Understanding Types of Evidence: A Guide for Educators

October 24, 2016

This guide describes four key types of evidence educators are likely to encounter and explains how to tell whether these types of evidence can provide strong support for claims about an educational technology's effectiveness. This document is meant to be a resource for districts seeking to evaluate the educational technologies being used in their school(s), though the lessons here can be adapted for other contexts or applied to other educational interventions. Understanding how to assess the quality of available evidence is an important step in making the best possible decisions regarding which educational technology to use to achieve the outcomes you want to see.

Gregory Chojnacki
Alexandra Resch
Alma Vigil
Ignacio Martinez
Steve Bates

Submitted by:

Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Alex Resch
Reference Number: 50183.01.502.485.000

MATHEMATICA
CENTER FOR
IMPROVING RESEARCH EVIDENCE

UNDERSTANDING TYPES OF EVIDENCE

When making decisions about which technologies to use, you need evidence about which options are effective to make the best possible use of your technology budget. Many information sources, from marketing material to peer reviewed studies published in prestigious journals, present evidence of product effectiveness. The quality of this evidence can vary widely. This guide describes four key types of evidence you are likely to encounter and explains how to tell whether these types of evidence can provide strong support for claims about effectiveness. The types of evidence in this guide are ordered from weakest to strongest, and each evidence description is accompanied by examples of information sources containing that type of evidence.

Anecdotal: Impressions from User Experience

Anecdotal evidence consists of personal descriptions or claims based on one (or more) person's own experience. This may include claims about a technology's effectiveness or other features that are not necessarily related to effectiveness, such as user experience. This type of evidence cannot provide strong support for claims about the effectiveness of a technology because it is based on subjective impressions. However, anecdotal evidence may provide an indication about the context in which a technology might be expected to be effective, or aspects of the user's experience that may enhance or reduce the technology's effectiveness. In general, anecdotal evidence can help identify products that are promising enough to warrant more rigorous research.

Common Source of this Evidence Type (follow link for example): marketing testimonials

Descriptive: Measures of Outcomes over Time

Descriptive evidence summarizes characteristics of program participants and their outcomes over a period of time. This type of evidence is commonly found in marketing materials and news articles. Since descriptive evidence does not include a comparison group, it is impossible to know what would have happened without the program over the same time period. Therefore, descriptive evidence alone cannot provide strong support for claims about a program's (or product's) effect on the outcome of interest.

For example, an infographic may claim that an educational technology "gets results" because student achievement is higher after using the technology than before. But several other factors, such as traditional teaching or the introduction of a new curriculum, might be driving improvements in achievement. This descriptive evidence does not provide evidence about the technology's true effectiveness, since we don't know what would have happened in these schools if they had not used the technology.

Common Sources of this Evidence Type: marketing materials, news articles

Correlational: Comparisons of Users and Non-Users

Correlational evidence can identify the relationship between an educational condition or initiative, such as using an educational technology, and a specific outcome, such as student math test scores. This type of evidence can be useful as a starting point when learning about a

technology, but cannot conclusively demonstrate that a technology gets results. This is because it cannot rule out other possible explanations for the differences in outcomes between technology users and non-users. Correlational evidence is often misinterpreted and used to demonstrate success.

For example, a correlational analysis might compare a small group that used a technology versus students in the school district as a whole. Even if students who used the technology had higher year-end test scores, on average, than those who did not, there may be other important differences between technology users and the rest of the district that explain differences in improvement. Often, schools or students chosen to pilot a technology are a special group; for example, they may be highly-motivated students who volunteered to participate in a new program, or they may be low-achieving students who have been selected to receive several additional supports.

Common Sources of this Evidence Type: blog posts or news articles

Less Common Source: grey literature

Causal: How to Accurately Measure Effectiveness

Causal analysis is the only way to determine effectiveness with confidence. This type of analysis compares "apples to apples" by ensuring the only difference between the group that received the program and a comparison group is the program itself. An otherwise identical comparison group tells us what would have happened without the program; we can then say that differences in outcomes between the groups were caused by the program. There are several ways to create the comparison group needed to generate causal evidence, but a strong causal analysis must show that the group receiving the technology and the comparison group are equivalent in characteristics such as previous test scores and demographic characteristics. This equivalence is what convinces the reader that we are comparing apples to apples.

