Empirical Education

RESEARCH REPORT

Toward School Districts Conducting Their Own Rigorous Program Evaluations

Final Report on the "Low Cost Experiments to Support Local School District Decisions" Project

Denis Newman
Empirical Education Inc.

September 30, 2008

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 210
Palo Alto, CA 94306
(650) 328-1734

Acknowledgements

This report represents a long term collaborative effort among researchers and educators. Those contributing ideas, evidence, solutions, and writing to this report include Andrew P. Jaciw, Gloria I. Miller, Kay Thomas, Jessica V. Cabalo, Xin Wei, Boya Ma, Minh Vu, Jane L. David, David Greene, and the many educators from district and state agencies who worked with us. We are grateful to all the people in school districts and state agencies who were willing to discuss and try out ideas about how to make research work in their settings. We are particularly grateful to Will Shadish, David Rindskopf, and Betsy Becker for the guidance and ideas in the early phases. The research was funded by a grant (#R305E040031) to Empirical Education Inc. from the U.S. Department of Education, which is not responsible for the content. The principal investigator, Denis Newman, takes responsibility for the errors that may remain in this report.

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.

Table of Contents

Introduction

Four years ago when Empirical Education began the investigations reported here, the convergence of the No Child Left Behind accountability, especially the scientifically based research provisions, and the rapid expansion of data warehouses for school districts led us to the following conjecture:

The combination of readily available student data and the greater pressure on school systems to improve productivity through the use of scientific evidence of program effectiveness could lead to a reduction in the cost of rigorous program evaluations and to a rapid increase in the number of such studies conducted internally by school districts.

When we began this work, the conjecture that school systems could conduct their own local program evaluations using rigorous methodologies was to a large extent, a new idea. The prevailing view of scientifically based research was that educators would be consumers of research conducted by professionals, particularly research that aimed at broad generalizations. There was also a prevailing view that rigorous research was extraordinarily expensive. This would mean that only the federal government or a few private foundations could afford to sponsor such research and these agencies would understandably not risk funds on researchers without a strong track record of publication of rigorous research; nor would they spend funds on program evaluations that were not meant to generalize beyond the local jurisdiction. In this context we proposed a research and development project called "*Low Cost Experiments to Support Local School District Decisions.*" The supposition behind our proposal was that the cost could be made low enough to allow experiments to be conducted routinely to support local decisions.

We were partially right. The technology for gathering and integrating electronic data has steadily improved and we have found that the costs of conducting rigorous research need not be as high as commonly assumed. But an increase in the use of local program evaluation for decision makers has not materialized for many reasons that we explore in this report.

As a summary of our work, this report is intended to answer three sets of questions about the project:

1. What were the successes of the project and its impact? What were the unanticipated outcomes or benefits from the project and what unanticipated barriers did we encounter?
2. Especially given the barriers encountered, how did our original ideas change as a result of this work? What would we now recommend as advice to educators interested in local experiments?
3. What approaches can now be proposed for continuing to build capacity to use experiments to support local decisions?

Findings

Our program of research and development, funded as a "goal two" or development program, resulted in a number of important insights about conducting local research. Our focus was exclusively on randomized control trials (RCT) or randomized experiments. While challenging in many ways, we believed it was the appropriate place to start. An RCT requires explicit planning for the evaluation because teachers or schools must be identified before any training or implementation begins. It can't be used for after-the-fact evaluations. Thus the method is more likely to come into play while decisions about whether the new program to be evaluated should be purchased on a larger scale are still on the table. It was a basic premise of the project that the purpose of conducting program evaluations using an RCT was to inform a local decision in the district central office. As we detail in this report, our first insight was that the use of research evidence in districts is not straightforward.

The project also contributed a number of methodological, analytic, and reporting approaches with potential to lower cost and make rigorous program evaluation more accessible to school district researchers. An important result of the work was our bringing to light the differences between research design aimed at answering a local question, where sampling is restricted to the relevant "unit of decision-making," and research design attempting to come to generalized conclusions about an

intervention. This work includes the development of a paired randomization approach in which units (classes, teachers, and schools were all used) are first paired based on salient similarities and then assigned to treatment and control conditions. The results included both a methodology in terms of a teacher workshop and a statistical analysis showing the technical benefit in terms of requiring a smaller sample as measured across a number of experiments. We also made headway in forms of reporting, attempting to render complex results more accessible. Interaction effects are an interesting case. Educational interventions do not work equally well for all students or all teachers and, in many cases, where there is only a small overall impact, that impact will be distinct for some subgroups, for example, students initially scoring low or teachers not certified in their subject. We made progress in constructing tables and graphs for displaying these effects. In doing so we came to understand that we had three target audiences: 1) educators with little training or interest in research methodologies but who are in decision-making positions, such as senior central office administrators, 2) district research managers with some, but often rusty or outdated, understanding of research methodology and statistical analysis, and 3) technical experts who may be called upon to review the evidence. A report constructed for a local audience has to address at least the first two audiences.

Challenges

The most intractable challenge we confronted during the project was not the efficient conduct of local research or the technical side of the communication; it was in the area of what we can call the “theory of action” for using rigorous evidence. The problem is not exclusive to the use of locally generated evidence, but arises from more general characteristics of decision-making approaches and priorities in school systems. Rigorous evidence, either reported in scientific journals or gathered from local experiments can be only one element of a larger set of considerations. The problem is not just that other considerations have higher prioritization, for example, teacher acceptance of a program drowns out the evidence. It is also that reports of evidence or the decision to conduct a rigorous study can be put to other purposes, for example, selectively reporting results to support to a predetermined decision. To take its place as one of several important factors in informing a decision, the school system’s decision process must have a slot into which the evidence is inserted and from which it can have an influence before the decision is made.

Our project did not discourage us from pursuing the idea that school systems can conduct useful local research cost effectively or dissuade us from pursuing the systemic changes that are called for in getting school system leaders to ask for rigorous evidence to inform their decisions. As a result of this project, we can report technical progress in improving the rigorous methodologies for low cost local experiments. We can also help to provide a stronger policy framework for considering how the evidence from such experiments can be used productively. Finally, we can recommend next steps in a program of research and development that can increase the use of evidence from rigorous studies in school systems.

The Narrow Focus of the Current Project

Our primary focus was on increasing the efficiency and lowering the cost of randomized experiments conducted by and for school systems. This focus was narrow in two ways.

First, we limited our investigation to randomized experiments. While this is the gold standard in terms of eliminating selection bias, it is not by any means the only methodology available for local program evaluations. Quasi-experimental designs, including interrupted time series, can be useful in local evaluations and have the advantage of not requiring the evaluation to be planned prior to selecting the participants in the new program.

The second way in which our focus was narrow was that the social or organizational context of research use was not initially an explicit topic of our research and development. As we progressed and especially as we worked with our outside formative evaluators, we came to understand the fundamental importance of how school system administrators understand the value and use of

research. The problem we have identified as a result of this project is the extent to which school systems are unprepared to use scientific evidence of any sort in decisions.

Our narrow focus was useful, however, in providing challenges to getting research conducted and used in school systems. The challenges are the basis for our main findings. In future work we will relax both these constraints. We now understand that the conduct and use of information from simpler, even descriptive, designs can be precursors to an advanced end point, which is the idealized evidence-based decision maker we began the work with.

Randomized Experiments Conducted as Part of the Project

Our three-year project included an ambitious task of designing, conducting, and, given time within the project, reporting back on randomized experiments conducted in school districts. We were able to initiate eight randomized experiments during the course of the funded project. Four were completed within the timeline of the project and, for three of those, our reporting back to the district was captured within the project. Two experiments were started but not completed. The following are brief synopses of these experiments with references to the relevant full reports:

1. Middle School Social Science (2004-2005)

Anticipating an upcoming adoption of a new history/social studies textbook, a California district conducted a randomized study to determine the impact of a new textbook (History Alive!) on student achievement. We reported the results of this experiment in March 2006. This experiment is reported in Cabalo, Newman, & Jaciw (2006).

2. Elementary Science (2005-2006)

The following year, the same district, working with project staff, began planning for a second experiment to compare two elementary science textbook candidates. Because of difficulties with cooperation of the publishers and their concern with a head to head comparison, however, the experiment was not completed.

3. Computer-based Teacher Support System (2004-2006)

A randomized experiment began in a small Eastern state with the introduction of a new computer-based World Languages assessment and teacher support system. The experimental treatment was the addition of a computer-based lesson development tool. Because insufficient numbers of participants took the posttest, the experiment was repeated with a new group of teachers the following year. Difficulties in using the online assessment system continued and the final report focused more on the adequacy of the test than on the professional development and support. This work is reported in Cabalo, Ma, & Jaciw (2007).

4. Cognitive Tutor for Algebra 1 (2005-2006)

As part of the required external evaluation of their National Science Foundation grant to create a Math-Science Partnership, a school district and its associated Community College began an experiment in 2005. This was designed to determine whether a new Algebra 1 program (Cognitive Tutor) would prove more effective than traditional Algebra 1. An unexpectedly low turnout of teachers for the random assignment meeting resulted in a within-teacher design. This experiment was presented at AERA (Cabalo & Vu, 2007) and reported in Cabalo, Jaciw, & Vu (2007).

5. Professional Development for Interactive Whiteboards (2005-2006)

A southern school district, which had previously invested in Interactive Whiteboards for all teachers, chose to test additional staff development for a small sample of teachers who were matched to those without the additional training. This experiment was reported in Cabalo, Ma, Jaciw, Miller, & Vu (2007) and presented as Cabalo & Miller (2007).

6. Cognitive Tutor for Pre-algebra (2006-2007)

This experiment took place in the same school system as the one for the Algebra 1 product and used the same experimental design. It was reported in Cabalo, Ma, & Jaciw (2007) and presented by Newman & Zacamy (2008).

7. Middle School Math Tutoring (2006)

This experiment was planned in summer and fall to evaluate a volunteer tutoring program. Although sufficient teachers were recruited, we called it off after an insufficient number of volunteers signed up to help in the classrooms.

8. State-wide Math and Science Initiative (2006-2010)

The origin of this experiment was unique in that it was initially discussed between the state department of education and SERVE, which at the time was proposing to operate the Southeastern Regional Education Lab (with Empirical Education as a subcontractor). However, the decision on funding and contracting for the REL was delayed beyond the point at which randomization had to be conducted (February 2006). Since the experiment represented a significant opportunity to work with and observe a locally initiated experiment aimed at informing a state policy decision, we took this on as a one-year experiment. Subsequently the REL was funded and current operations are being funded through that contract, while our R&D project has continued to track the policy and decision-making side of the activity.

Randomized Experiments Outside the Project

In parallel to the grant-funded project, Empirical Education has also been designing, conducting, and reporting commercially sponsored randomized experiments where the district participants were not the initiators of the study but, for one reason or another, were willing to host the experiment. While grant funds did not go to any of the operations or reporting of these experiments, we were able to use the data from several of these in theoretical statistical investigations that were part of the work of the project (Jaciw & Ma, 2008; Jaciw, Wei, & Ma, 2008; Jaciw & Wei, 2007; Jaciw, Wei, Ma, & Newman, 2007). For example, the experiments reported in Miller, Jaciw, Ma, & Wei (2007) and Hoshiko, Jaciw, Ma, Miller, & Wei (2007) were both large experiments where the randomization was blocked by school district so that the data from each site could be used as if it were a small local experiment.

Decision-Making Context for School District Research

This project proposed a new model for conducting research on the effectiveness of instructional and professional development programs. The rigorous methods advocated in federal legislation could apply to research conducted locally in the school district. This approach could not only provide highly relevant local evidence of effectiveness but, in the long run, provide an abundance of rigorous data points that would be the basis for broader generalizations about particular programs. We knew that a necessary condition for local use of randomized control designs was that they had to be less expensive and easier to design, conduct, analyze, and report than had previously been the case. At the time of our initial proposal five years ago, we did not have an inventory of other necessary conditions. Our primary discovery, made largely through unsuccessful attempts to have local experimental data used in decisions, was that the organizational conditions for using experimental evidence for decisions must also be in place. By organizational conditions we refer to a large number of considerations including the leadership style of the superintendent, the communication channels among research, IT, and curriculum departments, the external pressures on the district and, in general, the understanding of how data (not just research evidence) can be useful in planning and decision-making. In this section, we document some of the observations on which we base our finding that organizational conditions are necessary for the success of evidence use in school district decisions.

While our project worked both with district- and state-level education agencies, most of our discussion will focus on decisions in the school district central office, where decisions to implement new programs typically occur. This is not always the case, but it is common for state-level programs and policies to be adopted through a district decision. Individual schools also often have a discretionary budget but typically it is used for smaller programs about which an experiment is less likely to be warranted.

Background: Education versus the Medical Model

The use of randomized experiments in schools has become an explicit policy fairly recently. The idea of doing them in K-12 schools got a large push with the passage of the No Child Left Behind Act as well as from the Education Sciences Act the following year. NCLB famously calls for using interventions that are based on "Scientifically Based Research" (now reduced to the acronym SBR). The Education Sciences Act established the Institute of Education Sciences, which has promoted and funded the use of randomized experiments in research on effectiveness of instructional methods, interventions, products, and services. Randomization in the assignment of schools, teachers, or classes to the treatment or the control condition assures that there is no bias in the results by making the two groups essentially equivalent to start with. In medical research this has been the standard practice for some time. In education, the application has been controversial and even the belief that evidence from effectiveness research has a role in decisions about instructional or professional development programs is not widely held.

When NCLB introduced scientifically based research, some proponents made the case for a strong parallel between medicine 40 years ago and education today with respect to effectiveness research. The debate 40 years ago in medicine concerned the need for clinical trials to establish that a new vaccine, a new surgical procedure, or a new medication was more effective than the current practice. Some felt that the "cold logic of science should not replace the clinical judgment of the seasoned practitioner." Ethical objections were raised about withholding drugs or procedures from a randomly selected control groups when administering them could save a life. But now the consensus is that clinical trials, while not foolproof, have been beneficial for medical practice.

While analogies to medical research can be helpful to a point, the comparison is quickly exhausted as soon as we begin inspecting the differences between the two industries. In medicine, the Food and Drug Administration requires that new medications are shown to be effective and not dangerous prior to giving permission for marketing. NCLB suggests that instructional products should have been proven effective before purchased with federal funds. But an FDA-like requirement would be unworkable in education. Three differences in particular are the basis for a specific approach in education to experimentation and the application of evidence to decision-making.

Less Risk

First, in education, the difference between an effective and ineffective instructional program never carries the danger of illness or death. Because the danger is minimal, experimentation in schools can be conducted under the authority of school districts to improve instruction, choose new curricula, set new standards, and specify testing practices. In many cases the Family Educational Rights and Privacy Act (FERPA) rules allow schools to release student records and, frequently, institutional review boards grant waivers for obtaining written consent where the researcher is not interacting directly with students.

