RESEARCH REPORT

Effectiveness of *Measuring Up* as Preparation for the New York State Fourth Grade Test in English Language Arts: A Report of a Randomized Experiment in Mount Vernon City School District

June 30, 2005

Denis Newman, Ph.D.
Empirical Education Inc.

Andrew P. Jaciw
Stanford University

Saundra Young, Ph.D.
KnowledgeQuest

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 220
Palo Alto, CA 94306
(650) 328-1734

Acknowledgements

This research was performed by MarketingWorks, Inc. and Empirical Education Inc. under a contract to MarketingWorks from Peoples Publishing Group. We are grateful to the people in the Mount Vernon City School District for their cooperation and assistance in conducting this research.

About Empirical Education Inc.

Empirical Education Inc. was founded to help school districts, publishers, and the R&D community assess new or proposed instructional and professional development programs through scientifically based pilot implementations. The company draws on the expertise of world-class researchers and methodologists assuring that the research is objective and takes advantage of current best practice in rigorous experimental design and statistical analysis. The company's findings let educators quantify the value of programs and help them partner with providers to implement those most effective for their students.

Reference this report: Newman, D., Jaciw, A., & Young, S. (2005, June). Effectiveness of Measuring Up as preparation for the New York State fourth grade test in English Language Arts: A report of a randomized experiment in Mount Vernon City School District. Palo Alto, CA: Empirical Education Inc. Retrievable from https://www.empiricaleducation.com/past_research/

Executive Summary

We were asked by Peoples Publishing Group to find out whether their product, *Measuring Up* (*MU*), was more effective in helping a district prepare fourth-grade students for the New York State test of English Language Arts than materials the district already had in place. We conducted a randomized experiment during the 2004-2005 school year in Mt. Vernon, NY.

Intervention. *Measuring Up* (*MU*) is a supplementary, text-based product designed to help teachers in K-12 schools prepare students for their standards-based state exams. *MU* is customized for a range of state standards and assessment practices and goes well beyond simple test prep by systematically addressing the state standards and providing a range of pedagogical tools. After a half-day in-service session led by a PPG staff member, teachers were free to use *MU* as best suited their needs.

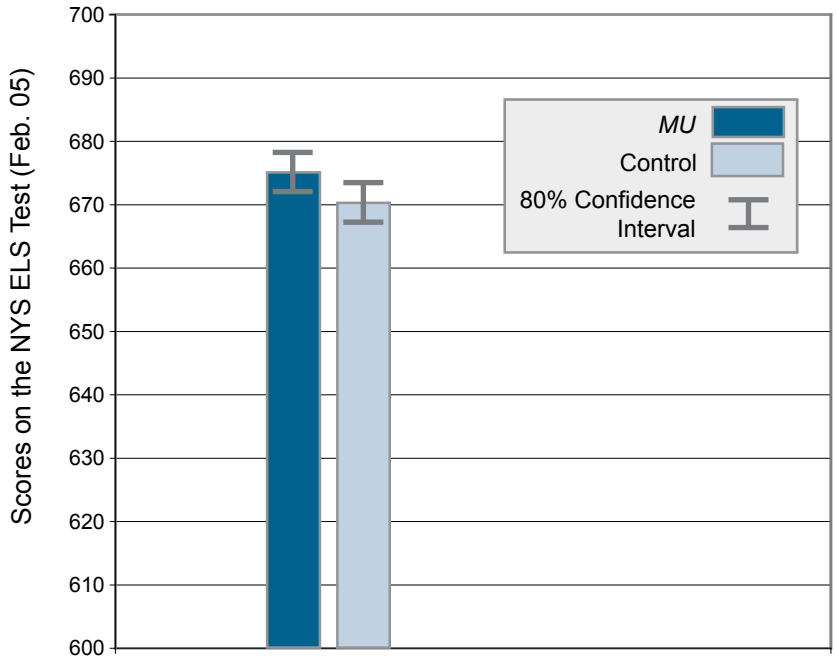
Setting. Our research site is a city of approximately 68,000 located in Westchester County just north of New York City, where the median household income is about $41,000. The district enrolls about 10,000 K-12 students, including 78% African American, 13% Hispanic, and 8% White students. Materials for preparing fourth graders for the ELA test had not been adopted and teachers used a wide variety of materials and methods to prepare their students: teacher-developed materials, enrichment activities, some supplementary products, and segments of their basal text.