For example, strong causal evidence of a technology program's effect on student achievement will examine differences in characteristics and test scores between students in the technology program and comparison groups before the intervention took place. This way, the reader can see whether the two groups are the same before the students began using the technology. If they are equivalent, differences in outcome scores between treatment and comparison students can be attributed to the technology. While a randomized controlled trial is often considered the "gold standard" in causal analysis, other methods can also be used to identify or create a comparison group.

Common Sources of this Evidence Type: independent evaluations

Less Common Source: news articles

Example of Anecdotal Evidence: Marketing Testimonials

These testimonials make different types of claims about DreamBox Learning® products based on anecdotal evidence.

"I was a huge supporter of bringing DreamBox to Stubbs Elementary after seeing a huge success with it while I was assistant principal at Oberle last year. We saw more than a 15% increase in our math scores in one year and the only thing we did differently was use DreamBox. Based on Stubbs' State assessment data, closing the achievement gap in math is a priority. I am excited to see the impact it will have on our students here."

-- Elementary School Assistant Principal

This testimonial indicates that the program raised test scores. A rigorous evaluation would be needed to make a strong conclusion about this. The assistant principal may not remember or recognize other changes that may have affected her students' achievement; these could include changes in the student body, teacher experience, or other recent reforms.

"My students love using Dreambox. They use it about 20 minutes a day. On average, my first grade class is working at a middle of second grade level."

-- First grade teacher

This statement indicates that the program is popular with students in this teacher's class. This anecdote might stimulate the reader's curiosity about the ideal amount of use per day, which could be assessed rigorously in a pilot. It is not clear how students' grade level of work is measured or where they started at the beginning of the year.

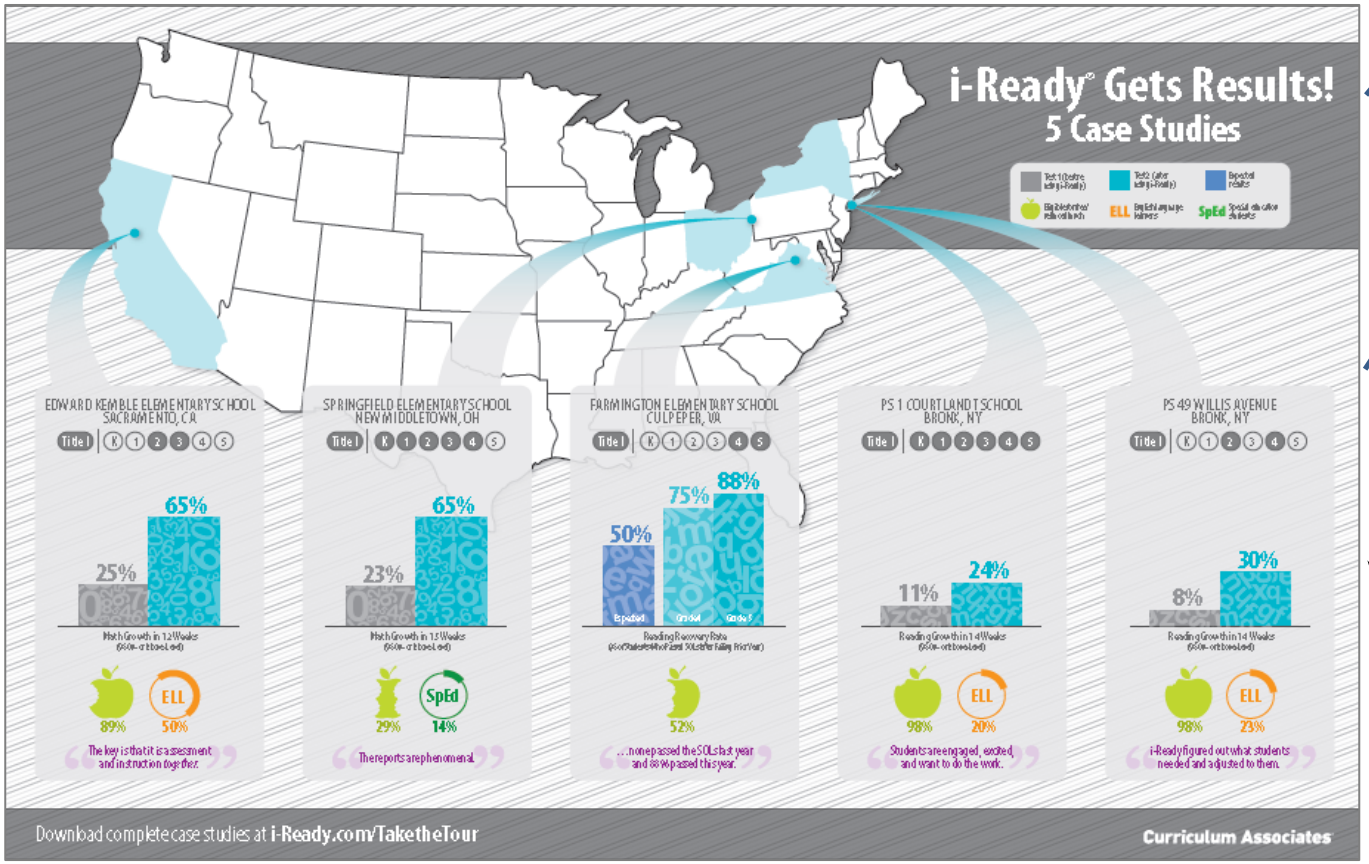
"The Common Core Report is my favorite. It helps me to see exactly what areas the students are working on and passing. I can also see where they are having difficulty and spending more time. I use this data for small group time where I can focus on the specific areas that each student needs help."