More Variability

Second, FDA-like regulations are unrealistic because enormous differences district-to-district make generalization from findings in one district to another difficult. While in medical trials the unit of testing is usually the individual, in education, the smallest reasonable unit is usually the classroom, the school, or even the district. While cell biology is relatively consistent from human to human, the conditions of implementation of instructional programs in schools demonstrate enormous variation. An instructional program cannot be "proven effective" by a single program of research. Unless decision makers can recognize their own situation in the site where the research was conducted, they have little reason to think that the same effect would occur for them. The best evidence will be

that collected locally. Accumulating evidence from dozens of districts may begin to provide generalizable findings about whether a program works and under what conditions.

Effectiveness Research In, Not Before, Practice

Third, the time frame for generating rigorous evidence in education is measured in school years rather than months; therefore, if the instructional program had to be proven effective before it was marketed, the cost of instructional programs would be much higher and available only from the very largest publishers who could afford the cost of years of effectiveness research in dozens of school systems. Innovation could not be supported on a small scale. For example, a school district would not be allowed to adopt the NSF-supported math program from a local university professor until the university had licensed it to a publisher large enough to afford to conduct official nationwide trials.

When it comes to effectiveness research, however, the evidence is currently in very short supply. New textbooks, software, professional development programs, and hundreds of other kinds of materials used in classrooms are continually coming to market, being revised and improved. Research and development of new technologies and curricula are being conducted and funded by many federal agencies. If we read NCLB as requiring that research on the effectiveness of each of these has to be conducted before schools may begin to use them, innovation will grind to a halt. One of the premises of our research and development project was that developers and publishers need the schools as partners in piloting new products to test out their effectiveness in real classroom settings.

The Education Model for Research

As we have reported (David & Greene, 2007) and will further elaborate in this report, much of our learning in this project has concerned the difficulty of inserting evidence from research into the decision-making process. Other differences we noted concern the greater difficulty in generalizing and it is this characteristic that leads us to the idea that local experiments have great utility in education.

Generalization and the Local Experiment

Within the educational research community, the randomized experiment is most often a tool used in one of two ways. It is used for answering questions that arise from an investigator's theoretically driven program of research. Or it is used by investigators who are conducting an evaluation of a new program or intervention and planning to report an estimate of the effect. In both cases, the particular locale in which the randomized experiment is conducted is not as important as the question that the investigator derived from theory or introduced with respect to the intervention. The goal is a generalization about the theory or about the effectiveness of the intervention.

But when the experiment is initiated by the school district itself, the immediate goal is not to generalize to other districts or to build generalized scientific knowledge. Instead, the idea is simply to generalize to other schools, teachers, and students in the district. Over time, the evidence from many of these small experiments may be useful as the basis for generalizable findings, but the individual randomized experiment stands on its own as a contribution to a local decision.

The Trapped Administrator

The premise behind local experiments is not new. Thirty years ago, Donald T. Campbell wrote an essay called "Reforms as Experiments" (Campbell, 1969) in which he painted a picture of the "trapped administrator" that fits very well the situation of many school district decision makers today. The trapped administrator has made a bold decision to invest considerable resources in a promising but unproven program. When the evaluators are called in, the only politically acceptable outcome is to show that the program worked. Weak research designs that are open to multiple interpretations are preferred over rigorous research that may determine unambiguously that the program had little effect. Negative evidence must be explained away, while the failed program is forgotten as quietly as possible. As long as the administrator is in the trap, an evaluation of an

initial implementation can provide little information to support continuous improvement or to guide later decisions.

Campbell's approach to opening the trap is to implement reforms literally as experiments. By phasing in a new program starting with a pilot implementation, the decision maker can generate local evidence of effectiveness. If the program provides an initial indication of success, the implementation can be broadened the following year. If it turns out to be ineffective, the decision maker can pull the plug.

Scientifically Based Research and NCLB

The No Child Left Behind Act's provision for scientifically based research points in the right direction but doesn't actually open the trap. A major reason for this is that the Act is most commonly interpreted as requiring that the scientific research has already been conducted. When we read that schools should use "effective methods and instructional strategies that … are based on scientifically based research and that have proven effective," we may assume the Act means that the methods and strategies must already have been proven effective and perhaps even reported in the scientific literature before they are first used. But with little or no existing evidence for the products currently on the market, and with questionable generalizability of the extant research, school decision makers are forced into non-compliance or accepting very weak and non-applicable evidence in order to demonstrate compliance.

A different way of reading the requirement is possible. One can take the provisions in NCLB to mean that the scientific evidence can be generated in the process of a school piloting new programs. In this reading of the Act, innovative or locally developed programs do not have to wait for researchers to prove them effective. The analysis of data does have to be rigorous as envisioned by NCLB because important decisions depend on its accuracy. Since phased pilot implementations make excellent environments for well controlled experiments, we reasoned that the school district's own research could be used in making a decision for a larger scale-up of the pilot program in the district. In this interpretation, the federal government is not setting up a central adoption process and school districts are not held back by the current lack of scientific evidence but, in fact, become proactive producers of scientific evidence. This approach could get the school district leaders out of the trap in which they often find themselves.

Conclusions for Policy

While, as we report here, we have learned a considerable amount through the course of this project, an idea that we started with still holds true. The research enterprise in schools aimed at the effectiveness of instructional and professional development programs cannot follow the model of medicine with a rather rigid centralized process for approving new treatments. The enormous diversity of schools as well as the rapid development and improvement of new products and programs means that a single clinical trial will not prove a product effective. It is difficult to generalize from an experiment in one district to the impact in another district, even where there is considerable similarity. A large number of experiments with a variety of control conditions, student populations, and resource constraints will begin to build a level of useful generalization. We do not see how the required volume of research can be produced if state and local school systems do not take the lead in initiating tests of the programs available to them. Changes in the research funding mechanisms, for example, to include rigorous evaluations as a requirement of federal grant programs to the states and local school systems would generate more evidence than is possible through direct federal grants and contracts to research institutions or through funding by the vendors themselves. Changes in or waivers from the relentless NCLB requirements for achievement improvements will also be necessary to provide the extra time required for a district to make a decision on the basis of locally collected evidence. Rigorous proactive evaluations by districts can be one element in a decision process that incrementally improves instructional methods and programs. Although widespread use of the methods described here face serious but surmountable technical and organization obstacles, this approach is probably the only route to providing the volume of research that is needed for schools.

Recruiting School Systems to Participate

We approached far more school districts and state systems to conduct research than we actually helped to conduct research. Most had at least one reason for not participating even after, in some cases, preparing letters of support and holding multiple meetings on the topic. Recruiting was conducted before the project began, resulting in letters of commitment, and after the project began, especially in the second of the three years, during which a more systematic process was put in place. In most cases, even where a commitment letter had been provided, there was a subsequent discussion to identify a particular program to evaluate or question to answer. The difficulties in recruiting provide an important window into the organizational barriers that constitute an important finding of the project.

Recruitment process

To identify likely recruits, we drew upon several sources: the NCES database, focusing on larger districts; Council of Great City Schools member list; contacts known from prior professional work; and referrals given to us by the contacts we made. In addition, the National School Boards Association published an appeal in one of its newsletters. We also worked with a foundation that was supporting school districts in improving their management through better use of data. In a number of cases, vendors offered to find candidate districts interested in piloting their product. We considered these on the condition that, through interviewing the district personnel, we could determine that the district actually wanted to know whether or not the program was effective.

Typically we contacted the official in charge of evaluation or a person with a similar role. Of approximately 250 districts contacted, we received approximately 50 positive responses representing a 20% success rate, which is quite favorable with respect to the success of the proposed project. Nevertheless, the 200 failures are worth considering. In many cases, we were unable to find anybody, such as a research director, in a position to consider the offer of a “free” randomized experiment—on a topic of their choice. In some cases, we were referred to a research office that assumed we wanted to conduct research in their district. But in fact the idea was that we were offering to help them conduct their own research. We disqualified districts that asked us to fill out applications to conduct research without initially discussing what a question of interest to them might be. Through our recruitment efforts, we learned that most district administrators either do not understand how a randomized experiment can have value or are specifically opposed to the method. We found also that many districts are willing to undertake an experiment but for reasons unrelated to using the evidence in decisions.

Reasons for Not Conducting an Experiment

Our procedure was first to discuss options by phone and, in many cases, a site visit to determine that there was interest and an experiment was feasible. This was followed by a short proposal outlining the topic, timeline, and methods for the research. Of the roughly 20% of the district contacts who gave us an initially positive response, and in spite of extensive discussions in many cases, as well as written proposals, most did not in the end conduct an experiment. For example, frequently a research director would be excited by the idea of being able to conduct this kind of research but, after discussions with the other central office administrators, had to decline the offer. Either there was no appropriate program to evaluate or there was simply no interest. Some specific cases where the discussion proceeded further before failing to result in an experiment were the following:

- Although the district supplied a letter of support, had our project staff attend a meeting on site, and prepared a written proposal for a two-year experiment on an elementary reading professional development program, it was determined that the district had already identified certain schools to receive the program and, in any case, that it was not able to control the rollout to specific teachers or schools. Although a quasi-experiment might still have been possible (depending on the way the schools were selected), our project was focused only on randomized experiments and could not provide assistance.

- We were invited by the new superintendent of a large urban district to participate in meetings in which solutions for reading were discussed and vendors presented their products. The opportunity to pilot one or more of these products using randomized experiments was clear. A proposal was written and discussed among the assistant superintendents. However, word came back that the superintendent did not want a pilot. He wanted to move directly to implementing a solution. He was interested in what amounted to an interrupted time series design so see whether an aggressive broad program made a difference over time. But moving forward on the program precluded taking time to pilot alternatives.
- In a large district that was implementing a complex formative assessment tool for teachers, we proposed to study a very expensive professional development alternative compared to a less expensive offering from the same vendor. Meetings with the head of research as well as the assistant superintendent, along with the vendor, were progressing well and a proposal had been provided. The discussions were abruptly cut off, however, when the district superintendent was replaced. Later we were informed that, as a matter of policy, no randomized experiments were permitted.

Our project was not focused on collecting and analyzing these occasions and our data collection was not systematic. Our external formative researchers did follow up on a number of our unsuccessful attempts and their analysis of the interactions is reported in David & Greene (2008). The point to be made in recounting the three anecdotes is simply that, even where significant progress is made toward introducing an experiment into a district, events can be outside the control of the external researcher as well as the district administrators. The problem is not as simple as a disagreement with the principle of randomization. In many cases, the program implementation moves forward, leaving the idea of a research agenda behind.

Reasons for Conducting an Experiment

Our idealized goal of a study conducted to inform a specific decision didn't happen, even in the cases where school systems wanted to conduct a randomized experiment. We had only a single case among all the experiments we conducted where it appeared that the motivation was to decide whether the program was worth further investment. This is a case where, unfortunately, the experiment was curtailed because insufficient volunteers were available. The following were other reasons we encountered for districts agreeing to conduct an experiment.

- In one case, the school district had received a grant for which an evaluation was expected. The kind of evaluation was not made clear and there was no requirement for a randomized experiment aimed at determining effectiveness. Working collaboratively with our project, however, met the requirement and made possible a more elaborate study. There was the perception that a rigorous study would pay dividends down the road. But the strong focus on quantitative results was not expected and an evaluation focusing on observations of implementation would have been preferred. In this case, their grant had already specified the use of the program under study and the idea that the evaluation might inform a decision to change course was not considered.
- In several cases, the fact that a rigorous study was being undertaken was considered relevant to judging the value of the program under study. While there is some reason to believe that researchers and their funders would focus resources on programs that have already shown some success, this short circuiting of the process was counter to the notion that, once in progress, the value of the research is in the outcome.
- As documented by David & Greene (2008), it was also common for educators to value participation in research as a benefit in itself. If research funds are available, especially if the study will be rigorous, a topic ought to be found that is suitable. Often extensive discussions with interested participants followed in which an appropriate topic was identified often with more concern with whether a randomized experiment could apply to the question than with whether the answer to the question could inform a future decision.

- We did find cases where the desire for research was driven by an administrator's need to influence a decision—not a decision of his or her own, but that of a higher-up (e.g., a school board that would be deliberating on future funding). At the level of the program implementer, a commitment to the program under study was already made and, as in the "trapped administrator" scenario, the correct result was predetermined. Thus the study's initiator was often confident that a positive result would be found.

The idealization with which this project began was not found anywhere in practice. Discrepancies were found both in the reasons for not conducting an experiment and the reasons for conducting one. A richer understanding of how research evidence is used in practice is essential before we can expect locally initiated research to help in decision-making.

Understanding How Evidence is Used

Our experience in recruiting and working with school systems is reflected eloquently by our external formative research team, Bay Area Research Group. Bringing the perspective of researchers who have investigated school reform and policy initiatives as well as a solid understanding of experimental research methods, they provided us with essential insights about how evidence is used.

Questions of Interest to School System Decision Makers

This section quotes extensively from a report by our external formative research team (Greene & David, 2007). Their observations and analyses became an essential part of our project, providing a perspective from the educator's point of view and gaining access to opinions that would not have been shared with the core development team. In the following passage, they contrast the researcher's perspective with that of the district decision maker.

> From the district perspective, however, the primary empirical questions about a new program typically concern its usability, feasibility, and fit into their classrooms. They assume that students will learn more if the program is well-implemented, so district leaders want an *implementation* study, not an impact study. District decision makers want to know, What does it take to make this program work? If the program turns out not to work as well as expected, they want to know why: What different conditions do teachers need? What can administrators do to make it work better? To the state and district leaders we interviewed, the critical success factor is not the program or textbook per se, but how teachers use it.
>
> In fact, the decision to pilot a new program or textbook is often tantamount to choosing the program. By the time district decision makers receive the results of a pilot, they have often already committed themselves to the program under study, unless the reaction of teachers and students convinces them to reject it. Just as drug companies run trials only on drugs expected to be effective, district decision makers "test" only programs they presume to be effective. (The main exception is when they conduct a pilot to choose between two competing programs, in which case they are already committed to using one or the other.)
>
> From the Empirical Education perspective, the main outcome of interest is student achievement as measured by local or state tests. *Effectiveness* is a function of treatment effect sizes. For many district decision makers, however, one particular test may not represent a good measure of the treatment under study. Also, district decision makers are typically more focused on the *practical* action implications of results than on effect size or statistical significance. For example, if the only sizeable effect is an interaction between the treatment and one subgroup, implications for action are clouded. Similarly, a small effect size, negative or positive, is unlikely to overwhelm other outcomes of interest, such as whether students are engaged and teachers' degree of enthusiasm for the program. In response to the RCT design at one site a state department official said: "I would rely more on the teacher and student

responses to the program." Finally, when the intervention is a supplement to a program, such as a professional development component, district decision makers do not necessarily expect to see results reflected in student test scores.

In spite of Empirical Education's ability to customize the research question and study design to each situation, our findings suggest that the differences in perspective described above carry implications for the likelihood that local decision makers will find the results of an RCT useful in their decision making.

Decision Making in the School District

Again we quote from the Greene & David (2007) report on what they found regarding the use of evidence from a randomized experiment in making decisions.

