NEPC
NATIONAL EDUCATION POLICY CENTER

REVIEW OF *KIPP MIDDLE SCHOOLS*

Reviewed By

Gregory Camilli

University of Colorado Boulder

April 2013

Summary of Review

Using two different approaches, researchers from Mathematica Policy Research conclude that Knowledge Is Power Program (KIPP) students scored higher than comparison students not attending KIPP schools by an amount equivalent to 11 months of additional learning in math and about eight months in reading. The impact was unevenly distributed across KIPP schools, and a number of factors were identified that were weakly related to this variation in effectiveness. The evaluation study was carefully planned and executed, and the results are about the same magnitude as those from other experiments in education. The KIPP outcomes may be substantial if found to persist into later grades. The benefits, however, appear to be overstated in the evaluation study for two reasons. First, translating educational outcomes into "months" of additional learning is an inexact science and can lead to absurd results if taken literally. Second, reported measures of effectiveness that take attrition into account are smaller than the estimates used to draw conclusions about the effectiveness of KIPP. In addition, the effect of KIPP on higher-order reasoning is less certain than is portrayed in the report. The latter topic requires additional empirical work to provide greater clarity.

Kevin Welner
Project Director

William Mathis
Managing Director

Erik Gunn
Managing Editor

National Education Policy Center

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: (802) 383-0058

Email: NEPC@colorado.edu
http://nepc.colorado.edu

Publishing Director: Alex Molnar

GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

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

Review of *KIPP Middle Schools*

Gregory Camilli, University of Colorado Boulder

I. Introduction

The report *KIPP Middle Schools: Impacts on Achievement and Other Outcomes*, conducted by researchers at Mathematica Policy Research (MPR), represents a thorough and ambitious attempt to evaluate the benefit to students of attending Knowledge Is Power Program (KIPP) middle schools from (primarily) grade 5 to grade 8.[1] The study was undertaken by a large and experienced staff, and is a follow-on to an earlier 2010 report.[2] The 2013 study has two main components. First, KIPP students in 41 schools were compared with a matched sample of students from non-KIPP schools (though only 38 schools provided enough information to be included in this component of the study). Data were available for between 100 and 800 students who attended the KIPP schools. Student scores on state tests were used to estimate the impact of KIPP in mathematics, reading, and science along with a number of measures collected in surveys administered to parents and schools. Second, a randomized experiment was carried out for over-subscribed schools. Students were assigned to KIPP schools based on a supervised lottery, while non-selected lottery participants were put on a waiting list. Ten KIPP schools had valid data for the lottery experiment analysis, and school sample sizes ranged from 536 in Year 1 to 441 in Year 2. Data from two different achievement tests were used to collect information on the impact of KIPP.

The 2013 report consists of 156 pages, including 86 pages of technical appendices. The basic findings of the report seem reasonable, if not common sense. The methodology is carefully thought through and executed. This is not to claim that its results are correct, but rather that there is no obvious reason to suspect major procedural errors in the MPR evaluation. With respect to the report, three key issues regarding KIPP benefits in reading and mathematics are the focus of this review.

II. Findings and Conclusions of the Report

The MPR researchers concluded that the effects of KIPP were "large enough to be educationally meaningful"; the KIPP impact translated into about "11 months of additional learning in math after three years. . . and approximately 8 months" in reading (p. xvii).

Across 38 schools, the proportion of effect sizes that were positive and significantly different from zero was 65% in mathematics and 55% in reading. Thus, substantial variation in effectiveness exists across schools: in some schools the effect was much larger than for others. This raises the question of *why* KIPP schools are effective. To address this issue, researchers from MPR examined a broad range of variables. In terms of statistically significant correlations, higher-achieving KIPP schools tended to have shorter school days (r =-.19), but more time spent on core subjects (r = .13). Principals in more effective schools tended to be more experienced (r = .02). Outcomes correlated positively (r = .08) with the presence of school-wide behavior plans (e.g., discipline policies, rewards for positive behavior, individualization).

