AROI

Academic Return on Investment

FOUNDATIONS AND SMART PRACTICES

VERSION 1 — NOVEMBER 2017 | A PUBLICATION FROM GOVERNMENT FINANCE OFFICERS ASSOCIATION

Government Finance Officers Association

Credits

This paper was written by Shayne C. Kavanagh, Senior Manager of Research at the Government Finance Officers Association, and Nate Levenson, President of the District Management Council.

To get involved in a community of education professionals who are pursuing academic return on investment and other techniques for aligning school district resources with student achievement goals, visit SmarterSchoolSpending.org.

GFOA would like to recognize the following people for contributing their knowledge to its research:

- Matthew Lenard, Director, Data Strategy and Analytics Data, Research and Accountability Department, Wake County Public School System
- Paul Soma, Superintendent, Traverse City Area Public Schools
- David Anderson, Laura and John Arnold Foundation, Director of Evidence-Based Policy
- Jeremy Ayers, Vice President of Policy, Results for America
- Kelly Hallberg, Managing Director, University of Chicago Urban Labs

GFOA would like to recognize the following people for helping to ensure the quality of the report:

- Ron Cabrera, Ph.D., Assistant Superintendent of Instructional Services and Equity, Boulder Valley School District
- J. Scott Gooding II, Executive Director, Budget and Business Services, Columbus City Schools
- Claire Hertz, Chief Financial Officer, Beaverton Public Schools
- Rodney Jackson, Director of Financial Services, Fayette County Public Schools
- Gayellyn Jacobson, Administrator for Fiscal Services, Beaverton Public Schools
- Jimmy Meadows, Jr., Director of School Improvement/Innovation, Fayette County Public Schools
- Arun Ramanathan, CEO, Pivot Learning

203 N. LaSalle Street, Suite 2700, Chicago, IL 60601
312-977-9700 www.gfoa.org

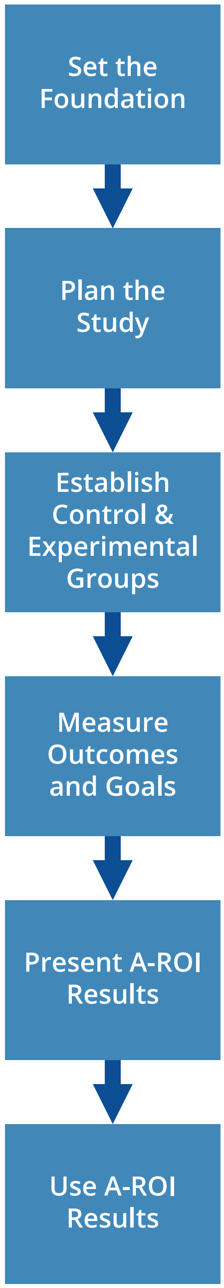
Table of Contents

Executive Summary

Academic return on investment (A-ROI) is the practice of scientifically evaluating the cost-effectiveness of academic programs and then deciding where to allocate resources accordingly. Put more simply, A-ROI is a structured approach to getting the most bang for the buck.

A-ROI has six conceptual foundations.

1. **Reconsider your knowledge of what really works.** A preponderance of research on A-ROI shows that professional judgment is a flawed tool for predicting which programs will be most cost-effective, regardless of the forecaster's experience or degree of specialization.
2. **Define the problem before seeking its solution.** A-ROI requires that we first define what we hope to accomplish through an educational intervention, and then consider the relative costs and benefits of the means to accomplishing those ends.
3. **Follow the scientific method.** The essence of the scientific method is to: 1) form a hypothesis; 2) do an experiment to test the hypothesis; 3) analyze the data; and 4) draw a conclusion. At the core of the scientific method is experimentation, which is the surest way to find out if a program really works.
4. **Seek out the greatest net benefit.** A district will nearly always have multiple options for how to achieve a given student-learning goal. It should choose the option that will provide the greatest gain in student learning for each dollar spent.
5. **Ignore sunk costs.** Sunk costs are the resources that have already been spent on a program. Only the future benefits and costs should be considered when making a decision on whether to fund a program. Investments that were made into an existing program are gone and can't be retrieved (i.e., they are sunk), so they are irrelevant to the decision.
6. **Pay attention to opportunity costs.** Opportunity costs are the benefits that are given up by electing not to undertake an alternative course of action. Paying attention to opportunity costs highlights the benefits that are surrendered when funds are put toward programs that are not cost-effective.



Practitioners, including school district leaders and professional education researchers, have learned a great deal about how to be successful with A-ROI. This paper divides their lessons into six categories that represent the stages of progression through A-ROI, as shown in the diagram to the right.

This paper presents 25 "smart practices" across these six steps. Below are some important themes from across these smart practices.

Establish your own principles of A-ROI. A-ROI is a logical way to make decisions. However, emotions are an essential part of how decisions are actually made. Therefore, A-ROI must speak to hearts and minds. When a district establishes principles, it is deciding what kind of organization it wants to be. This speaks to the passions and values held by the members of the organization.

Recognize that not all forms of evidence are equal. There are many ways to measure the academic impact of a program. This paper discusses some of the key distinctions among methods, but the main point is to have a clear standard of comparison beyond just comparing students' present and past performance. This is because many factors can influence educational attainment other than the program itself, so we must be able to separate out the impact of the program.

Be meticulous about the research question and outcomes. A-ROI analysis should be preceded by careful thought. First, the district should have a clear sense of its student achievement goals. Assessments should be performed on large programs that are closely related to the district's most important goals. High-quality assessments take time and effort, so it is best to focus time and effort where it will matter most. Once the district has determined its goals, it should develop a thoughtful hypothesis about how it might reach them. The hypothesis can then be tested and changed in response to what has been learned.

Make sure the program is implemented well. Too often, school districts get disappointing results from a program because the implementation did not adhere to the original design. Conducting an in-depth A-ROI analysis of a program with a seriously flawed implementation is not usually a good use of time and energy.

Make the results resonate. The results of an A-ROI analysis are necessarily quantitative and technical, but that doesn't mean that the audience can't engage with them. Keys to making the results resonate include telling a story with the data, providing examples of individual students who exemplify the results, involving program staff in the presentation, and making A-ROI a positive, forward-looking experience about finding what works (and not about assigning blame for ineffective programs).

Specify the outcome you are measuring and how it will be measured. Make sure everyone knows exactly how "success" will be defined and what data would be considered proof of success. This can help avoid acrimonious debates later, after the outcome has been measured.

Avoid common decision-making pitfalls. A common pitfall when using A-ROI is "narrow framing," which is framing the choice as an either/or prospect — i.e., either keep the program or get rid of it. Often, districts have many other good options available. Another pitfall is letting short-term emotional considerations crowd out longer-term considerations. Districts can pose questions to decision makers that ask them to step outside of the pressures associated with short-term situations and help them look past those considerations.

Introduction

Academic return on investment (A-ROI) is the practice of scientifically evaluating the cost-effectiveness of academic programs and then deciding where to allocate resources accordingly. The rationale for A-ROI is simple enough: by comparing the learning gains students have achieved from a program with the cost of that program, school districts can get the most bang for the buck with their budgets and do the most good for the greatest number of children.

However, this simple explanation may not be entirely satisfactory. We would be hard-pressed to find many educators who would not agree that it is a good idea to do the most good for the greatest number of children, and most would probably also agree that it is better to do more good, instead of less good, with a given amount of money.

Why, then, do we need a new mode of evaluating and decision making, complete with its own acronym, to reach these seemly uncontroversial goals? The answer to this question has three parts.

The first part of the answer is that A-ROI really does help school districts produce results. Traverse City Area Public Schools (TCAPS), in Michigan (approximately 10,000 enrollment), found that its elementary students were not doing as well in math as they could. TCAPS' initial investigation into the problem suggested the curriculum was a root cause. However, as TCAPS board members put it, buying a new curriculum in the conventional way is like buying a building based on blueprint — full of promise, but also full of uncertainty. TCAPS, therefore, decided to do a year-long pilot study of three new curricula, including a control group that would continue with the old curriculum. Each of the new curricula was backed by research that suggested it would be an improvement on the existing curriculum, and the pilot would reveal just how much improvement TCAPS might expect, and at what cost. TCAPS found that two of the curricula produced statistically significant improvements over the old curricula, and they could compare the prospective costs for making those gains.

The most surprising finding was the enthusiasm the study generated. One board member said, "For the first time in my board tenure, I feel that decisions have been rooted in objective information." The associate superintendent called it "the best experience of my career." A teacher at a TCAPS elementary schools said, "I knew all kids could learn, but I never expected this!" In fact, the first meeting to introduce the new math curriculum to the teaching staff took place during the beginning of summer break and was so well attended that TCAPS found itself short on both seats and handouts. The public also recognized the work TCAPS had done. According to an editorial in the *Traverse City Record- Eagle*, TCAPS' A-ROI analysis "shows commitment to students, parents, and taxpayers."[1]

TCAPS 4th Graders:
"We love math!
We love math!"

The most important fans of the study, though, were the students, as indicated by the new chant at one of TCAPS' 4th grade elementary schools: "We love math! We love math!"

TCAPS' experience also brings us to the second part of our answer as to why A-ROI is needed. The conventional approach to selecting a curriculum is for district leaders to listen to sales presentations from vendors and then use their personal judgment to predict which curriculum will be most beneficial. Personal/professional judgment plays a similarly prominent role in the selection of any other instructional strategy, as well. However, research has shown that human judgment, including that of experts in their fields, is subject to serious limitations when making predictions. One landmark study took place over 15 years and asked 284 experts, in many different fields, to assign probabilities to one of three possible outcomes for questions germane to their fields.[2] The three available choices covered persistence of the status quo, a change in one direction (e.g., growth), or a change in the opposite direction (e.g., shrinkage). The results did not reflect well on expert judgment. A *New Yorker* review of the study put it memorably: "The experts performed worse than they would have if they had simply assigned an equal probability to all three outcomes—if they had given each possible future a 33% chance of occurring. Human beings who spend their lives studying the state of the world, in other words, are poorer forecasters than dart-throwing monkeys, who would have distributed their picks evenly over the three choices."[3] Further, these disappointing results were consistent regardless of area of expertise, experience, or degree of specialization. In other words, greater expertise did not lead to better projections. Later in this paper we will review some of reasons why expert judgment falls short for projecting unknown qualities, but for now, suffice to say that expert judgment is inadequate, by itself, for predicting which programs will provide the most bang for the buck.

This is not an adequate tool for decision making.