> From Empirical Education's perspective, the ultimate goal of conducting state or district RCTs is to yield better decisions about major investments. Each RCT is intended to provide rigorous results to answer the question: Will *this* intervention work in *this* site? From the local perspective, however, the "formal data" type of evidence from a pilot or experiment usually plays a marginal role at best. District decisions to adopt or expand a program are made for a multitude of reasons, some based on various types of evidence; however, from what our interviewees said, rarely is demonstrated impact on achievement a factor.
>
> Instead, districts typically adopt or expand a program based on the availability of a grant or other outside funding (such as a donation), or the recurrence of a textbook adoption cycle. As one district leader said: "Just as many factors influence the implementation of a program, many factors influence the choice of a program. We sell our soul to get a [federal] grant or a [foundation] grant that [dictates] what we do." Other instigators we identified include personal relationships, such as with a publisher's representative, or a school board member's preference, or lobbying by influential teachers. When evidence is presented in support of a particular intervention, it commonly takes the form of personal anecdotes or testimonials. As one publisher put it: "It's much more convincing from a sales perspective to have a superintendent or teacher say [the program] is great—much more effective than showing them the data."
>
> Even when decisions to choose a particular program are not predetermined by a grant or other non-empirical source, local decision makers report that they pay attention to a set of criteria not related to measures of effectiveness. In our small sample, decision makers cited several such examples. Most important is evidence that the materials adequately align with topics and standards that the state requires. Second is evidence that the program is deemed "appropriate" for their particular students (for example, does it support English language learners) and instructional orientation (for example, degree of emphasis on basic skills). Third is evidence of feasibility, including costs of materials and training, availability of needed equipment (such as computers), and usability and compatibility from the teacher perspective. Only when specifically prompted do district decision makers acknowledge in principle the relevance of evidence of program effectiveness as measured by achievement test scores.
>
> For example, a textbook adoption decision in one district was ultimately made by the Board in July 2006, ratifying the earlier recommendation of a committee of teachers consulting a matrix of state adoption criteria provided by the county. The teachers constituting this committee had piloted two candidate textbooks between January and April 2006. Their judgments about the programs fitting well with their students and meeting their own needs for instruction and assessment weighed heavily in the decision process. The relative effectiveness of the two textbooks by achievement

measures was not considered. In fact, effectiveness is not even a factor in the matrix of this state's criteria for textbook adoptions.

Indeed, although Empirical Education had conducted an RCT in this district with one of the textbook candidates during the prior year, district decision makers made the choice not to share the results of this study with the adoption committee. Subsequent interviews with the district administrators revealed two basic reasons for this choice. First, they did not think the report findings were "clear-cut" enough to base a decision on them. In fact, the principal RCT finding was presented as a statistical interaction: a positive impact of treatment for bottom quartile students only, with no advantage for the average student. The presenter was hesitant to claim clear practical guidance from these data. The administrators' second reason for withholding these findings was that, "we didn't think the data were really important to the [adoption] decision." In the same interviews, these decision makers freely acknowledged and appreciated Empirical Education's status as a prestigious external research group, and volunteered that the mere fact that "research was done" carried the message that the candidate curriculum was a worthy choice.

In another district outside our sample where an RCT was conducted, even unequivocally clear achievement results did not sway decision makers or teachers, according to one researcher involved in the study. A new program tested in half the schools clearly outperformed traditional instruction. Nevertheless, the culture of teacher choice prevailed, and those preferring traditional instruction stuck with it. When the new program's advocate left the district, its usage dropped even further.

In the sites we studied, the one example where results were used was an instance of using research findings to justify decisions made on other grounds, an established use of research results by decision makers. In this case, a prior external non-experimental evaluation of the program had yielded positive results, which contributed to a significant increase in state funding for program expansion two years ago.

On closer inspection we learned that the research results had made a difference precisely because they amplified the appeal of a program which already had very strong, broad-based support. The positive results helped to justify the increase in funding, but would not have been sufficient were the strong political support not already present. Conversely, negative results could have had— and could still have— a negative impact on future funding, because advocates on either side of the decision will predictably use research results to support their position. As one state leader said of the planned RCT: "We know we're taking a risk on this, but we think our program is that good, that it can stand up to it."

Decision Processes

It became clear in our attempts to initiate randomized experiments, as well as in our observations of the process and of the use made of the findings, that we were confronting barriers that demand approaches beyond the methodological improvements on which most of our work in this project focused and which are reported on in the next section of this report. Here we consider prior research and conclusions about decision processes.

We are not alone in making these observations. Honig & Coburn (2008) recently published an article that is important for understanding how data and evidence are used at the school district level. This article was especially interesting because it specifically addressed decisions at the district central office. Much of the data-driven decision making research concerns the classroom level; teachers get immediate and actionable information about individual students. But data at the district level are more complicated and, as the authors document, infused with political complications. When district leaders are making decisions about products or programs to adopt, evidence of the scientific sort is at best one element among many.

Honig & Coburn review three decades of research and, after eliminating purely anecdotal and advocacy pieces, they found 52 books and articles of substantial value. What they document parallels our own experience in many respects. That is, rigorous evidence, once it is gathered through either reading scientific reviews or conducting local program evaluations, is never used "directly." It is not a matter of the evidence dictating the decision. They document that scientific evidence is incorporated into a wide range of other kinds of information and evidence. These may include teacher feedback, implementation issues, past experience, or what the neighboring district superintendent said about it—all of which are legitimate sources of information and need to be incorporated into the thinking about what to do. This "working knowledge" is practical and "mediates" between information sources and decisions. Inevitably, this involves the organizational or political context of evidence use. In many cases the decision to move forward has been made before the evaluation is complete or even started; thus the evidence from it is used (or ignored) to support that decision or to maintain enthusiasm. As in any policy organization or administrative agency, there is a strong element of advocacy in how evidence is filtered and used. While Honig & Coburn are referring largely to evidence from research conducted outside the district, we have found that many of the same processes apply also to local experiments.

Conclusion: Avoid the Dichotomy Between Process and Rigorous Evidence

One might conclude from our work and the review by Honig & Coburn that rigorous evidence has little value for district decisions, given the cognitive/organizational reality that "mediates" between evidence and policy decisions. Honig & Coburn contrast this reality with the position they attribute to federal policy makers and the authors of NCLB that scientific evidence ought to be used "directly" or instrumentally to make decisions. In fact, they see the federal policy as arguing that "these other forms of evidence are inappropriate or less valuable than social science research evidence and that reliance on these other forms is precisely the pattern that federal policy makers should aim to break" (p. 601). But we recognize a danger of creating a false opposition. To posit that a contrast must exist between the idea of practical knowledge mediating between evidence and decisions and the idea that evidence should be used directly is in some ways to set up a straw man. There are certainly researchers and research methodologists who do not study and are not familiar with how evidence is used in district decisions. But not being experts in decision processes does not make them advocates for a particular process called "direct." The federal policy is not aimed at decision processes. Instead, it aims to raise the standards of evidence in formal research that claims to measure the impact of programs so that, when such evidence is integrated into decision processes and weighed against practical concerns of local resources, local conditions, local constraints, and local goals, the information value is positive. Federal policy is not trying to remove decision processes, it is trying to remove research reports that purport to provide research evidence but actually come to unwarranted conclusions because of poor research design, incorrect statistical calculations, or bias.

There may also be a temptation to mistake descriptions of decision processes for evidence of deep and unchangeable human cognitive tendencies. If we take a developmental point of view, it is reasonable to expect that district decision makers can learn to be better consumers of research, to distinguish weak advocacy studies from stronger designs, and to identify whether a particular report can be usefully generalized to their local conditions. We can also anticipate an improvement in the level of the conversation between districts' evaluation departments, curriculum departments, and IT people so that local evaluations are conducted to answer critical questions and to provide useful information that can be integrated with other local considerations into a decision.

While our attempts at inserting randomized control into district decisions confronted the same phenomena that Honig & Coburn document in their review, we don't come to the same conclusion as to the immutability of the central office decision context. As we outline in the final sections of this report, there may be ways of reducing the apparent mismatch between the activity of conducting rigorous experimental research and the purposes and motivation of district decision makers. We conclude with a hypothesis that the gap can be bridged through a fuller understanding of a developmental sequence of data use.

Developing Methods for Conducting and Reporting Local Experiments

Our work with districts, including recruiting, conducting, and reporting experiments, pointed us to the importance of the organization context for the use of experimental evidence. While this was not the primary focus of our research, we nevertheless find that it is an important "lesson learned" from the project. At the same time, we conclude that improving the quality of program evaluations and lowering the cost of quantitatively rigorous studies continues to be an important research and development agenda. While such evidence may be used only incidentally, if at all, and may serve mainly to support the position of program advocates, it is still the case that evidence based on good design and leading to warranted conclusions is better than reports of studies that provide biased information. Even if the evidence is used selectively, it is still better to have solid evidence to use.

In this section we summarize and illustrate our findings from the development side of the project that was focused on improving methods for local randomized experiments. We report what we have learned in designing, conducting, and reporting research. These methods are also described and illustrated in each of the research reports. Here we provide an overview and conclusions about the usefulness of these techniques, especially with respect to lowering the cost of providing useful evidence to the local decision maker.

Designing for Local Questions and Local Resources

As we began going into districts to set up randomized experiments, we were confronted with an important fact about local experiments. We were limited by the number of teachers that the district could recruit. It made little sense to go outside the district to find additional teachers because the point was to understand how well the program worked with the local population. In fact, in one experiment we did go to the neighboring district to recruit additional teachers but found that, although their classes had a similar percentage of English learners, their ethnicity was different and the pattern of results was also different. We judged that the original district would find evidence on their own population more useful than results from an experiment that had marginally greater statistical power. This illustrates a cluster of issues in designing local experiments.

Limited Generalizability

The most obvious characteristic of a local experiment is that there is no need to generalize beyond the local district. For local decisions, an average result from a national sample is not useful unless the decision maker can discern the local conditions, for example in sub-samples.

The question that the district has is specific to their programs and their population of teachers and students. As we noted earlier, it is seldom evident to a district decision maker what kinds of questions can be answered using a randomized experiment. But as we homed in on questions that were of interest and were answerable, they often took on a local flavor, either because the program itself was home grown or because a particular sub-population of interest.

Another characteristic of the local experiment is that we often may more readily and easily characterize what the control group is doing than in an experiment where the sample is drawn from multiple districts. Ideally, this makes the result directly relevant—the experiment measures the difference between the district's current program and a new one being tried out. Normally, however, districts do not have clearly quantified achievement gaps or, when such needs have been identified, don't view a rigorous pilot as a means to address them.

Members of the larger research community may view this localization as unproductive in generating more widely generalizable evidence. In fact, however, there is a distinct advantage in using a meta-analysis of many local experiments that were conducted for the purpose of determining whether a new program will fill an achievement gap left by the current program. Doing so assures that the effect size measured is a gauge of improvement across multiple cases where the local district was primarily interested in whether a new instructional or professional

development program would be more effective than a current program with which they are not satisfied. A district that has just purchased a new high school literacy program would have little interest in testing another one. But a district that is dissatisfied has a valid control group to offer the research community. Their situation would offer at least one example of the difference between an unsatisfactory program and the new program being tested.

In other respects, the local relevance of an experiment is more difficult to use in research reviews because the program itself may be unique to the district or the issues of greatest concern (e.g., the achievement of a particular subgroup) may not have been captured in other local experiments. The value of the local experiment ultimately is in its specificity.

Identifying Subgroups of Interest

School districts differ from one another in their demographic composition and their issues with respect to achievement gaps between subgroups. They also may differ in the level of training or accreditation in their teaching staff. Factors such as these can make a difference to how successful the new program is. They also represent issues that are salient to decision makers, both because accountability provisions call out these subgroups and because the composition of the district often makes it distinctive. In many of our experiments we find that a program is more effective for students starting out with low achievement. English learners may respond differently than English fluent students. We have found a new program that led experienced teachers to be no more effective than their inexperienced colleagues. These interactions between the student or teacher characteristics and the treatment condition are frequently brought into the plan for the experiment by the district decision makers. Results, however, are not always actionable. For example, finding that a new textbook program gives an advantage to lower achievers within the range represented by the district population may not be relevant if the goal is to adopt a textbook for the whole district. On the other hand, finding that a new math program provides needed support to inexperienced teachers may help a school principal target the program for those teachers. Where there are policies that can be affected, a local experiment that highlights the subgroups of particular concern can be uniquely valuable. The value, however, is strongly tied to a decision maker's hypothesis that can be supported by the evidence and has an action associated with it.

Resource Limitations

In most of the experiments we conducted in this project, we were limited by the number of teachers who were available and willing to participate in the study. In most cases, piloting a new program requires extra effort and districts often do not have funds to pay the honoraria typical of government or foundation sponsored research. Where the program itself is a professional development opportunity, participating teachers are often compensated for their time. However, the time required to fill out surveys, conduct any additional testing, and perform other activities related to the study add to effort for teachers in both program and control groups.

A researcher can calculate the number of teachers or other participants needed in order to come to a conclusion that the differences observed in an experiment were not just a matter of chance. These calculations require making assumptions about relevant values. In many cases, the local decision makers do not have the information available to provide these values on the basis of local conditions and other factors such as, for example, the level of certainty demanded by the school board for findings of the proposed research. Ultimately, a decision as to whether or not to go forward with an experimental evaluation may depend on the number of teachers available locally and rough calculations as to whether the experiment can discern important difference.

Calculating How Many Teachers are Needed

Each of the values that go into the calculation can be problematic. The following two are tied to local conditions and represent areas that need further exploration and better tools for researchers working on local experiments

How big a difference does the new program have to make?

District decision makers may have identified areas of concern such as middle school algebra or third-grade reading and may even have identified subgroups particularly at risk. It is less common to have explicitly quantified this need in terms of the size of the difference a new program has to make to solve the problem or close the gap. But this is what is called for. The researcher needs to know the smallest gain that would be considered useful or worth going to the trouble and expense of implementing a new program in order to achieve. The smaller the difference, the more teachers are needed in the experiment.

Tolerance for risk of coming to the wrong conclusion

Two kinds of wrong conclusions are possible for research studies. First, one may conclude that the program had an effect when it actually did not. Second, one may conclude that it did not have an effect when it actually did. The scientific convention is to be much more stringent with the first kind of error than the second. Conventional "statistical significance" is set at .05 or a 5% probability of making the first kind of error. The scientific convention sets the second kind of error at .2 or 20%. But for any decision context, the decision maker may have a greater concern about rejecting a worthwhile program than going with a program that does not work as well as advertised. Or the decision maker may be equally concerned with both kinds of error. The greater the tolerance for risk of either kind, the fewer teachers are needed in the experiment. We believe that increasing error rates beyond conventionally held levels is justifiable so long as doing so is acceptable to the district and that levels are specified before the experiment begins.

The area of acceptable risk merits further investigation. It is likely that a large number of factors influence this, for example, the size of the investment, the urgency of the problem, or even a desire to influence a decision one way or another. In the reporting format we developed under this grant, we have taken a middle road of reporting in terms of a range of confidence from strong to limited. Our solution, however, does not address the problem of communication with decision makers about their own tolerance in terms of which our confidence range can be understood.