-- Second grade teacher

This observation highlights one possible way the program could be used – to diagnose areas of difficulty in order to plan individualized instruction.
A rigorous rapid-cycle evaluation could evaluate whether students of teachers who pair program use with daily small-group instruction outperform students of teachers who also use daily small-group instruction but without tools developed by DreamBox Learning®.

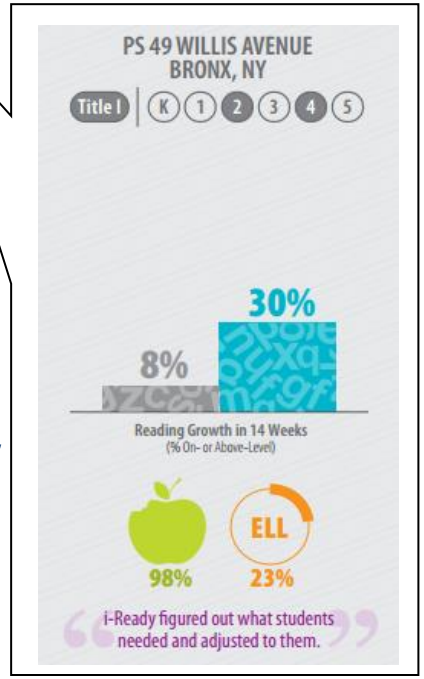
Testimonials drawn from the DreamBox Learning® website.

Example of Descriptive Evidence: Marketing Infographic



This infographic makes a claim about i-Ready's effectiveness. What type of evidence is presented?

As shown in the pop-out below, each case study compares student achievement before and after using i-Ready.



Because the case studies do not include a similar comparison group, they are not able to provide information on what would have happened to student achievement without i-Ready.

Factors other than the use of i-Ready may have caused the changes in student achievement presented in these case studies. Therefore, the case studies do not provide strong evidence of i-Ready's effectiveness.

Because they do not include a comparison group, these are descriptive analyses rather than correlational or causal analyses.

Infographic drawn from the i-Ready

Example of Descriptive Evidence: News Article

EDUCATION WEEK

Published Online: July 25, 2007
Published in Print: July 17, 2007, as **Software Improves Reading**

Software Improves Reading

By **Kathleen Kennedy Manzo**

This article includes some evidence on the effectiveness of a literacy software program– but what type of evidence?

Like Ms. Lebron, school leaders in the 55,000-student Paterson district, and their counterparts across the nation, are learning the benefits of incorporating computer-based features into the reading curriculum to help teachers address their students' varying skills and experience.

Teachers describe the perceived learning benefits of software, such as the components included in the READ 180 program.

During the 90-minute English/ language arts block at Eastside High, for instance, each of the 15 students in the remedial class gets a chance at using a computer to strengthen basic skills, including decoding, reading fluency, vocabulary, and comprehension. An audio feature allows students to record themselves reading or listen to a taped version of the text. The activities bolster group lessons on grammar, writing conventions, and literature, and equip students for tackling grade-level reading assignments independently, educators here say.