The third and final part of the answer to our question about why we need A-ROI is that, in many districts, there is a disconnect between financial decision making and academic decision making. For example, one survey found that only 26% of school district CFOs were involved in decisions about allocating and prioritizing instructional resources.[4] It would be very difficult for a school district to get the most bang for the buck under these conditions. A-ROI requires a partnership between the academic leader ("bang") and a finance leader ("buck"), with both parties working together to build a financial plan and budget that best aligns resources with student achievement goals. For example, the finance and academic officers in Beaverton School District, Oregon, worked together to plan and fund a series of pilots for a summer learning program. This allowed the district to evaluate the program before committing to a full implementation.

The rest of this paper is divided into two major sections. First, we will review the foundations of A-ROI. When you understand the foundations, you can better internalize the A-ROI way of thinking. Second, we will review A-ROI "smart practices." Smart practices are what practitioners have learned about doing A-ROI though hard-won experience, and they will help you put A-ROI into practice in your own district.

The Foundations of Academic Return on Investment

A-ROI has two components: evidence-based decision making and cost-benefit analysis. Evidence-based decision making is the practice of using scientifically rigorous evidence of programs' academic impact to decide which programs have the greatest potential to improve student learning.

Cost-benefit analysis says that the program with the greatest net benefit (benefit minus cost) should be chosen from the set of available programs.[5] These two disciplines are better together than separately. For example, imagine that the evidence tells us that program A produces 1.25 years of reading improvement in one year, while program B produces 1.40 — a 12% greater gain. With just that information, B is the obvious choice. However, what if we also said that B costs 75% more than A? Perhaps the choice is not so clear-cut now, because the money saved by selecting A could be put toward some other worthy goal, like help for students who are struggling with math.

In this section of the paper, we will review the foundations of both evidence-based decision making and cost-benefit analysis. Learning the foundations can change your thought process, helping you make evidence-based decisions.[6]

Foundations of Evidence-Based Decision Making

1. Reconsider your knowledge of what really works. Earlier in this paper we reviewed a study that demonstrated the fallibility of expert judgment in projecting unknown qualities. Why is judgment so unreliable? Much of the answer has to do with what are called cognitive biases. A cognitive bias is a deviation in judgment that is inherent to the way the human mind works and that leads people to draw irrational conclusions from their observations. The challenge with cognitive biases is that they operate unconsciously and their influence is often subtle.

Consider, for example, the overconfidence bias. According to Tali Sharot, a neuroscientist who specializes in this topic, most people overestimate their own capabilities and the chances of good things happening — we are more optimistic than realistic.[7] For example, 70% of people rate themselves as above average in leadership ability, and only 2% rate themselves below average. People routinely underestimate their chances of getting divorced and losing their jobs.

However, cognitive biases can actually confer important benefits. For example, overconfidence helps reduce your stress about undertaking a life-changing event. You will feel better about changing jobs or getting married if you are overconfident about how well these changes will work out. However, when it comes to clear-eyed evaluation of the cost-effectiveness of programs, cognitive biases aren't as beneficial. For example, overconfidence bias might lead us to overestimate what we know about how a program works and its efficacy. Exhibit 1 lists just some of the cognitive biases (and other common logical fallacies) that can cause us to misjudge a program's effectiveness.

Exhibit 1 — Some Cognitive Biases and Other Logical Fallacies

Name	Description	Example
Overconfidence Bias	Unjustified belief in good outcomes or one's own personal efficacy	Overestimating how effective a program is or what we know about how it works
Familiarity Effect[8]	The more people are exposed to a stimulus, the more they will like it (as long as they didn't dislike it to start)	Staff who work closely with a program on a day-to-day basis are likely to think it is effective
Confirmation Bias[9]	Tendency to look for evidence that confirms a hypothesis or failing to look for disconfirming evidence	Interpreting a student's behavior in a way that suggests an intervention is having the hoped-for effect
Representative-ness Heuristic[10]	Assuming a causal explanation for an event, if we can point to an event that resembles it	One instance of a student benefiting from a program is assumed to be representative of all students' experience
Loss Aversion	Weighing a potential loss much more heavily than an equally sized gain.	Giving up an existing program that produces modest gains is difficult, even if its replacement program produces larger gains
Post Hoc Ergo Propter Hoc	Assuming that, because one event preceded another, the preceding event was the cause	A student is given an intervention and the student improves, so it is assumed that the intervention caused the improvement
Apophenia	Developing a seemingly reasonable explanation to make a connection between unrelated objects or ideas	Developing a seemingly plausible explanation for why a program improves student achievement and taking it at face value

These biases can even work together to compound decision-making problems. For example, the familiarity effect and loss aversion combine to create a strong preference for status quo conditions.[11]

2. Define the problem before seeking its solution. A survey of public school superintendents and private employers asked respondents to rate the most important cognitive capacities in the workforce. The superintendents listed "problem solving" first. Their private-sector counterparts listed "problem identification" first, and "problem solving" eighth.[12] Early 20th century education reformer John Dewey said that "a problem well put is half solved"; however, these survey results suggest that some public educators today might be prone to jump to solutions too early, before carefully analyzing the problem and the available choices. A-ROI requires that we clearly define what we hope to accomplish through an educational intervention as a first step, and then consider relative costs and benefits of different means to accomplish those ends.

Essence of the Scientific Method

1. Form Hypothesis

2. Do Experiment

3. Analyze Data

4. Draw Conclusion

To illustrate the importance of defining the problem, consider the experience of an actual school district where low graduation rates made dropout prevention a top priority.[13] Under the assumption that a disadvantaged home life and a lack of academic skills were to blame, the district launched a multitude of programs to remedy the situation. But they remained puzzled when, year after year, the problem lingered with no discernable improvement. They then decided do a detailed analysis of the root causes and found that their initial solutions had come up short because they were predicated on a false, yet seemingly plausible, premise. (See "Apophenia" in Exhibit 1.) This premise was based, in large part, on a few of the district leaders' personal experiences with students (see "Representativeness Heuristic," in Exhibit 1). In fact, the main cause was that incoming freshmen did not realize that failing classes meant an extra year to graduate, unlike their experience in middle school, which had automatic promotion. Armed with the new and accurate understanding, new dropout prevention efforts cut the five-year dropout rate from 5.2% to 1.9% within four years, a 63% reduction.

3. Follow the scientific method. The essence of the scientific method is: 1) form a hypothesis; 2) do an experiment to test the hypothesis; 3) analyze the data; and 4) draw a conclusion. At the core of the scientific method is experimentation, where an intervention is tested on a sample of the student body (the "treatment group") and the results are compared to a sample of students who did not receive the intervention (the "control group"). The advantages of the scientific method are many. First, working with samples reduces the risk of a district spending too much of its money and students' time on an intervention that doesn't work. The risk that an intervention will be ineffective is not a trivial one: an analysis by the Coalition for Evidence-Based Policy found that of 90 rigorous evaluations of educational interventions conducted since 2002, 90% had weak or no positive effects.[14] Second, the presence of a control group is the only way to know whether an intervention really does work. Many factors can influence learning, including teacher quality, student skill, and learning environment. Without a control group, it is much harder to rule out the influence of factors outside the program as a possible cause for improved student achievement. With a control group, we can see if the treatment group outperformed the control group, not just if the treatment group outperformed where they, themselves, started from. For example, consider language development in very young children. If children in a language development program show improvement over the course of a year, without a control group, we can't know if that improvement indicates a successful program or if that improvement was simply due to the natural improvement in language development that most young children would show over a year.

Critically, the scientific method also includes steps for analyzing the data and drawing the conclusion. Seldom is the conclusion from an A-ROI study so straightforward that you would use this alone to justify keeping or abandoning a program. Careful analysis of the data provides a district with the ability to identify more nuanced options and make more strategic decisions. We will address this in detail later in the paper.

Foundations of Cost-Benefit Analysis

1. Seek out the greatest net benefit. A district will nearly always have multiple options for how to achieve a given student-learning goal. It should choose the option that will provide the greatest gain in student learning for each dollar spent — in other words, the most bang for the buck. This means school districts should do the greatest amount of good for the most children.

2. Ignore sunk costs. While our first foundation was self-explanatory, our second is a bit counter-intuitive. Sunk costs are the time, effort, and money that have already been spent on a program in the past. However, only the likely future benefits and costs of a program—not the sunk costs—should be considered when making a decision on whether to fund that program in the future. Investments that were made in existing programs or even pilots are gone and can't be retrieved (i.e., they are sunk), so they are irrelevant to the decision. This foundation is similar to the adage "don't cry over spilled milk."[15]

© CartoonStock

People routinely let sunk costs influence their decisions because they feel they should get some benefit from the time, effort, and/or money they have already spent. So common is this phenomenon that researchers have given it a name: "the sunk cost fallacy." Here are some everyday examples:

- Holding on to a losing stock, when it would be better to sell and invest whatever is left in another security that is more likely to go up in value.
- Staying in a theater to watch a 2.5-hour movie after determining it will be terrible 30 minutes in.

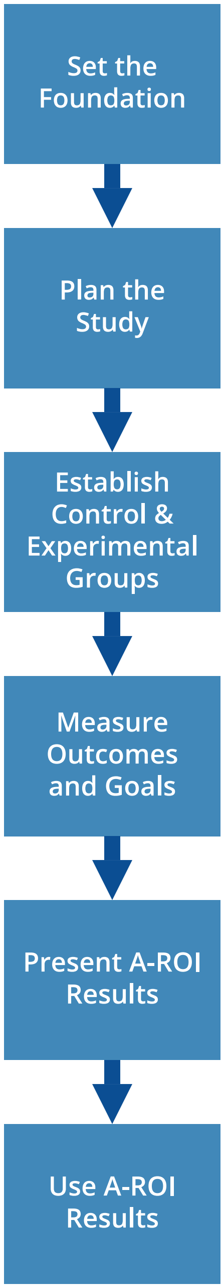
In a school district, an example might include going forward with a program even after a pilot produced disappointing results because funds had already been spent on materials and on training for staff to provide the program. Letting sunk costs influence decisions leads to sub-optimal resource allocation.

3. Pay attention to opportunity costs. Opportunity costs are the benefits that are given up by electing *not* to undertake an alternative course of action. Paying attention to opportunity costs is essential to doing the most good for the most children because it highlights the benefits sacrificed when funds are put toward programs that are not cost-effective.

Reading Recovery, a reading intervention program, is a striking example of opportunity cost. This much praised and often-successful reading intervention has a legion of fans in many districts. The results can be very impressive. But the model limits instruction to one-on-one support for only first graders. Accordingly, this program is very expensive and often consumes a district's entire reading intervention budget. This leaves few resources for other grades and, often, even many first graders go underserved because resources are stretched too thin. The opportunity cost here is the reading help that underserved first graders and students in other grades are not getting because Reading Recovery has consumed those resources.