Level of randomization

So far this discussion has treated teachers as the unit of randomization and this is appropriate for many interventions. But to a large extent, the size of an experiment is related to how many "units" are randomly assigned to either treatment or control. Over the last three years, we have successfully used school, grade-level (the team of teachers at a single grade level within a school), teacher, and class-level randomizations. School-level randomization is required where the unit of implementation is the school or where teacher collaboration within the school makes it difficult to distinguish the conditions, as we found in one of our experiments. Teacher-level randomization has worked for interventions that were quite specific or complex and especially where technology determined who had access. In such cases, we can be assured that teachers not trained in the new techniques or without access to the software will not be able to share the intervention with control teachers in their school.

In two of our experiments, the randomization was by class. This initially came about because the number of teachers who turned up at the recruitment meeting was far fewer than the district had expected. Since these were middle and high school algebra teachers, each one taught up to four classes. Rather than call off the experiment due to the small turnout, we got the teachers to agree to teach half of their classes using the new program and half using the old program. This was an added burden on the teachers but one that they were willing to undertake as a condition to working with the new program. While the number of students in the experiment stayed the same (809), we increased the randomized units from 12 teachers to 32 classes giving the experiment sufficient statistical power to detect differences of potential interest.

We believe that school district decision makers will need a considerable amount of experience in using data from their own district to reach the point of being able to state what the size of the impact is that

they need. Tools for considering the cost of decisions, weighed against the possible benefits, are also needed to support the reasoning called for in determining the tolerance for risk in coming to the wrong conclusion from an experiment. As we illustrated with the example of moving from teacher to class-level randomization, multiple factors are at play in deciding whether an experiment is even feasible. The decision to conduct a local experiment can be motivated, as we've seen, by considerations that have little to do with informing a decision about whether to move forward with adopting or scaling up a new program. The considerations that go into deciding whether an experiment will be feasible or useful assume that there is a decision to be made for which there are goals and risks. They assume a familiarity with needs that can be quantified and with existing programs and their costs. Where district decision makers are not already steeped in the district data and without tools for understanding the costs and risks in introducing new programs, moving toward planning an experiment—even a simpler non-randomized comparison—represents a fairly large leap.

Conducting the Local Experiment

This section addresses the challenges we encountered in conducting experiments in districts and reports on some of the solutions we arrived at. Our goal in the development of methods was to lower cost and, to that end, we developed "standard operating procedures" that could be repeated across a number of experiments when appropriate. We've also identified challenges, particularly with respect to timeliness, that must be solved for local experiments to serve the function we envisioned at the beginning of the project.

What Makes Randomization Difficult

We start by addressing the issue that is often held up as the major barrier to experimentation and often associated with enormous expense. We see no reason for randomization to be inherently expensive. When we find resistance, it is usually at higher levels of an organization where fear of political backlash drives what appears to be a principled objection. And to some people, the coin toss (or other lottery method) just doesn't seem right. Any number of other criteria could be suggested as a better rationale for assigning the program: some students are needier, some teachers may be better able to take advantage of it, and so on. But the whole point is to avoid exactly those kinds of criteria and make the choice entirely random. The coin toss itself highlights the decision process, creating a concern that it will be hard to justify, for example, to a parent who wants to know why his child's school didn't get the program.

But in fact, randomized control works quite well where the district is doing a small pilot and has only enough materials for some of the teachers, where resources call for a phased implementation starting with a small number of schools, or where slots in a program are going to be allocated by lottery anyway. Our own experience with random assignment has not been so negative. Most districts will agree to it, although some do refuse on principle. When we begin working with the teachers face-to-face, there is usually camaraderie about tossing the coin, especially when it is between two teachers paired up because of their similarity on characteristics they themselves have identified as important.

The main problem we find with randomization, if it is being used as part of a district's own local program evaluation, is the pre-planning that is required. Typically, decisions as to which schools get the program first or which teachers will be selected to pilot the program are made before consideration is given to doing a rigorous evaluation. In most cases, the program is already in motion or the pilot is coming to a conclusion before the evaluation is designed. At that point in the process, the best method will be to find a comparison group from among the teachers or schools that were not chosen or did not volunteer for the program (or to look outside the district for comparison cases). The prior choices introduce selection bias that we can attempt to compensate for statistically; still, we can never be sure our adjustments eliminate the bias. In other words, in our experience the primary reason that randomization is harder than weaker methods is that it requires that the evaluation design and the program implementation plan are coordinated from the start.

The issue returns us to the organizational constraints on program implementation. A randomized experiment necessarily puts the researcher in between the district program person and the

program. Where the goal is to assure the most successful and expeditious implementation, an experiment simply stands in the way. In a standard researcher-initiated experiment, the implementation is done for the purpose of the experiment. In any local district-initiated experiment the program is the priority even where there is a strong motivation to pilot the program as a means to determine its suitability and effectiveness. A randomized experiment is not inherently expensive but it does require putting priority on the research at the expense of the implementation and, in many cases, it is simply not possible for the research to control assignment to the program.

Paired Randomization

Early on, at the recommendation of a research advisor, we started using a technique for conducting a randomization that involves finding pairs of teachers (or schools or classes) that are maximally similar on characteristics that should have an impact on the outcome and tossing a coin between the pair, assigning one teacher to the new program and the other to control, that is, to continuing with "business as usual." We developed what amounted to a workshop format for this activity, usually conducting it after school with the group of teachers who came to participate (or at least to hear more about the study). The meeting included a presentation by the researcher explaining the way the randomization would be done and providing a motivation in terms of an explanation of the sources of bias that would otherwise distort the findings and make them less useful. Usually a representative of the company whose product was being tested would also do a short presentation and explanation of the training. And a district administrator who had agreed to be the point of contact for the research would attend and often speak to the district's interest in finding out how well the program worked.

After all the questions were answered, consent and background information forms were distributed. Teachers could either fill them out and sign them or leave if they did not want to participate. A. discussion ensued in which the remaining teachers identified characteristics that would likely make a difference for this program in their district. Often, factors not anticipated by the researchers were raised, such as an unusual scheduling at some schools that would impact the pacing and implementation of the program. Once factors were put up on the flip chart, they were clustered and prioritized. In some cases, the factors were obvious, such as teachers representing different grades or some schools being Title I. These factors were then used to divide the teachers up into blocks, for example teachers in Title I schools at one end of the room and others at the other end. These groups were further subdivided by grade and then by years of experience. Using the factors agreed to, the clusters became pairs whose consent forms were stapled together and the coin tossed to establish the assignment.

This method was efficient in two ways. First, it was very easy for the teachers to understand how the groups were equivalent, both in the sense that the pairing equalized the two groups (e.g., equal numbers of highly experienced teachers, teachers from Title I schools, etc.) and in realizing that the coin toss arbitrarily placed one of them in the program and the other in control. Standing with their pair, some teachers might suggest that one or the other would be better piloting the program. When that happened, we made it clear that the two groups resulting from random assignment would have teachers with equal levels of interest in the program—and, of course, on any other factor that nobody had thought about or measured. Second, paired randomization turned out to be efficient statistically. Although some methodologists were concerned, especially with small numbers as we often had, that more would be lost by blocking on pairs than was gained through the equalization, our own analyses using a large number of small experiments showed that in most cases the impact was either neutral or provided a small advantage (Jaciw, Wei, Ma, & Newman, 2007; Jaciw & Ma 2008; Jaciw, Wei, & Ma, 2008). In local experiments where the number of units is often a concern, we showed that the method was efficient. The greater cost savings may be in the level of cooperation from the control group teachers, who understood through participating in the process that both groups were essential to the study.

Data Warehouses

One of our initial hypotheses was that the deployment of sophisticated data warehouses in districts would substantially lower the cost of conducting research. Over the last several years, commercial development and marketing of data warehouse systems have continued to move forward. The US Department of Education has funded experiments in growth models for student records that call for longitudinal data systems, at least at the state level. The Data Quality Campaign has also been tracking state-level data systems and providing definitions of more and less sophisticated techniques. It is unclear how the work at the state level is impacting work at the district level. In one of our experiments, we found that the state had selected a single vendor to provide data warehousing for all the districts. While this enhanced the ability of the state to process data by assuring a high level of consistency, we also found that administrators in the individual districts often did not have the training required to conduct their own queries of the local system that had been selected for them.

Although our project did not conduct a survey of the school district data systems, our experience in the small sample we worked with indicated a growing awareness. Still, a majority of systems lacked a systematic way to, for example, link teachers, class rosters, and student scores and demographics across years. In many cases, the district staff members were able to conduct queries and put data together from multiple internal sources but in other cases, matching of records had to be done at Empirical Education. We developed a warehouse structure that allowed us to create a systematic view of the relevant subset district data to provide to our statistical analysis function. It is clear, however, that if districts are to conduct their own studies, their internal data warehouses will have to be more comprehensive than those we typically encounter. It is likely that the development of such systems will be motivated initially by simpler investigations such as identifying problem areas and systematic needs analyses. The same basic data system that would support such data mining would also support the analysis of experiments.

Standardizing on and Partially Automating a Few Key Statistical Methods

One way to lower the cost of conducting randomized experiments was to settle on several "standard operating procedures" that could apply to any experiment and could, to some extent, be automated within the statistical analysis environment. The statistical analysis environment we had available for this project was SAS and we found advantages of using this one environment consistently. Because our procedures could also be programmed in other environments, there was no particular dependence on this environment. While our project did not conduct any market research as to penetration of SAS in school districts, we were aware of districts that used SAS tools both for data warehousing and for statistical calculations suggesting that, in the future, use of our techniques by school districts would be feasible.

As we worked on the statistical analysis of several experiments, it became clear that some of the steps in running these analyses could be automated through the use of macros within the SAS environment and templates for capturing the output of SAS. The following procedures exemplify the choices that the project made in developing the standard procedures:

- The analysis plan is specified before the outcome dataset is inspected. This is standard procedure for any experiment and helps to avoid mining for results post hoc. However, we held open the possibility of a change in design during the conduct of the experiment, for example, depending on the units available or other measures that may or may not have been provided by the district.
- Pairs of randomized units are treated as fixed effects. This is the important level above the level of randomization to include in the model.
- Pretest measures are always used. It is well known that the use of a pretest improves statistical power. We made the availability of a pretest a pre-condition for the feasibility of an experiment in any case. In schools this often restricted the grade range that could be included (e.g., where state testing starts in 3rd grade, we would not have a pretest for the 3rd graders).

- We always investigate the interaction of treatment with pretest. It was common enough to find such interactions that we built this in as part of our standard analysis procedures.
- Other moderators, identified prior to inspecting the outcome, were included, each in its own model. This model, for example, would include the main effect for pretest and the moderator by treatment interaction. Each moderating effect is investigated separately. These were often requested by the district based on their specific questions (e.g., a concern for specific ethnicities) or their theory of action (e.g., that the program would assist inexperienced teachers more than experienced teachers).
- Standard methods for accounting for attrition were developed and applied consistently.
- A templated "laboratory notebook" was developed to guide statistical analysts through the standard sequence of steps, including invoking the appropriate SAS macros. While rhis procedure provided for recording the steps and the output in case questions arose later, its value in lowering cost was in its use as a scaffold for the consistent steps in the analysis.

Survey Techniques to Capture Implementation and Mediators

From the outset we were concerned with the cost of conducting observations that were sufficiently structured, frequent, and time intensive to yield statistically usable quantifiable results. These seemed beyond the resources likely to be available to a school district researcher. We did conduct walk-throughs with administrators and, in some cases, with the vendor support staff who were concerned with the success of the implementation. We used a standard procedure for conducting interviews with the participating teachers. We also consistently observed the training and, where possible, the support events. The primary result of these observations and interviews was qualitative descriptions of the problems and issues in the implementation. For example, in one case, our observations brought to light a network server resource by which program teachers were systematically sharing the curriculum modules they developed as part of the professional development with the control teachers. Once discovered, it was obvious that a school-level rather than teacher-level randomization would have been required (although not feasible given the number of schools available). Nevertheless, the observation allowed the research team to explain the paradoxical result of the control teachers apparently benefiting more from the treatment than the treatment teachers. We found also that interviews very early in the process helped disambiguate classroom arrangements (such as team teaching or the use of resource rooms and specialists) that were not evident from the official rosters. Our conclusion was that this kind of awareness of how the implementation is progressing and unstructured discussion with the participants is essential to frame the overall context and interpretation of the results.

Our primary tool for gathering information from teachers was web-based surveys. We found the almost universal access to email and to the web made this method available for most school districts conducting research. We used a commercial survey tool that would also be available to districts at a relatively low cost. The expense, however, was in our follow-up to achieve response rates above 95%. Achieving such a rate requires a level of persistence that a district staff person would have to adopt, but certainly work that could be done at a secretarial level. Our main innovation over the usual use of surveys was to survey multiple times—once a month or even more frequently. By asking the same questions concerning amounts of time spent on various relevant instructional practices, resource availability, professional development occasions, and so on, we were able to sample across the whole year. Because the question was phrased, "last week how much time…." the teacher did not have to rely each time on long term memory to provide answers. This technique did require complicated aggregation of the responses across multiple surveys but we believe that the technique was adequate and effective as a substitute for multiple observations.

Implementation vs. Mediation

While not always pleasing to strict experimentalists, school district decisions are heavily weighted toward implementability of the new program. This begins with affordability of costs for professional development, additional infrastructure, and the vendor support, and includes the acceptance by the school staff and how well the instructional program fits with the standards. As we illustrated in the earlier section of this report, having quantitative evidence of comparative effectiveness of the new program is not the first consideration. If the program simply will not work or is too costly, it does not make it into the running. Moreover, negative findings are easily explained by poor implementation or by the fact that treatment impact was measured over a single school year where a second year may show a different impact as teachers get accustomed and more skillful at the new approaches.

Many of these factors that go into a decision can be determined without a control group. The traditional pilot provides a level of exposure to the new program that will uncover many of them. In some cases, we pushed back on a district request to highlight implementation difficulties and move the negative findings in the comparison with the control group to the background. In part we were concerned that the implication of poor implementation tied with poor results implied that there would be positive results with strong implementation. For example, we did not want to substitute correlational results for the experimental ones. Finding that teachers who implemented "with fidelity" also got better achievement results does not show that proper implementation results in effectiveness. It only shows that teachers who are enthusiastic and interested enough to implement fully also achieved good results. But this is confounded since the same enthusiasm may have led to success regardless of the program. In fact, we simply do not know whether there are conditions that would lead to positive results. We came to two conclusions from these observations.

Early implementation reports

First, we see that implementation difficulties can be observed and reported considerably prior to the quantitative achievement results being available. An interim report just on implementation can be provided as initial guidance. The advantage to the district is that a decision to move forward with the program or to begin a major scaling up can be influenced simply by the difficulties experienced by the program teachers. In our experiments, for example, it was known early on in one case that the vendor was unable to deliver the books in a timely manner and that some of the schools were not sufficiently well equipped with technology. These observations do not require an experiment but if one is in progress, an interim report of this sort can be very useful. By providing this report prior to a report on achievement impact, there is no temptation to justify decisions on the basis of confounded data correlating implementation and achievement.

