort of sixth-grade students. The authors examined differences separately for each school, cohort, and grade, and reported positive and statistically significant differences between the *KIPP* students and comparison students for each contrast. The WWC aggregated the findings across schools and cohorts by grade, and found the effect for each grade to be positive and statistically significant. The WWC characterizes these study findings as a statistically significant positive effect.

Thus, for the mathematics achievement domain, four studies, one of which meets WWC group design standards without reservations, showed a statistically significant positive effect. This results in a rating of positive effects, with a medium to large extent of evidence.

Summary of effectiveness for the English language arts achievement domain

Table 4. Rating of effectiveness and extent of evidence for the English language arts achievement domain

Rating of effectiveness	Criteria met
Positive effects *Strong evidence of a positive effect with no overriding contrary evidence.*	In the four studies that reported findings, the estimated impact of the intervention on outcomes in the *English language arts achievement* domain was positive and statistically significant, and one of these studies meets WWC group design standards without reservations. No studies show statistically significant or substantively important negative effects.
Extent of evidence	**Criteria met**
Medium to large	Four studies that included 20,804 students reported evidence of effectiveness in the *English language arts achievement* domain.

Four studies that met WWC group design standards with or without reservations reported findings in the English language arts achievement domain.

Tuttle et al. (2015, Middle School, RCT) examined one outcome in the English language arts achievement domain: a state standardized reading test score from each state included in the sample. The authors reported, and the WWC confirmed, a positive and statistically significant difference between *KIPP* students and the comparison group for the middle school lottery sample measured in the third follow-up year after *KIPP* entry. The WWC characterizes this study finding as a statistically significant positive effect.

Tuttle et al. (2015, Middle School, QED) examined one outcome in the English language arts achievement domain: a state standardized reading test score from each state included in the sample. The authors reported, and the WWC confirmed, a positive and statistically significant difference between incoming fifth- or sixth-grade students in *KIPP* schools and the comparison group in the fourth follow-up year after *KIPP* entry. The WWC characterizes this study finding as a statistically significant positive effect.

Tuttle et al. (2015, High School) examined three outcomes in the English language arts achievement domain: a state standardized general literacy test score from each state included in the sample, the TerraNova reading test score, and the TerraNova English language arts test score. The authors reported, and the WWC confirmed, a positive and statistically significant difference on the general literacy state standardized test score between students who entered *KIPP* schools for the first time in grade 9 and the comparison group. The authors reported a positive and statically significant effect of *KIPP* on the third year follow-up TerraNova reading between between students who entered *KIPP* high schools for the first time in grade 9 and the comparison students; however, after applying a correction for multiple comparisons, the WWC found this result was no longer statistically significant.

The authors reported, and the WWC confirmed, no statistically significant effects of *KIPP* on the third year follow-up TerraNova English language arts test score between students who entered *KIPP* high schools for the first time in grade 9 and the comparison students. Taken together, the WWC characterizes this study finding as a statistically significant positive effect.

Woodworth et al. (2008) examined one outcome in the English language arts achievement domain: the California Standards Test English/Language Arts score. The authors reported findings on the first follow-up year after entry at *KIPP* for two cohorts of fifth-grade students and one cohort of sixth-grade students. The authors examined differences separately for each school, cohort, and grade, and reported positive and statistically significant differences between the *KIPP* students and comparison students for each contrast. The WWC aggregated the findings across schools and cohorts by grade, and found the effect for each grade to be positive and statistically significant. The WWC characterizes these study findings as a statistically significant positive effect.

Thus, for the English language arts achievement domain, four studies, one of which meets WWC group design standards without reservations, showed a statistically significant positive effect. This results in a rating of positive effects, with a medium to large extent of evidence.

Table 5. 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 two studies that reported findings, the estimated impact of the intervention on outcomes in the *science achievement* domain was positive and statistically significant.
Extent of evidence	**Criteria met**
Medium to large	Two studies that included 18,712 students reported evidence of effectiveness in the *science achievement* domain.

Summary of effectiveness for the science achievement domain

Two studies that met WWC group design standards with reservations reported findings in the science achievement domain.

Tuttle et al. (2015, Middle School, QED) examined one outcome in the science achievement domain: a state standardized test score for science from each state included in the sample. The authors reported, and the WWC confirmed, a positive and statistically significant difference between incoming fifth- or sixth-grade students in *KIPP* schools and the matched comparison group in eighth grade. The WWC characterizes this study finding as a statistically significant positive effect.

Tuttle et al. (2015, High School) examined one outcome in the science achievement domain: a state standardized test score for science from each state included in the sample. The authors reported, and the WWC confirmed, a positive and statistically significant difference between students who entered *KIPP* schools for the first time in grade 9 and comparison students. The WWC characterizes this study finding as a statistically significant positive effect.

Thus, for the science domain, two studies reported a statistically significant positive effect. This results in a rating of potentially positive effects, with a medium to large extent of evidence.

Summary of effectiveness for the social studies domain

Table 6. Rating of effectiveness and extent of evidence for the social studies achievement domain

Rating of effectiveness	Criteria met
Potential positive effects *Evidence of a positive effect with no overriding contrary evidence.*	In the two studies that reported findings, the estimated impact of the intervention on outcomes in the *social studies achievement* domain was positive and statistically significant for one study, and neither statistically significant nor large enough to be substantively important for one study.
Extent of evidence	**Criteria met**
Medium to large	Two studies that included 10,363 students reported evidence of effectiveness in the *social studies achievement* domain.

Two studies that met WWC group design standards with reservations reported findings in the social studies achievement domain.

Tuttle et al. (2015, Middle School, QED) examined one outcome in the social studies achievement domain: a state standardized history test score from each state included in the sample. The authors reported, and the WWC confirmed, a positive and statistically significant difference between incoming fifth- or sixth-grade students in *KIPP* schools and the matched comparison group. The WWC characterizes this study finding as a statistically significant positive effect.

Tuttle et al. (2015, High School) examined one outcome in the social studies achievement domain: a state standardized social studies test score from each state included in the sample. The authors reported, and the WWC confirmed, no statistically significant difference between students who entered *KIPP* schools for the first time in grade 9 and matched comparison students. According to WWC criteria, the effect size was not large enough to be considered substantively important (that is, an effect size of at least 0.25). The WWC characterizes this study finding as an indeterminate effect.

Thus, for the social studies achievement domain, one study showed a statistically significant positive effect, and one study showed an indeterminate effect. This results in a rating of potentially positive effects, with a medium to large extent of evidence.

Summary of effectiveness for the student progression domain

Table 7. Rating of effectiveness and extent of evidence for the student progression domain

Rating of effectiveness	Criteria met
No discernible effects *No affirmative evidence of effects.*	In the one study that reported findings, the estimated impact of the intervention on outcomes in the *student progression* domain was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Small	One study that included 852 students reported evidence of effectiveness in the *student progression* domain.

One study that met WWC group design standards with reservations reported findings in the student progression domain.

Tuttle et al. (2015, High School) examined one outcome in the student progression domain: high school graduation within 4 years of grade 9 entry. The authors reported, and the WWC confirmed, no statistically significant difference between students who entered *KIPP* schools for the first time in grade 9 and comparison students. According to WWC criteria, the effect size was not large enough to be considered substantively important. The WWC characterizes this study finding as an indeterminate effect.