A-ROI Smart Practices

Practitioners, including school district leaders and professional education researchers, have learned a great deal about how to be successful with A-ROI. This section of the paper divides their lessons into six categories that represent the stages of progression through A-ROI, as shown in the diagram below.



At various points throughout the smart practices section, we will offer illustrations from two school districts: Wake County Public School System (WCPSS) in North Carolina (160,000 students) and Traverse City Area Public Schools (TCAPS, discussed earlier). Both of these districts have taken distinct and successful paths toward using data on cost-effectiveness to inform their planning and budgeting. WCPSS, the more experienced of the two, has followed a systematic approach to analyzing its programs and using the results in its budget process since 2013. It has evaluated more than a dozen programs or policies since that time, including several particularly large and comprehensive randomized control trial studies. TCAPS started more recently. Its first A-ROI analysis was to help the district pick a new math curriculum for elementary students (as mentioned earlier in this paper). This analysis was concluded successfully, and the enthusiasm it generated among the staff and board led TCAPS to launch a new A-ROI study to help pick a new English curriculum.

Set the Foundation before Measuring Anything

Before gathering any evidence or determining any costs, there are number of things that you should do to lay the foundation for a successful A-ROI experience.

Establish Your Principles

A-ROI does not make the hard decisions about where to allocate limited budgets any easier. This is because A-ROI is a means of introducing more objectivity into decision making; however, subjective emotion is at the core of how people actually make decisions. Therefore, it can be easy for emotion to outweigh objective considerations if the two are at odds. Therefore, a district needs to get emotions on the side of A-ROI. The way to do this is to establish core principles that answer questions like: What kind of district do we want to run? What kind of district leaders do we want to be? These are emotional questions that speak to passions and values.[16]

Before proposing to use A-ROI to pick a new math curriculum, TCAPS' superintendent established the principles below with the school board. The plain English, non-technical way in which the principles were stated allowed the superintendent to communicate them clearly and often.

- **Education priorities should drive the budget.** Though it might seem like this principle should go without saying, the superintendent pointed out that, in practice, the budget often drives educational priorities. In many districts, the budget process has a way of freezing in place decisions about curriculum and instruction made years ago. This is because each year's budget is often largely based on historical precedent. Instead, TCAPS should be the kind of district where budget intentionally reflects the most current strategies for providing a world-class education to its learners.
- **You can't be all things to all people.** Again, this principle at first seemed to be a truism to some people within TCAPS, but as they thought more deeply about it they realized that, in many cases, they were trying to be all things to all people. There is probably a tendency for many school districts, as democratic institutions, to try to please as many people as possible. However, becoming a district that delivers world-class education at an affordable cost demands focus.
- **Examine academic return on investment.** Finally, and, in some ways, following from the first two principles, was A-ROI. The A-ROI principle articulated the aspiration to make a practical connection between academic and financial decision making.

WCPSS has a board policy that emphasizes the importance of running experiments to determine the true academic impact of programs. WCPSS' fruitful history with evaluating programs led the school board to adopt a policy to institutionalize it. The policy encourages practices that support A-ROI, such as always running small pilots as a precursor to the full implementation of a new program. WCPSS' policy is available in Appendix 1 of this document.

Recognize That Not All Forms of Evidence Are Equal

Randomized control trials (RCTs) are considered the gold standard for measuring program impact. In a randomized control trial, a group of students who are eligible to participate in an academic program are identified. A portion of those students is randomly selected to actually participate in the new program. The remaining students remain under the same conditions as before. The district then determines which group performs better, and by how much.

An RCT has two key features that make it the gold standard. First, participants are randomly selected. The unique value of random assignment is that that it helps you determine whether the program itself, as opposed to other factors, causes the observed outcomes. For example, imagine that a school district wishes to evaluate the cost-effectiveness of an after-school tutoring program. It starts with a pilot and asks for students to volunteer to come to the tutoring. It is not difficult to imagine that this non-random sample of volunteers could very well be different from the general student body in important ways, such as their motivation to improve. This problem is called "selection bias." It occurs whenever the members of the treatment group are different from the general population in some important way and are, therefore, not a good representation of how the program will perform in general. Selection bias is especially problematic when the members of the treatment group are volunteers or handpicked.

The second feature is the existence of both a control group, the condition of which isn't changed, and a separate group that receives the program (the treatment group). The experimenters can then see if the group that got the program is any better off than the control group. Without a control group, we can't know if any observed improvement is due to the program or to other factors.

RCTs are the most easily understandable and least complicated of the scientifically rigorous forms of evidence. Essentially, an RCT takes two equivalent groups of students, gives the program to one of them, and then checks to see if the students who got the program are better off than those who didn't. This doesn't mean that RCTs are always easy to design and administer, but it does mean that the results are relatively easy to explain to a non-expert audience.

If the RCT gets the gold medal, then quasi-experimental design takes the silver. When RCT isn't possible, the power of statistics and thoughtful program evaluation design is a good alternative, if designed properly. For example, you can search for "natural experiments," or experiments hiding in your existing data. To illustrate, many schools provide a reading program for struggling students (i.e., those scoring below a given threshold, say 300, on a standardized test). Students who score 299 would get extra help, but students who score 301 wouldn't. Since the margin of error on the test is far greater than a few points, students scoring 299 and 301 are actually very similar. An analysis comparing reading growth of students who were just above the cut-off score with those students who were just below it the cut-off score can provide great insight into the impact of the reading intervention.

Another example of a quasi-experiment would be to compare the growth in learning of last year's students to the growth in learning of this year's students, if this year's students were given a new reading program. This second example would provide weaker evidence than the first (reading cut-off scores) because a number of factors in addition to just the program could influence student performance in one year versus another.

The advantage of a quasi-experimental design is that it could be less expensive than an RCT because you are finding quasi-experiments within activities that the district was doing anyway. The disadvantages are that, first, the lack of true random assignment might make the results less valid. In our example of comparing the performance of two successive cohorts of students, we can't rule out some dissimilarity in the environment as a cause of any difference in performance. For example, perhaps a particularly cold winter or bad flu season meant that more students missed school in one of the years. Second, you might have less flexibility in the design, resulting in less information about program effectiveness than

A-ROI and the Law

The gold, silver, and bronze levels of evidence described in this paper correspond to the three levels of evidence that define "evidence-based" in the Every Student Succeeds Act (ESSA).

you'd prefer. In our example of students near the cut-off score, the quasi-experiment would give useful information about the performance of students near the threshold, but not much useful information about how the program affects students further below the threshold.

Because a quasi-experimental design might rely on some clever use of existing data, it might not be as easily understandable to non-experts as an RCT. However, it can still produce compelling findings. For instance, consider this observation from social psychology researcher Richard Nisbett: "Children of parents with little education, and who are therefore at risk for low academic achievement themselves, are likely to have a poor elementary school outcome if their first-grade teacher, judged by observers, is in the bottom third of teaching effectiveness. If they're lucky enough to get a teacher in the top third of effectiveness, their performance is likely to nearly equal the performance of middle-class children.[17] This finding constitutes a natural experiment. If children were to be randomly assigned to classrooms with teachers of different judged competence, we would have a true RCT experiment. Meanwhile, what parent would be indifferent to teacher effectiveness after hearing about the result of the natural experiment?"[18]

The next level is bronze, or correlational studies. A correlational study gathers historical data on "independent" or "explanatory" variables and looks to correlate change in these variables with change in the "dependent" variable, which is the outcome of interest (i.e., student achievement). For example, there have been many attempts to correlate class sizes with improved student achievement using a technique called multiple regression analysis, where researchers see if the data show that students in smaller classes tend to show higher achievement. A correlational study attempts to control, through statistical methods, for all other variables that could provide a competing explanation for a change in the dependent variable. For example, studies that attempt to correlate smaller class sizes with higher student achievement control for variables like average income of families in the district, size of school, and city size.[19] The problem is that, in practice, it is very difficult to control for all of the factors that could affect that independent variable because the researcher must be able to identify them all and then develop measures of the variable that are sufficiently accurate enough to statistically control them. Random assignment to groups avoids this problem entirely because, with a large enough sample, we can safely assume that important differences among individuals are averaged out between the two groups.

The major advantage of correlational studies is that they can be performed using historical data, so they don't necessarily require that the study be designed before the program occurs. The major disadvantages are that correlational studies are of lower validity than RCT and, sometimes, quasi-experiments. Also, the amount of statistical analysis required means that correlational studies can be more difficult to explain to non-experts.

So, What Did the Correlational Studies Say?

The correlational studies on the relationship between class size and student achievement showed no relationship between the two variables.[20] However, more recently, RCTs have shown a modest increase in learning for class sizes of fewer than approximately 20 students.[21] This illustrates two things. First, because RCTs are a superior form of evidence over correlational analysis, we should give the RCT studies more credence. Second, just because there is some gain in student achievement from small class sizes does not mean that all districts should seek to lower class sizes to fewer than 20 students. To do so would be very expensive for many districts, which is why cost-benefit is integral to A-ROI. The question becomes: Is lowering class size the best way to increase student achievement, given a district's available options?

The fourth tier of evidence after gold, silver, and bronze is lead: relying on gut decisions, anecdotes, and or personal observation. Eyewitness accounts are not valid sources of evidence because they are highly vulnerable to the cognitive biases and logical fallacies reviewed earlier in this paper.

Finally, we should mention "data dashboards," a technology solution that allows a district to access the student-learning data more easily by compiling it in a database and providing user-friendly interfaces (e.g., graphics). Data dashboards do not earn a gold, silver, or bronze A-ROI medal, but they do at least get a district into the A-ROI game. They do not earn a medal because they do not provide a rigorous standard for the evaluation of program performance. For example, if reading scores have been improving, we can't know for sure if it is due to a new reading program or to some other factor. In fact, if reading scores are flat or declining, we can't even know if the new program might have prevented the results from being even worse. Our gold, silver, and bronze winners provide a standard for knowing these things (e.g., a control group, statistical controls).

But data dashboards get a district into the game because they are an improvement over gut-level decision making. For example, in one district, cognitive biases had led the superintendent to conclude that a math program to help struggling middle schoolers was ineffective, and that an English program for middle schoolers was effective. He was going to cancel the math program and provide more funding for the English program, but the district had adopted a principle of looking at the data before making budget decisions. When he did, he got a surprise: the math program was effective and the English program was ineffective! He then changed his budget decision accordingly. You can read the full case study in Appendix 2.