Reports on mediation

Second, toward the end of this project, we began exploring the application of mediation analysis. This is a way to systematically use the impact of the program on teacher practices in the analysis of the impact of the program on student achievement. This technique calls for a theory associating observable processes with both the program and the student outcome. Unlike in the reports of implementation issues, which apply only to the program group, mediation analysis looks at impact on teacher practices and other process variables across both program and control. A survey of time on task, teachers' perception of student engagement, or amount of project-based teaching can apply to both groups and differences on these can also be measured as impacts of treatment, just as test scores can. Implementation measures that can apply to both the treatment and control groups also have a value in helping to explain how the treatment had its effect.

Because technical issues about how to do this kind of analysis in the multi-level context are still being worked out, we have not fully integrated this into our standard models. But the idea is that the difference between treatment and control causes changes in both the student outcome and a mediating variable (such as time spent, student engagement, and so on). Insofar as there is a positive impact of treatment on the mediator and a positive correlation of the mediator and the outcome, the mediator helps to explain how treatment had its effect. By including the mediator

analysis, the randomized experiment provides better insight and a more persuasive case for or against the treatment program.

Using our survey system to get multiple samples of these classroom process measures allows us to detect differences resulting from the program intervention. These provide a more refined look at whether the program is likely to have an effect. Essentially, we are measuring the mediating links or "active ingredients" hypothesized by the theory of action. In so far as these process measures are available for analysis before the student outcome data, this approach also gives us an early indication of the impact on students. Providing, of course, that the theory of action is correct, this experimental result based on both the program and control groups tells us whether we can anticipate an impact at the student level when that analysis is complete.

Timeliness of the Results

Making use of preliminary results based on implementation success and changes in instructional processes is a partial solution to the need to provide the evidence before it becomes irrelevant. For evidence based on impact on student achievement, there is a greater challenge. If the district is going to introduce a new program at the beginning of the next school year, a final decision must be made in the May-June timeframe at the very latest. This provides very little time between spring testing and the report. We have had some success in speeding up the delivery of results but more work is needed.

Quick turn-around of the results is one critical requirement for using evidence in decision-making. The issue is deeper than that. If the district personnel are going to use the evidence (even preliminary implementation or process results), they must have already agreed to defer a decision for a year while the experiment takes place. In other words, a decision to base a later decision, in part, on the evidence produced by an experiment must be made perhaps 18 months prior to the implementation—that is, it would best be made in the spring of the previous year. Extending the normal "sales cycle" also requires that the funding be committed but not spent, something that is often impossible either because the funds disappear on an annual basis, or the demands of the board or community to take action puts pressure to spend without a delay. An approach that is compatible with a randomized experiment is for the district to plan and budget for a rollout in stages over two or three years. The results of the first stage feed into a decision whether to expand the rollout in the second phase. This arrangement will work out only, of course, if the evidence can be produced, reported, explained, and integrated into a decision within a very short time of the outcome test. But the core issue is that, from the beginning of identification of the problem, selection of candidate solutions, and negotiating with vendors, the district must be proactive in planning the evaluation.

Reporting Local Experiments

A challenge from the beginning of this project was to strike the right balance between a report that was understandable and useful within the school district and a report that contained sufficient information for technical review. This is a problem with many layers. It is not just a technical description of methods and the common research jargon that must be made accessible, there is the larger question of how to understand the findings in relation to the questions and motivations that led to getting involved with researchers in the first place. Our project has made progress in addressing these issues. Improvements continue to be made, although we have settled on an overall framework based on an identification of the audience. Further work is needed to understand the level of technical expertise required of the audience to correctly interpret and use experimental findings. We suspect that experience with simpler forms of mining of the district's own data warehouse for needs analysis and other projections will make some of the underlying data processes clearer and open the way to what amounts to a rather complex reporting of experimental results.

Choosing an Audience

Our first step was to visualize the audience for our reports. We imagined three levels. First, there was the school board member, superintendent, or central office curriculum director with limited or no technical understanding of experimental design or statistical analyses. Second, we visualized a school district director of research or evaluation who has had some background in research methods but for whom concepts such as hierarchical linear modeling or fixed and random effects would not be familiar. Finally, we anticipated that there would be sophisticated technical reviewers, such as journal editors or consultants who would want to understand the design and analysis decisions that require detail and distinctions that would not be useful to other readers.

Our basic strategy was to focus on the second audience and write a report that introduced and explained in accessible language concepts such as effect size, random effects, interactions, and the like. For example, we tried to avoid jargon that is familiar to researcher but would not be familiar to district people whose research training was perhaps a decade or so in the past. As we explain and illustrate in this section, we attempted to present the logic of what we were doing so that if our hypothetical district staff person with a mid-level of research background were to read it carefully, the concepts and methods that are currently in use in experimental educational research would be explained. For the sophisticated audience, we provide footnotes often with a fair amount of technical detail. While the report itself takes the form of a standard journal article with sections for methods, results, and discussion, the report is otherwise not typical of reports aimed at the narrow audience of the peer-reviewed journal.

We recognize that the primary audience for the evidence, if not the report, is the first level that we visualized—the actual decision makers. For these executives, we did provide an executive summary as we describe below. However, the use of the evidence, as we discussed in the previous section, is complex and a summary even as brief as two pages, focused on one or two main findings may not by itself be useful beyond evidence that a study was conducted. This is not meant to denigrate central office executives. Rather it is to point out that scientific evidence of impact is one of many considerations that must be juggled. Ultimately, we believe that processes for using data for needs analysis and in understanding important trends will provide a context for the use and interpretation of evidence. The document, by itself, will not result in effective use of evidence. In what follows we illustrate the approach we developed for reporting our results.

The Executive Summary

We have provided as an illustration a summary provided to one of the districts in reporting on their experiment with the Cognitive Tutor program for pre-algebra called “Bridge to Algebra.” All our executive summaries take the same form, which is to begin with a paragraph explaining the context and purpose of the experiment. We then provide the results, including what we judge to be the most salient finding, given the district’s interest. This text generally is based on the “Discussion” section of the larger report. Details on method and other findings follow. Restricting this to two pages allows it to be distributed at a meeting on a single sheet of paper, which we believed would improve its usefulness.

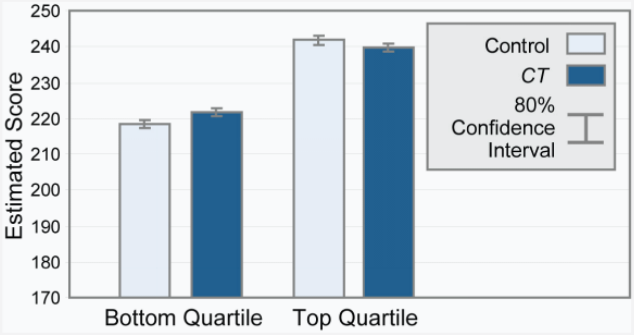
Executive Summary

Introduction. Under the *Math Science Partnership Grant*, the Maui Hawaii Educational Consortium sought scientifically based evidence for the effectiveness of Carnegie Learning's *Cognitive Tutor® (CT)* program as part of the adoption process for pre-Algebra programs. During the 2006-2007 school year, we conducted a follow-on study to a previous randomized experiment in the Maui School District on the effectiveness of *CT* in Algebra I. In this second year, the focus was on the newly developed *Bridge to Algebra* program for pre-Algebra. Maui's choice of *CT* was motivated in part by previous research showing substantially positive results in Oklahoma (Morgan & Ritter 2002). Our previous findings in Maui—less positive results for *CT* overall and somewhat negative results for certified teachers—called for additional study with the unique locale and ethnic makeup of Maui.

The research question was whether students in classes using *CT* materials score higher on standardized math assessment, as measured by the Northwest Evaluation Association (NWEA) General Math Test, than those in a control classroom using the pre-Algebra curricula currently in place. The district was also interested in learning whether *CT* is a teacher-friendly tool that could be used feasibly in their setting, whether there would be a differential impact on specific ethnic groups, and whether uncertified teachers would gain more from *CT* than certified teachers.

Findings. We found that most students in both *CT* and control groups improved overall on the NWEA General Math Test. We did not find a difference in student performance in math between groups. Our analysis of the Algebraic Operations sub-strand revealed that many students in both groups did not demonstrate growth in this scale, again with no discernible group differences.

However, for Algebraic Operations outcomes, we found a significant interaction between the pre-test and *CT*: students scoring low before participating in *CT* got more benefit from the program's algebraic operations instructions than students with high initial scores (see bar graph). Moreover, we noted an indication of a differential impact favoring Filipino students over White students on the Algebraic Operations sub-strand. Since the groups of interest (Filipino and Hawaiian/part-Hawaiian students) overall had lower average pretest scores, the results suggest that *CT* may help to reduce the achievement gap between those groups and others.



Differences between *CT* and Control Algebraic Operations Outcomes: Median Pretest Scores in Top and Bottom Quartiles

The district was also interested in *CT's* effectiveness for students taught by certified teachers versus non-certified teachers. In the previous year's study of *Cognitive Tutor for Algebra I*, control students of certified teachers had outperformed control students of non-certified teachers. But the program appeared to have a detrimental effect for certified teachers and no effect for non-certified teachers, both for math overall and for algebraic outcomes. In this experiment on pre-Algebra, we find certified and non-certified teachers performing about the same in their control classes. For math overall (but not the Algebraic Operations sub-strand) we find that *CT* gave the non-certified teachers an advantage.

Our goal was to provide the Maui School District with useful evidence for determining the impact of *CT* within the local setting. Considered as a district pilot, the study adds to the information available on which to base local decisions. Although our study did not provide evidence of a positive impact of *CT* on student math achievement in general, we found some positive effects. Overall, despite the repeated challenges teachers faced in implementation, *CT* was successful in raising student engagement in math and demonstrating, on the algebra-related sub-strand, gains for previously lower-

performing students. The program also appeared to be particularly beneficial for non-certified teachers. Because a small number participated in the study, we consider these conclusions for teachers suggestive but not conclusive.

Design and analysis. The design of our Maui experiment was similar to the Oklahoma study, in that pre-Algebra classes were randomly assigned to *CT* or to control. We used a coin toss to assign 32 classes in five Maui schools to use the *CT Bridge to Algebra* program or to continue using the pre-Algebra program currently in place. Each of the 12 teachers involved in the experiment had equal numbers of *CT* and control classes. In their *CT* classes, they used *Bridge to Algebra* for eight to nine months until the NWEA math posttest was administered in May 2007.

The research for this experiment encompasses a multiple methods approach. We collected pre- and posttest math scores from NWEA, and class rosters and demographic information on students and teachers from the district. To measure and document implementation factors and student and teacher interactions with the materials, we also collected qualitative data through classroom observations, phone interviews, and web-based surveys from teachers.

Because our findings differed from those in Oklahoma, this small study illustrates a general caution in interpreting findings from isolated experiments and demonstrates the importance of conducting multiple replication trials of any application in varying contexts and conditions. Large numbers of trials will begin to build the confidence we can have about the product and, more importantly, they will provide the multiple examples of its functioning with different populations and conditions. Then users of the research will not only have evidence of the product's average impact, but they will also be able to find contexts that are very similar to their own in order to obtain more specific guidance of its likely impact under their conditions. Here, it is important to interpret the results in relation to what teachers were using in control classes and to the usage patterns, implementations, and applications in *CT* classes. It is also relevant that half the *CT* teachers were using *CT* for the first time; and their initial unfamiliarity may have affected implementation. Finally, the size of this experiment precluded detection of small differences.

Overall teacher impressions. Our qualitative data sources revealed that teachers experienced similar resource challenges in implementing *CT* as in the Algebra study: lack of classroom computers, access to the computer lab, and *CT* materials. Another challenge related to the misalignment between *CT* content and state math standards in middle and high school. Despite these challenges, teachers (and students) reported a generally positive attitude about *CT* overall. Teachers were particularly pleased with how engaged their students were with the *CT* software and the *CT* approach to collaborative learning. (It must be noted that 45% of the teachers reported that this approach affected instructional practices in their control classes. We were not able to determine whether this contamination made a difference in outcomes.)

Explaining Our Methods

The following segment from our methods section of the report on the Cognitive Tutor program illustrates our attempt to explain some of the core issues and approaches in statistical analysis. It is important to note that, at the end of this section, we address the question of confidence in the results in relation to statistical significance. Consistent with our concern that the conventions in social science research may not apply to practical decisions based on local data, we provide a way of interpreting *p* values in terms of levels of confidence. Implicitly, we allow a relaxation to a *p* value of .2 while warning the reader that this represents a very limited level of confidence. As you will see in the examples where results are reported, as well as in the figure shown in the executive summary, we provide an 80% confidence interval—again allowing the decision maker to see differences that in the research literature might be reported as "non-significant trends." Also, in reporting interactions between a moderator and the treatment, we provide graphs to illustrate any effect with a *p* value lower than .2. Again, with proper explanation of confidence and with these procedures established prior to inspection of the results, we avoid any *post hoc* fishing for results.

Statistical Equations and Reporting on the Impact of *CT*

Setting Up the Statistical Equation[1]

We put our data for students, teachers, and classes into a system of statistical equations that allow us to obtain estimates of the direction and strength of relationships among factors of interest. The primary relationship of interest is the causal effect of the program on a measure of achievement. We use SAS PROC MIXED (from SAS Institute Inc.) as the primary software tool for these computations. The output of this process are estimates of effects as well as a measure of the level of confidence we can have that the estimate is true of the population to which the experiment is meant to generalize.

Program Impact

A basic question for the experiment was whether, following the intervention, students in *CT* classrooms had higher math scores than those in control classrooms. Answering this is not as simple as comparing the averages of the two groups. The randomization gave us two groups that are equivalent to each other on average in every way, except that one receives *CT* and the other one does not. But as we saw in the section on the formation of the experimental groups, in a single randomization we expect chance imbalances. Adjusting for these random differences gives us a more precise measure of the program's effect. It is also essential that we understand how much confidence we can have that there really is a difference between the two groups, given the size of the effect estimate that we obtain. To appropriately estimate this difference, our equation contains a term for *CT* as well as terms for other important factors such as the student pretest score. The student's prior score, is of course, an important factor in estimating his or her outcome score. By including pretest as a term in the statistical equation, we are able to improve the precision of this estimate because it helps to explain a lot of the variance in the outcomes and makes it easier to isolate the program impact. We also have to account for the fact that students are clustered by classes and teachers. We expect outcomes for students who are in the same class or who have the same teacher to be dependent as a result of shared experiences. We have to add this dependency to our equation or else our confidence levels about the results will be artificially high.