Research design. Our research design was a randomized experiment (or randomized controlled trial). This type of study is the best way to assure that the new product and not some characteristic of the teachers or students caused the differences observed between groups. Teachers who volunteered to participate were assigned by coin toss to the *MU* group or to the control group. The process of randomly assigning teachers to conditions assured that classes from each school were approximately evenly distributed between conditions and that the distinct populations were represented in each condition. The outcome measure was the state test of English Language Arts. The pretest measure was the TONYSS (Riverside Publishing).

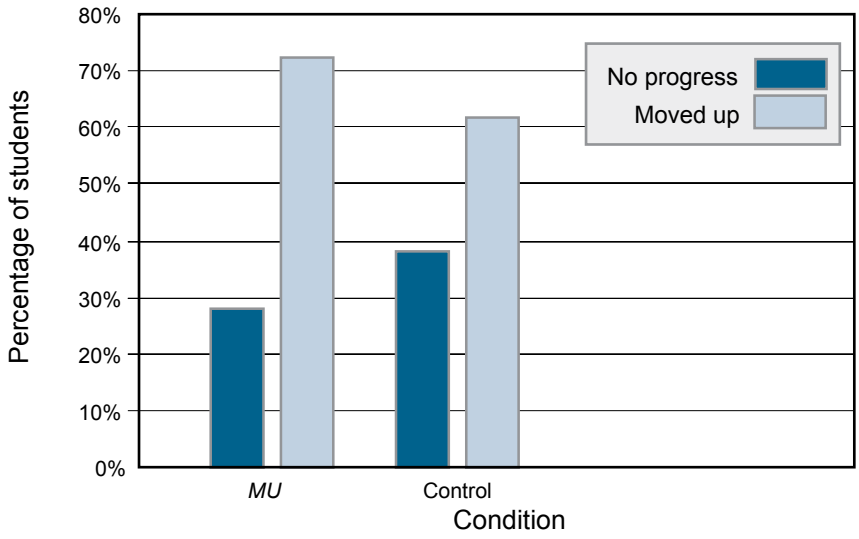
Participants. A total of 375 fourth-grade students and 19 teachers participated in the study. Random assignment resulted in two groups containing students who were evenly split on socioeconomic status. Control group students scored slightly higher on the pretest than *MU* students, a difference that was controlled statistically.

Statistical analysis. Our method for drawing a quantitative conclusion from the data considered what was known about student demographics and their incoming level of ELA as well as outcome differences between *MU* and control group student scores. We used an analysis of covariance combined with a multi-level analysis to increase the precision of the estimate of *MU*'s impact and to account for the clustering of students in classes. We also tested whether the impact depends on students' incoming skills and other important factors.

Results. Teachers in both the control group and the group piloting *MU* reported spending approximately the same amount of classroom time preparing for the state test. Comparison of the means for *MU* and control groups revealed a difference of 5.46 points. Translated into a standardized measure that takes into account the distribution of the scores, we find an effect size of .15. Effect sizes in this range are often found to be educationally meaningful. The bar graph shows this difference between *MU* and control group scores. The bars represent the score that would be predicted for a student performing at the average level in the two groups. At the top of each bar, we have indicated the 80% confidence interval. There is a probability of 80% that the true values for groups lie within their respective confidence intervals.



Viewed in terms of the four proficiency levels established by New York State, we find that a significantly larger portion of students in the classes with *MU* moved to a higher proficiency level compared to those in the control classrooms.



Conclusion. For the students and teachers in Mt. Vernon, *MU* was generally more effective than the other products in use for helping to improve student achievement. We can conclude that *Measuring Up* is a valuable option for supplementing English Language Arts instruction where the goal is higher achievement on the New York State test.