Make Use of Third-Party Evidence

Besides conducting your own study, you can consult studies performed by others. Third-party studies offer a number of advantages.

First, they reduce the cost of A-ROI by limiting the need for a district to perform its own studies. For example, TCAPS reviewed third-party studies on the effectiveness of elementary math curricula before conducting its own pilot study. This way, TCAPS was able to limit its test to curricula that had shown positive results elsewhere. Combining third-party studies with your district's own cost information might even allow your district to perform a low-cost version of A-ROI. Popular sources of publicly available information on program effectiveness include What Works Clearinghouse and Visible Learning. In addition, some private research firms can provide information on program effectiveness based on information they've gained from working with districts. However, districts must beware of the dangers of relying purely on third-party research, which we will cover on the next page.

Second, third-party studies might also highlight existing programs where the research does not show the program to be cost-effective. For example, co-teaching is a common way to help those students who need it most. The hypothesis is reasonable: A student who struggles academically and has special needs requires both a teacher who knows the content, such as math, and a teacher who knows special ways of teaching students who learn differently. Hence, schools should have two teachers teach together, bringing together the required expertise. However, the data are disappointing. Nationally, special education students in co-taught classrooms perform slightly worse than those taught by a single teacher,[22] yet co-teaching costs considerably more than other interventions (since there are two teachers). Hence, looking at the studies on co-teaching might inspire a district that uses co-teaching to think about the possibility of finding more cost-effective ways to improve learning for special-education students.

Third, high-quality third-party studies have a strong methodological foundation. This means they will provide decision makers with a good introduction to what solid evidence looks like.

A final advantage is that seeing the evidence on what works expands your understanding of what can be achieved through public education. When you only know what your own programs have achieved, your understanding of what is possible may be narrowed. When you can see the evidence of what others have been able to achieve, it may inspire you to think differently.

Relying on third-party studies also has its limitations. First, third-party research will never be comprehensive; it does not comprise all possible programs or even all available evidence on popular programs. For example, WCPSS compiles its own private database of third-party studies because it has found that publicly available databases, though helpful, are not sufficiently robust to include all available information.

The second limitation of third-party studies is that they are not necessarily the final word on program effectiveness because of the challenges of transferability and faithful implementation. "Transferability" means that the program may have been studied under conditions that differ significantly from conditions in other districts, calling into question the applicability of the study. An example is the "small schools" approach to improving outcomes for high school students living in poverty. Studies showed that small schools that emphasized knowing and building relationships with their students, tailoring instruction to their interests, and setting high expectations had far better results than traditional large, comprehensive high schools. This inspired many districts to create their own small schools, but unfortunately, the model didn't transfer well. The schools in the study were created by charismatic principals who had handpicked the teachers, who were committed to small schools. In other districts, sitting principals were ordered to implement the model, and the existing teaching staff was assigned to the small schools, regardless of whether they believed in, or even understood, the new model. "Faithful implementation" refers to the likelihood of an evidence-based program failing if the implementation is not performed correctly. For example, one study of health promotion programs for children and adolescents found that programs that were implemented correctly achieved two to three times the effects of programs with a flawed implementation.[23]

Build Relationships between Program Staff and the Analysts

Though many districts will need to employ specialized analysts to help conduct A-ROI, the cooperation of program staff is still essential for high-quality research. For example, program staff must not give the program to students in the control group. A-ROI has its best chance for success if it is framed as joint inquiry where the analyst and the program staff work together to find out what is working and what isn't, so that student learning can be improved.

One of TCAPS' elementary principals led its A-ROI study.

It is fundamentally important to set expectations properly. Given that academic programs take years to realize their full potential, a study can't be expected to show that a program has dramatic results right away. Hence, program staff must have realistic expectations for what the research will show. Also, be clear about how the central office will use the information. If program staff perceives A-ROI as a power grab or a budget-cutting tool, cooperation will suffer. All of this means that district leadership will need to be very intentional about how they communicate A-ROI to others in the district. GFOA's *Best Practice in School Budgeting: Identify Communication Strategy* can help district leaders develop their approach to communicating A-ROI.

Districts should also get the right program staff involved in planning the study at an early stage. In fact, TCAPS had one of its elementary school principals lead its math curriculum test, with its analyst providing background support. This provided a powerful advantage for communicating the A-ROI analysis.

Make a Connection between Resource Allocation Decisions and A-ROI

Similar to how expectations should be clarified with program staff as to how A-ROI will affect their work, district leadership should clarify amongst themselves how A-ROI information will be used. For example, WCPSS has a standard form that is completed whenever someone wants to propose a new or expanded program. The form asks the user to show how they have connected the proposed spending to a demonstrated need (i.e., that they've defined the problem), what third-party evidence they've consulted that suggests that the proposed spending will be effective to address the need, and preliminary information about the population served in order to support the design of WCPSS's own study of the program's effectiveness. The form is available at this link.

TCAPS' approach was customized to the situation it was facing — the need to procure a new math curriculum. There was widespread agreement that the conventional way of curriculum selection (i.e., listening to vendors' sales pitches and then a committee picks the one it feels is best) was suboptimal.

Instead, TCAPS would take a more scientific approach by testing out three proven curricula and comparing the results to their existing curricula, as well as comparing the cost of the three options (as discussed previously).

A View from TCAPS School Board

"For the first time in my board tenure, I feel that decisions have been rooted in objective information."

— Megan Crandall, TCAPS Board Vice President

Consider a Program Inventory of Districtwide Programs

A program inventory is simply a list of all districtwide programs. Knowing the universe of programs a district offers can provide some context for evaluation. For example, it helps the district see which programs are the largest relative to others. It can then focus its evaluation efforts on the programs that consume the most resources. It might also be useful to compare the programs in the inventory to third-party research, to see if any of them have proven effective or ineffective elsewhere, or even if there just is a lack of evidence to determine whether or not a program is effective. Programs within school buildings can also be inventoried, but this might be too much work, especially if the district is just starting out with A-ROI.

WCPSS has found that its program inventory is essential to making better decisions. WCPSS staff can use the inventory to determine what supports are in place, whether they are adequate, and whether they are distributed equitably. Gaps and duplications can be identified. WCPSS also thinks about which program to evaluate rigorously based on program size, cost, and the length of time is has been in existence. School site leaders also look across WCPSS to see what kind of interventions other schools have tried. An excerpt from Wake County's program inventory is available in Appendix 3.[24]

Plan the Study

After putting the foundation in place, the next step is to plan a study of a program's effectiveness. The smart practices in this section address ways to pick a program for evaluation and the essential design elements of a good study.

Make Sure the Implementation of the Program Is of High Quality

A-ROI is concerned with comparing a program's impact on student achievement to the cost of the program. However, experience has shown that, all too often, a program that *should* improve student achievement fails to do so because the implementation of the program has not faithfully adhered to key elements of the program's design. Spending the time and effort to research the impact of a poorly implemented program is a waste. In fact, WCPSS considers an evaluation of the quality of program implementation an essential part of its approach to researching program effectiveness, and will not evaluate outcomes without first examining implementation quality.

This smart practice has two specific implications. First, districts should set up systems to measure implementation quality on a continuous basis. For example, they should be able to measure the quantity/quality of inputs and outputs of a program. To illustrate, one district purchased a well-respected intervention, READ 180, which has a track record of success in many districts. Yet this district did not see the expected gains, and a review of the implementation effort revealed why: The required 90 minutes per day for the program had been shortened to 45 because of scheduling constraints; teachers hired after the initial rollout didn't get trained; and students who didn't meet the target profile were still assigned to the program because no other alternatives existed in the district.[25] Simple measures likes "hours of instruction provided" or "percentage of teachers trained" might have revealed that the program was unlikely to produce the anticipated results.

The second implication is that a new program might start out with a very small pilot to make sure the district is capable of implementing the program with fidelity. The overconfidence bias might lead a district to think implementation is practical, when experience might prove otherwise. For example, in the READ 180 pilot, the district would have been able to find out if they could accommodate the time and staff training requirements for the program. A small and well-planned pilot allows the district to work out any implementation hitches before investing in a larger implementation and an assessment of the impact of the program.

Be Meticulous about the Research Question and Outcomes

The decision to assess the impact of a program should be preceded by careful thought. First, the district should have a clear sense of its student achievement goals. Assessments should be performed on large programs that are closely related to the district's most important goals. High-quality assessments take time and effort, so it is best to focus time and effort where they will matter most.

Once the district has determined its goals, it should develop a thoughtful hypothesis about how it might reach them. For example, TCAPS was lagging behind the state average in math scores. Preliminary research into the problem provided strong clues that the curriculum could be a primary cause. For example, children who did well in other subject areas didn't do well in math. In the classes where teachers were having success in achieving better math scores, they had developed their own materials to work around the standard district curriculum. These clues strongly suggested that testing whether or not a new curriculum would result in significant growth in math scores over the existing curriculum would be time well spent for TCAPS.

The research questions should concern the outcomes of greatest importance to students. For example, a study of a remedial reading program should measure reading comprehension, not just the ability of participants to sound out words.[26] This is because intermediate products of the program (e.g., sounding out words) may or may not predict the outcome of real importance (e.g., reading comprehension).

Conduct Forward-Looking Studies

It will almost always be best to conduct A-ROI by setting up a research design first, then capturing data about program performance, and, finally, drawing conclusions. It takes patience to set up a study and then wait for

How to Pick A-ROI Candidates

In addition to the considerations described in this section, if you can answer yes to many of the questions below for a given program, then it might be a good candidate for A-ROI.

- Does the program consume a lot of staff time or money?
- Are the necessary data readily available?
- Are there plans to substantially expand the program?
- Does the program serve a large number of people?
- Is it politically feasible to make changes?
- Is there uncertainty about the program's effectiveness?

the results to play out. For example, TCAPS' math curriculum pilot took an entire year, and WCPSS often evaluates its programs over a multiyear period. However, it is natural for decision makers and other people to want to know how an existing program performs under a rigorous assessment, and to satisfy their curiosity right away by looking at historical data. However, the big problem with historical studies is that, unless a natural experiment happens to be available within the data, the best that can usually be done is a correlational study. As we saw earlier, correlational studies are severely limited by the researcher's inability to assign students to a treatment group or a control group. Instead, the researcher must attempt to control for misleading correlations through statistical means, which is complex and not often possible to accomplish completely, especially when historical data do not contain the necessary information (which they often don't).