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

References

Study that meets WWC group design standards without reservations

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Understanding the effect of KIPP as it scales: Volume I, Impacts on achievement and other outcomes. Final report of KIPP's Investing in Innovation grant evaluation* [Middle School; RCT]. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560079

Additional sources:

Gleason, P. M., Tuttle, C. C., Gill, B., Nichols-Barrer, I., & Teh, B. (2014). Do KIPP schools boost student achievement? *Education Finance and Policy, 9*(1), 36–58. Retrieved from https://eric.ed.gov/?id=EJ1016285

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP middle schools: Impacts on achievement and other outcomes, final report.* Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED540912

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Going to scale: As KIPP network grows, positive impacts are sustained* (InFocus brief). Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560043

Tuttle, C. C., Teh, B., Nichols-Barrer, I., Gill, B., & Gleason, P. (2010). *Student characteristics and achievement in 22 KIPP middle schools: Final report.* Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED511107

Tuttle, C. C., Teh, B., Nichols-Barrer, I., Gill, B., & Gleason, P. (2010). *Supplemental analytic sample equivalence tables for student characteristics and achievement in 22 KIPP middle schools: A report from the National Evaluation of KIPP Middle Schools.* Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED511108

Studies that meet WWC group design standards with reservations

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Understanding the effect of KIPP as it scales: Volume I, Impacts on achievement and other outcomes. Final report of KIPP's Investing in Innovation grant evaluation* [Middle School; QED]. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560079

Additional sources:

Gleason, P. M., Tuttle, C. C., Gill, B., Nichols-Barrer, I., & Teh, B. (2014). Do KIPP schools boost student achievement? *Education Finance and Policy, 9*(1), 36–58. Retrieved from https://eric.ed.gov/?id=EJ1016285

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP middle schools: Impacts on achievement and other outcomes, final report.* Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED540912

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Going to scale: As KIPP network grows, positive impacts are sustained* (InFocus brief). Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560043

Tuttle, C. C., Teh, B., Nichols-Barrer, I., Gill, B., & Gleason, P. (2010). *Student characteristics and achievement in 22 KIPP middle schools: Final report.* Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED511107

Tuttle, C. C., Teh, B., Nichols-Barrer, I., Gill, B., & Gleason, P. (2010). *Supplemental analytic sample equivalence tables for student characteristics and achievement in 22 KIPP middle schools: A report from the National Evaluation of KIPP Middle Schools.* Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED511108

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Understanding the effect of KIPP as it scales: Volume I, Impacts on achievement and other outcomes. Final report of KIPP's Investing in Innovation grant evaluation* [High School]. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560079

Additional source:

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Going to scale: As KIPP network grows, positive impacts are sustained* (InFocus brief). Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560043

Woodworth, K. R., David, J. L., Guha, R., Wang, H., & Lopez-Torkos, A. (2008). *San Francisco Bay Area KIPP schools: A study of early implementation and achievement. Final report.* Menlo Park, CA: SRI International.

Studies that do not meet WWC group design standards

Crosby, K. D. (2015). *An evaluation of KIPP charter high schools utilizing the framework for academic quality.* Available from ProQuest Dissertations and Theses database. (UMI No. 3685859) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Understanding the effect of KIPP as it scales: Volume I, Impacts on achievement and other outcomes. Final report of KIPP's Investing in Innovation grant evaluation* [Elementary School]. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560079 The study does not meet WWC group design standards because it is a randomized controlled trial in which the combination of overall and differential attrition rates exceeds WWC standards for this area, and the subsequent analytic intervention and comparison groups are not shown to be equivalent.

Additional source:

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Going to scale: As KIPP network grows, positive impacts are sustained* (In Focus brief). Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560043

Studies that are ineligible for review using the Charter Schools Evidence Review Protocol

Abell Foundation. (2006). *Baltimore's "New" middle schools: Do KIPP and Crossroads Schools offer solutions to the city's poorly-performing middle schools?* Baltimore, MD: Author. Retrieved from https://eric.ed.gov/?id=ED546158 The study is ineligible for review because it was out of the scope of the protocol.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2012). Who benefits from KIPP? *Journal of Policy Analysis and Management, 31*(4), 837–860. Retrieved from https://eric.ed.gov/?id=EJ988158 The study is ineligible for review because it was out of the scope of the protocol.

Additional sources:

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2010). Inputs and impacts in charter schools: KIPP Lynn. *American Economic Review, 100*(2), 239–243.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2010). *Who benefits from KIPP?* (NBER Working Paper No. 15740). Cambridge, MA: National Bureau of Economic Research. Retrieved from https://eric.ed.gov/?id=ED508549

Betts, J. R., & Tang, Y. E. (2011). *The effect of charter schools on student achievement: A meta-analysis of the literature.* Seattle, WA: Center on Reinventing Public Education. Retrieved from https://eric.ed.gov/?id=ED526353 The study is ineligible for review because it does not use an eligible design.

Brown, K. J., & Holt, C. R. (2014). Factors influencing the improved academic success in literacy at the Knowledge Is Power Program (KIPP) schools in the Delta Region according to adult perceptions. *Education Leadership*

Review of Doctoral Research, 1(2), 23–39. Retrieved from https://eric.ed.gov/?id=EJ1105745 The study is ineligible for review because it does not use an eligible design.

Furgeson, J., Knechtel, V., Sullivan, M., Tuttle, C. C., Akers, L., Anderson, M. A., ... Nichols-Barrer, I. (2014). *KIPP leadership practices through 2010–2011* (Technical Report). Washington, DC: Mathematica Policy Research. The study is ineligible for review because it does not use an eligible design.

Knechtel, V., Anderson, M. A., Burnett, A., Coen, T., Sullivan, M., Tuttle, C. C., & Gleason, P. (2015). *Understanding the effect of KIPP as it scales: Volume II, leadership practices at KIPP - Final report of KIPP's Investing in Innovation grant evaluation.* Washington, DC: Mathematica Policy Research. The study is ineligible for review because it does not use an eligible design.

Lopez, R. E. (2015). *School climate, developmental assets, and academic success in KIPP Hispanic students* (Doctoral dissertation). Retrieved from http://scholarworks.waldenu.edu/ The study is ineligible for review because it does not use an eligible design.

Macey, E., Decker, J., & Eckes, S. (2009). The Knowledge Is Power Program (KIPP): An analysis of one model's efforts to promote achievement in underserved communities. *Journal of School Choice, 3*(3), 212–241. Retrieved from https://eric.ed.gov/?id=EJ864717 The study is ineligible for review because it does not use an eligible design.

Mac Iver, M. A., & Farley-Ripple, E. (2007). *The Baltimore KIPP Ujima Village Academy, 2002-2006: A longitudinal analysis of student outcomes.* Baltimore, MD: Johns Hopkins University, Center for Social Organization of Schools. Retrieved from https://eric.ed.gov/?id=ED505119 The study is ineligible for review because it was out of the scope of the protocol.

Maddin, B. W. (2011). *KIPP student mobility: Whether and when students make non-promotional exits from KIPP middle schools* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 10099042) The study is ineligible for review because it does not use an eligible design.

Miron, G., Urschel, J. L., & Saxton, N. (2011). *What makes KIPP work? A study of student characteristics, attrition, and school finance.* Kalamazoo: Western Michigan University, College of Education and Human Development. The study is ineligible for review because it does not use an eligible design.

Nichols-Barrer, I., Gleason, P., Gill, B., & Tuttle, C. C. (2016). Student selection, attrition, and replacement in KIPP middle schools. *Educational Evaluation and Policy Analysis, 38*(1), 5–20. Retrieved from https://eric.ed.gov/?id=EJ1089924 The study is ineligible for review because it was out of the scope of the protocol.

