What Works Clearinghouse™

ies INSTITUTE OF EDUCATION SCIENCES

November 2013

WWC Review of the Report "KIPP Middle Schools: Impacts on Achievement and Other Outcomes, Final Report"[1,2]

The findings from this review do not reflect the full body of research evidence on the *Knowledge is Power Program (KIPP)*.

What is this study about?

The study examined whether attending a *Knowledge is Power Program (KIPP)* middle school improved students' academic performance for up to 4 years following enrollment.

The authors used two study designs, one experimental and one quasi-experimental, to examine student achievement outcomes in reading, math, science, and social studies.

The experimental portion of the study included about 1,000 students who applied to attend 13 *KIPP* middle schools that conducted admissions lotteries. The intervention group was comprised of students who won a lottery for a slot in a *KIPP* middle school, and the comparison group was composed of those who did not win the lottery and instead enrolled in other middle schools in the area. The authors estimated the impact of the offer of a slot at a *KIPP* school through fall of third year after the admissions lottery.[3]

In the quasi-experimental portion of the study, the authors used baseline achievement and demographic characteristics to match 15,916 students in 41 *KIPP* middle schools with similar students who had attended non-*KIPP* public middle schools in the same school district in the previous year. They followed these students for up to 4 years.

Features of the *Knowledge is Power Program (KIPP)*

Founded in 1994, *KIPP* operates a charter school network to improve the education of low-income children. Since then, the network has expanded to serve 125 *KIPP* charter schools in 20 states during the 2012–13 school year.

The *KIPP* model rests on what are described as "Five Pillars":

- High expectations for academic achievement
- Choice and commitment of students and families to college preparatory education
- More time spent learning, both in academic and extracurricular activities
- Power to lead for school principals, who are given freedom in budgeting, personnel, and other decisions
- Focus on results by regularly assessing student learning and driving accountability

What did the experimental design find?

The experimental portion of the study found positive and statistically significant impacts on mathematics achievement both 1 and 2 years after the lottery (effect sizes of 0.11 and 0.22, respectively) and in fall of the third year after the lottery (effect size of 0.20); however, reading impacts were not statistically significant at any follow-up period.

WWC Rating of the Experimental Design

The research described in the experimental portion of the study presented in this report meets WWC evidence standards without reservations for the 1-year follow-up and meets standards with reservations for the later follow-ups

Strengths: The analysis sample for 1-year outcomes was based on a randomized controlled trial with low attrition. The analysis samples for later outcomes contained intervention and comparison groups that were baseline equivalent, despite the study experiencing high attrition after the first year.

Cautions: The study experienced high sample attrition at the second-year follow-up and on the TerraNova outcomes in fall of the third follow-up year. The authors demonstrated baseline equivalence on achievement and demographic characteristics for these outcomes, so they meet WWC evidence standards with reservations.

What did the quasi-experimental design find?

The quasi-experimental portion of the study concluded that, for all 4 years examined, students enrolled in *KIPP* middle schools scored statistically significantly higher on state assessments in mathematics (effect sizes ranging from 0.15 to 0.36) and reading achievement (effect sizes ranging from 0.05 to 0.22) than similar students who attended non-*KIPP* public middle schools.

Science and social studies achievement were also statistically significantly higher for students attending *KIPP* schools, as measured 3 to 4 years after enrollment, than similar students in non-*KIPP* schools (effect sizes of 0.33 and 0.25, respectively).

WWC Rating of the Quasi-experimental Design

The research described in the quasi-experimental portion of the study presented in this report meets WWC evidence standards with reservations

Strengths: Intervention and comparison students were well-matched on baseline achievement and demographic characteristics, and the analysis included appropriate statistical controls.

Cautions: Although the authors matched *KIPP* students to traditional public school students on a number of observable characteristics, it is possible that there were other differences between the two groups that were not accounted for in the analysis and that could have influenced student achievement.

Appendix A: Study details

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.