III. The Report's Use of Research Literature

The research literature is frugally cited with respect to related issues, such as experiments on the effects of school voucher plans. Since the MPR report is an evaluation of a particular program rather than a research study, a more inclusive literature review was neither expected nor necessary.

IV. Review of the Report's Methods

Three areas of methodological interest are examined in this section: the choice of impact estimate, the educational significance of the impact estimate, and the key question of what makes some KIPP schools more effective than others.

ITTs and TOTs—Choice of Impact Estimate

To provide some background for the following remarks on the interpretation of impact, consider the following statements excerpted from the 2013 report:

> For the subset of KIPP middle schools in which randomized lotteries created viable treatment and control groups, we present two sets of impact estimates: (1) intent-to-treat (ITT) estimates that rely on treatment status as defined by the random lotteries to estimate the impact of being offered admission to a KIPP middle school and (2) treatment-on-the-treated (TOT) estimates that represent the impact of attending a KIPP middle school.
>
> Our benchmark experimental model is the ITT model, comparing outcomes for the experimental treatment and control groups. (p. 16)

I surmised from this passage that ITT estimates were preferred to TOT estimates as "benchmarks." It is important to understand why this is the case and what is meant by these terms.

With ITT, the goal is to estimate the effect of all students originally in the experiment. Attrition and other factors (such as missing data) that narrow the original population of students must be controlled. In other words, the KIPP impact group is defined relative to all students originally randomized to treatment and control conditions, regardless of whether they remained in the program over the study period or had usable data. Although

> *Given the small size of the correlations of impact with program variables, the reasons that some KIPP schools are more effective than others are not presently understood*

at first blush this seems impossible, modern methods of statistics (e.g., missing value imputation) can compensate to some degree for the lack of information, to the degree that sufficient background information was originally collected for all students. Sample attrition is an important issue for any large-scale study.

With TOT, on the other hand, the goal is to estimate a treatment effect for those students who actually received the treatment. This is an important distinction from the ITT approach. Causal effects are obtained from ITT estimates. This is equivalent to asking the degree to which a randomly selected student performs better in a KIPP school than in a non-KIPP school. In contrast, the TOT approach is targeted to a slightly different question: Is the treatment beneficial to those who actually received the treatment? For this reason, TOT estimates are typically preferred in program evaluations. The bottom line here is that if a population receiving the treatment differs substantially from the population initially assigned to treatment conditions, then the TOT estimate of treatment effect may be influenced by extraneous variables such as the degree of attrition and missing data. For this reason, the ITT estimate of effectiveness is often the standard in clinical trials in medical and pharmaceutical experiments: it provides primary evidence, whereas the TOT provides supportive evidence in arriving at a judgment of efficacy.[3]

It might be somewhat confusing to readers of the MPR report that the primary estimates of impact (see Tables IV.1 and IV.2) are described as having been obtained with the "benchmark" model. Yet these are not ITT estimates as originally stated on p. 16; rather, they are TOT estimates (e.g., see p. 42). The effect of this confusion is to reverse the roles of primary and secondary evidence, if indeed the ITT approach is the benchmark model. The MPR researchers may have swapped benchmark models because ITT estimates were obtained from only 10 schools, a subset that may be less representative of KIPP than the 38 schools in the matched sample. The case is not explicitly made for changing the preferred estimate of causal impact, though this point is indirectly addressed by the claim that the ITT estimates of effect are "similar" to the TOT estimates.