Furthermore, the adage "haste makes waste" applies to decision making. Slowing down can lead to better decisions, and A-ROI requires a slower, more thoughtful analysis. Here are some of the reasons why fast decision making can lead to lower-quality decisions:[27]

- We tend to focus on outliers rather than real trends.
- We become blind to longer-term considerations.
- We reach for the first available solution rather than the best solution.
- We are not sufficiently attuned to unintended consequences of the decision.
- We do not build and test hypotheses, so learning does not occur.

Follow the Law of Large Numbers

The "law of large numbers" says that as the number of observations increases, the average value of those observations will get closer to the expected average value for the whole population. To illustrate, the chance of flipping a coin and getting heads is 50%. Hence, theoretically, after flipping a coin any even number of times you should end up with 50% of the flips being heads. However, the chance of getting results much different from a 50/50 split of heads and tails is pretty good with a small number of observations. If you flipped the coin twice, your chance of getting two heads or two tails is not that small — there is a 50% chance of getting either 100% heads or 100% tails. However, if you flipped the coin hundreds of times, your chance of getting results that are drastically different from 50/50 are much smaller. For example, there is a vanishingly small chance of getting 100% heads or tails — far less than a 1% chance, in fact.

Applied to the assessment of educational programs, this means that you will get a much more reliable result from research when you have a sufficiently large number of students participating in a study. There is no rule of thumb about the number of students necessary to get a "sufficiently large sample" because the right number will depend on the nature of the program being evaluated and the degree of uncertainty in the results that the district is willing to live with. That said, often, hundreds of participants will be necessary to get the level of certainty that many districts desire, given the sorts of programs they typically evaluate.

In many cases, it will be more practical to organize an experiment by school site instead of by student. For example, it might be more practical to train the staff in the treatment group to deliver the program if they are all located at the same school. Here, simply flipping a coin to place an entire school building in the control of experimental groups might leave the study open to risk because the vast majority of districts do not have enough school sites to be confident that differences between school sites would average out in a simple randomization process. For example, WCPSS wanted to do an RCT for a differentiated instruction program with 32 participating schools. Simply randomly placing 16 schools in a control group was not sufficient. So, after making sure that the schools understood what an RCT was and that each school had an equal chance of being in the treatment or control group, WCPSS sorted these schools by their existing level of student achievement and matched each one up with another school that had a similar level of existing achievement. Next, one of the schools in each of these pairs was randomly selected to be the control and the other tested the program. This helped to ensure that the control and treatment groups were roughly equivalent at the start of the experiment.

Don't Let the Perfect become the Enemy of the Good

A district should design its research to accommodate the realities encountered on the ground. An uncompromising attitude about methodological rigor might result in an analysis method that is too expensive and/or complicated to ever get off the ground. For example, for the pilot test of its math curriculum, TCAPS had different school sites volunteer to test a different curriculum, with those that were left serving as the control group. This non-random sample had the potential to skew the results. However, the district believed that the enthusiasm this approach generated for its first foray into A-ROI was worth the compromise in rigor, and TCAPS was hopeful that its attempts to design some equivalencies between the volunteer pools (e.g., assign similar schools different curricula) would balance out the non-random assignment to some extent. In the end, TCAPS' A-ROI process received very positive reviews from both the school board and staff. They felt that the decision was far better than it would have been under the traditional model of curriculum selection.

TCAPS' experience illustrates that an A-ROI evaluation does not need to be of Nobel Prize-winning rigor to help a district make substantially better decisions. In essence, an evaluation just needs to follow these three steps: get two equivalent groups (ideally randomly assigned), give one group the program, and see which group does better. That said, districts should remain mindful of the smart practices for designing high-quality studies, or they may find themselves with studies that are too flawed to provide reliable information.

Reduce the Burden That A-ROI Places on Program Staff

Participating in a rigorous evaluation of program effectiveness will often create more work for program staff, particularly to initiate a new program. For example, they might need training on how to apply the intervention. A district should give thought to what can come "off the plate" of program staff when participation in an A-ROI study gets put on the plate. This is important for the evaluation to be a positive experience for program staff. For example, TCAPS stopped pulling teachers out of the classroom in order to provide professional development and, instead, trained each teacher on the math curriculum they would be testing by using live coaching in the classroom, modeling, and co-teaching. TCAPS also worked with teachers to better manage their collaborative planning time. Because the teachers would be spending a lot of time participating in professional development on math, spending more time on math during the collaborative planning time probably would not produce much additional benefit. Focusing planning time on other subjects helped teachers maximize the value of all their available time.

Consider a Partnership with Third-Party Research Organizations

Many districts can benefit by forming a partnership with third-party research organizations to conduct rigorous studies of program effectiveness. For example, some universities have programs that help school districts conduct studies, such as the University of Chicago's Urban Lab or Harvard's Strategic Data Project. Grants may also be available to help districts establish these partnerships.[28] Though the technical ability of universities or other professional program evaluators can usually be taken as a given, districts will need to take more care in ensuring that their partner can communicate the product of a sophisticated analysis in plain English and with actionable conclusions. Establishing an ongoing relationship, rather than contracting for a one-time study, will usually be more successful because each party is incentivized to meet the other's needs over a longer term.

Establish Control and Experimental Groups

After planning the study, the next step is to make random assignments to either the control group or the treatment group, if a district is doing an RCT. If this gold standard of research is not possible, then it is still important that a district establish some standard of comparison.

Address Staff Concerns about Random Assignment Head On

To put it bluntly, the idea of randomly assigning some children to get a potentially helpful program while other children don't get it might not sit well with some district staff. There are a number of strategies that can be used to help make people more comfortable with the prospect of random assignment. The first is to draw an analogy to the way new medications are approved for use by the public. Certainly, no one would want unproven medications to be given to children, no matter how well intentioned the pharmacologist. Similarly, unproven educational interventions could turn out not to provide any benefit, or even to set learning back. Earlier we described how an analysis by the Coalition for Evidence-Based Policy found that of 90 rigorous evaluations of educational inventions conducted since 2002, 90% found weak or no positive effects.[29] Hence, it cannot be assumed that most educational initiatives will necessarily work. If educators want to make a positive impact on children, they will need to relentlessly search for what works and discard what doesn't. Random assignment is the most reliable way to find out what works. Another argument for randomization is the funding pressures that most districts are under. Random assignment in the context of a pilot test allows the district to quickly and efficiently determine the best use of its resources. Also, because the district may not have the money to provide a new program to all students right away, a pilot test with a lottery-like system of random assignment is a fair way to decide who experiences a new program.

A different concern staff might have relates to the potential for wasted time due to a failed test. In other words, if the intervention doesn't succeed, will staff have wasted their time? When some teachers at TCAPS raised this concern, TCAPS pointed out that the teachers in the experimental group were getting extra professional development and being exposed to new ideas that the control group was not. So, even if the intervention did not work out, the teacher would be enriched from their experience. Also, as the Thomas Edison quote below illustrates, failure is often necessary to reach success.

The foregoing are logical arguments for random assignment. However, logical arguments will only carry an idea only so far. There needs to be an emotional component. One possibility might be to appeal to the identity of educators as representatives of fields, like math and science, which value rigorous investigation into what works. For example, the principal and teachers and TCAPS were invested in TCAPS pilot study of math curricula because they were closely involved in helping to distribute and collect surveys and to even analyze the student performance data. Because of her background as a math teacher, the idea of using statistics to resonated with the school principal who led the pilot study.

"If you want to succeed, double your failure rate."
—Thomas Edison

Another way to bring emotions on to the side of rigorous evaluation is to allow program staff to put a handful of students with the greatest need in the program. This at least partially addresses their concern about the neediest students getting help and might even engender positive feelings about the potential of A-ROI studies to help needy students. The program's effectiveness would then be evaluated *without* counting the results from these handpicked students. This strategy preserves the rigor of the evaluation, while addressing emotional concerns about randomization.

Make Sure Assignments Are Truly Random

Beyond simply reaching agreement that, in principle, random assignment is desirable, a district must take steps to ensure that random assignment is carried out in practice. After all, randomly assigning students to a treatment or a control group is an extra administrative step that would not otherwise need to occur. Also, well-intentioned program staff may balk when they come face to face with the prospect of a deserving student being assigned to the control group instead of the group that will receive a potentially beneficial program.

The best way to make sure random assignment is actually carried out is to remove the responsibility for making assignments to a control or experimental group from the hands of program staff. For example, after conducting a very small non-randomized pilot to ensure that it could faithfully implement Achieve3000, a differentiated instruction program for early literacy, WCPSS wanted to do an RCT for the program to see if it had an impact on student achievement. Thirty-two of the district's schools expressed strong interest in the program. WSCPSS' central office randomly selected the schools for the control and treatment groups, but invited the program staff to observe how they did it. This helped give the program staff more confidence in the process.

What is "Random"?

Popular culture has made "random" a synonym for "strange" or "motley," so to say that a "random" selection of students participates in a program may not mean the same thing to a professional researcher as it might to non-experts. Hence, the definition may need to be clarified to make program staff and others more comfortable with random assignment. A good definition of random is that students are chosen by chance. An easy way to conceptualize this is that for each of the students in the study, a coin is flipped. Heads means that the student gets the program; tails means that the student does not. The district then finds out which group does better, and by how much.

Measure Outcomes and Costs

The next step in A-ROI is to measure the outcomes of the program and cost of the program. Both of these measures allow calculation of academic return on investment.

Pre-Specify the Outcome You Are Measuring and How It Will Be Measured

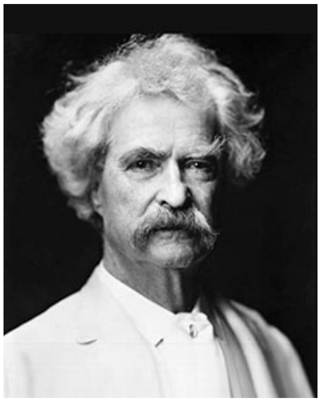
Before measuring, be sure there is a common definition of the outcome that the district is looking for from the program. For example, when evaluating a reading program for struggling students, one district learned that students in the program made eight months of progress in a year. No one debated the figure, but raging disagreements ensued just the same. Some felt that making less than one year's growth in one year was proof of failure because the students ended further behind their classmates than when they started. Others felt it showed progress because some gain is better than no gain at all. Ultimately, they decided that interventions are only successful if the students made much more than a year's gain in a year's time, but a lot of time was wasted and hard feelings were created in the process.[30]

Additionally, districts should decide how they will measure success. For example, WCPSS began an RCT for Nurturing for a Bright Tomorrow, which is a program to increase the rate of gifted identification among minority students. Hence, at every stage of the RCT WCPSS emphasizes that the RCT will measure the extent to which the program increases the rate of gifted identification. As of this writing, the RCT is still ongoing, but everyone is clear on the goal of the RCT and how it will be measured.