Covariates and Moderators at the Student and Teacher Level

In addition to estimating the average impact, we also include in the equation other variables (called covariates) associated with characteristics of the teachers and students, which we expect to make a difference in the outcomes for the students. For example, as was described above, we add the pretest score into almost all our statistical equations in order to increase precision. In addition, we consider whether there is a difference in the effect of the intervention for different levels of the covariates. For example, we consider whether the program is more effective for higher-performing students than for lower-performing students. We estimate this *difference* (between subgroups) *in the difference* (between the program and control groups) by including an interaction term in the statistical equation. This term multiplies together the variable that indicates whether the student is in the intervention group, and the covariate. We call covariates, that are included in such analyses, potential "moderators" because they may

---

[1] The term 'statistical equation' refers to a probabilistic model where the outcome of interest is on the left hand side of the equation and terms for systematic and random effects are on the right hand side of the equation. The goal of estimation is to obtain estimates for the effects on the right hand side. Each estimate has a level of uncertainty which is expressed in terms of standard errors or *p* values. The estimate of main interest is for the treatment effect. In this experiment, we model treatment as a fixed effect. With randomized control trials, the modeling equation for which we are estimating effects, takes on a relatively simple form: Each observed outcome is expressed as a linear combination of a treatment indicator, one or more covariates that are used to increase the precision of intervention effect, and usually a series of fixed or random intercepts, which are increments in the outcome that are specific to units. As a result of randomization, the other covariates are distributed in the same way for both the treatment and control groups. For moderator analyses we expand these basic models by including a term that multiplies the treatment indicator with the moderator variable. The coefficient for this term is the moderator effect of interest.

moderate—either increase or decrease—the effect of the program on student outcomes. The value for the interaction term is a measure of the moderating effect of the covariate on the effect of the program.

Fixed and Random Effects

The covariates in our equations measure either 1) fixed characteristics that take on a finite set of values (e.g., there are only two levels of gender); or 2) a set of characteristics that is assumed to have a distribution over a population and where we treat the values that we measure as though they were a random sample from that larger population. The former are called "fixed effects", the latter, "random effects". Random effects add uncertainty to our estimates because they account for sampling variation, or the changes we would observe in the outcomes if we re-sampled units from the same population. Fixed effects produce less uncertainty but also limit the extent to which we can generalize our results.

We usually treat the units that were randomized as "random effects", so that in the statistical equations, our estimates reflect the degree of uncertainty that comes if we were to draw a different sample of units from the same population[1]. This allows us to argue for the generalizability of our findings from a sampling perspective. Treating the units that were randomized as fixed, forces us to use other arguments if our goal is to generalize.

Using random or fixed effects for participating units serves a second function—it allows us to more accurately represent the dependencies among cases that are clustered together (e.g., students in classes.) All the cases that belong to a cluster share an increment in the outcome--either positive or negative—that expresses the dependencies among them. An appropriate measure of uncertainty in our estimate of the program's effectiveness takes into consideration whether there is more variation *within* the larger units or *between* them. All of our statistical equations include a student-level error term. The variation in this term reflects the differences we see among students that are not accounted for by all the fixed effects and other random effects in our statistical equation.

The choice of terms for each statistical equation is not rigid but depends on the context and the importance of the factors for the question being addressed. The tables reporting the estimates resulting from the computation will provide an explanation of these choices in table notes where necessary for technical review.

---

[1] Although we seldom randomly sample cases from a broader population, and in some situations we use the entire population of cases that is available, we believe that it is still correct to estimate sampling variation (i.e., model random effects). It is entirely conceivable that some part or the whole set of participants at a level end up being replaced by another group (for whatever reason) and it's fair to ask how much change in outcomes we can expect from this substitution.

Reporting the Results

When we run the computations on the data, we produce several results: among them are effect sizes, the estimates for fixed effects, and *p* values. These are found in all the tables where we report the results.

Effect sizes

We translate the difference between program and control groups into a standardized effect size by dividing the average group difference by the amount of variability in the outcome. The amount of variability is also called the "standard deviation" and can be thought of as the average distance of all the individual scores from the average score (more precisely, it is the square root of the average of squared distances.) Dividing the difference by the standard deviation gives us a value in units of standard deviation rather than units of the scale used by the particular test. This standardized effect size allows us to compare the results we find with results from other studies that use different measurement scales. In studies involving student achievement, effect sizes as small as 0.1 (one tenth of a standard deviation) are sometimes found to be important educationally. When possible we also report the effect size of the difference after adjusting for pretest score and other fixed effects, since that adjustment provides a more precise estimate of the effect by compensating for chance differences in the average pretest of the program and control groups. Theoretically, with many replications of the experiment, these chance differences would wash out so we would expect the adjusted effect size on average to be closer to the true value.

Estimates

We provide estimates to approximate the actual effect size. Any experiment is limited to the small sample of students, teachers, and schools that represent a larger population in a real world (or hypothetical) setting. Essentially we are estimating the population value. When we report an estimate in a table, the value refers to the change in outcome for a one-unit increase in the associated variable. For example, since we code participation in the control group as 0, and participation in the program group as 1, the estimate is essentially the average gain that we expect in going from the control to the program group (while holding other variables constant).

p values

The *p* value is very important because it gives us a gauge of how confident we can be that the result we are seeing is not due simply to chance. Specifically, it tells us what the probability is that we would get a result with a value as large as—or larger than—the absolute value of the one observed when in fact there is no effect. Roughly speaking, it tells us the risk of concluding that the intervention has had an effect when in fact it hasn't. This mistake is also known as a "false-positive" conclusion. Thus a *p* value of .1 gives us a 10% probability of drawing a false-positive conclusion. This is not to be confused with a common misconception about *p* values: that they tell us the probability of our result being true.

We can also think of the *p* value as the level of confidence, or the level of belief we have that the outcome we observe is not simply due to chance. While ultimately depending on the risk tolerance of the user of the research, we suggest the following guidelines for interpreting *p* values:

1. We have a high level of confidence when $p \leq .05$. (This is the level of confidence conventionally referred to as "statistical significance.")
2. We have some confidence when $.05 < p \leq .15$.
3. We have limited confidence when $.15 < p \leq .20$.
4. We have no confidence when $p > .20$.

In reporting results with *p* values higher than conventional statistical significance, our goal is to inform the local decision-makers with useful information and provide other researchers with data points that can be synthesized into more general evidence.

Reporting Quantitative Results

The next several panels illustrate our approach to reporting the results of our impact analysis. You should note that footnotes are used to explain technical details of interest only to the technical reviewer. All our reports follow a very similar format, reflecting the fact that, as part of our attempt to improve efficiency, we have standardized on a sequence of reporting. What is illustrated here is a report on the key “Algebraic Operations” scale from the outcome measure. This is the graph previously provided in the executive summary.

Overall effect size

We start with a table representing the entire sample and the effect size that we found overall. This summary of the sample is useful in reminding the readers just how many units the results are based on and the actual means that we refer to. We provide both the unadjusted and adjusted results with a full explanation of what the adjustment consists of.

This table is followed by a figure that helps visualize the nature of the adjustment inherent in the analysis of covariance that we use. The left two panels (used when we have pre- and posttests on the same scale) show growth for both treatment and control groups. We found that readers were often interested in whether the treatment students showed growth. However, we felt that just showing the growth in the treatment would imply that the treatment caused that growth. In this format, we can see that both groups improved and by about the same amount. Our convention for bar graphs is to always set the bottom of the bar at approximately the lowest point in the distribution of the outcome scale (excluding outliers). In this way, we avoid the distortion of showing a truncated bar that makes the difference between treatment and control conditions appear much greater than it really is.

Algebraic Operations

Our next set of calculations addresses Algebra achievement as measured by the Algebraic Operations sub-strand of the test. Table 28 provides a summary of the sample we used in the analyses and the results for the comparison of the *CT* and control groups. The "Unadjusted" row gives information about all the students in the original sample for whom we have pretest and posttest scores. This shows the means and standard deviations as well as a count of the number of students, classes, and teachers in that group. The last two columns provide the effect size, which is the size of the difference between the means for *CT* and control in standard deviation units. Also provided is the *p* value, indicating the probability of arriving at a difference as large as—or larger than—the absolute value of the one observed when there truly is no difference. The "Adjusted" row is based on the same sample of students, but uses the effect estimate from an equation that adjusts for the effect of the pretest as well as fixed effects for group membership above the level of randomization (e.g., pairs). In other words, the adjusted effect size is based on an equation that contains the standard effects that are used in most of the analyses that follow.

Table 28. Overview of Sample and Impact of *CT* on the Algebraic Operations Sub-strand

	Condition	Means	Standard deviations[a]	No. of students	No. of classes	No. of teachers	Effect size	*p* value[b]	Percentile standing
Un-adjusted	Control	225.75	15.28	237	14	11	0.08	.57	3.19%
	CT	227.10	14.03	239	14	11			
Adjusted	Control	225.75	15.28	The same sample is used in both calculations			0.04	.65	1.60%
	CT	226.52[c]	14.03						

[a] The standard deviations used to calculate the adjusted and unadjusted effect sizes are calculated from the scores of the students in the sample for that row.

[b] The unadjusted effect size is Hedges' *g* with the *p* value adjusted for clustering. The *p* value is computed using a model that figures in clustering of students in classes but does not adjust for any other covariates. The adjusted effect size is the impact estimate from PROC MIXED divided by the estimate of the pooled standard deviation. The *p* value for the adjusted effect size is computed using a model that controls for clustering and that includes both the pretest and, where relevant, indicators for upper-level units within which the units of randomization are nested. The *p* value is for the effect estimate from PROC MIXED.

[c] For the adjusted effect size, separate intercepts are modeled for levels of the blocking variable, therefore leading to different estimates for control group performance for each block. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

Figure 3 provides a visual representation of specific information in Table 28. The bar graphs represent average performance using the metric of the NWEA test of Algebraic Operations.

The panel on the left shows average pre- and posttest scores for the control and *CT* groups. The pre- and posttest bars show that both the *CT* and control groups on average improved their Algebraic Operations scores.

The panel on the right shows estimated performance on the posttest for the two groups that includes an adjustment for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled "Adjusted" in Table 28). The overall impact on the Algebraic Operations sub-strand as an effect size (standard deviation units) is .04, which is equivalent to a gain of about 1.6 percentile points for the median control group student if the student had received *CT* . The high *p* value gives us no confidence that the observed difference occurred for reasons other than chance.

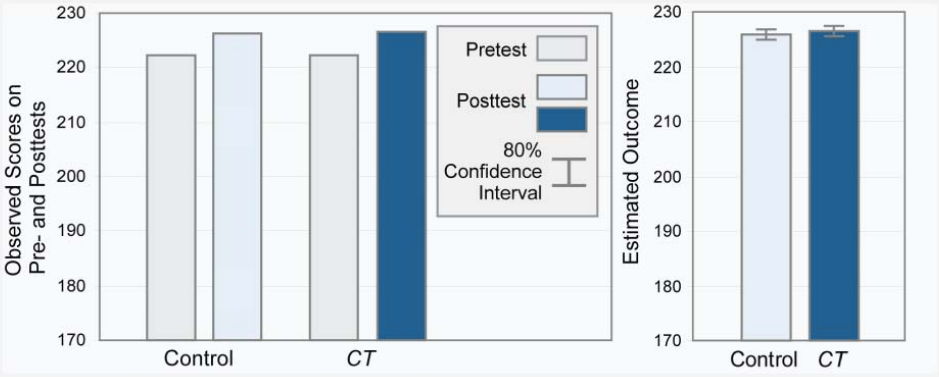


Figure 3. Impact on Algebraic Operations: Unadjusted Pre- and Posttest Means for Control and *CT* (Left); Adjusted Means for Control and *CT* (Right)

Impact Tables and Pretest Interaction Effects

We next report the impact estimates including the interaction with pretest. The arrangement of the rows of the table and the labels are meant to represent the logic of the derivation of the estimates. For example, the first row is frequently labeled "Intercept" in statistical output. However, in terms that would be familiar to our readers, we label it as the outcome for a control student with an average pretest. This value gives us the starting point for the other values that are added or subtracted from it.

Table 29 shows the estimated impact of *CT* on students' performance on the Algebraic Operations sub-strand. For a student with an average score on the pretest, there is roughly a .85-point advantage to being in the *CT* group. The high *p* value of .53 suggests that the observed advantage is easily a chance result; that is, we have no confidence that there is a true advantage. However, we also observe a low *p* value (.02) for the change in the effect of *CT* for each unit-increase on the pretest. This means that the value of *CT* cannot be understood without considering how *CT* and the pretest score work together. Specifically, the higher the pretest score, the lower the impact of *CT*. In other words, there is a diminishing return to the impact of *CT* as the pretest score increases.[1]

Table 29. Impact of CT on Student Performance on the Algebraic Operations Sub-strand of the NWEA General Math Test

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for a control student with an average pretest	229.36	2.43	13	94.23	<.01
Change in outcome for a control student for each unit-increase on the pretest	0.69	0.05	443	13.66	<.01
Effect of *CT* for a student with an average pretest	0.85	1.32	13	0.65	.53
Change in the effect of *CT* for each unit-increase on the pretest	-0.16	0.07	443	-2.37	.02
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Class mean achievement	5.79	4.9		1.18	.12
Within-teacher variation	102.95	6.92		14.88	<.01

[a]Pairs of classes were modeled as a fixed factor but are not included in this table.

[b]Classes were modeled as a random factor.

[1] In the model used, intercepts are modeled as random at the student and class levels and as fixed at the pair level; however, slopes are not modeled as random; the interaction of pretest with treatment and the corresponding *p* value do not reflect uncertainty due to the re-sampling of classes or teachers.

Representing the Interaction Effect

Whenever the interaction effect has a *p* value less than .2, we provide a set of explanatory figures. The scatterplot is a fairly intuitive display. On top of this we present the lines predicted by our model and representing the interaction. In this case, the negative value in the table is shown as a crossover with the line representing the treatment above the control line at the lower end of the pretest scale.

This interaction is most readily interpreted through inspection of graphs. As a visual representation of this result, we present a scatterplot in Figure 4, which graphs student growth over the school year in terms of Algebra achievement as measured by the Algebraic Operations sub-strand of the NWEA General Math Test. This graph shows where each student fell in terms of his or her starting point (horizontal x-axis) and his or her outcome score (vertical y-axis). Each point plots one student's post-intervention score against his or her pre-intervention score. The darker points represent CT students; the lighter points, control students. The shaded area in the lower right of the graph is the area of negative change (i.e., where students lost ground). We are unable to discern an average difference between the two conditions on the posttest. The interaction is evident in the crossing of the prediction lines: it indicates that with these classes and teachers, lower-performing students benefit from treatment; that is, they are helped more by *CT* than are higher-performing students.[1]

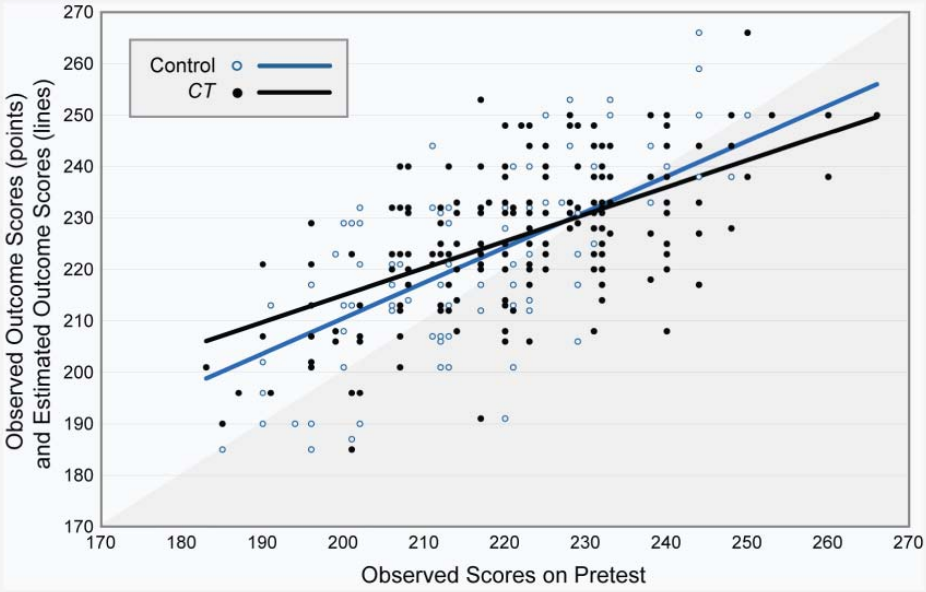


Figure 4. Comparison of Estimated and Actual Algebraic Operations Outcomes for Control and *CT* Students

[1] The apparent dip of the regression lines below the "no growth line" (i.e., into the gray area) towards the top of the pretest scale, and more generally, the non parallelism between the no growth line and the regression lines is an artifact of the regression process (high-performers on average tend to do less well when retested, and low-performers tend to do better) and has no connection to treatment.