One teacher describes large learning gains among students who used the program, with nearly all students advancing two or more grade levels in reading.

In the course of the school year, Ms. Valenz said, nearly all the students advanced two grade levels or more in reading, and most had mastered 9th grade work, skills that have carried over to their other schoolwork.

Is this description strong evidence of the software's effectiveness?

No. Other factors may have caused the gains. The article cites changes in reading level over time but the changes could be due to many factors besides the program. The lack of a comparison group that did not receive the software program prevents us from knowing what would have happened without it.

This article presents descriptive, rather than correlational or causal, evidence.

Excerpted from a news article on the Education Week

Example of Correlational Evidence: Blog Post

EDUCATION WEEK

Study: Struggling Math Students Gain Using Personalized, Blended Program

By Michelle Davis on December 4, 2014 10:29 PM

This blog post includes some evidence on the effectiveness of "School of One" – but what type of evidence?

Middle school students participating in a personalized, blended-learning math program showed increased gains in math skills—up to nearly 50 percent higher in some cases—over the national average, according to a new study from Teachers College, Columbia University.

The post cites a study that compares students who use School of One to national average test scores on the Measures of Academic Progress (MAP) test.

During the 2012-13 school year, students using Teach to One: Math gained math skills at a rate about 15 percent higher than the national average. In the second year of the program's implementation students made gains of about 47 percent above national norms, even though some of those students were still in their first year of using Teach to One: Math.

Students using the program showed substantially higher gains than the average student nationally.

Is this conclusive evidence of the technology's effectiveness? No. Other factors may have caused some of the gains. Since the comparison is <u>not</u> between groups constructed to be very similar, this is a <u>correlational</u>, rather than <u>causal</u>, analysis.

The follow-up study mentioned below is a more rigorous quasi-experimental study designed to provide a stronger answer about the program's effect on learning.

Ready cautioned that the data in the study did not allow him to conclude definitively that Teach to One: Math caused the skills improvements. However, New Classrooms Innovation Partners plans a more definitive trial over the coming two years in the Elizabeth, N.J. public schools, Rush said. New Classrooms, in partnership with the Elizabeth district, received a $3 million federal Investing in Innovation Fund grant to do that work.

Excerpted from a blog post on the Education Week website.

Example of Correlational Evidence: Grey Literature

Year Two Results: Evaluation of the Implementation and Effectiveness of SuccessMaker During 2002-2003

Charleston County School District

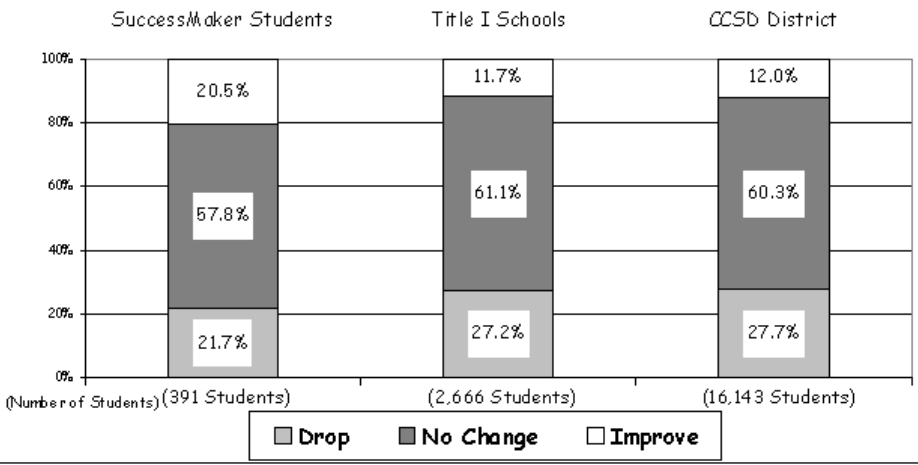
This correlational study, conducted by a school district, presents information on the computer-based instructional program "SuccessMaker." Is this information strong evidence of effectiveness?

performance category. For both ELA and Math, Below Basic students who worked on SuccessMaker were more likely than Below Basic students in the comparison groups to improve their PACT performance category (in ELA, 36% of the Below Basic SuccessMaker

The study reports that students who used the program were more likely to improve on the state test (the PACT).