Effectiveness of *Measuring Up* as Preparation for the New York State Fourth Grade Test in English Language Arts:

A Report of a Randomized Experiment in Mount Vernon City School District

Contents

Peoples Publishing Group (PPG) is supporting an independent program of research with the goal of producing scientific evidence of the effectiveness of *Measuring Up* (*MU*), a product designed to help teachers prepare students for state exams. Our research focused on the *MU* product that prepared fourth graders for the English Language Arts test in New York.

We conducted a research experiment in 19 fourth-grade classrooms in Mt. Vernon, NY. We randomly assigned ten teachers to use *MU* and nine to continue their standard practice (the control group). Teachers used the materials for five months from the beginning of the school year, then administered the New York state test in February.

The question we addressed is whether *MU* is more effective, as measured by the state mandated achievement test, than the exam preparation program the district already has in place. We analyzed our results taking into account student demographics and their English Language Arts (ELA) proficiency.

In our quantitative study of overall effectiveness of *MU*, we used a variety of data collection methods, including surveys, interviews, and student test scores. The researchers did not directly observe classrooms or analyze qualitative differences in classroom implementation. Although we examined quantitative measures of implementation, our focus was on differences in test scores between classes of teachers using *MU* and existing materials.

Our experimental design reflects the requirements of the No Child Left Behind Act, which directs schools to consult reports of rigorous research to guide their adoptions of instructional programs. The US Department of Education has been explicit in interpreting this requirement in terms of randomized experimentation (also called “randomized controlled trials”) for determining effectiveness.

Methods

Research Design

This research is a comparison of outcomes for classes taught using *Measuring Up* and classes taught with other materials. We randomly assigned approximately equal numbers of teacher volunteers to *MU* and control groups. The outcome measure was student test scores of English Language Arts on the New York State test. The Test of New York State Standards (TONYSS) by Riverside Publishing was used as a pretest. The design includes two levels: the unit of random assignment is the teacher and the unit for the outcome data is the individual student. Within a multi-level experiment, analysis of covariance (ANCOVA) is used to increase the precision of the estimated treatment impact; factors such as incoming ELA proficiency and ethnicity can also be modeled to determine whether the treatment impact depends on the levels of these covariates.

Materials

Measuring Up is a supplementary, text-based product designed to support teachers in K-12 schools in preparing their students for their standards-based state exams. *MU* is customized for the standards and assessment practices in each of the 12 states where it is sold. Teachers in our study used materials addressing the New York State fourth-grade test of ELA. Although *MU* includes practice tests, it goes well beyond a simple test prep product in that it systematically addresses the state standards for content areas and provides a range of pedagogical tools for that purpose. Teachers received a half-day in-service training session led by a PPG staff member. Beyond the initial training, teachers were free to make use of

the materials as best suited the needs of their classroom and students.

Site Description

Our research site, Mount Vernon, NY, is a city of approximately 68,000 located in Westchester County just north of New York City. The median household income is about $41,000. The city's population is predominantly African American (about 60%), with White Non-Hispanic (24%) and Hispanic (10%) making up most of the remainder. The Mount Vernon City School District enrolls approximately 10,000 K-12 students. The ethnic make-up of the district includes 78% African American, 13% Hispanic, and 8% White students. The district had not adopted a standard set of materials for the purpose of preparing for this test; thus teachers were using a wide variety of materials and methods to prepare their students, including teacher-developed materials, enrichment activities, some supplementary products, and segments of their basal text.

Sample and Randomization

The site was initially identified as a district interested in the product and willing to conduct a structured pilot with a subset of classes. Researchers met with this district's administrative staff to explain the procedures. Principals invited interested teachers to an after-school meeting at which we introduced the *MU* product and held a discussion about the research procedures.