The important question then is whether the separation of the two lines is more important at one end of the scale than the other. For this we present a figure representing the difference between the lines and hyperbolae showing the areas of confidence. In this case, we can see that the median student in the bottom quartile has a value that is significantly above zero. The median students for the other quartiles are not significantly different from zero.

Finally, the bar graph that is reported in the executive summary is explained as a representation of the same information found in the prior figure.

Figure 5 displays an alternative visual of the results reported in Table 29 by graphically showing the predicted difference between the *CT* and control groups. The graph is a representation of this separation as a difference, that is, the predicted outcome for a *CT* student minus the predicted outcome for a control student. Around the difference line, we provide gradated bands representing confidence intervals. These confidence intervals are an alternative way of expressing uncertainty in the result. The band with the darkest shading surrounding the dark line is the "50-50" area, where the difference is considered equally likely to lie within the band as not. The region within the outermost shaded boundary is the 95% confidence interval; here, we are 95% sure that the true difference lies within these extremes. Between the 50% and 95% confidence intervals we also show the 80% and 90% confidence intervals. Consistent with the results in Table 29, there is evidence of a positive impact for lower performing students, and little or no impact at the higher end of the pretest scale. (The 95% confidence interval does not cross the horizontal axis for the median student in the first quartile, indicating the presence of an effect; in contrast, the 80% confidence interval crosses the horizontal axis for the median student in the top quartile.)

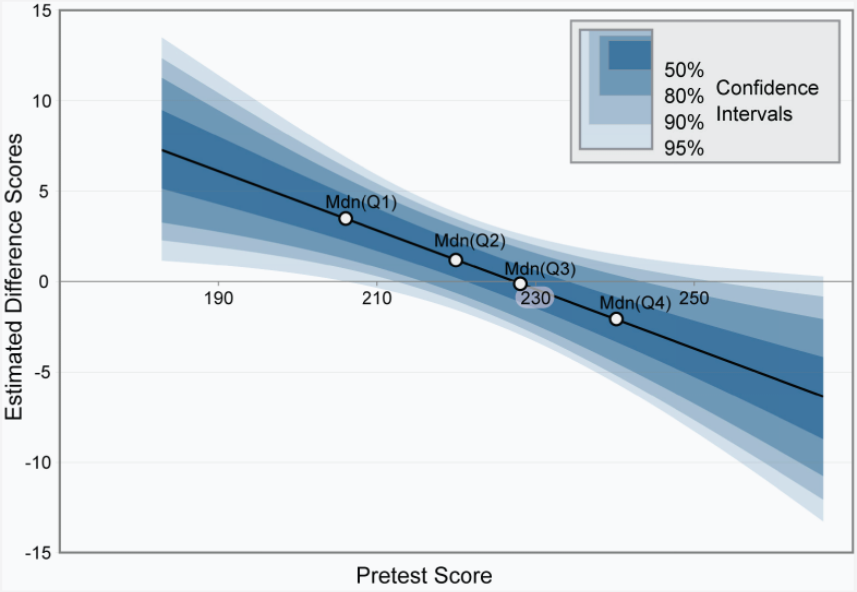


Figure 5. Differences between *CT* and Control Algebraic Operations Outcomes: Median Pretest Scores for Four Quartiles Shown

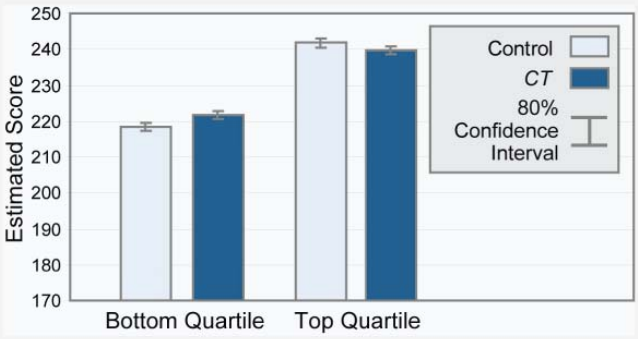


Figure 6. Differences between *CT* and Control Algebraic Operations Outcomes: Median Pretest Scores in Top and Bottom Quartiles

An alternative way of understanding the information in Figure 4 is to represent the result using a bar graph specifically for the students at the median of the top and bottom quartiles of the pretest. Figure 6 presents the estimated difference between *CT* and control for the median student at the two extreme quartiles. Figure 6 indicates that there is an advantage to being in *CT* for the median student in the bottom quartile. The small overlap in confidence intervals for the median student in the top quartile means we have no confidence that such a student would perform differently in the two conditions.

Reporting Implementation

Although we understand the importance to the district of getting timely reports of implementation, we have made less progress in standardizing characteristics to report and reporting formats. In part this is because of the qualitative nature of some of the information from observations and interviews that are pertinent to district decisions. Also, the number of questions that can be presented to teachers is more open-ended than is the case with the student impact results where limited numbers of outcomes are collected. This also gets back to the fact that implementation of the program itself may be better reported separately and earlier, more in the manner of formative feedback to the district. Mediation analysis, which we have only begun to use, provides an interesting alternative to the purely formative value (and potentially confounded interpretation) of implementation measures. The mediation analysis calls for a theory of the "active ingredient" in the processes set in motion by the intervention. Engaging the district decision makers in reasoning about why a program would have an impact may be an important and useful avenue in getting the resulting evidence to be used.

The Unsolved Problem: Using Evidence for Decisions

We started with the premise that school districts could conduct their own local experiments if the cost of doing so were low enough. Our work points to several challenges–layers of challenges, in fact—to be addressed before an improved methodology for conducting low cost experiments can be utilized for local decisions. Because new programs are often specified externally (for example, by a funding source), the decision making focuses on whether to apply for the grant rather than the impact of the program. In many cases where consideration is being given to piloting a new program, a decision is made to go forward before the pilot begins. The pilot serves to get the kinks out of the implementation process. Even when the results of a pilot are reviewed before a final decision is made, questions about ease of implementation, alignment with standards, and teacher acceptance may play a larger role than effectiveness compared to the district's current solution measured in terms of quantifiable gains. Moreover, as Honig & Coburn (2008) document, when scientific evidence about a program is available, it is incorporated into existing decision processes in a variety of ways, including selective reporting to support a program already under way. Our observations and case studies conducted as part of this R&D project add credence to the point of view that understanding district central office decision processes is a necessary step in promoting the use of scientific evidence.

Our project made progress in conducting and reporting experimental program evaluations of local initiatives and in understanding the adaptations of methods appropriate to the context in which the range of generalization is narrowed and the resources are limited to those available within the jurisdiction serving as the unit of decision making. Still, given the layers of challenges to using scientific evidence, a systemic approach that looks more broadly at data-driven decision making as applied to central office decisions is needed.

Data-Driven Decision Making (D3M) at the District Level

We can think of our R&D project as being about data-driven decision making (D3M) at the school district central office level. While we examined a very specific form of study design and data analysis (randomized experiments), the general idea of using local evidence is consistent with what has become a popular topic (Mandinach & Honey, 2008). With the onset of the accountability provisions of NCLB, the growing focus has been on organizing and integrating such school district data as test scores, class rosters, and attendance. While the initial motivation may have been to provide the required reports to the next level up, there continues to be a lively discussion of functionality within the district. The idea behind data-driven decision making is that educators can make more productive decisions if based on this growing source of knowledge.

The major difference between our project and other D3M work is that we are focused on the central office level, whereas most others have focused on teachers using student data to make instructional decisions for individuals. At practitioner conferences such as Consortium for School Networking (March 9-11, 2008 in Washington DC) with its sizable representation by Chief Information Officers from school districts, one prominent speaker asserted that teachers' use of data for classroom decisions was the true meaning of D3M; uses at the district levels to inform decisions were at best of secondary importance. However, at a full-day workshop sponsored by US ED and attended by representatives of most of the Regional Education Laboratories (July 23, 2008 in Washington DC), the focus turned to state-level D3M with some consideration of the questions that can be asked at the state level, given the availability of warehouses of rich data. The parallel between the more common use of D3M and the application to district or state decisions is intriguing.

We are now involved in two experimental studies on D3M as applied to classrooms, but there is little evidence yet that giving teachers access to warehoused testing data is effective in improving achievement. Nevertheless, implementations of this D3M technology is widespread and it is quite reasonable to expect that, with several waves of data available during the year, teachers can become action researchers, working through the following steps: 1) determining where specific students are having trouble, 2) trying out intervention techniques with these individuals or groups, and 3) examining the results within a few months (or weeks). Thus interventions would be based not just on teacher impressions, but also on assessments that provide a measurement of student growth relative to standards and to the other students in the class. If an intervention technique isn't working, the teacher will move to another. And the cycle continues.

Although D3M can be used in similar three-step process at the district level, this practice is much rarer. At the district level, D3M is currently most often used diagnostically to identify areas of weakness. For example, D3M can help district administrators to identify schools performing worse than they should or to identify achievement gaps between categories of students. This is similar to the first step in the teacher D3M. District planners may then make decisions about acquiring new instructional programs, providing professional development to certain teachers, replacing particular staff, and so on. This corresponds to the teacher's second step. What we observe far less frequently at the district level is the teacher's third step: looking at the results so as to measure whether the new program is having the desired effect. In the district decision context this step requires a certain amount of planning and research design. Experimental control is not as important in the classroom because the teacher will likely be aware of any other plausible explanations for a student's change. On the scale of a district pilot program or new intervention, research design elements are needed to distinguish any difference from what might have happened anyway or to exclude selection bias. Also, where the decision potentially impacts a large number of schools, teachers, and students, statistical

calculations are needed to determine the size of the difference and the level of confidence the decision makers can have that the result is not just a matter of chance. “Evidence-based decision making” may be a better term than D3M in referring to data use at the district level where more sophisticated data analysis and even research design are called for.

Hypotheses for Next Stages: Evidence-based Decision Making

Our conclusion from this discussion of data and evidence use in school districts is that the strategy we adopted initially is unlikely to succeed in promoting the use of randomized experiments in school districts. While we can lower the cost, we cannot place the RCT into a context in which it will flourish. A way to understand the problem is that we, as external researchers, attempted to introduce a concept that was foreign in many ways into the processes already in place in the districts. Our discussion of D3M, however, suggests a different approach, which we can characterize more as growing out from the inside. Instead of starting with a method that is poorly aligned with school district procedures, we believe that starting “where the districts are” and moving in logical steps to the idea that data collected and organized in a warehouse can be used descriptively to look at differences—for example, in success rates—and then to the idea that these descriptive statistics can suggest areas of need. The next steps would be to identify resources that may address those needs, and then finally to draw from the same datasets used to identify the problem to track the success or failure of the innovation. At that point, it is possible to introduce experimental design and begin to anticipate an experimental evaluation in the earlier stages of the decision process.

We can posit eight layers or stages in a developmental sequence:

1. **Standardized Assessments**. Consistent testing using standardized measures: high stakes tests, formative assessments, or both.
2. **Student Data Linked to Class Rosters.** Development of a data system that links student results and other demographics with class roster information so that teacher linkages are preserved. Most often development of these systems is motivated by the goal of providing teachers with timely information about their students.
3. **Longitudinal Data Warehouse.** Extending the data system to a longitudinal data warehouse—a step that requires extensive data cleaning and regularization of the data across multiple systems.
4. **Descriptive Needs Assessment.** Use of the data system by district central office administrators to identify areas of weakness through data mining that is done without statistical analysis. Results are then used to allocate resources or to suggest areas where new interventions could improve performance. The areas of interest may be teacher practice, student achievement scores, dropout rates, teacher retention, and other outcomes made available by the data warehouse.
5. **Needs Assessments Based on Statistical Trends**. Going beyond simple data mining based on averages of prior performance to using trends to predict future success or to conduct value-added analysis to identify stronger or weaker teachers or schools. This is a more sophisticated version of the needs analysis in the prior stage.
6. **Quantitative Program Evaluation.** Use of comparative data to track the success of resources shifts and new interventions put in place to address identified problem areas. As program evaluation, this stage represents a qualitative shift because the question is whether the intervention caused a change in the measure of interest. While the data mining in earlier stages was descriptive of areas of strength and weakness, now the district has conducted an “experiment” in trying out a new intervention and it is important to know what would have happened without the intervention. How much of the change is caused specifically by the intervention? At this point, quasi-experimental research designs as well as appropriate statistical techniques are required.
7. **Evidence-based Decision Making.** A next step, which is not always obvious or possible, is to use the evidence from the evaluation to take action to improve the intervention, to feed into

decisions about expanding the program, or even to decide to scrap it and try something different.

8. **Proactive Evaluation.** The final stage is what we call "proactive evaluation," which involves designing the evaluation as part of the implementation plan of the intervention. Here the intent from the beginning is to use the evidence from the evaluation to guide the subsequent steps in the intervention's implementation. It is only at this stage that a randomized experiment can be planned and conducted. The reason is that, unless the evaluation is planned as part of the implementation, the participating schools or teachers would have been identified before a random selection could be conducted.

We view these steps as developmental stages, each building on the prior stage. When we talk about evidence-based decision making at the district central office level, we are referring only to the final two stages.

Our project attempted to move directly to stage 8 without the district itself moving through any of the prior stages. We believe that this accounts for our difficulty in identifying districts with interventions amenable to randomized experiments and, when districts were willing and interested in conducting an experiment, for the fact that their reasons for doing so were not connected to the evaluative purpose of the experiment. It also accounts for why the research effort essentially from beginning to end was taken on by the external researchers (i.e., Empirical Education project staff funded by the grant). As we observed, this often included providing a definition of the district's question that was amenable to a randomized experiment. But included also in conducting a randomized experiment is collecting and compiling the data, conducting surveys and interviews, running the statistical analysis, and preparing a research report. Since our experiment was not building on a district's experience with using data (and in many cases, there was not even a stage 2 data system), the district personnel became spectators with no commitment to using the results in a stage 7 action. Understanding the preconditions for the kind of proactive use of randomized experiments or other research designs that can provide evidence for decision making helps us now to understand an R&D agenda that can begin to develop the internal capacity to use cost-effective methodologies for districts to conduct their own research.