Additional sources:

Nichols-Barrer, I., Gill, B., Gleason, P., & Tuttle, C. C. (2012). *Student selection, attrition, and replacement in KIPP middle schools* (Working Paper). Princeton, NJ: Mathematica Policy Research.

Nichols-Barrer, I., Gill, B. P., Gleason, P., & Tuttle, C. C. (2014). Does student attrition explain KIPP's success? *Education Next, 14*(4), 62–70. Retrieved from https://eric.ed.gov/?id=EJ1039347

Rose, C. P. (2013). *The academic impacts of attending a KIPP charter school in Arkansas* (Doctoral dissertation). Retrieved from https://eric.ed.gov/?id=ED558890 The study is ineligible for review because it was out of the scope of the protocol.

Upton, K. M. (2009). *Seeking quality in the face of adversity: Fordham Sponsorship Accountability Report, 2008-09.* Dayton, OH: Thomas B. Fordham Foundation. Retrieved from https://eric.ed.gov/?id=ED508050 The study is ineligible for review because it does not use an eligible design.

Yeh, S. S. (2013). A re-analysis of the effects of KIPP and the Harlem Promise Academies. *Teachers College Record, 115*(4), Article No. 040301. The study is ineligible for review because it does not use an eligible design

Appendix A.1: Research details for Tuttle et al. (2015, Middle School, RCT)

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Understanding the effect of KIPP as it scales: Volume I, Impacts on achievement and other outcomes. Final report of KIPP's Investing in Innovation grant evaluation* [Middle School; RCT]. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560079

Additional sources:

Gleason, P. M., Tuttle, C. C., Gill, B., Nichols-Barrer, I., & Teh, B. (2014). *Do KIPP schools boost student achievement?* Education, 9(1), 36–58. Retrieved from https://eric.ed.gov/?id=EJ1016285

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP middle schools: Impacts on achievement and other outcomes, final report*. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED540912

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Going to scale: As KIPP network grows, positive impacts are sustained* (InFocus brief). Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560043

Tuttle, C. C., Teh, B., Nichols-Barrer, I., Gill, B., & Gleason, P. (2010). *Student characteristics and achievement in 22 KIPP middle schools: Final report*. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED511107

Tuttle, C. C., Teh, B., Nichols-Barrer, I., Gill, B., & Gleason, P. (2010). *Supplemental analytic sample equivalence tables for student characteristics and achievement in 22 KIPP middle schools: A report from the National Evaluation of KIPP Middle Schools*. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED511108

Table A1. 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	455 students	+7	Yes
English language arts achievement	458 students	+6	Yes

Setting

This analysis includes students and schools in multiple states and districts in the United States where *KIPP* charter schools operate. The study took place in 43 middle schools in the *KIPP* network in 20 cities across the following 12 states and the District of Columbia: Arkansas, California, Colorado, Georgia, Maryland, Massachusetts, New Jersey, New York, North Carolina, Pennsylvania, Tennessee, and Texas.

Study sample

The study used a lottery-based randomized controlled trial design, where *KIPP* applicants that won admission to the *KIPP* middle school through the lottery formed the intervention group, and those whose lottery draw led to their not being offered admission formed the comparison group. Of the 60 *KIPP* middle schools open in 2011–12, 16 were sufficiently oversubscribed to conduct a lottery and be included in the RCT analysis. The sample after random assignment included 891 students, with a intervention group of 459 students offered admission and a comparison group of 432 students not offered admission.

Among students in the intervention condition in the analysis sample, 53% were female, 51% were Hispanic, 44% were Black, 86% were eligible for free or reduced-price lunch, 51% lived in bilingual homes or homes where a language other than English was the main language, 44% had mothers with a high school education or less, and 35% were from single-parent households. Among students in the comparison condition in the analysis sample, 51% were female, 47% were Hispanic, 44% were Black, 86% were eligible for free or reduced-price lunch, 44% lived in bilingual homes or homes where a language other than English was the main language, 50% had mothers with a high school education or less, and 27% were from single-parent households.

Intervention group

Students in the intervention condition were offered admission to a *KIPP* middle school, and 72% of the intervention students attended a *KIPP* middle school.

Comparison group

Students in the comparison condition were not offered admission to a *KIPP* middle school. The majority of comparison students attended non-*KIPP* middle schools, though 5% attended a *KIPP* middle school at some point during the follow-up period.

Outcomes and measurement

Outcomes were the statewide reading and mathematics assessments for each state, measured in the third year of exposure. For a more detailed description of these outcome measures, see Appendix B.

Supplemental findings include the statewide reading and mathematics assessments for each state, for the first and second years of exposure. The supplemental findings do not factor into the intervention's rating of effectiveness.

Support for implementation

The study did not provide information about implementation support; however, authors noted that staff at *KIPP* schools had considerable autonomy in the implementation process to set the direction of the school.

Appendix A.2: Research details for Tuttle et al. (2015, Middle School, QED)

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Understanding the effect of KIPP as it scales: Volume I, Impacts on achievement and other outcomes. Final report of KIPP's Investing in Innovation grant evaluation* [Middle School; QED]. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560079

Additional sources:

Gleason, P. M., Tuttle, C. C., Gill, B., Nichols-Barrer, I., & Teh, B. (2014). *Do KIPP schools boost student achievement? Education, 9*(1), 36–58. Retrieved from https://eric.ed.gov/?id=EJ1016285

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP middle schools: Impacts on achievement and other outcomes, final report*. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED540912

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Going to scale: As KIPP network grows, positive impacts are sustained* (InFocus brief). Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560043

Tuttle, C. C., Teh, B., Nichols-Barrer, I., Gill, B., & Gleason, P. (2010). *Student characteristics and achievement in 22 KIPP middle schools: Final report*. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED511107

Tuttle, C. C., Teh, B., Nichols-Barrer, I., Gill, B., & Gleason, P. (2010). *Supplemental analytic sample equivalence tables for student characteristics and achievement in 22 KIPP middle schools: A report from the National Evaluation of KIPP Middle Schools*. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED511108

Table A2. Summary of findings **Meets WWC group design standards with reservations**

Outcome domain	Sample size	Study findings	
		Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	13,624 students	+11	Yes
English language arts achievement	14,551 students	+6	Yes
Science achievement	17,413 students	+10	Yes
Social studies achievement	9,762 students	+9	Yes

Setting

This analysis includes students and schools in multiple states and districts in the United States where *KIPP* charter schools operate. The study took place in 43 middle schools in the *KIPP* network in 20 cities across the following 12 states and the District of Columbia: Arkansas, California, Colorado, Georgia, Maryland, Massachusetts, New Jersey, New York, North Carolina, Pennsylvania, Tennessee, and Texas.

Study sample

The study used a matched-student quasi-experimental design, where the intervention group consisted of students who attended 37 *KIPP* middle schools, and the comparison group was a sample matched based on student baseline characteristics: baseline reading and math test scores; gender, race, special education, limited English proficiency, and free or reduced-price lunch status; and whether the student repeated a grade in the baseline year.

Sample characteristics for the analysis samples with non-imputed baseline data, on which the WWC based the intervention's effectiveness rating, are not reported.[13]

Intervention group

Students in the intervention condition attended a *KIPP* middle school at some point over the period 2001–13.

Comparison group

Students in the comparison condition attended non-*KIPP* middle schools.