Twenty-one qualified teachers attended the kick-off meeting for the experiment on September 22, 2004. After a question-and-answer period, teacher volunteers engaged in a discussion of the important district factors that they believed would affect the results. They expected that researchers would find differing impacts across schools because of differences in student population and teacher experience. To help assure that the pilot and control groups were made up of approximately the same number of experienced and inexperienced teachers as well as the same number of classrooms from lower and higher SES schools, teachers grouped themselves into school teams and, within each team, formed pairs based on teaching experience. Once maximally similar pairs were established, we tossed a coin to decide which member of each pair joined the *MU* group and which one joined the control group. (When the group had an uneven number of members, we tossed a coin to decide the assignment of the unpaired member.) Matching teachers in pairs helps to improve our precision, and the coin toss assures that our estimate of the effect is unbiased. These assignments were recorded on information sheets, along with relevant teacher characteristics such as years of teaching experience.

Teachers volunteered to participate. As a sample of the district teachers, volunteers may differ from the other eligible teachers who did not volunteer in being more motivated or in some way having a greater ability to take advantage of the features of a product such as *Measuring Up*. This condition may restrict the generalizability of the findings because, if *MU* were adopted district wide, the same advantage may not be evident among teachers who were not inclined to volunteer.

Data Collection

Test Scores

The outcome measure for our study was the student scores on New York's English Language Arts assessment, the Test of New York State Standards (TONYSS). Pretest scores and demographic information came from the district's administration of the test at the beginning of the year. The data returned from the publisher (Riverside) reported scores for individual students and provided a breakdown of students by schools and classrooms. It also provided some of the basic demographic

data used in the study. Additional demographics (ethnicity and free/reduced-price lunch status) were supplied as a paper report by the district.

Surveys and Interviews

Researchers tracked the use of the *MU* materials in pilot classrooms and the alternative products in control classrooms, as well as potential contamination, through periodic web-based surveys and telephone interviews of teachers. Telephone interviews have provided a barometer of teacher morale, their level of cooperation, and a gauge of product acceptance. Three monthly surveys were conducted of all the teachers. The survey data measured how much the teachers used the materials in terms of progress through the topics and time spent.

Statistical Analysis

As noted, our primary outcome measure was the New York State ELA test for fourth graders. The basic question for the statistical analysis was whether students in classrooms using *Measuring Up* had higher ELA scores than those in control classrooms. Use of multilevel models in the analysis helps to account for the clustering of students in classes, providing a more accurate, and often more conservative, assessment of the confidence we should have in the findings. An analysis of covariance (ANCOVA) was performed to increase the precision of the estimated treatment impact. The ANCOVA involves including covariates such a student's pretest score in the analysis. Recognizing that factors other than whether or not a teacher was piloting *MU* influenced the results, we developed a more complex statistical model to determine whether the estimated treatment impact changes, depending on the levels of the covariates. Although we inspected multiple models involving combinations of the variables of initial interest, we report here only the model that we believe provides the best and most parsimonious account of the results. We use SAS PROC MIXED (SAS Institute, Inc.) as the primary tool for this work.

Results

Formation of the Experimental Groups

The randomizing process ensures that our estimates from the experiment are unbiased, but does not guarantee that the pilot and control groups will be perfectly matched on all characteristics. It is important to inspect the two groups to determine whether any significant differences occurred that would have to be controlled statistically. The following tables address the nature of the groups.

Table 1 shows the distribution of students between control and pilot conditions and the distribution of classrooms in schools. The classrooms are well distributed between control and pilot groups with nine control classrooms and ten pilot classrooms.

Table 1: Distribution of Schools, Teachers and Students

Condition	Schools	Teachers	Students
Measuring Up	6	10	188
Control	7	9	169

We also compared the control and pilot students on variables that may be relevant to the analysis, as shown in the following tables.

Table 2: Distribution of Pretest Scores between Pilot and Control

Descriptive statistics: ELA-pretest outcomes	Raw Group Means	Standard Deviation	Number of Students	Standard Error	Effect size
MU	48.612	16.247	188	1.185	-0.071
Control	52.178	16.360	169	1.259	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*MU* – control)	3.566		355	2.06	0.0398

Table 2 compares the groups on the pretest score and indicates that the control group was slightly more proficient to start with. As measured by the effect size, this difference is very small.