Agenda Moving Forward

Our next logical step is to embark on an R&D program to create a toolkit for district central office administrators that will assist them, over time, in moving through the eight stages. We anticipate the following components.

Documentation of the Developmental Stages

Our eight-stage model of the development of evidence-based decision making is a hypothesis that can be verified descriptively by identifying districts that appear to be at each of the eight specified stages and conducting interviews with the research, IT, and curriculum departments as well as with the superintendent. Tracing the history of their use of standardized test data would provide case study evidence that the stages play out as posited or that they do not. Many alternatives may be discovered in this process, including skipping whole steps or perhaps even following a different order. This part of the agenda provides a basis for the parts that follow, since the pieces of a toolkit must fit into a coherent framework.

We expect that it will be useful to restrict the investigation to districts with populations greater than 15,000 because the application of many research techniques locally within a district requires an adequate number of units (schools and teachers) and because districts of that size are more likely to include a research or evaluation department. It is also important to consider the cost of interventions that may be decided upon. In larger districts, decisions often cost in the millions of dollars and, at that scale, investment in data systems and research is more readily justified.

There are approximately 535 such districts in the U.S. We expect the number of districts at each stage to decrease as we move up the stages, so that locating districts at stage 8 may be a matter of discussions with the leadership of organizations such as the Division H of AERA. Sampling

techniques may be appropriate for getting the proportion of these districts that are at the earlier stages.

Workshops for District Administrators

A series of workshops on evidence-based decision making geared for organizations at different stages of progress toward that goal can provide a roadmap overview as well as details. With NCLB we can assume all districts will have attained stage 1. But districts at stage 2 would need a different presentation than those at stage 6. Districts at stages 7 and 8 may still benefit from workshops, both to introduce more advanced designs and statistical methods and to assist the organization in communicating between departments such as curriculum and research so as to effect proactive evaluation. At the same time, in developing the workshops, researchers from districts at stages 5 through 8 could also serve as advisors and co-developers of the research content.

The workshop conducted in April 2008 by Dr. Mark Lipsey is an example of a workshop aimed at districts having reached at least stage 5. The Institute of Education Sciences RFA it was intended to support also presupposed that respondents' districts had reached at least stage 6 and, more likely, stage 7. Our development work can build on workshops such as this one already developed. By expanding the workshops based on a theory of the development of these capabilities, we would start people where they are and draw them along the path without jumping directly to the end point.

It is important that these workshops also address the organizational aspects of the eight developmental stages. For example, districts at stage 1 may have a research department that only conducts testing and plays no role in discussions of new programs. Its information technology (IT) department may be primarily occupied with collecting data and forwarding them to state-level administrators. By stage 4, however, the research department may begin to play a role in examining patterns in the data and the IT department may have a much more active role in developing specifications for data systems. For these, conversations with the curriculum department would begin to make the connection to the educational issues and how they might be addressed.

Developing Technical Approaches Appropriate for Local Investigations

Our project made progress in the technical domain with respect to randomized experiments. As reported earlier, the shift in focus from a design that begins with a question and a power analysis and then recruits units across a wide range of jurisdictions faces issues different from those faced by a design constrained to the resources available in a particular unit of decision making. Improving our understanding of using and analyzing paired randomizations and other progress in identifying appropriate analytic and reporting approaches has allowed some standardization and efficiencies. However, the current project was limited to randomized experiments and, except for our formative research toward understanding the social and organizational constraints on research, did not develop solutions for any of the stages below the final stage 8, as we now understand the developmental process.

The following are some of the technical questions that must be addressed in future research and development work addressing the needs of school districts approaching stage 5 and beyond:

- Designs and models for quasi-experimental evaluations in situations typical of school district studies. These will include recommendations for matching approaches, conditions under which matches outside the district can be used, and conditions for interrupted time series studies. These are the basic tools needed for the stage 6 program evaluation.
- Designs and models for projections and possibly residual-based value-added analysis of schools within a district as support for a stage 5 type of study. The importance of these tools is not just in their value for supporting more sophisticated needs analyses, but also in the fact that many of the same basic tools of regression analysis come into play in stage 6.
- Methods for cost benefit analyses that can make effect size estimates meaningful in terms of the costs of the intervention and the severity of the problem being addressed.

- Methods for establishing sensitivity to false-positive and false-negative errors on the part of the decision makers so that significance thresholds and expectations can be set appropriately.
- Methods for using formative assessments that are given multiple times during the year. This includes understanding the psychometric requirements of formative tests, the statistical methods for multiple test points, and the appropriate ways that preliminary findings can be reported based on, for example, winter testing that can inform decisions that have to be made prior to the final testing results being available.
- Methods for cost-effective measures of implementation including simple teacher surveys and walk-through methods that can be readily implemented by district personnel. To impact decisions that are on a tight timeframe, quick turnaround reports showing a weak implementation can provide interim information that an impact on students is unlikely to be forthcoming.

Each of these investigations would have to go through phases of literature search, collecting input from educators, and formative testing in the context of school district decision making (including application to actual datasets), and then integration into written documents. Specialized documents on particular topics can be published as working papers and technical reports that can be used in the workshops we envision as well as constituting part of the toolkit we intend to create.

Conclusion

Our project demonstrated that it is feasible to conduct useful randomized experiments on a relatively small scale within a medium-size school district. We also demonstrated that many school districts are not prepared to do so as a means for generating evidence to inform a later decision. When we did get randomized experiments to happen, the expectation often was that the results would be positive and that they would thereby support a direction that was already firmly set. We also encountered an attitude that research was valuable in its own right and worth participating in. While our project focused exclusively on randomized experiments conducted locally, we expect the same would apply to evidence from quasi-experiments as well as to evidence from experiments conducted outside the district. Local randomized experiments are uniquely problematic in that they call for planning the experiment while simultaneously planning the implementation of the intervention.

Our original proposal did not anticipate the widespread disconnect between having evidence based on local data and making decisions based on the evidence. While our experience in many ways reflects the conditions that Honig & Coburn (2008) document through their review of decades of case studies, we do not share a belief that there are deeply ingrained modes of thought that will inevitably distort the use of evidence toward supporting pre-existing positions. Instead we posit developmental stages that begin where most school districts are, collecting the test data required for compliance with federal law, and progress in incremental steps to a stage where randomized experimentation for the purpose of generating evidence to inform a decision is at least feasible.

We see in this sequence of stages one critical transition: going from stage 5 to stage 6. At stage 5 the district is using data to identify needs and has introduced statistical tools. At stage 6, the same tools are applied to following the success of interventions put in place to address the identified needs. This important next step is not inevitable and requires a comparative research design. But we posit that the investment of effort in identifying a need and then putting in place an intervention to address those needs greatly reduces the size of the next step, which uses many of the same tools to continue following the cohort receiving the intervention to see what happens.

We recognize that moving from stage 6 to stage 7, that is, to using evidence to modify the path of the intervention's implementation, is also far from trivial. But this stage depends more on increasing the organizational cooperation between the research and curriculum departments than on technical capabilities. Technical barriers do remain in moving from stage 6 to stage 7, not the least of which

is the ability of the research process to provide timely information. If a decision to expand the implementation must be made in February but the research report is not ready until July, it simply cannot be used as part of the decision process. Fast turnaround and provision of interim reports is essential.

The quantitative research report, even when provided in time to inform a decision, will always be one of many considerations. We expect that quantitative research that uses strong methods and is understood to be credible will have more weight, and ultimately do more good, than reports of research that are inappropriately designed and incorrectly analyzed. Thus, even in the "marketplace" of competing considerations, strong research will provide more value. Considerable work is needed to provide the research tools and the communication tools that can bring strong research closer to the decision process. Ultimately, integrating the research process into the plan for the rollout of a new intervention, which we describe as proactive evaluation, will provide information for incremental improvements, for better targeting of the intervention for the students and teachers who will benefit the most, and for a clear measure of the value it is providing. In this sense, rigorous experimentation can be viewed as a continuous formative process. The task of fully effecting that integration must build on the precursor technologies and activities so that districts can be helped to move from where they are to using evidence to improve the education they provide.

References

Cabalo, J.V., & Ma, B., & Jaciw, A. (2007). *Comparative Effectiveness of Carnegie Learning's Cognitive Tutor Bridge to Algebra Curriculum: A report of a randomized experiment in the Maui School Distric*t. (Empirical Education Rep. No. EEI_EdCT2-06-FR-Y2-O.1). Palo Alto, CA: Empirical Education Inc.

Cabalo, J.V., Jaciw, A., & Vu, M. (2007). *Comparative Effectiveness of Carnegie Learning's* Cognitive Tutor *Algebra I Curriculum: A report of a randomized experiment in Maui School District.* Empirical Education Research Reports, Palo Alto, CA: Empirical Education Inc. May 29, 2007

Cabalo, J.V., Ma, B., & Jaciw, A. (2007). *Comparative Effectiveness of Professional Development and Support Tools for World Language Instruction: A report on a randomized experiment in Delaware.* Empirical Education Research Reports, Palo Alto, CA: Empirical Education Inc. March 1, 2007

Cabalo, J.V., Ma, B., Jaciw, A., Miller, G.I., & Vu, M. (2007) *Effectiveness of Ongoing Professional Development on Interactive Whiteboard Use: A report of a randomized experiment in Forsyth County Schools.* Empirical Education Research Reports, Palo Alto, CA: Empirical Education Inc. January 25, 2007

Cabalo, J.V., & Miller, G.I. (2007). *Technology Emerging Evidence: A Small Study on Activboard Use.* Paper presented in a symposium at the Promethean Summit, Riverside, CA.

Cabalo, J.V., & Vu, M. (2007). *Effectiveness of Carnegie Learning's* Cognitive Tutor *Algebra I Curriculum: A report of a randomized experiment in Maui School District*, Technology Research. Paper presented in a paper discussion at the annual meetings of the American Education Research Association Conference, Chicago, IL.

Cabalo, J.V., Newman, D. & Jaciw, A. (2006). *Effectiveness of TCI's* History Alive! *for Eighth Graders: A report of a randomized experiment in Alum Rock Union Elementary School District.* Empirical Education Research Reports, Palo Alto, CA: Empirical Education Inc. March 2006

Campbell, D.T. (1969) Reforms as experiments. *American Psychologist*, **24**: 409-429.

David, J.L., & Greene, D. (2008) Conducting randomized experiments to inform district decisions: Challenges and lessons. (Being prepared for submission to a refereed education journal).

Greene, D., & David, J.L. (2005). *Implementing Low-Cost RCTs to Support School District Decisions: Formative Evaluation Report for Year One*. Bay Area Research Group, Palo Alto, CA. May 31, 2006

Greene, D., & David, J.L. (2006) *Implementing Low-Cost RCTs to Support School District Decisions: Formative Evaluation Report for Year Two*. Bay Area Research Group, Palo Alto, CA. November 2006

Greene, D., & David, J.L. (2007, April). Challenges in Implementing Randomized Experiments to Support School District Decision Making. In J. David (Chair), *Increasing the Rigor and Relevance of District Evaluations through Federal Policy: Do Randomized Experiments Fill the Bill?* Symposium conducted at the annual meetings of the American Education Research Association Conference, Chicago, IL.

Honig, M. I., & Coburn, C. (2008). Evidence-Based Decision Making in School District Central Offices. *Educational Policy*, 22(4), 578-608.

Hoshiko, B., Jaciw, A., Ma, B., Miller, G.I., & Wei, X. (2007, December). *Comparative Effectiveness of the Texas* Instruments TI-Navigator™*: Year 2 Report of Randomized Experiments in the East Side Union High School District and San Diego Unified School Districts.* (Empirical Education Rep. No. EEI_TI-05-FR-Y2-0.2). Palo Alto, CA: Empirical Education Inc.

Jaciw, A., & Ma, B. (2008, March). *Effects of Pairing on Moderator Estimates: Lessons from Nine Group Randomized Trials.* Poster session presented at the annual meeting of the American Education Research Association Conference, New York, NY.

Jaciw, A., & Newman, D. (2006, April). Experiment-based Evaluations of Pilot Programs in Education: Trade-Offs Between Sample Size and Other Factors. In S. Schellenberg (Chair), *Examining Issues Related to Program Evaluation and Student Achievement.* Symposium conducted at the annual meetings of the American Educational Research Association, San Francisco, CA.

Jaciw, A., & Wei, X. (2007). Heterogeneity of Impact Estimates in Group Randomized Trials: Sensitivity of Inferences about Differential Treatment Impacts to Modeling Assumptions. In A. White (Chair), *Research Design Meets Real-World Educational Contexts.* Symposium conducted at the annual meetings of the American Education Research Association Conference, Chicago, IL.

Jaciw, A., Wei, X., & Ma, B. (2008, March). *Matched-Paired Designs and Standard Errors of Impact Estimates: Lessons from 10 Experiments*. Paper presented in a paper discussion at the annual meeting of the American Education Research Association Conference, New York, NY.

Jaciw, A., Wei, X., Ma, B., & Newman, D. (2007). *Paired Randomization for Greater Precision in Local RCTs*. Poster presentation at the Institute of Education Sciences 2007 Research Conference, Washington, D.C.

Mandinach, E., & Honey, M. (2008). *Data-Driven School Improvement: Linking Data and Learning.* New York: Teachers College Press.

Miller, G.I., Jaciw, A., Ma, B., & Wei, X. (2007, March). *Comparative Effectiveness of* Scott Foresman Science*: A report of randomized experiments in five school districts.* (Empirical Education Rep. No. EEI_PEdSFSci-05-FR-Y1-O.1). Palo Alto, CA: Empirical Education Inc.

Newman, D. (2007). *Generalization and the Unit of Decision Making.* Empirical Education Research Reports, Palo Alto, CA: Empirical Education Inc. March 2007

Newman, D. (2007, April). Generalization and the Unit of Decision Making. In A. F. Feldman (Chair), *The Use of Experimental Designs in Education Research: Current Examples from the Regional Educational Laboratories.* Symposium conducted at the annual meetings of the American Education Research Association Conference, Chicago, IL.

Newman, D. (2007, April). The District Motivation and Design Constraints of Experimental Evaluations. In J. David (Chair), *Increasing the Rigor and Relevance of District Evaluations through Federal Policy: Do Randomized Experiments Fill the Bill?* Symposium conducted at the annual meetings of the American Education Research Association Conference, Chicago, IL.

Newman, D., & Zacamy, J. (2008, October). *Comparative Effectiveness of Carnegie Learning's* Cognitive Tutor *Algebra I Curriculum.* Poster session presented at the annual meeting of the Consortium for Research on Educational Accountability and Teacher Evaluation Conference, Wilmington, DE.