Setting

The experimental portion of the study was conducted in 13 *KIPP* schools located in California, the District of Columbia, Georgia, Massachusetts, New York, and Texas. The quasi-experimental portion of the study was conducted in 41 *KIPP* schools located in Arkansas, California, Colorado, the District of Columbia, Georgia, Indiana, Louisiana, Massachusetts, New York, North Carolina, Oklahoma, Pennsylvania, Tennessee, and Texas.

Study sample

For the experimental portion of the study, each of 13 *KIPP* schools used a lottery to randomly assign students either to receive an offer to attend the *KIPP* school or to not receive an offer. The schools included in the analyses varied depending on the availability of outcome data, and a list of schools used in each analysis was provided by the authors after an inquiry by the WWC. Specifically, the analysis of state assessments in reading and mathematics included 725 students (260 intervention and 465 comparison) entering fifth or sixth grade who applied to attend 10 of the *KIPP* schools that used a lottery. The follow-up sample in these schools included 536 students (202 intervention and 334 comparison) 1 year after random assignment and 441 students (181 intervention and 260 comparison) 2 years after random assignment. The experimental portion of the study also included the administration of the TerraNova reading and mathematics assessment at 10 schools, some of which were different than the 10 schools included in the analysis of state assessments. The sample for this portion of the study included 1,016 students (431 intervention and 585 comparison) at the beginning of the study and 590 students (272 intervention and 318 comparison) at the time of the TerraNova assessment, which was administered in fall of the third follow-up year.

For the quasi-experimental portion of the study, students in 41 *KIPP* schools that were established before or during the 2009–10 school year were matched to comparison students who had never attended a *KIPP* middle school. *KIPP* students enrolled in a *KIPP* school in the fifth or sixth grade, and they were matched to non-*KIPP* students enrolled in the same district who had similar demographic characteristics and prior achievement scores using nearest neighbor propensity score matching without replacement. Between two and 10 cohorts of students per school were included in the study, depending on data availability, and outcome data were drawn from the 2001–02 through 2010–11 school years. The study included 31,832 students in the investigation of reading and math outcomes in year one (half were *KIPP* students and half were non-*KIPP*). Students were matched once and then followed over time and across subjects. Therefore, study sample sizes in later analysis years and for science and social studies outcomes vary depending on the availability of outcome data for the originally matched sample.

Intervention group

The intervention groups for both the experimental and quasi-experimental portions of the study attended *KIPP* schools, which are designed to engage students and parents in the educational process, expand the amount of time dedicated to learning, reinforce students' social competencies and positive behaviors, and improve academic achievement. The *KIPP* model rests on the "Five Pillars": (a) high expectations for academic achievement; (b) choice and commitment of students and families to college preparatory education; (c) more time spent learning, both in academic and extracurricular activities; (d) power to lead for school principals, who are given freedom in budgeting, personnel, and other decisions; and (e) focus on results by regularly assessing student learning and driving accountability.

Comparison group

In the experimental design, 62% of students in the comparison group attended traditional public schools, 20% attended non-*KIPP* charter schools, 14% attended *KIPP* schools, and 4% attended private schools. Students in the quasi-experimental comparison group attended non-*KIPP* middle schools in the feeder school district.

Outcomes and measurement

Both the experimental and quasi-experimental studies measured state assessments in math and reading, which are typically administered in spring of the school year. These outcomes were measured for 4 follow-up years in the quasi-experimental study and for 2 follow-up years in the experimental study. In addition, science and social studies state exams were included in the quasi-experimental study. These outcomes were measured by the latest available middle school score in each jurisdiction, which was typically eighth grade (i.e., 3 to 4 years post-enrollment). The TerraNova reading and mathematics exams were administered in the experimental design only, in fall of the third follow-up year. For a more detailed description of these outcome measures, see Appendix B.

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.

Reason for review

This study was identified for review by receiving media attention.