Before the A-ROI evaluation starts, it is important to define the academic outcome that will be measured and how it will be measured, in order to ensure the integrity of the A-ROI evaluation. The risk to integrity is the temptation to slice-and-dice all the data associated with a program to see if significant statistical relationships can be found. As a hypothetical, imagine that WCPSS does not find a significant increase in gifted identification, but maybe sifting through the data would show a statistically significant increase in reading scores for the participants. It might be tempting to then declare the program a success for improving participants' literacy.

In the sciences, this practice of sorting through data to find statistical relationships is disparagingly known as "p-hacking." A "p-value" is a statistic used to judge the statistical significance of a finding. For example, imagine that you are evaluating a reading program where the control group scored a 70 on a reading assessment and the treatment group scored a 75. These scores look close, so you might reasonably wonder if the treatment group's superior score was due to the program or just chance. You find that the p-value for the comparison of the two groups is 0.05, which means that there is only a 5% chance that the difference in test scores as large as the one you have observed could occur simply by chance. A p-value of 0.05 is generally considered to be a benchmark for minimally acceptable statistical significance for scientific inquiries, but higher p-values are sometimes considered acceptable.

With a sufficiently large data set it is often possible to find some combination of the data that works out to produce a p-value of the desired size. One group of statisticians showed the potential for p-hacking to produce statistically significant findings for some very questionable propositions: for instance, they found a significant correlation (p-value of 0.0043) between drinking iced tea and believing that *Crash* didn't deserve to win the Best Picture Oscar.[31] Agreeing to what will be measured and how it will be measured eliminates the possibility of p-hacking.



"Lies, Damned Lies, and Statistics"

The line above was popularized by Mark Twain, so it has long been recognized that statistics can be used to mislead.[34] *"Big data" has introduced the potential to mislead with statistics via p-hacking. One might reasonably ask, though: if the numbers show a significant relationship, why is p-hacking considered to be cheating? The reason is that p-hacking is the act of sorting through data, and then only reporting the good (i.e., statistically significant) relationships while leaving out the bad (i.e., not significant). The reported results are too good to be true.*[35] *The scientific method calls for a hypothesis to be formulated and then to find out if the data support the hypothesis. P-hacking tempts us to reverse this process, by finding relationships and then imagining why the relationship might be legitimate.*

Use "Good-Enough" Program Cost Estimates

The accounting methods used to determine the cost of programs can get very elaborate and complex. However, much like a game of horseshoes, close counts with program cost estimates. The estimate does not have to be perfect to enable a better decision — it just needs to be close enough to enable a better decision. Below is a five-step method for estimating program costs using data commonly available from a line-item budget. This method is particularly useful if you have developed a program inventory, as was described in the "set the foundation" step earlier in this paper.[36]

- **Step 1: Distinguish between recurring and non-recurring costs.** The first step is to categorize each line item in the budget as a recurring or non-recurring cost. Examples of recurring costs are salaries, benefits, insurance, office supplies, and materials. One-time costs might include capital improvements and special projects. Differentiating between these two categories allows you to estimate a reliable ongoing cost for a program. Including one-time costs could inflate the perceived cost of a program. Of course, if a program has significant one-time costs to start it up those should be considered,[37] but it is also important to know the difference between start-up cost and ongoing operating costs.

- **Step 2: Distinguish between personnel and non-personnel costs.** Next, line items are further categorized as personnel-related versus non-personnel costs. Any cost that is directly associated with an employee (e.g., salaries, health care benefits, pensions) is a personnel cost. Because personnel comprise the vast majority of the cost for most school districts' programs, just estimating the full cost of the personnel that provide the program will go a long way toward accurately estimating program costs.

- **Step 3: Associate personnel with the programs they provide.** Since people are the largest cost for most programs, the next step is to link each person (or position) with the program they support. An individual might support multiple programs throughout the year, so positions could be divided across more than one program. Many districts do not have records describing how employees allocate their time to different programs. A simple survey of the employee or the employee's direct supervisor can be sufficient to get a serviceable estimate.

- **Step 4: Allocate non-personnel costs to programs.** Non-personnel costs, like equipment, facilities, and information technology, are usually a relatively minor component of total program costs. Therefore, we don't want to use overly elaborate methods of allocating non-personnel costs. In some cases, allocating costs by the number of employees in a program might be good enough. For example, if a given program consumes 25% of the personnel costs for the English department, we might assume that program also accounts for 25% of non-personnel costs. Of course, a district might also choose a more precise method. In any event, the allocation method should bear some relation to the actual resources consumed by the program, as well as being transparent and generally regarded as fair.

- **Step 5: Account for any revenues associated with programs.** After determining the costs of a program, an optional step is to make the connection between the program and revenues it generates for the district, with grant revenue being a leading example. The great benefit of taking into account revenues generated by the program is to enable a decision based on the true cost of the program. The pitfall is that because grants are often temporary, offsetting the short-term cost of a program by its grant revenue may give misleading representation of the long-term cost of the program.

If a school district only wishes to estimate the cost for one or a small number of programs, a method to consider is the "ingredients" method, which entails determining the ingredients required to implement a program and the cost of those ingredients.[38] Typical ingredients include staff time, materials and equipment.

When estimating program costs one must also think about how broadly to define "costs." For example, a broad definition might include the cost of the facilities that classes are located in, the cost of the central office staff that supports the staff who works directly with students, the cost of transporting students to school, etc. The right definition of cost depends on the nature of the decision before the district. For example, in the TCAPS math curriculum test there would be very little value in calculating the cost of the facilities or transportation because TCAPS will be paying the same amount for these items regardless of which direction it might choose to go with its curriculum. Hence, for A-ROI analysis a district will usually be best served by a definition of cost that focuses on the direct cost of providing a new service. Of course, this would include the new, out-of-pocket costs a district would incur to implement the program, such as the cost of purchasing new curriculum materials or hiring consultants to conduct professional development. Districts should also take account of time that existing staff would need to spend implementing the new program (like the cost of teachers' time to participate in training). This is because there is an opportunity cost to the use of staff time. For instance, time spent being trained on a new math curriculum is time teachers are not spending helping students in other ways.

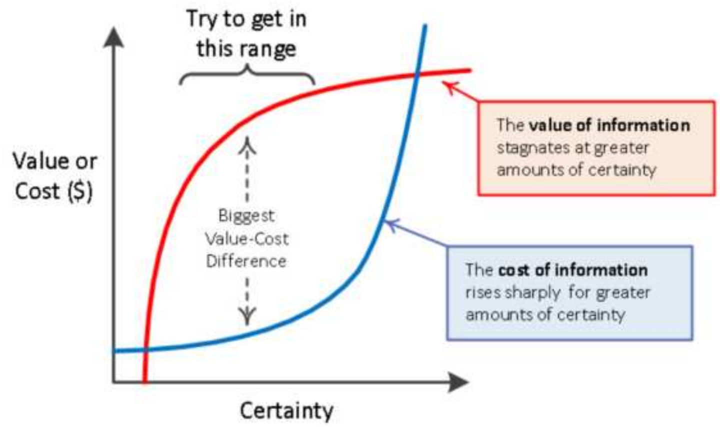
Either of the cost-estimation methods described in this section can be implemented at increasing levels of sophistication to get more precise estimates of program costs. However, districts should be mindful of the increasing expense of generating increasingly precise cost figures and the eventual reduced rate of increase in decision-making value that increasing precision will produce. This is the subject of the next smart practice.

Know the Value of Information: The Yardstick versus the Micrometer

Districts will sometimes need to gather new data to calculate A-ROI. A leading example is program costs, as many districts use line-item budgets that only track costs by objects of expenditure (e.g., salaries, benefits, materials, etc.). However, this could also apply to academic programs. Even for programs that naturally produce outcome data, these are not a perfect representation of program effectiveness — for example, there are margins of error in any standardized test. Collecting extremely precise data, especially where no data existed before, is an expensive proposition. Furthermore, after a certain point, improved precision does little to improve the quality of decisions. For example, would a 2% versus a 3% margin of error on a test make a big difference to the quality of a decision based on the test results?

Hence, a key to making A-ROI analysis practical and affordable is to understand the value of information. More and better information

The Value of Information



Source: Douglas W. Hubbard. How to Measure Anything: Finding the Value of Intangibles in Business. Wiley. 2014

increases the certainty that decision makers can feel about a decision. However, at some point, further increases in certainty add little value to decision making — but it costs a lot of money to make those incremental gains in certainty. This phenomenon is illustrated in the chart above.[39] Hence, school district leaders need to strike a balance between the degree of certainty they are willing to live with and the cost of obtaining that degree of certainty. In many cases, an approximate measure will provide sufficient information to make a better decision.

Beware the Flaw of Averages

Summarizing the results of the hundreds of students who participate in a program in a single average number is an easy way to get a handle on the results of an A-ROI study. However, an average number obscures the variation that might be occurring between the students in the study. [40] This "flaw of averages" might mask that some students do quite well under the program while others make no progress at all, or even go backwards. For instance, one math intervention program studied showed a very impressive and cost-effective *average* gain of 18 months learning. The district could have ruled the program an unqualified success, but a deeper look at the results showed that many students made closer to two years' gain, but one group made just six months' progress, which means they fell further behind. It turned out that for students just one to three years behind the intervention was a success, but not so for students three or more years behind. The district, therefore, learned that it could make even greater student learning gains by limiting the program to students that were three years behind or less, and then using the resources it saved to provide a different intervention to students who were more than three years behind. This example demonstrates the need to look beyond the average results for all participants in the study in order to find out what can be learned from variation between groups.

Present A-ROI Results

A-ROI is, obviously, a quantitative decision-making tool. Not everyone can easily include quantitative information in their decision making, so the results of an A-ROI analysis need to be presented carefully. Below are smart practices in presenting A-ROI results. To illustrate these smart practices, we have developed a model presentation of A-ROI analysis that you can download. We encourage you to compare the slide show that you download to the smart practices. The paper will note particularly useful points, and you might wish to examine the model presentation.

Prepare the Groundwork

Before getting in front of an audience, you should take steps to increase the chances that the presentation will be positively received.

Create a receptive environment for A-ROI. People are strongly influenced by their environment. Hence, the environment can impact how they receive a presentation about A-ROI. Establishing principles that support A-ROI early on is a good first step to creating a welcoming environment. Another powerful step is to establish a budget process where academic priorities, rather than historical precedents, are the driving force behind how resources are allocated. This way, decision makers see A-ROI analysis as essential to making good budgetary decisions. You might recall that TCAPS had "education priorities should drive the budget" as one of its principles, thereby hitting both of these points. They key is to establish a decision-making environment where A-ROI information is integral to making the decision and not just supplementary. The Smarter School Spending process of planning and budgeting is one such approach to doing this.