Table 3: Distribution of SES Categories between Pilot and Control

Condition	In Lunch Program		Totals
	No	Yes	
MU	88	85	173
Control	78	85	163
Totals	166	170	336
Chi-square statistics	**DF**	**value**	***p* value**
	1	0.3051	0.5807

Table 3 shows that the experimental groups were very well matched in terms of socio-economic status (as indicated by participation in the free/reduced-price lunch program).

Attrition

Of the 26 teachers who attended the initial meeting, five were not considered qualified. Three were reading specialists who served students from several classes, one was a special education teacher who served four fourth-grade students in a non-mainstream program, and one was a teacher who team-taught with another teacher, so did not have a separate class. Of the remaining 21 teachers, pretest data for two classes were lost in transport. Both teachers had been assigned to the control group. We continued the analysis on the assumption that the loss of data was unrelated to the experimental assignment. No other teacher attrition occurred.

We began with 365 students among the 19 participating classes, and 324 students took the posttest, for a loss of 41 students or about 11%. A Chi-square test did not indicate differential attrition between *MU* and control students.

Program Implementation

Periodic surveys of the teachers in both pilot and control groups provided us with additional quantitative results as well as answers of a more qualitative nature. In addition, all of the teachers participating in the experiment were interviewed by telephone at the end of November 2004. The quantitative findings from these sources are reported here.

Teachers' self-reports on the number of hours spent (calculated over several surveys) indicate that both pilot and control teachers spent approximately equal amounts of time preparing their students for the test (139 hours and 138 hours respectively). *MU* teachers reported that about 37% of that time involved using the *MU* materials. Both *MU* and control teachers counted the entire literacy program to be preparation for the test.

Another quantifiable measure was the level of engagement of the students. Using a scale where 1 is "significantly not engaged" and 5 is "significantly engaged," teachers answered the following question: "Compared to the test preparation activities you and your students have used in the past, how engaged were your students with the activities?" *MU* teachers reported an average level of 3.5, whereas control teachers reported a slightly higher level of engagement—an average level of 4.

Statistical Models for the Outcome Measure

Our outcome measure was the score on the state test. Our primary covariate for use in the ANCOVA was their score on the TONYSS, a test designed to parallel the NY State test. We were also interested in whether *MU* might be more effective depending on incoming achievement level, so we included in our initial models an interaction between condition and the student's level of ELA achievement. In addition to looking at incoming ability, we were interested in whether the treatment impact depended on levels of other covariates that were identified ahead of time as being important, including SES. Because the district was fairly homogeneous with respect to ethnicity and English language development, those were not used in the models. The model we report here did not include those other factors beyond pretest scores, since they did not contribute significantly to the explanation of the outcomes.

Results for ELA Outcome Controlling for Pretest Score

Table 4 displays both the descriptive statistics, including the raw means for the two conditions and the analysis of these results, using the statistical model that includes the pretest. The bottom segment of the table presents technical information on how the model accounts for clustering of students into classes. Of interest here is the line for condition. Being in a classroom using *MU* provides an advantage for the average student of about 5.46 points on the state test. The *p* value of .20 indicates that there is a 20% chance of getting a difference with an absolute value this large or larger just by chance. This is a small difference, given the scale used in the state test, which has a 345-point range (between 455 and 800). A metric often used to evaluate such differences is to express the difference as a portion of a standard deviation, or what is called an effect size. In this case the effect size is 0.149 of a standard deviation, which, for educational interventions, is often found to be educationally meaningful. (This calculation is based on the unbiased estimated mean difference adjusted for prior score.).