Appendix B: Outcome measures for each domain

Mathematics achievement	
State assessments	State assessments in mathematics achievement were typically administered in spring of the school year. Up to 2 years of mathematics assessments were included in the experimental portion of the study, and up to 4 years of mathematics assessments were included in the quasi-experimental portion of the study. Z-score transformations were made that standardized test scores by subject, grade, and year within a given district (i.e., *KIPP* students' Z-scores were standardized relative to all other students in the same district). This allowed students to be compared across schools and jurisdictions in different states.
TerraNova	The TerraNova 3, Math Survey Exams, Level 17, Form G was administered to students in the experimental portion of the study in fall of the third follow-up year. For students promoted on time, 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).
Reading achievement	
State assessments	State assessments in reading achievement were typically administered in spring of the school year. Up to 2 years of reading assessments were included in the experimental portion of the study, and up to 4 years of reading assessments were included in the quasi-experimental portion of the study. Z-score transformations were made that standardized test scores by subject, grade, and year within a given district (i.e., *KIPP* students' Z-scores were standardized relative to all other students in the same district). This allowed students to be compared across schools and jurisdictions in different states.
TerraNova	The TerraNova 3, Reading Multiple Assessment, Level 17, Form G was administered to students in the experimental portion of the study in fall of third follow-up year. For students promoted on time, 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).
Social studies achievement	
State assessments	State assessments in social studies achievement for students in the quasi-experimental portion of the study were typically administered in spring of the school year. Because social studies exams were not administered every year for a given cohort, impacts on these outcomes were measured by the latest available middle school score in each jurisdiction, which was typically eighth grade (i.e., 3 to 4 years post-enrollment). Z-score transformations were made that standardized test scores by subject, grade, and year within a given district (i.e., *KIPP* students' Z-scores were standardized relative to all other students in the same district). This allowed students to be compared across schools and jurisdictions in different states.
Science achievement	
State assessments	State assessments in science achievement for students in the quasi-experimental portion of the study were typically administered in spring of the school year. Because science exams were not administered every year for a given cohort, impacts on these outcomes were measured by the latest available middle school score in each jurisdiction, which was typically eighth grade (i.e., 3 to 4 years post-enrollment). Z-score transformations were made that standardized test scores by subject, grade, and year within a given district (i.e., *KIPP* students' Z-scores were standardized relative to all other students in the same district). This allowed students to be compared across schools and jurisdictions in different states.

Table Notes: As part of the experimental portion of the study, student and parent surveys were administered 2 years after the admissions lotteries. These outcomes cover four domains of attitudes and behaviors: (a) student engagement and effort in school, (b) educational aspirations and expectations, (c) student well-being and behavior, and (d) satisfaction with and perceptions of school. These outcomes were not included because attitudes and behaviors are outside the scope of the single study review protocol.

Appendix C: Study findings for each domain, experimental design

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Mathematics achievement								
State assessments (RCT)	1-year posttest	536 students	nr	nr	nr	0.11	+4	< 0.05
State assessments (RCT)	2-year follow-up	441 students	nr	nr	nr	0.22	+9	< 0.01
TerraNova (RCT)	3-year follow-up	589 students	nr	nr	nr	0.20	+8	< 0.01
Domain average for mathematics achievement						**0.18**	**+7**	**Statistically significant**
Reading achievement								
State assessments (RCT)	1-year posttest	535 students	nr	nr	nr	0.02	+1	> 0.05
State assessments (RCT)	2-year follow-up	441 students	nr	nr	nr	0.09	+4	> 0.05
TerraNova (RCT)	3-year follow-up	590 students	nr	nr	nr	0.08	+3	> 0.05
Domain average for reading achievement						**0.06**	**+2**	**Not statistically significant**

Table Notes: For 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 student outcomes, representing the average change expected for all students who are given the intervention (measured in standard deviations of the outcome measure). The effect sizes presented here were reported in the original study and are calculated by dividing the regression-adjusted impact estimate by the standard deviation for the comparison group. The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is given the intervention. nr = not reported. RCT = Randomized controlled trial.