Outcomes and measurement

Outcomes included the statewide assessments in each state for reading, mathematics, science, and social studies, measured in the fourth year of exposure. For a more detailed description of these outcome measures, see Appendix B.

Supplemental outcomes included the statewide assessments in each state for reading and mathematics measured in the first, second, and third years of exposure, as well as a subgroup analysis restricted to intervention students in new *KIPP* middle schools. Supplemental outcomes also included outcomes from the 2013 study, statewide assessments for each state for reading, mathematics, science, and social studies, measured in the first year of exposure, and outcomes from the 2010 study, statewide assessments for each state for reading and mathematics in the second, third, and fourth years of exposure. The supplemental findings do not factor into the intervention's rating of effectiveness.

Support for implementation

The study did not provide information about implementation support; however, authors noted that staff at *KIPP* schools had considerable autonomy in the implementation process to set the direction of the school.

Appendix A.3: Research details for Tuttle et al. (2015, High School)

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Understanding the effect of KIPP as it scales: Volume I, Impacts on achievement and other outcomes. Final report of KIPP's Investing in Innovation grant evaluation* [High School]. Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560079

Additional source:

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... Goble, L. (2015). *Going to scale: As KIPP network grows, positive impacts are sustained* (InFocus brief). Washington, DC: Mathematica Policy Research. Retrieved from https://eric.ed.gov/?id=ED560043

Table A3. 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	1,928 students	+8	Yes
English language arts achievement	2,260 students	+6	Yes
Science achievement	1,299 students	+13	Yes
Social studies achievement	601 students	+1	No
Student progression	852 students	+5	No

Setting

The two analyses included in the report include students and schools in multiple states and districts in the United States where *KIPP* charter schools operate. The study took place in 18 high schools in the *KIPP* network.

Study sample

The study used two designs. The first, for the analysis of new *KIPP* entrants (new *KIPP* student analysis), is a matched-student quasi-experimental design, where the intervention group consisted of students who attended 14 *KIPP* high schools, and the comparison group was a sample matched based on student baseline characteristics: baseline reading and math test scores; gender, race, special education, limited English proficiency, and free or reduced-price lunch status; and whether the student repeated a grade in the baseline year.

The second design, for the analysis of middle school *KIPP* students transitioning to high schools (continuing *KIPP* student analysis), the intervention group included *KIPP* middle school students who had the option to attend the local *KIPP* high school after completing grade 8. These students attended eight *KIPP* high schools (including four that were in the new *KIPP* student analysis). The comparison group consisted of *KIPP* students in grade 8 (in the same year) from middle schools in regions with no *KIPP* high school open at the time.

The comparison group was chosen from *KIPP* middle schools that most resembled the intervention middle schools on the basis of average school-level characteristics. Within that matched set of schools, a comparison sample of students was matched based on baseline reading and math test scores; gender, race, special education, limited English proficiency, and free or reduced-price lunch status; and whether the student repeated a grade in the baseline year.

For the new *KIPP* student analysis, sample characteristics for analysis samples with non-imputed baseline data are not reported.

For the continuing *KIPP* student analysis, 55% of students in the intervention condition were female, 49% were Black, and 38% were Hispanic. Among students in the comparison condition, 54% were female, 49% were Black, and 45% were Hispanic.

Intervention group

For the new *KIPP* student analysis, students in the intervention condition entered the *KIPP* network for the first time in grade 9. For the continuing *KIPP* student analysis, students in the intervention condition attended *KIPP* middle schools in grade 8 and had the option to attend *KIPP* high schools in grade 9. The majority of the students in the intervention condition attended *KIPP* high schools.

Comparison group

For the new *KIPP* student analysis, students in the comparison condition were from non-*KIPP* middle schools who remained at non-*KIPP* public schools in their high school years. For the continuing *KIPP* student analysis, students in the comparison condition attended *KIPP* middle schools in grade 8 and did not have the option to attend *KIPP* high schools in grade 9 because no local *KIPP* high schools were open at the time. Students in the comparison condition attended a wide variety of non-*KIPP* high schools.

Outcomes and measurement

Outcomes included the statewide assessments in each state for reading, mathematics, science, and social studies, measured in the second year of exposure; the TerraNova mathematics, reading, and language tests, measured in the third year of exposure; and high school graduation, measured in the fourth year of exposure. For a more detailed description of these outcome measures, see Appendix B.

Supplemental findings included a subgroup analysis of students who also attended a *KIPP* middle school, with statewide assessments in each state for reading, mathematics, science, and social studies, measured in the second year of exposure; and high school graduation, measured in the fourth year of exposure. The supplemental findings do not factor into the intervention's rating of effectiveness.

Support for implementation

The study did not provide information about implementation support; however, authors noted that staff at *KIPP* schools had considerable autonomy in the implementation process to set the direction of the school.

Appendix A.4: Research details for Woodworth et al. (2008)

Woodworth, K. R., David, J. L., Guha, R., Wang, H., & Lopez-Torkos, A. (2008). *San Francisco Bay Area KIPP schools: A study of early implementation and achievement. Final report.* Menlo Park, CA: SRI International.

Table A4. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size	Study findings	
		Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	3,535 students	+20	Yes
English language arts achievement	3,535 students	+13	Yes

Setting

The study was conducted in two unnamed school districts in the San Francisco Bay Area in California. Three *KIPP* schools were included in the intervention group in the analysis.

Study sample

The study used a matched-student quasi-experimental design, where the intervention group was comprised of students at five *KIPP* middle schools, and the comparison group was a sample of students matched on baseline reading and math test scores; gender, race, special education, limited English proficiency, and free or reduced-price lunch status; and whether the student repeated a grade in the baseline year.

There were 263 fifth-grade *KIPP* students included in the analytic sample. On average, 11% were Latino, 78% were African American, 8% were English learners, 14% were special education students, 81% were eligible for free or reduced-price lunches, and 49% were female. The average age in years of these students was 10.3. Among the 810 sixth-grade students (70 *KIPP* students and 740 comparison students) included in the analytic sample, 26% were Latino, 56% were African American, 28% were English learners, 7% were special education students, 86% were eligible for free or reduced-price lunches, and 59% were female.

Intervention group

The intervention consisted of 1 year of attendance at one of three *KIPP* schools in the Bay Area in California. Only students who attended the full year were included in the intervention group sample. Teachers joining *KIPP* schools in the sample generally came from highly selective colleges, were alternatively certified, and had a median of 3 years of classroom experience. The sixth-grade analysis includes students who joined one of the three study schools during their sixth-grade year.

Comparison group

Students in the comparison group experienced business-as-usual instruction at other schools in the district. Students who attended a *KIPP* school but transferred to another school in the district are excluded from the comparison group.

Outcomes and measurement

Outcomes included the California Standards Test for English Language Arts and Mathematics. The authors reported results separately by school and cohort, and the WWC aggregated results across schools and across both cohorts (2003 and 2004), separately for grade 5 and grade 6. For a more detailed description of these outcome measures, see Appendix B.

Support for implementation

The Bay Area *KIPP* schools raise between $400,000 and $700,000 each year to cover the gap between operating costs and the money they receive from state and local funds. The *KIPP* Foundation provides support by helping with teacher recruitment, fundraising, and other logistics. *KIPP* school leaders have substantial control over teacher hiring and their schools in general.