The study compares SuccessMaker users to students at other Title I schools and to the district as a whole.

Figure 4: Changes from 2002 to 2003 in PACT English/Language Arts Performance Categories for SuccessMaker, Other Title I Schools, and CCSD Students in Grades 3-8



However, the study does not include enough information on whether program users and their schools were similar to non-users in the comparison groups. Although both groups were formed from "Below Basic" students, differences in other characteristics may exist. Additionally, it is not clear whether the background information provided in the text applies to the sample in Figure 4.

Differences in improvement may be due to SuccessMaker or other factors. This study does not provide strong evidence of effectiveness.

Excerpted from a report on the Charleston County School

SRI International

This report provides experimental evidence on the impact of the DreamBox Learning® Math program on kindergarteners' and first-graders' math achievement.

Evaluation of Rocketship Education's Use of DreamBox Learning's Online Mathematics Program

Haiwen Wang
Katrina Woodworth

This was an independent evaluation using a Randomized Controlled Trial (RCT) design. RCTs are the gold-standard for establishing causal effects.
This means they can provide strong evidence on a program's effectiveness

Exhibit 4 presents the means and standard deviations of the pre- and posttest scores (NWEA mathematics test scores in September 2010 and in January/February 2011) for the treatment and control students. The differences in pretest scores were in general less than 3 points, all within .2 standard deviations of the scores for the entire sample, and none of the differences were statistically significant at a .05 significance level, meeting the What Works Clearinghouse (WWC) standards for a balanced sample.

Exhibit 4
Pre and Post NWEA Math Test Scores by Treatment and Control Condition

	Treatment					Control				
	Pretest			Posttest		Pretest			Posttest	
	N	Mean	*SD*	Mean	*SD*	*N*	Mean	*SD*	Mean	*SD*
Math overall	446	146.0	18.0	159.0	16.6	111	144.7	15.0	156.2	15.1
Problem solving	444	147.0	19.3	161.4	16.3	109	144.7	17.1	159.8	15.2
Number sense	444	146.9	20.0	159.6	18.9	109	143.4	16.6	157.0	17.2
Computation	438	147.5	22.4	163.0	20.7	108	147.0	19.8	158.8	19.5
Measurement and geometry	441	144.5	18.9	155.5	18.3	109	144.8	18.4	151.8	18.1
Statistics and probability	443	145.5	19.3	156.3	18.9	109	145.1	15.6	154.1	17.6

The strength of the evidence on DreamBox's impact relies on the fact that students in the study who used the program were very similar to those who did not – in other words, the sample was "balanced" across the user and comparison groups. The paragraph and table to the left show that this study met widely accepted standards for balance. In particular, they found that students had similar scores on a baseline version of the test they used to measure outcomes – this is generally considered the most important aspect of balance.

Exhibit 7
Summary of Regression Results for the ITT Effects on NWEA Mathematics Scores