Build your own credibility. The credibility and trustworthiness of the presenter is essential for the audience to heed the message. However, we often overestimate how trustworthy others perceive us to be. (Remember the overconfidence bias.) Even if you don't have a credibility "problem," it is never a bad thing to have more credibility. The way to increase credibility is to be seen as someone who produces valuable results, who is honest, and who is dependable.[42]

A good starting point is to be familiar with the concerns of the audience. For example, they might be concerned about the impact of the findings on particular school sites, on a particular segment of the student population, or on the job prospects of the staff running the program. They might have an emotional attachment to the program under evaluation and might be concerned about the repercussions of a less-than-positive finding. Taking the time to talk with audience members before the presentation to find out what their interests are prepares the presenter to address those concerns.

Another important strategy for convincing the audience that the A-ROI analysis is valuable is to reference external credentials. TCAPS' close involvement of school principals and teachers in administering and presenting the A-ROI study was a powerful way to build credibility by showing that A-ROI analysis was supported beyond the central office. Another strategy would be to reference authoritative standards like the GFOA Best Practices in School Budgeting or the Every Student Succeeds Act, which advocate that school districts use rigorous standards of cost-effectiveness.

Being perceived as honest starts with examining biases that you might have and making sure that those do not influence how the results are presented. Also, avoid presenting A-ROI information in overly technical terms that audience will not understand, which could be perceived as obfuscating.

Finally, being dependable means making sure that the study results are presented on schedule and that the district's A-ROI policies and practices are applied consistently across the district.

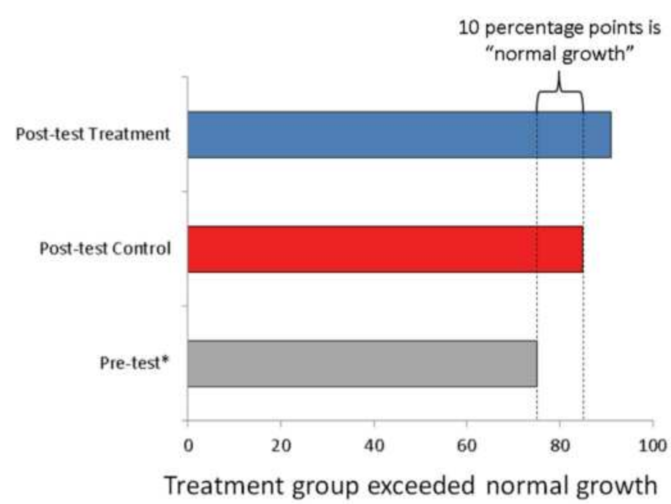
Make the Results Understandable

Because A-ROI presents estimates of student achievement using statistical techniques there is a substantial risk that the presentation of the results of the assessment might get bogged down in statistical jargon. The five-part presentation outline below can help you avoid this problem:

Present the research question. Describe the problem that prompted A-ROI in the first place, the goal of the A-ROI assessment, how the district came to select the program as a candidate for A-ROI analysis, and the time frame of the study.

Show the difference in student achievement between the experimental and control groups. Display the difference in student achievement graphically. The most basic presentation would show the average outcome for the control group versus the average outcome for the experimental group, as depicted in the chart below. A problem with

Average Growth in Proficiency Rates All Schools



**Combined pre and post test group results* the chart is that non-experts may be tempted to interpret any difference in the averages between the groups as proof that the program does (or does not) work. The next point addresses this challenge.

Address the significance of the difference. Just because there is a difference in the average performance of control and experimental groups does not mean that the program is necessarily effective. This is because the difference could have occurred due to chance. A p-value can be used to describe the probability that a difference as large as the one being observed could occur just by chance. Earlier, we mentioned that a p-value of 0.05 is often used as a benchmark by scientists. School districts may wish to settle for a less stringent benchmark. For example, a p-value of 0.10 would provide for a 10% probability that a positive program impact was found by chance. However, a district should not take the decision to accept a higher p-value lightly. Accepting a higher p-value raises the risk of a "false positive" — that is, the district runs a greater risk of concluding that a program has a positive impact when it actually doesn't.

While showing the p-value is necessary, it may not be sufficient because the p-value is not an intuitive concept for many people. It may be wise to supplement the p-value with a graphical representation of the range of outcomes produced by the control and experimental group. The average outcome obscures the variation inherent to the individual students. Showing the entire distribution using a histogram can provide the audience with a better sense of how different the performance of the two groups really was. If the histograms show a lot of overlap, even if the means are different, it demonstrates that the performance of the two groups was more similar than a comparison of the means might suggest. Such a presentation also helps to counteract the "flaw of averages," which was discussed earlier. The model presentation of A-ROI analysis shows an example of histograms.

Address the magnitude of the difference. The significance of the difference does not necessarily bear a relationship to whether the difference between treatment and control groups is a big or small difference. Significance only tells us if any observed difference, big or small, is likely due to chance or not. Decision makers will want to know if student achievement has increased a lot or a little under the program, not just if the observed effect was likely due to chance or not. TCAPS showed the growth in math scores for each of the treatment groups and compared that to the control group. The ability to compare growth under each of the new curricula to the existing curricula was instructive. WCPSS uses standardized effect size statistics.[43] Though these statistics take some sophistication to calculate and interpret, they allow WCPSS to more easily compare the impact of programs across different studies and even across different outcome variables. For example, when evaluating its Achieve3000 early literacy program, WCPSS was able to compare its standardized effect size scores to the results obtained by other researchers in other districts in order to see if the effects of the program were comparable.

Address what it all means. Even with the steps above, the implications of the assessment might not be totally self-evident. The presentation should suggest the next steps, which might not always be as straightforward as full implementation or cancellation of the program. Later, this paper will address the options that an A-ROI analysis would typically present decision makers with.

Also, figures like means and effect sizes do not tell the audience "why" a program did not work. Hence, the quantitative information should be supplemented with a qualitative explanation of why the district got the results that it did.

Make the Results Resonate

The preceding smart practice addressed making a logical presentation of results. However, logic alone does not make a compelling presentation. The presentation must address both the mind and the heart. Here are some strategies for touching the heart:

Tell a story with the data. The last section was about getting across the technicalities of the analysis in an understandable way. However, to really resonate, the presentation should tell a story about the A-ROI study and what it found. For example, the model presentation of A-ROI analysis addresses a hypothetical personalized learning program for 3rd grade reading, "MyGrade3." The story is that MyGrade3 results in improved learning over what the district had been getting, but the district still isn't quite meeting the goal. The district's implementation of MyGrade3 wasn't flawless, which suggests that the district could get more out of MyGrade3 if they address their implementation weaknesses. One of the schools that participated in the pilot has done quite well with implementation compared to the others, so perhaps there are lessons from that school's experience that can be transferred to the others.

Make A-ROI an optimistic, forward-looking experience. Having assumptions about a program's effectiveness proved wrong can be a humbling experience for those that held such assumptions. Focus the presentation on the positive things the district has learned and can do in the future as a result. Do not dwell on where the district had gone wrong in the past. This is not to say the district shouldn't learn from missteps, but it doesn't necessarily need to focus on them when presenting the results of the study.

Present results in a way that will be meaningful to the audience. Imagine that a board member reads an evaluation of the effects of a vocabulary-building program on the reading ability of fifth graders, in which the primary outcome measure was the CAT/5 reading achievement test. The mean post-test score for the treatment group was 718 compared to 703 for the control. According to the report, this difference is statistically significant, but is this a big effect or a trivial one? Do the students who participated in the program read a lot better now, or just a little better? If they were poor readers before, is this a big enough effect to now make them proficient readers? If they were behind their peers, have they now caught up?

The point of this example, adapted from the Institute of Educational Sciences,[44] is that the numbers used to measure student achievement are sometimes not easily interpreted and will not be adequate to tell the audience what they want to know. Hence, the key to helping the audience find meaning in A-ROI analysis is to understand the questions the audience will want an answer to and then present information that answers those questions. Continuing our 5th grade vocabulary program example above, a district could compare the growth in pre-test to post-test for both groups. We already know that the treatment group gained 15 more points than the control. If both groups started at 700 then the treatment group's results are much greater relative to the control: the control group improved by less than half a percentage point, while the treatment group went up 2.4%, an approximately sixfold difference. If both groups started at 600, then the treatment group only outperforms the control group by 15%. A district could also compare these scores to those of students in other districts to get a sense of relative progress. Another possibility would be to show a benchmark score for what is considered reading-proficient on this test. The model presentation of A-ROI analysis shows a variety of methods to put student-learning measures into context.

Whenever possible, involve program staff in the presentation. When program staff help present the results it can help make the presentation more credible. For example, staff in other programs might be less apprehensive about participating in an A-ROI study if they have seen other staff have a positive experience. Also, staff in the program being assessed will likely take more ownership of the results if they are part of the process, rather than passive observers. This was precisely the case at TCAPS, where a school principal led the presentation of its math curricula pilot to the board. The principals and teachers who participated in the pilot were excited by the level of trust shown by the central office and their experience helped inspire the district's staff to undertake a similar pilot for TCAPS' English curriculum.

Principal Jessie Houghton presented the A-ROI results to the TCAPS board.

Provide personal examples that typify the findings. For many people, stories and individual experiences resonate more than numbers. Consider highlighting a few individual students who serve as archetypes of the broader findings of the study. These examples do not necessarily have to be actual students, but could be composites that represent the larger finding. This can help those who are not quantitatively inclined to understand and remember the study results. It might also be a good way to provide insight into why the A-ROI scores came out the way they did. The model presentation of A-ROI analysis provides an example of a student archetype.

Appeal to the audience's identity as educators. Psychological research shows that we perceive others to be more motivated by baser interests (power, status) and less motivated by higher-order interests (doing the best thing possible for students) than they are in reality.[45] The message here is not to underestimate the audience's receptiveness to a message about doing the most good with the money. The presenter's job is to show them how A-ROI can do that.

Put Cost Information in Context

Just as we need some point of comparison to evaluate academic impact (e.g., a control group), we need a point of comparison to evaluate the cost. The best point of comparison is the cost of other programs that are considered to be reasonable substitutes. TCAPS, for example, was able to compare the costs of the various math curricula it was testing. Even if a district is not testing multiple, substitutable programs at the same time, then this method of comparison can still work. For example, if a district was analyzing the A-ROI of a reading recovery program, then it could compare the actual cost of the program with the estimated cost of small group reading instruction led by a certified reading teacher. Even a cost estimate should be close enough to provide useful context. The model presentation of A-ROI analysis provides an example of this type of cost analysis.