Appendix B: Outcome measures for each domain

Mathematics achievement	
Statewide mathematics assessments (z-score)	Statewide assessments are collected through administrative records requested from each state or district with students in the sample. Test scores are standardized within each year and state, with a mean of 0 and a standard deviation of 1 (as cited in Tuttle et al., 2015).
TerraNova Mathematics Test	This standardized assessment (Form G, Level 21/22) measures students' performance with *z*-scores that were standardized to capture student achievement relative to that of a nationally representative norming population. Test scores are standardized within each year, with a mean of 0 and a standard deviation of 1 (as cited in Tuttle et al., 2015).
TerraNova 3: Math Survey Exams	The TerraNova 3: Math Survey Exams, Level 17, Form G was administered to students in the fall of the third follow-up year. This one-time test was administered in fall of seventh grade (to lottery applicants for fifth grade) and fall of eighth grade (to applicants for sixth grade) (as cited in Tuttle et al., 2013).
California's State Standardized Assessment in Mathematics	School-level average sores from the California Standards Test (CST) are publicly available and were collected from the California Department of Education's (CDE's) website (as cited in Woodworth et al., 2008).
English language arts achievement	
Statewide assessment of reading achievement (z-score)	Statewide assessments are collected through administrative records requested from each state or district with students in the sample. Test scores are standardized within each year and state, with a mean of 0 and a standard deviation of 1 (as cited in Tuttle et al., 2015).
Statewide assessment of general literacy achievement (z-score)	Statewide assessments are collected through administrative records requested from each state or district with students in the sample. Test scores are standardized within each year and state, with a mean of 0 and a standard deviation of 1 (as cited in Tuttle et al., 2015).
TerraNova reading assessment (z-score)	This standardized assessment (Form G, Level 21/22) measures students' performance with *z*-scores that were standardized to capture student achievement relative to that of a nationally representative norming population. Test scores are standardized within each year, with a mean of 0 and a standard deviation of 1 (as cited in Tuttle et al., 2015).
TerraNova language assessment (z-score)	This standardized assessment (Form G, Level 21/22) measures students' performance with *z*-scores that were standardized to capture student achievement relative to that of a nationally representative norming population. Test scores are standardized within each year, with a mean of 0 and a standard deviation of 1 (as cited in Tuttle et al., 2015).
California Standards Test English/ Language Arts (CST-ELA)- scaled score	School-level average sores from the California Standards Test (CST) are publicly available and were collected from the California Department of Education's (CDE's) website (as cited in Woodworth et al., 2008).
Science achievement	
Statewide science assessments (z-score)	Statewide assessments are collected through administrative records requested from each state or district with students in the sample. Test scores are standardized within each year and state, with a mean of 0 and a standard deviation of 1 (as cited in Tuttle et al., 2015).
Social studies achievement	
Statewide social studies assessments (z-score)	Statewide assessments are collected through administrative records requested from each state or district with students in the sample. Test scores are standardized within each year and state, with a mean of 0 and a standard deviation of 1 (as cited in Tuttle et al., 2015).
Student progression	
High school graduation	This outcome is an indicator for whether or not a student graduated from high school within 4 years of grade 9 entry. Students who transfer to another high school in the district are included, but those who transfer to a private high school or a school in another district are classified as non-graduates (as cited in Tuttle et al., 2015).

Appendix C.1: Findings included in the rating for the mathematics 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
Tuttle et al. (2015, Middle School, RCT)[a]								
Statewide mathematics assessments (z-score)	Middle school: lottery sample; Year 3	455 students	0.01 (nr)	–0.17 (nr)	0.18	0.18	+7	< .01
Domain average for (Tuttle et al., 2015, Middle School, RCT)						**0.18**	**+7**	**Statistically significant**
Tuttle et al. (2015, Middle School, QED)[b]								
Statewide mathematics assessments (z-score)	Middle school: matched-student sample; Year 4	13,624 students	0.14 (nr)	–0.13 (nr)	0.27	0.27	+11	< .01
Domain average for mathematics (Tuttle et al., 2015, High School)						**0.27**	**+11**	**Statistically significant**
Tuttle et al. (2015, High School)[c]								
Statewide mathematics assessment (z-score)	High school: new *KIPP* student analysis; Year 2	1,416 students	0.24 (nr)	–0.04 (nr)	0.27	0.27	+11	< .01
TerraNova Mathematics test	High school: continuing *KIPP* student analysis; Year 3	512 students	0.07 (nr)	–0.07 (nr)	0.14	0.14	+5	.22
Domain average for mathematics (Tuttle et al., 2015, High School)						**0.20**	**+8**	**Statistically significant**
Woodworth et al. (2008)[d]								
California's State Standardized Assessment in Mathematics	Grade 5, Year 1	2,725 students	331.91 (74.29)	298.76 (59.55)	33.15	0.54	+21	< .01
California's State Standardized Assessment in Mathematics	Grade 6; Year 1	810 students	349.91 (82.46)	311.29 (76.02)	38.62	0.50	+19	< .01
Domain average for mathematics (Woodworth et al., 2008)						**0.52**	**+20**	**Statistically significant**
Domain average for mathematics across all studies						**0.29**	**+12**	**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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. na = not applicable. nr = not reported.

[a] For Tuttle et al. (2015, Middle School, RCT) third year follow-up mathematics outcome of the middle school lottery sample, 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. This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0) p. 26.

[b] For Tuttle et al. (2015, Middle School, QED), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented was reported in the original study. This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[c] For Tuttle et al. (2015, High School), 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 were reported in the original study. This study is characterized as having a statistically significant positive effect because at least one measure is positive and statistically significant and no effects are negative and statistically significant, accounting for multiple comparisons. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[d] For Woodworth et al. (2008), the impact estimates were reported separately by school and cohort. The WWC aggregated the findings across schools and cohorts. The *p*-values presented here were calculated by the WWC based on the aggregated findings. This study is characterized as having a statistically significant positive effect because at least one measure is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons. 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 English language arts domain

			Mean (standard deviation)		WWC calculations			
Outcome measure	**Study sample**	**Sample size**	**Intervention group**	**Comparison group**	**Mean difference**	**Effect size**	**Improvement index**	***p*-value**
Tuttle et al. (2015, Middle School, RCT)[a]								
Statewide assessment of reading achievement (z-score)	Middle school lottery sample; Year 3	458 students	–0.13 (nr)	–0.28 (nr)	0.14	0.14	+6	.01
Domain average for English language arts (Tuttle et al., 2015, Middle School, RCT)						**0.14**	**+6**	**Statistically significant**
Tuttle et al. (2015, Middle School, QED)[b]								
Statewide assessment of reading achievement (z-score)	Middle school: matched-student sample, Year 4	14,551 students	0.08 (nr)	–0.09 (nr)	0.16	0.16	+6	< .01
Domain average for English language arts (Tuttle et al., 2015, Middle School, QED)						**0.16**	**+6**	**Statistically significant**
Tuttle et al. (2015, High School)[c]								
Statewide assessment of general literacy achievement (z-score)	High school: new *KIPP* student analysis; Year 2	1,748 students	0.11 (nr)	–0.07 (nr)	0.18	0.18	+7	< .01
TerraNova Reading (z-score)	High school: continuing *KIPP* student analysis; Year 3	512 students	0.26 (nr)	0.10 (nr)	0.16	0.16	+6	.03
TerraNova Language (z-score)	High school: continuing *KIPP* student analysis; Year 3	512 students	0.07 (nr)	–0.05 (nr)	0.12	0.12	+5	.15
Domain average for English language arts (Tuttle et al., 2015, High School)						**0.15**	**+6**	**Statistically significant**

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
Woodworth et al. (2008)[d]								
California Standards Test English/Language Arts (CST-ELA)	Grade 5, Year 1	2,725 students	325.47 (44.58)	316.39 (42.33)	9.08	0.21	+8	< .01
CST-ELA	Grade 6, Year 1	810 students	339.85 (54.10)	313.33 (62.84)	26.52	0.43	+17	< .01
Domain average for English language arts (Woodworth et al., 2008)						**0.32**	**+13**	**Statistically significant**
Domain average for English language arts across all studies						**0.19**	**+8**	**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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. na = not applicable. nr = not reported.