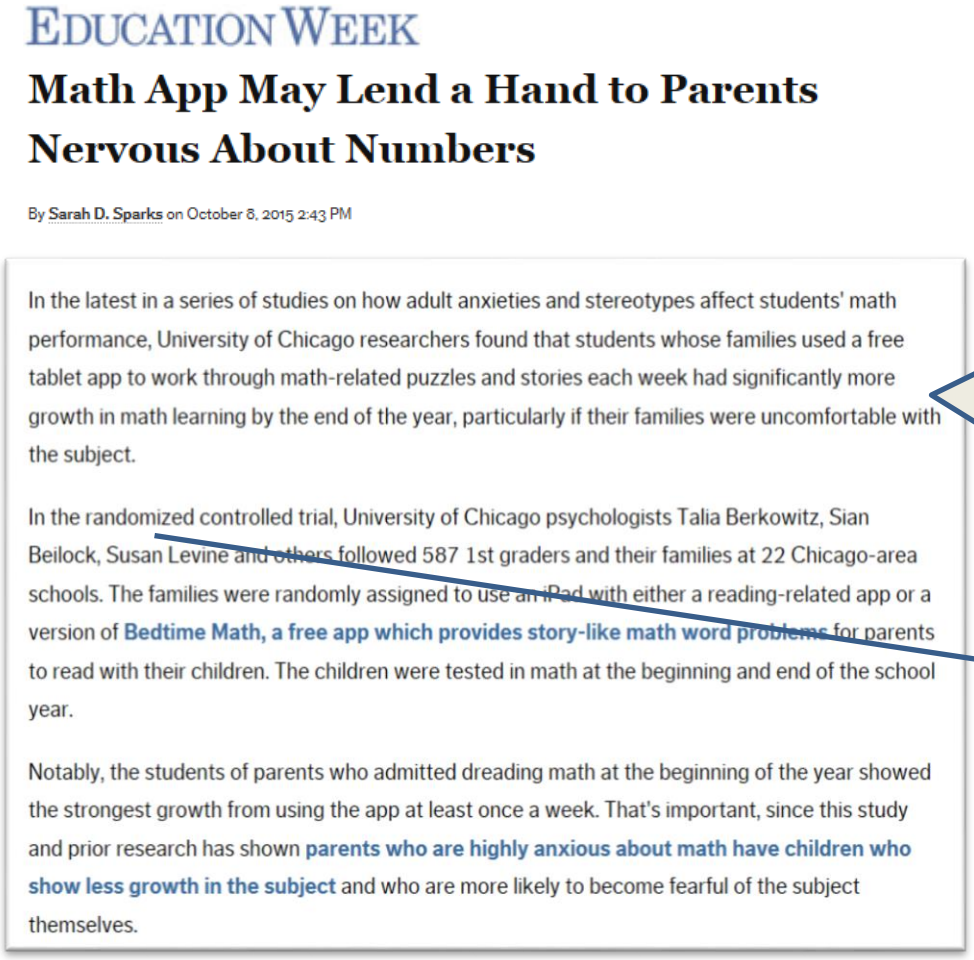
	Math Overall	Problem Solving	Number Sense	Computation	Measurement and Geometry	Statistics and Probability
Effect on RIT scale score	2.30**	1.02	1.53	2.68	2.91*	2.20
SE	(0.83)	(1.11)	(1.23)	(1.41)	(1.23)	(1.36)
Effect size	0.14	0.06	0.08	0.13	0.16	0.12

*$p < .05$

As shown in the first row of Exhibit 7, DreamBox Learning® Math had a positive and statistically significant impact on tests of overall math skills and of measurement and geometry. The statistically significant impacts, marked with an asterisk, indicate that it is very unlikely that those differences in outcomes are due to chance.

Excerpted from a report available on the DreamBox Learning®

Example of Causal Evidence: Blog Post



EDUCATION WEEK

Math App May Lend a Hand to Parents Nervous About Numbers

By Sarah D. Sparks on October 8, 2015 2:43 PM

In the latest in a series of studies on how adult anxieties and stereotypes affect students' math performance, University of Chicago researchers found that students whose families used a free tablet app to work through math-related puzzles and stories each week had significantly more growth in math learning by the end of the year, particularly if their families were uncomfortable with the subject.

In the randomized controlled trial, University of Chicago psychologists Talia Berkowitz, Sian Beilock, Susan Levine and others followed 587 1st graders and their families at 22 Chicago-area schools. The families were randomly assigned to use an iPad with either a reading-related app or a version of **Bedtime Math, a free app which provides story-like math word problems** for parents to read with their children. The children were tested in math at the beginning and end of the school year.

Notably, the students of parents who admitted dreading math at the beginning of the year showed the strongest growth from using the app at least once a week. That's important, since this study and prior research has shown **parents who are highly anxious about math have children who show less growth in the subject** and who are more likely to become fearful of the subject themselves.

This blog post presents information on the effectiveness of a technology called "Bedtime Math" – what type of evidence is presented?

"Students whose families used a free tablet technology to work through math-related puzzles and stories each week had significantly more growth in math learning by the end of the year."

The article reports results from a randomized controlled trial – the gold standard in causal analysis. Students who used the technology were randomly selected, so the group of students who were not selected should be very similar to the group who was. Because we would expect these groups to be equivalent prior to the trial, any difference in outcomes can be considered the effect of the technology.

Therefore, this article – and the study it reports on – present strong evidence on the effectiveness of this technology among these Chicago-area students.

Excerpted from an article on the EdWeek

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ TUCSON, AZ ■ WASHINGTON, DC

MATHEMATICA
Policy Research