Table 4: Multi-level mixed model for ELA—results controlling for pretest

Descriptive statistics: ELA reading outcomes	Raw Group Means	Standard Deviation	Number of Students	Number of Classes	
MU	669.415	35.212	164	10	
Control	670.731	38.253	160	9	
Mixed model: Fixed factors related to CST reading outcomes	**Estimate of Coefficient**	**Standard Error**	**DF**	***t* value**	***p* value**
Intercept	667.18	6.335	6	105.32	<.0001
Pretest score (centered at the mean)	1.229	0.083	297	14.89	<.0001
Condition (MU = 1; control = 0)	5.460	4.260	297	1.28	0.2009
Mixed model: Fixed factors related to CST reading outcomes	**Estimate of Variance Coefficient**	**Standard Error**		***z* value**	***p* value**
Class mean achievement	49.215	35.698		1.45	0.412
Within class variation	461.08	37.908		12.16	<.0001

We display this finding graphically in Figure 1, which shows the difference between *MU* and control for the average student as a bar graph. The bar graphs show our best estimate of the effect. As with any statistical estimate, there is a level of uncertainty, depicted by the markers at the top of each bar. These markers represent what is called a confidence interval and, in this case, they are set to 80%. (That is, we are 80% sure that the true heights of the bars lie within their respective intervals, and there is a one-in-five chance that the height of one or both bars is actually outside this range.)

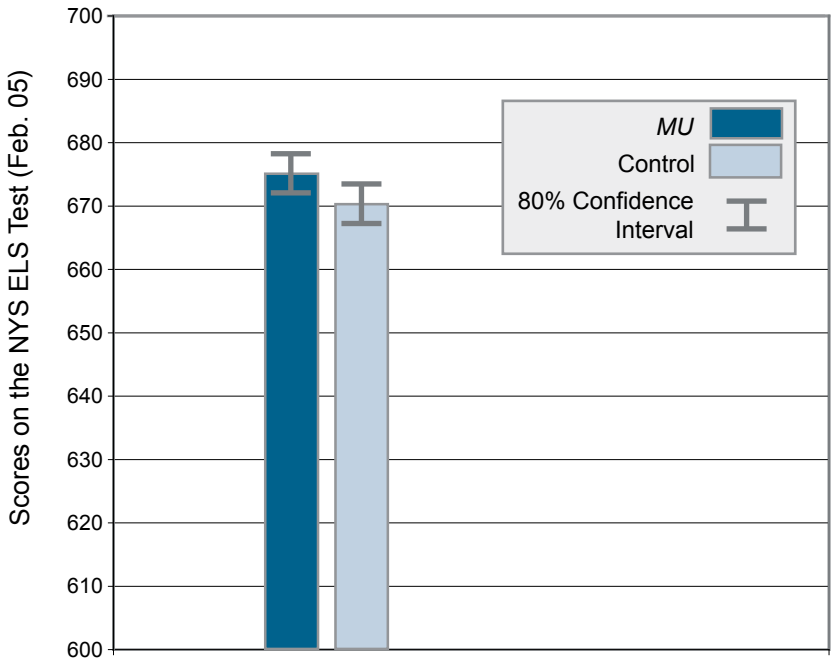


Figure 1: Bar graphs representing the means for *MU* and control groups adjusted for the pretest score

Results in Terms of Change in Proficiency Level

Results in Terms of Change in Proficiency Level

Another way to compare the changes in *MU* and control groups is to look at the proficiency levels established for the test by the state of New York. There are four proficiency levels: at the bottom level, students are considered to have serious academic deficiencies, whereas at the top level, they are considered to exceed the standards. The average student in our sample fell into level 3, "meets standards." The publisher provided the equivalent cut points for each of these levels for the TONYSS. For each student, we coded whether he or she changed levels—that is, moved down a proficiency level (or failed to progress to a higher level) or moved up one or more levels—from the pretest to the posttest.

Table 5: Chi-square table of the distribution of students moving up a proficiency category between *MU* and control classes

Condition	Stayed Same or Moved Down	Moved Up One or More Levels	Totals
Measuring Up	46	118	164
Control	61	99	160
Totals	107	217	324
Chi-square statistics	**DF**	**value**	***p* value**
	1	3.718	0.054

Table 5 compares *MU* and control groups with respect to the number of students who did not progress versus the number who did progress one or more proficiency levels. Proportionally more *MU* students made progress. The statistical test indicates that this difference is unlikely to have occurred by chance. Figure 2 shows this comparison in terms of the percentage of students in the "no progress" and "moved up" categories. The likelihood of making progress was greater for students in the classrooms using *MU*.