A number of ITT and TOT estimates of impact are reported for both state tests and the TerraNova (third edition), which is a major standardized achievement test battery used nationally. The estimates are reported in the metric *effect size*, which places the treatment benefit on a scale that can be compared with other studies. Whether the ITT and TOT are

actually similar is a matter of judgment. The ITT estimates of effects of KIPP on student outcomes on the TerraNova 3 after two years were $d = 0.22$ for mathematics and $d = 0.09$ for reading. The corresponding TOT estimates are $d = 0.36$ for mathematics and $d = 0.15$ for reading. Using the MPR method of translating effect sizes into months, the TOT estimate is larger than the ITT estimate in mathematics by about 4 months over the course of 2 years, and the corresponding difference in reading about 3 months. Is this difference negligible?

Educational Significance of Impact

The impact of a treatment can be thought of as the benefit a control-group student would have obtained *if* he or she had been assigned to treatment. This *counter-factual* interpretation is common thinking in modern experimental studies. The average difference between outcomes measures in the treatment (in this case, KIPP) and comparison groups is an estimate of the benefit or impact of an intervention in this counter-factual sense. Though informally many refer to impact as "gain," this latter term can be misleading because treatment impact is not necessarily a measure of growth over time. With this in mind, educational interventions in experimental studies with large samples typically have impacts corresponding to effect sizes in the 0.2 to 0.4 range. For any particular study, however, it is necessary to find more appropriate ways to benchmark an effect size in order to communicate its practical significance. The effect sizes for the MPR report are summarized in Table 1.

Table 1. Reported Effect Sizes after 2 Years

Group	Assessment	After 2 Years*	
		ITT	TOT
Lottery	***State***		
	math	0.22	0.36
	read	0.09	0.15
	TerraNova		
	math	0.20	0.35
	read	0.08	0.12
Matched	***State***		
	math	n/a	0.27 (0.36)
	read	n/a	0.14 (0.21)

* (Third-Year Outcomes in Parentheses)

The impact in mathematics in terms of effect size is similar to that reported in Success for All prior to middle school,[4] and in other studies cited by the authors. The MPR researchers attempted to provide more concrete ways to convey the educational impact of KIPP, however. To arrive at the conclusion of 11 months of gains for mathematics, a study was

used that provided the norms for annual growth across grades in effect size units as shown in the first column of Table 2.[5] For example, the KIPP impact in mathematics was found by the MPR analysts to be 0.36, and normative growth from grades 7-8 is 0.32 from the first column of Table 2. Dividing 0.36 by 0.32, and multiplying by the length of a school year (about 10 months), results in 11.25 months. To a large degree, effect size norms for annual growth rely heavily on a number of foundational assumptions, and in particular, that a scale can be created in which a 10-point interval, for example, on a test means the same thing in different grades. This is the "equal interval" assumption, which is viewed as unwarranted by many measurement experts.[6] It is not clear that any test score scale can be used to measure ability across a range of grades in the same way a tape measure can be used to represent distance across a Persian carpet in terms of feet and inches.

Table 2. Growth in Reading (Mathematics) across Grades in Effect Size Units

	Effect Size		
Annual Gain	Bloom*	Kim	Kolen
Grade 3 -4	0.36 (0.52)	0.97 (0.87)	0.75 (0.69)
Grade 4 -5	0.40 (0.56)	0.88 (0.41)	0.76 (0.39)
Grade 5 -6	0.32 (0.41)	0.63 (0.41)	0.61 (0.40)
Grade 6 -7	0.23 (0.30)	0.74 (0.34)	0.76 (0.37)
Grade 7 -8	0.26 (0.32)	0.58 (0.26)	0.71 (0.31)

*Note: In the Bloom norms, the gain from grade K-1 is 1.14 effect size units.

In addition to this issue, other estimates of growth (in effect size units), shown in the second[7] and third[8] columns of Table 2, have been obtained that diverge substantially from those in the first column.