[a] For Tuttle et al. (2015, Middle School, RCT) third year follow-up reading outcome of the middle school lottery, 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. This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[b] For Tuttle et al. (2015, Middle School, QED), 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. This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[c] For Tuttle et al. (2015, High School), 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 were reported in the original study. This study is characterized as having a statistically significant positive effect because at least one measure is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[d] For Woodworth et al. (2008), the impact estimates were reported separately by school and cohort. The WWC aggregated the findings across schools and cohorts. The *p*-values presented here were calculated by the WWC based on the aggregated findings. This study is characterized as having a statistically significant positive effect because at least one measure is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons. 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 science 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
Tuttle et al. (2015, Middle School, QED)[a]								
Statewide science assessments (z-score)	Middle school: matched-student sample, Year 4	17,413 students	0.08 (nr)	–0.17 (nr)	0.25	0.25	+10	< .01
Domain average for science (Tuttle et al., 2015, Middle School, QED)						**0.25**	**+10**	**Statistically significant**

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Tuttle et al. (2015, High School)[b]								
Statewide science assessments (z-score)	High school: new *KIPP* student analysis, Year 2	1,299 students	0.11 (nr)	–0.22 (nr)	0.33	0.33	+13	< .01
Domain average for science (Tuttle et al., 2015, High School)						**0.33**	**+13**	**Statistically significant**
Domain average for science across all studies						**0.29**	**+11**	**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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. na = not applicable. nr = not reported.

[a] For Tuttle et al. (2015, Middle School, QED), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented was reported in the original study. This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[b] For Tuttle et al. (2015, High School), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented was reported in the original study. This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. 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 social studies domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Tuttle et al. (2015, Middle School, QED)[a]								
Statewide social studies assessments (z-score)	Middle school: matched-student sample, Year 4	9,762 students	0.11 (nr)	–0.13 (nr)	0.24	0.24	+9	< .01
Domain average for social studies (Tuttle et al., 2015, Middle School, QED)						**0.24**	**+9**	**Statistically significant**
Tuttle et al. (2015, High School)[b]								
Statewide social studies assessments (z-score)	High school: new *KIPP* student analysis, Year 2	601 students	–0.13 (nr)	–0.15 (nr)	0.02	0.02	+1	.80
Domain average for social studies (Tuttle et al., 2015, High School)						**0.02**	**+1**	**Not Statistically significant**
Domain average for social studies across all studies						**0.13**	**+5**	**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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. na = not applicable. nr = not reported.

[a] For Tuttle et al. (2015, Middle School, QED), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented was reported in the original study. This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[b] For Tuttle et al. (2015, High School), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented was reported in the original study. This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix C.5: Findings included in the rating for the student progression domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Tuttle et al. (2015, High School)[a]								
High school graduation	High school: new *KIPP* student analysis; Year 4	852 students	0.71 (na)	0.67 (na)	0.04	0.11	+5	0.36
Domain average for student progression (Tuttle et al., 2015, High School)						**0.11**	**+5**	**Not statistically significant**
Domain average for student progression across all studies						**0.11**	**+5**	**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. Some statistics may not sum as expected due to rounding. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. na = not applicable.

[a] For Tuttle et al. (2015, High School), 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. This study is characterized as having an indeterminate effect because the estimated effect for the outcome in the student progression 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 D.1: Description of supplemental findings for the mathematics domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Tuttle et al. (2015, Middle School, RCT)[a]								
Statewide mathematics assessments (z-score)	Middle school: lottery sample; Year 1	607 students	–0.12 (nr)	–0.22 (nr)	0.10	0.10	+4	.05
Statewide mathematics assessments (z-score)	Middle school: lottery sample; Year 2	555 students	–0.01 (nr)	–0.25 (nr)	0.24	0.24	+9	< .01
Tuttle et al. (2015, Middle School, QED)[b]								
Statewide mathematics assessments (z-score)	Middle school: matched-student sample; Year 1	34,938 students	–0.05 (nr)	–0.11 (nr)	0.06	0.06	+2	< .01
Statewide mathematics assessments (z-score)	Middle school: matched-student sample; Year 2	27,736 students	0.09 (nr)	–0.14 (nr)	0.23	0.23	+9	< .01
Statewide mathematics assessments (z-score)	Middle school: matched-student sample; Year 3	21,926 students	0.17 (nr)	–0.12 (nr)	0.29	0.29	+11	< .01
Statewide mathematics assessments (z-score)	Middle school: matched-student sample (new *KIPP* middle schools); Year 1	2,366 students	–0.19 (nr)	–0.23 (nr)	0.04	0.04	+2	.07
Statewide mathematics assessments (z-score)	2013 Full sample: Year 1	31,832 students	nr (nr)	nr (nr)	0.15	0.15	+6	.01
Statewide mathematics assessments (z-score)	2010 Full sample: Year 2	8,020 students	nr (nr)	nr (nr)	0.35	0.35	+14	< .01
Statewide mathematics assessments (z-score)	2010 Full sample: Year 3	5,439 students	nr (nr)	nr (nr)	0.41	0.41	+16	< .01
Statewide mathematics assessments (z-score)	2010 Full sample: Year 4	2,576 students	nr (nr)	nr (nr)	0.35	0.35	+14	< .01
Tuttle et al. (2015, High School)[c]								
Statewide mathematics assessments (z-score)	Cumulative middle and high school matched-student sample: Year 2	2,930 students	0.34 (nr)	0.00 (nr)	0.34	0.34	+13	< .01

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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. nr = not reported.

[a] For Tuttle et al. (2015, Middle School, RCT), 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.

[b] For Tuttle et al. (2015, Middle School, QED), corrections for 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.

[c] For Tuttle et al. (2015, High School), 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.

Appendix D.2: Description of supplemental findings for the English language arts domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Tuttle et al. (2015, Middle School, RCT)[a]								
Statewide assessment of reading achievement (z-score)	Middle school: lottery sample; Year 1	608 students	–0.23 (nr)	–0.26 (nr)	0.03	0.03	+1	.58
Statewide assessment of reading achievement (z-score)	Middle school: lottery sample; Year 2	563 students	–0.16 (nr)	–0.34 (nr)	0.18	0.18	+7	< .01
Tuttle et al. (2015, Middle School, QED)[b]								
Statewide assessment of reading achievement (z-score)	Middle school: matched-student sample; Year 1	34,915 students	–0.11 (nr)	–0.11 (nr)	0.01	0.01	0	0.43
Statewide assessment of reading achievement (z-score)	Middle school: matched-student sample; Year 2	27,758 students	–0.01 (nr)	–0.11 (nr)	0.11	0.11	+4	< .01
Statewide assessment of reading achievement (z-score)	Middle school: matched-student sample; Year 3	22,155 students	0.06 (nr)	–0.09 (nr)	0.15	0.15	+6	< .01
Statewide assessment of reading achievement (z-score)	Middle school: matched-student sample (new KIPP middle schools); Year 1	2,360 students	–0.22 (nr)	–0.27 (nr)	0.05	0.05	+2	.03
Statewide assessment of reading achievement (z-score)	2013 Full matched sample; Year 1	31,832 students	nr (nr)	nr (nr)	0.05	0.05	+2	.01
Statewide assessment of reading achievement (z-score)	2010 Full sample: Year 2	8,041 students	nr (nr)	nr (nr)	0.14	0.14	+6	< .01