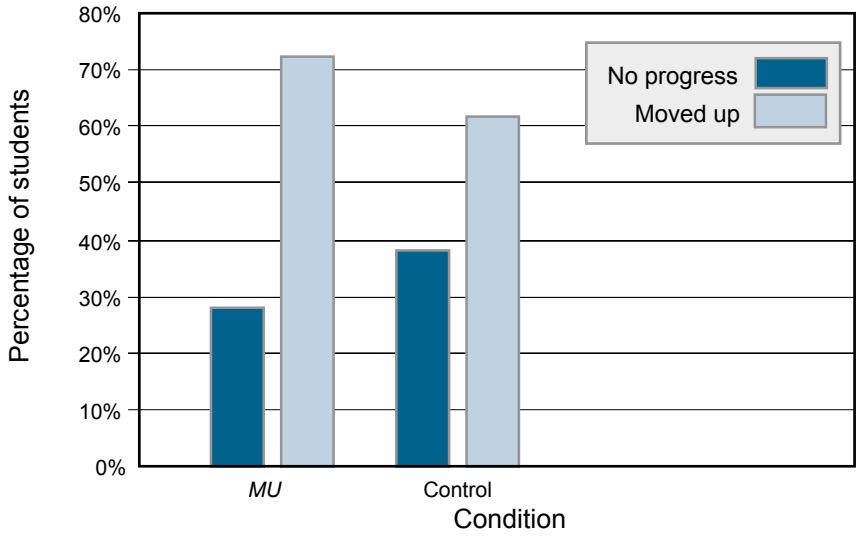


Figure 2: Comparison of *MU* and control on percentages of students moving up a proficiency level

Discussion

Under the conditions found in this school district, *Measuring Up* performed better than the other approaches used for test preparation. Control group teachers also helped their students prepare for the test, using teacher-developed materials, enrichment activities, some supplementary products, and portions of their basal text. The variety of materials they used makes it difficult to point to specific contrasts between *MU* and the control group that account for *MU*'s advantage. Nevertheless, the difference is based on a rigorous research methodology—random assignment—which assures that participating teachers were not biased in favor of or against the new product or in other ways.

We must be cautious in generalizing our findings to other districts that may have different populations, resources, or instructional methods in place. It is important to point out also that the conclusions apply to a particular subject matter (English Language Arts), at one grade level (fourth), and in the context of New York's standards and assessments. As a basis for a decision about broader implementation of *MU* in this district, generalization may be limited by the fact that the teachers were volunteers. The same impact may not be found for teachers who were not willing to try out a new product. Still, the measures of implementation showing both groups devoting similar amount of time to prepare for the test and showing that the control teachers perceived their students as slightly more engaged, counter the alternative explanation that the positive impact for *Measuring Up* was a result of motivational differences between teachers in the two groups.

Further research involving additional classrooms, schools, and districts could examine related process questions such as the independent impact of school-level, classroom-level, and individual-level effects. As the technically trained reader may have noticed, the random effects estimates from the analysis show that there is significant variation from class to class in the outcome that remains to be accounted for. Further work may involve the use of more elaborate multilevel models that include higher level covariates, including average school-wide SES, to account for the unexplained variability.

The current study allows us to draw conclusions with a reasonable degree of confidence. The statistical test comparing the means of *MU* and control groups was stringent. However, our conclusion is based on a *p* value corresponding to a one-in-five chance that the difference was a result of chance factors. The comparison based on the number of students moving up in proficiency levels corroborates the conclusion that students benefit from being taught with *MU*. As a basis for a local decision in this district, the evidence can be considered sufficiently favorable to move forward with a broader implementation of the product. In more general terms, we see evidence that *Measuring Up* may be a good option for supplementing English Language Arts instruction where the goal is higher achievement on the New York State test.