As can be seen, the annual reading growth estimates in the second column of Table 2 are more than twice those in the first column. In the third column, estimates similar to the first column were again obtained in mathematics, but a constant increase across grades was obtained for reading. Using the results of the third column would result in an estimate of impact one-third the magnitude of the reported 8-month impact in reading. This suggests that "months of benefit" is not a precise way to characterize educational benefit. Not only do different studies lead to different months of gain for the same effect size, but other head-scratchers also arise. For example, in mathematics the KIPP impact is equivalent to approximately one fourth of the increase from kindergarten to first grade. Moreover, a procedural translation of effect size into *months of gain* is highly prone to misinterpretation. For example, one might surmise that the KIPP impact in mathematics

of 11 months indicates that an average student in grade 8 could perform adequately in terms of a grade 9 curriculum. This interpretation is incorrect.[9] To their credit, the MPR researchers also assessed the practical importance of the KIPP impacts in terms of the Black-White achievement gap. Accordingly, the impact represented 40% of the gap in mathematics and 26% in reading. Such benefits would be very important educationally if they persist beyond middle school--and this research is apparently a part of an ongoing MPR study on scale-up (p. 69).

On p. xii, it is claimed that "KIPP produces similar [relative to state tests] positive impacts on the norm-referenced test, which includes items assessing higher-order thinking." The latter items are described as "constructed (open-ended) response item[s]" (p. 45). Such items require students to provide written responses to test items rather than merely selecting an option as in multiple-choice questions. Further, the MPR researchers state on p. 45, "if KIPP affects only students' basic skills, we would expect estimated impacts on TerraNova scores to be closer to zero than the estimated impacts on state test scores." The intent is to present evidence that KIPP impact was not the result of "teaching to the test." Three points are relevant here. First, the TerraNova (third edition) Survey was used to assess mathematics. There are no constructed response items on this instrument; such items only appear on the Multiple Assessments reading subtest. Second, the TerraNova is not simply a test of higher-order reasoning: it is a standardized test with items and content pitched at a range of levels. Some of the questions are quite basic—as they should be in a test designed to evaluate a range of knowledge and skills. Yet no information is reported to demonstrate that the impact for KIPP students was similar on constructed—and selected—response items. The lower benefit of KIPP in reading as seen in Table 1 suggests that performance on constructed-response items might fall below the effect size range of 0.22-0.36. This is merely a hypothesis, but one that could be investigated empirically. Third, constructed-response items do not necessarily tap higher-order thinking.

What Makes KIPP Effective?

KIPP schools differ from other schools in ways that might explain the overall impact. Namely, according to ITT estimates, KIPP schools required about 22 minutes more homework per day, as reported by students (parents reported an additional 32 minutes of homework). Tempering this difference somewhat was the finding that KIPP schooling is associated with a greater amount of undesirable behavior such as lying to or arguing with parents and a higher proportion of students who get into trouble at school (Table V.3). Nonetheless, students and parents report higher satisfaction with KIPP schools, and parents report less often that the school is too easy (Table V.4)

V. Review of the Validity of the Findings and Conclusions

It is common sense that longer homework policies and longer school days (about 9 hours) will have an effect on achievement, though the current research on homework "leaves

much to be desired."[10] Given the small size of the correlations of impact with program variables, however, the reasons that some KIPP schools are more effective than others are not presently understood. As noted by the MPR researchers, "The factors that drive the success of KIPP schools could not easily be determined in our analysis" (p. 68). This appears to be ongoing work at MPR, so the story here may not be over.

Evidence in support of the efficacy of KIPP is positive. An impact is reported that is similar in effect size magnitude to other large-scale interventions. This impact is substantial relative to the achievement gap, but overstated as additional months of education. Two methodological points also deserve some attention. First, attrition and missing data pose challenges in any experimental field study. The ITT approach to assessing impact addresses these issues more adequately than the TOT approach. Thus, the ITT effect sizes provide a safer, though more conservative, choice for stakeholders and policymakers. Second, measurement theory should be an explicit aspect of intervention research. It is important to understand how estimates of impact depend on data transformations,[11] vertical-scale assumptions, and item formats such as constructed-response and multiple-choice.