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Statewide assessment of reading achievement (z-score)	2010 Full sample: Year 3	5,442 students	nr (nr)	nr (nr)	0.23	0.23	+9	< .01
Statewide assessment of reading achievement (z-score)	2010 Full sample: Year 4	2,570 students	nr (nr)	nr (nr)	0.16	0.16	+6	< .01
Tuttle et al. (2015, High School)[c]								
Statewide assessment of reading achievement (z-score)	Cumulative middle and high school matched-student sample, Year 2	4,001 students	0.38 (nr)	0.09 (nr)	0.30	0.30	+12	< .01

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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. nr = not reported.

[a] For Tuttle et al. (2015, Middle School, RCT), 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.

[b] For Tuttle et al. (2015, Middle School, QED), corrections for 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.

[c] For Tuttle et al. (2015, High School), 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.

Appendix D.3: Description of supplemental findings for the science domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Tuttle et al. (2015, Middle School, QED)[a]								
Statewide science assessments (z-score)	2013 Full sample matched comparison sample; Year 3	8,699 students	nr (nr)	nr (nr)	0.33	0.33	+13	.02
Tuttle et al. (2015, High School)[b]								
Statewide science assessments (z-score)	Cumulative middle and high school matched-student sample; Year 2	3,582 students	0.42 (nr)	0.00 (nr)	0.42	0.42	+16	< .01

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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. nr = not reported.

[a] For Tuttle et al. (2015, Middle School, QED), 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.

[b] For Tuttle et al. (2015, High School), the WWC did not need to make corrections for multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study.

Appendix D.4: Description of supplemental findings for the social studies domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Tuttle et al. (2015, Middle School, QED)[a]								
Statewide social studies assessments (z-score)	2013 Full sample matched comparison sample; Year 3	6,904 students	nr (nr)	nr (nr)	0.25	0.25	+10	.02
Tuttle et al. (2015, High School)[b]								
Statewide social studies assessments (z-score)	Cumulative middle and high school matched-student sample: Year 2	1,495 students	0.18 (nr)	–0.09 (nr)	0.27	0.27	+11	< .01

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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis. nr = not reported.

[a] For Tuttle et al. (2015, Middle School, QED), 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.

[b] For Tuttle et al. (2015, High School), the WWC did not need to make corrections for multiple comparisons or to adjust for baseline differences. The *p*-value presented here was reported in the original study.

Appendix D.5: Description of supplemental findings for the student progression domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Tuttle et al. (2015, High School)[a]								
High school graduation	Cumulative middle and high school matched-student sample: Year 4	2,033 students	0.79 (na)	0.65 (na)	0.14	0.41	+16	< .01

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. Corrections for clustering were not needed, as the unit of assignment (student) is the same as the unit of analysis.

[a] For Tuttle et al. (2015, High School), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented was reported in the original study.

Endnotes

[1] The descriptive information for this intervention comes from the *Knowledge Is Power Program (KIPP)* website at http://www.*KIPP*.org, and from Furgeson et al. (2014). The What Works Clearinghouse (WWC) requests developers review the intervention description sections for accuracy from their perspective. The WWC provided the developer with the intervention description in April 2017; however, the WWC did not receive a response. Further verification of the accuracy of the descriptive information for this intervention is beyond the scope of this review.

[2] The literature search reflects documents publicly available by February 2017. The WWC released a single study review of Tuttle et al. (2013) in November 2013. In addition to the quasi-experimental analytic sample that currently meets WWC group design standards with reservations, the single study review also reported on the lottery-based analytic sample. The previous review used a liberal boundary for attrition for the lottery-based analytic sample, resulting in a rating of meets WWC group design standards without reservations. The current review is based on the Charter School review protocol (version 3.0), which uses the conservative boundary for attrition. The lottery-based analytic sample had high attrition, and in response to the WWC's request for baseline data to establish equivalence, the authors stated that these data were not available. Therefore, this analysis of the lottery-based analytic sample does not meet WWC group design standards in this review.

Reviews of the studies in this report used the standards from the WWC Procedures and Standards Handbook (version 3.0) and the Charter Schools review protocol (version 3.0). The evidence presented in this report is based on available research. Findings and conclusions could change as new research becomes available.

[3] Absence of conflict of interest: This intervention report includes studies conducted by staff from Mathematica Policy Research, Inc. Because Mathematica is one of the contractors that administers the WWC, staff members from a different organization reviewed the study. The lead methodologist, a WWC quality assurance reviewer, and an external peer reviewer reviewed this report.

[4] Please see the effectiveness summary in this report or the Charter Schools review protocol (version 3.0) for a list of all outcome domains.

[5] For criteria used to determine the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 32. These improvement index numbers show the average and range of individual-level improvement indices for all findings across the studies.

[6] Please see the Charter Schools review protocol (version 3.0) for details on the types of interventions that are eligible for review. A study of the effectiveness of an individual charter school is eligible to be included in a review of the evidence of the effectiveness of an individual charter school, but is not eligible to be included in a review of the evidence of the effectiveness of a named CMO or charter network, like *KIPP*.

[7] Tuttle et al. (2015) includes analyses at three school grade ranges: elementary, middle, and high school. These are treated here as four separate studies, with the elementary study labeled [Elementary]; two middle school studies labeled [Middle School, RCT] and [Middle School, QED]; and high school labeled [High School]. Tuttle et al. (2015) [Elementary] was a randomized controlled trial that included eight elementary *KIPP* schools that were operating in spring 2011. The study compared prekindergarten and kindergarten students who were offered admission to a *KIPP* elementary school in the 2011–12 school year with a comparison group of students who were not offered admission to a *KIPP* school. The elementary school-focused randomized controlled trial had high attrition but did not demonstrate the required baseline equivalence for the intervention and comparison groups, so this analysis does not meet WWC standards.

Tuttle et al. (2015, Middle School, QED), involves three related reports: Tuttle et al. (2010), Tuttle et al. (2013), and Tuttle et al. (2015). The WWC combined its review of these three reports rather than reviewing them separately because there is substantial overlap in the sample of schools included in the three reports. Tuttle et al. (2010) is a matched-students quasi-experimental design including 22 *KIPP* middle schools in the analysis. Tuttle et al. (2013) expands on the 2010 study, including additional schools (41 *KIPP* middle schools in the analysis), outcomes, cohorts, and years. Tuttle et al. (2015) included analyses of both middle school-based randomized controlled trials and quasi-experimental designs, included 37 *KIPP* middle schools, 25 of which were in the 2013 study, and included additional outcome measures, cohorts, and years.

[8] Tuttle et al. (2013) also presented results from a randomized controlled trial that assigned students to a *KIPP* middle school based on an admissions lottery. The results based on the randomized controlled trial had high attrition and did not demonstrate the required baseline equivalence for the intervention and comparison groups; therefore, the analysis does not meet WWC standards and did not contribute to the study's effectiveness rating.