Study Notes: No corrections for clustering or multiple comparisons and no difference-in-differences adjustment were needed. The *p*-values presented here were reported in the original study. The experimental portion of the study is characterized as having a statistically significant positive effect on mathematics achievement because the effect for at least one measure within the domain is positive and statistically significant, and no effects are negative and statistically significant. This portion of the study is characterized as having an indeterminate effect on reading achievement because none of the estimated effects within the domain was statistically significant. For more information, please refer to the WWC Standards and Procedures Handbook, version 2.1, page 96.

Appendix D: Study findings for each domain, quasi-experimental design

Domain and 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
Mathematics achievement								
State assessments (QED)	1-year follow-up	31,832 students	nr	nr	nr	0.15	+6	< 0.01
State assessments (QED)	2-year follow-up	22,819 students	nr	nr	nr	0.27	+11	< 0.01
State assessments (QED)	3-year follow-up	16,218 students	nr	nr	nr	0.36	+14	< 0.01
State assessments (QED)	4-year follow-up	8,262 students	nr	nr	nr	0.31	+12	< 0.01
Domain average for mathematics achievement						**0.27**	**+ 11**	**Statistically significant**
Reading achievement								
State assessments (QED)	1-year follow-up	31,832 students	nr	nr	nr	0.05	+2	< 0.01
State assessments (QED)	2-year follow-up	22,819 students	nr	nr	nr	0.14	+6	< 0.01
State assessments (QED)	3-year follow-up	16,218 students	nr	nr	nr	0.21	+8	< 0.01
State assessments (QED)	4-year follow-up	8,262 students	nr	nr	nr	0.22	+9	< 0.01
Domain average for reading achievement						**0.16**	**+6**	**Statistically significant**
Social studies achievement								
State assessments (QED)	3–4 years post-enrollment	6,904 students	nr	nr	nr	0.25	+10	< 0.01
Domain average for social studies achievement						**0.25**	**+10**	**Statistically significant**
Science achievement								
State assessments (QED)	3–4 years post-enrollment	8,699 students	nr	nr	nr	0.33	+13	< 0.01
Domain average for science achievement						**0.33**	**+13**	**Statistically significant**

Table Notes: For 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 regressions controlled for 2 years of baseline test scores as well as demographic characteristics and use robust standard errors that are clustered at the student level. The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is 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 the study's domain average was determined by the WWC. nr = not reported. QED = quasi-experimental design.

Study Notes: No corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study. The quasi-experimental portion of the study is characterized as having a statistically significant positive effect because the effect for at least one measure within each domain is positive and statistically significant, and no effects are negative and statistically significant. For more information, please refer to the WWC Standards and Procedures Handbook, version 2.1, page 96. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the average change expected for all students who are given the intervention (measured in standard deviations of the outcome measure). The effect sizes presented here were reported in the original study and are the average of equally-weighted impact estimates from regressions of Z-scores that were performed separately for each of the 41 KIPP middle schools in the sample.

Endnotes

[1] Single study reviews examine evidence published in a study (supplemented, if necessary, by information obtained directly from the author[s]) to assess whether the study design meets WWC evidence standards. The review reports the WWC's assessment of whether the study meets WWC evidence standards and summarizes the study findings following WWC conventions for reporting evidence on effectiveness. This study was reviewed using the single study review protocol, version 2.0. A quick review of this study was released on March 27, 2013, and this report is the follow-up review that replaces that initial assessment. The WWC rating applies only to the results that were eligible under this topic area and met WWC standards without reservations or met WWC standards with reservations, and not necessarily to all results presented in the study.

[2] Absence of conflict of interest: This study was conducted by staff from Mathematica Policy Research. Therefore, Mathematica reviewers were not involved in the WWC review of this study.

[3] There were four outcomes included in the experimental portion of the study that are not described in this WWC report. See the table notes in Appendix B for more information.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, November). *WWC review of the report: KIPP middle schools: Impacts on achievement and other outcomes, final report.* Retrieved from http://whatworks.ed.gov

Glossary of Terms

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

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

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

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

Domain A domain is a group of closely related outcomes.

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

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

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

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

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

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

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

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

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

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

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

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