Missing information may deserve more attention, though the MPR researchers appear to have treated this issue with great care. Rates of missing values (e.g., no test score) were found to be about the same across KIPP and comparison groups, and the estimates of effectiveness do not appear to depend on the statistical adjustments made for missing data. On a school-by-school basis, however, differential rates of missing values could have had an effect. Thus, it would be important to show that an effect size for a KIPP school does not depend on the proportion of missing test scores in that school.

VI. Usefulness of the Report for Guidance of Policy and Practice

The report's results are not intended to be extrapolated beyond KIPP schools. Nonetheless, the pedagogical methods used in KIPP schools, or in any school, are of interest to the degree that they provide information about what might work on a larger scale. As noted above, the variation in KIPP outcomes is poorly understood at present, thus limiting the potential for extrapolation. Positive achievement outcomes were slightly correlated with a few organizational features of schools and with some behavioral issues. (Oddly, difference in homework time was not used to explain the variability of effectiveness across KIPP sites.) The correlations are too low, however, to be taken as guidance for instructional policy. On the other hand, the overall positive outcomes are consistent with the results of other educational interventions and should not be downplayed relative to the variability in effectiveness among KIPP schools. Future work evaluating the persistence of KIPP impact will be key to drawing a conclusive judgment of the educational significance of KIPP schooling.

Notes and References

1 Tuttle, C.C. *et al.* (2013, February) . *KIPP middle schools: Impacts on achievement and other outcomes*. Washington, DC: Mathematica Policy Research. Retrieved April 26, 2013, from http://www.mathematica-mpr.com/publications/pdfs/education/KIPP_middle.pdf/.

2 Tuttle, C.C., Teh, B-r., Nichols-Barrer, I,. Gill, B.P., & Gleason, P. (2010, June). *Student characteristics and achievement in 22 KIPP middle schools*. Washington, DC: Mathematica Policy Research.

3 Gupta, S.K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2(3), 109–112.

4 Borman, G.D., Slavin, R.E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal*, 44, 701-731.

5 Growth estimates in effect size units are averaged across across seven tests in Bloom, H., Hill, C., Black, A.R., & Lipsey, M. (2008, October). Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions. Retrieved March 28, 2013, from http://www.mdrc.org/sites/default/ files/full_473.pdf/.

6 Briggs, D. (in press). Measuring growth with vertical scales. *Journal of Educational Measurement*.

7 Kim, J., Frisbie, D.A., Kolen, M.J., & Kim, D-I. (2007, April). *A comparison of calibration methods and proficiency estimators for creating irt vertical scales*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

8 Tong, Y., & Kolen, M.J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227-253.

9 This incorrect interpretation is similar to misconceptions of the grade equivalent scale. See pp. 235-236 of Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., 221-262). New York: Macmillan.

10 Patall, E.A., Cooper, H., & Allen A.B. (2010). Extending the school day or school year: A systematic review of research (1985–2009). *Review of Educational Research*, 80, 401-436.

11 To combine results across different state tests, variables were standardized in the MPR analyses. It should be recognized, however, this is not equivalent to equating the tests and may conceal important population differences.

DOCUMENT REVIEWED:	**KIPP Middle Schools: Impacts on Achievement and Other Outcomes**
AUTHORS:	Christina Clark Tuttle, Brian Gill, Philip Gleason, Virginia Knechtel, Ira Nichols-Barrer, Alexandra Resch
PUBLISHER/THINK TANK:	Mathematica Policy Research
DOCUMENT RELEASE DATE:	February 27, 2013
REVIEW DATE:	April 30, 2013
REVIEWER:	Gregory Camilli, University of Colorado Boulder
E-MAIL ADDRESS:	g.camilli@colorado.edu
PHONE NUMBER:	(303) 492-8391

SUGGESTED CITATION:

Camilli, G. (2013). *Review of "KIPP Middle Schools: Impacts on Achievement and Other Outcomes."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-KIPP-middle-schools/.