[9] Within each study, the primary findings that the WWC considered for the effectiveness rating are those measured at the latest follow-up period. Findings from earlier follow-up periods and earlier cohorts are considered supplemental.

[10] Tuttle et al. (2015, High School) also analyzed matched high schools, comparing students at five *KIPP* high schools in their first year of operation with a comparison group of students from the same *KIPP* middle schools in adjacent cohorts. Because the comparison group is from a prior cohort, the effect of the intervention cannot be isolated from the effect of events that may have occurred in the two different time periods. For this reason, this analysis does not meet WWC standards, and did not contribute to the study's effectiveness rating.

[11] Woodworth et al. (2008) estimated impacts separately by cohort and school for fifth-grade students, and by school for sixth-grade students. The WWC aggregated impacts across cohorts and schools for fifth-grade students and across schools for sixth-grade students.

[12] Throughout the report, the WWC applies adjustments for multiple comparisons when there are multiple outcomes with the same domain and years of exposure. The WWC does not make adjustments for clustering, as students choose to enter or exit charter schools individually, so the unit of assignment is at the same level (student) as the unit of analysis, per the topic area protocol.

[13] The WWC does not allow imputed baseline data to be used to assess the equivalence of the intervention and comparison groups at baseline, as noted in the WWC Procedures and Standards Handbook (version 3.0), p. 18.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2018, January). *Charter Schools intervention report: Knowledge is Power Program (KIPP).* Retrieved from https://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 without reservations, 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 without reservations, 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 subjects initially assigned to the intervention and comparison groups. If a randomized controlled trial (RCT) or regression discontinuity design (RDD) study has high levels of attrition, the validity of the study results can be called into question. An RCT with high attrition cannot receive the highest rating of *Meets WWC Group Design Standards without Reservations,* but can receive a rating of *Meets WWC Group Design Standards with Reservations* if it establishes baseline equivalence of the analytic sample. Similarly, the highest rating an RDD with high attrition can receive is *Meets WWC RDD Standards with Reservations.*

For single-case design research, attrition occurs when an individual fails to complete all required phases or data points in an experiment, or when the case is a group and individuals leave the group. If a single-case design does not meet minimum requirements for phases and data points within phases, the study cannot receive the highest rating of *Meets WWC Pilot Single-Case Design Standards without Reservations.*

Baseline A point in time before the intervention was implemented in group design research and in regression discontinuity design studies. When a study is required to satisfy the baseline equivalence requirement, it must be done with characteristics of the analytic sample at baseline. In a single-case design experiment, the baseline condition is a period during which participants are not receiving the intervention.

Clustering adjustment An adjustment to the statistical significance of a finding when the units of assignment and analysis differ. When random assignment is carried out at the cluster level, outcomes for individual units within the same clusters may be correlated. When the analysis is conducted at the individual level rather than the cluster level, there is a mismatch between the unit of assignment and the unit of analysis, and this correlation must be accounted for when assessing the statistical significance of an impact estimate. If the correlation is not accounted for in a mismatched analysis, the study may be too likely to report statistically significant findings. To fairly assess an intervention's effects, in cases where study authors have not corrected for the clustering, the WWC applies an adjustment for clustering when reporting statistical significance.

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 method by which intervention and comparison groups are assigned (group design and regression discontinuity design) or the method by which an outcome measure is assessed repeatedly within and across different phases that are defined by the presence or absence of an intervention (single-case design). Designs eligible for WWC review are randomized controlled trials, quasi-experimental designs, regression discontinuity designs, and single-case designs.

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 analytic sample groups are similar on observed characteristics defined in the review area protocol.

Extent of evidence An indication of how much evidence from group design studies supports the findings in an intervention report. The extent of evidence categorization for intervention reports focuses on the number and sizes of studies of the intervention in order to give an indication of how broadly findings may be applied to different settings. There are two extent of evidence categories: small and medium to large.

- **small:** includes only one study, or one school, or findings based on a total sample size of less than 350 students and 14 classrooms (assuming 25 students in a class)
- **medium to large:** includes more than one study, more than one school, and findings based on a total sample of at least 350 students or 14 classrooms

Gain scores The result of subtracting the pretest from the posttest for each individual in the sample. Some studies analyze gain scores instead of the unadjusted outcome measure as a method of accounting for the baseline measure when estimating the effect of an intervention. The WWC reviews and reports findings from analyses of gain scores, but gain scores do not satisfy the WWC's requirement for a statistical adjustment under the baseline equivalence requirement. This means that a study that must satisfy the baseline equivalence requirement and has baseline differences between 0.05 and 0.25 standard deviations *Does Not Meet WWC Group Design Standards* if the study's only adjustment for the baseline measure was in the construction of the gain score.

Group design A study design in which outcomes for a group receiving an intervention are compared to those for a group not receiving the intervention. Comparison group designs eligible for WWC review are randomized controlled trials and quasi-experimental designs.

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 An adjustment to the statistical significance of results to account for multiple comparisons in a group design study. The WWC uses the Benjamini-Hochberg (BH) correction to adjust the statistical significance of results within an outcome domain when study authors perform multiple hypothesis tests without adjusting the p-value. The BH correction is used in three types of situations: studies that tested multiple outcome measures in the same outcome domain with a single comparison group; studies that tested a given outcome measure with multiple comparison groups; and studies that tested multiple outcome measures in the same outcome domain with multiple comparison groups. Because repeated tests of highly correlated constructs will lead to a greater likelihood of mistakenly concluding that the impact was different from zero, in all three situations, the WWC uses the BH correction to reduce the possibility of making this error. The WWC makes separate adjustments for primary and secondary findings.

Outcome domain A group of closely-related outcomes. A domain is the organizing construct for a set of related outcomes through which studies claim effectiveness.

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 For group design research, the WWC rates the effectiveness of an intervention in each domain based on the quality of the research design and the magnitude, statistical significance, and consistency in findings. For single-case design research, the WWC rates the effectiveness of an intervention in each domain based on the quality of the research design and the consistency of demonstrated effects. The criteria for the ratings of effectiveness are given in the WWC Rating Criteria on p. 32.

Regression discontinuity design A design in which groups are created using a continuous scoring rule. For example, students may be assigned to a summer school program if they score below a preset point on a standardized test, or schools may be awarded a grant based on their score on an application. A regression line or curve is estimated for the intervention group and similarly for the comparison group, and an effect occurs if there is a discontinuity in the two regression lines at the cutoff.

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.

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.

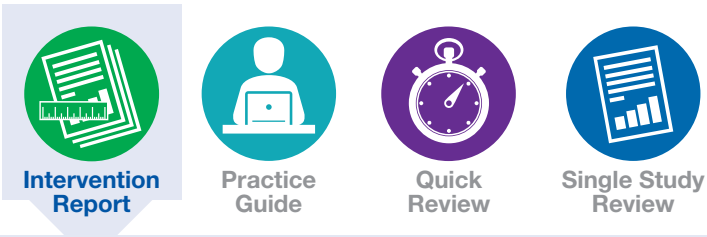
Statistical significance Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ($p < .05$).

Study rating The result of the WWC assessment of a study. The rating is based on the strength of the evidence of the effectiveness of the educational intervention. Studies are given a rating of *Meets WWC Design Standards without Reservations, Meets WWC Design Standards with Reservations,* or *Does Not Meet WWC Design Standards,* based on the assessment of the study against the appropriate design standards. The WWC has design standards for group design, single-case design, and regression discontinuity design studies.

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.