In some cases, it might not be as easy to identify a reasonable substitute or estimate costs as in our examples above. The district might only have the total cost of the program to go by. If this is the case, break the total down into "per unit" figures. For instance, maybe a dropout prevention program costs $10,000 per student who participated in the program. This information is helpful, but is still not as informative as it could be. The district might go further by calculating the per-unit cost per dropout prevented. For example, imagine that, in the control group, for every ten students at risk of dropping out, five actually did. Now imagine that in the treatment group that only three dropped out per ten at-risk students. If the program costs $10,000 per student in the program, then it cost $100,000 for ten students. If the program results in a net gain of two dropouts prevented per ten students (five in the control minus three in the treatment group), then it costs $50,000 per dropout prevented ($100,000 to treat ten students divided by a net prevention of two dropouts per ten students). These per-unit figures help put the cost-effectiveness of the program in perspective.

A ratio that compares cost to the benefit of program, like cost per dropout prevented, can be calculated for almost any intervention. For example, TCAPS could have developed a ratio that shows the cost per point gained on the state math test for each curriculum. However, such ratios might be too abstract for some members of the audience, especially if the way in which the benefit is measured is not intuitive (recall the example of measuring the reading ability of fifth graders with the CAT/5 reading achievement test). One way to address this would be to transform the benefit into some more meaningful scale. For TCAPS elementary math curricula, perhaps a meaningful common standard would be to compare growth made during the year under a new curriculum to the growth made under the existing curriculum. We could express the gains as the number of years' worth of growth. Comparing costs across this metric might be more understandable to the audience.

Use A-ROI Results

Once the results of A-ROI analysis have been presented, they need to be used. Establishing core principles that support the use of A-ROI in decision making in the "set the foundation" step will be essential to the rubber successfully meeting the road in this step. Below are other smart practices that support using A-ROI to make better decisions.

Don't Associate A-ROI with Cut-Back Budgeting

The smart practice of establishing core principles to support A-ROI was about connecting A-ROI to positive, aspirational emotions. This smart practice is about disconnecting A-ROI from negative emotions. Many districts find themselves short of financial resources and, as such, are regularly looking for ways to save money. In such an environment, it would be quite understandable for some staff to assume that the goal of A-ROI is to lower expenditures by eliminating programs. Of course, this is not the goal of A-ROI — if a district simply wished to cut programs, there are easier ways to go about it. In fact, A-ROI might result in a funding increase for a program if it shows results. Regardless, a district should assure staff that A-ROI is not a budget-cutting tool (and thus avoid the emotional baggage such a perception would bring). For example, make sure that presentations on A-ROI results come far enough in advance of budget discussions that they are not influenced by cost-cutting pressures. A district might also adopt a formal policy stating that no district employee will lose their jobs as a result of the findings of an A-ROI analysis.

Avoid "Narrow Framing" of Your Decision

Organizational psychologists Chip Heath and Dan Heath describe "narrow framing" of decisions as one most insidious enemies of good decision making.[46] Narrow framing is when we excessively limit the options under consideration. The most common manifestation of this is when choices are framed as "either/or." In case of A-ROI, this would translate to: "we either keep the program or we get rid of it." Presenting such a stark choice can not only make the decision more difficult than it needs to be, it might cause the district to miss some valuable alternatives. An A-ROI analysis usually allows decision makers at least five different choices:

- **Expand the program.** If the program is shown to work very well for a reasonable cost, there is a good case for expanding to wider audience.
- **Keep as-is and continue to monitor.** The A-ROI analysis might show that the program is already serving the right audience and is doing it well.

- **Focus the program where it works best.** The A-ROI analysis might show that the program works reliably well for some types of students, but not others. If so, the district can provide the program only to the types of students for whom it has been shown to work and redirect the remaining resources elsewhere.
- **Fix what isn't working and retest.** Hopefully, a district will have been able to verify that it can faithfully adhere to the technical details of a program's implementation protocols before conducting an A-ROI analysis. However, it is possible that the A-ROI analysis will uncover some flaw or other mitigating circumstance that suggests that a disappointing academic impact can be remedied.
- **Abandon the program.** In some cases, the program simply may not deliver the desired results for the right price.

Attain Distance before Deciding

Psychologists have found that that when we make decisions about our own circumstances we try to account for all of the complexities and nuances inherent in a decision. This can lead to getting lost in the details and letting short-term emotional considerations cloud our judgment. However, when we give advice to others we tend to zero in on the most important factors involved in the decision and to overlook short-term emotions. We can duplicate for ourselves the advantages of giving others advice by taking the perspective of a third party. For example, recall the example of the program that cost $50,000 per dropout prevented. Here are a couple of ways to change the perspective:

- Imagine a vendor offered to charge the school district $50,000 for every dropout their program prevents. What would you do?
- Imagine you are retiring or taking a new job elsewhere. What would your successors do with this program?
- Imagine a friend who works for another district is deciding on what to do with this program. What you advise them to do?

Where to Go from Here

Thank you for reading our paper on Academic Return on Investment. If you would like to learn more about A-ROI or to join with like-minded school districts in using A-ROI, please visit www.smarterschoolspending.org. You will find a variety of tools for A-ROI and can join a community of school districts that are on the journey toward optimizing the alignment between their student achievement goals and financial resources.

Appendix 1 –

Wake County Board of Education Policy on Program Evaluation

Roles and Responsibilities

Data & Accountability Department

The Data and Accountability (D&A) Department will, in conjunction with district leadership, develop and maintain a list of district-sponsored programs to be evaluated under this policy. For the purposes of this policy, programs are defined as all educational initiatives funded and managed at the district level which impact students or staff. Programs may include initiatives currently in operation or initiatives being considered for implementation. This list will be reviewed by D&A staff and district leadership on a regular basis to ensure its accuracy and completeness.

D&A staff will assign programs from that list to different evaluation scenarios based on multiple criteria, including the program's alignment with district goals and objectives, cost, scope, the extent to which data and other structural features of the program are able to support an evaluation, the timing of program implementation (i.e., new vs. existing), and available resources to support evaluation activities. D&A staff will support the following evaluation scenarios depending on how program rates on those criteria, including:

- **Supporting self-monitoring by program staff.** For programs selected for self-monitoring, D&A staff will ensure that program staff is equipped with data collection methods and procedures to support self-monitoring activities. Data collection and reporting activities will be the responsibility of program staff, and should provide actionable management information throughout all phases of the program life cycle. Data from self-monitoring evaluations will be used by department and district leadership to optimize program effectiveness and to inform future decisions about the program.

- **Conducting implementation evaluations.** In collaboration with program staff, D&A staff will conduct implementation evaluations of new and emerging programs. Implementation evaluations will usually occur during the early stages of a program's life cycle. Data from implementation evaluations will result in a written report for program staff that will include recommendations for any adjustments needed to optimize program effectiveness. Data from implementation evaluations will result in a written report for district leadership and will contain an assessment of program status and recommendations regarding possible alterations for improvement. A list of current implementation program evaluation projects will be provided annually to the Board.

- **Conducting impact evaluations.** In collaboration with program staff, D&A staff will conduct impact evaluations of implemented programs. Impact evaluations will occur once a program has had sufficient time to mature such that evidence of the programs ultimate value can be reliably ascertained. Data from impact evaluations will result in a written report for district leadership and the Board, and will contain recommendations regarding whether to continue, alter, expand, or discontinue the program. A list of current impact program evaluation projects and their anticipated reporting timelines will be provided annually to the Board.

D&A staff will ensure that methods utilized for program evaluation are technically sound and consistent with professional standards in educational research and evaluation. In cases where external contractors may conduct program evaluation work on behalf of the district that work will be coordinated and supervised by the Assistant Superintendent of Data, Research, and Accountability or her/his designee. External contractors hired to provide evaluation work may not be affiliated with the service provider whose work is being evaluated to maintain impartiality.

Central Services Departments and School Staff

For programs evaluated via self-monitoring staff responsible for implementing those programs will also be responsible for evaluation with assistance from D&A staff as needed.

For implementation and impact evaluations, Central Services and/or school staff will:

- Establish a clear theory of action including measurable, time-bound goals and an implementation framework which will serve as the basis for the evaluation.
- Facilitate access to data pertaining to relevant program activities, records, and personnel for all district-sponsored program evaluation activities; and
- Provide feedback on report drafts to ensure that evaluations are accurately reflecting program features addressing the key questions of interest and providing actionable information.

When new district programs are being considered for implementation, program staff will consult with D&A staff during the design phase to maximize opportunities to evaluate program effectiveness. To the extent possible, new program adoptions will incorporate random assignment strategies staggered implementations (i.e., “pilot” programs), or other techniques to help support efficacy determinations. New program adoptions will also ensure that the program budget provides adequate support for evaluation activities.

Exceptions

This policy does not affect predetermined evaluation reporting requirements for programs funded by external grants. It also does not apply to programs funded and managed entirely at the individual school level.

Adopted: December 6, 1999
Revised: February 12, 2009
Revised: December 15, 2014

Appendix 2 –

Relying on Your Gut Creates Issues

The English progam looked good. Surely it was succeeding...

A superintendent was presented with an idea by the school district's English Department for a program to help middle-school students who were struggling with writing their English classes. The program would feature small-group learning, with two students per teacher, to provide focused help for students who were a year or two behind their peers in their writing skills, but who otherwise seemed well-suited to catch up. The superintendent loved the idea — not only did it fill a pressing need to improve students' writing skills, but it also it aligned well with his theory on how the school district could best help children, which was to provide intensive, targeted support for struggling students by highly skilled teachers. Therefore, the superintendent and the district made a substantial commitment to this idea: the program was given a dedicated room, complete with new computers, new carpeting, and a new paint job — at a cost of $40,000. Further, four full-time equivalent teachers were dedicated to run the program. Besides the financial commitment, the superintendent showed his personal commitment. On his regular visits to the school buildings, he would make a point of stopping by this program to see how it was going — and he liked what he saw. Students were engaged in orderly and concentrated study, with teachers by their sides.

At the same time, a much smaller investment was made in a math program to help struggling students, which was a hybrid of multiple teaching and learning styles. At any one time, one-third of the class would participate in group lecture with the teacher, a third would work independently, and a third would work on computer-aided lessons. However, this program was only offered as a concession to the math department, who loudly voiced their displeasure with the disproportional amount of resources going to English. Needless to say, the superintendent was not personally invested in this program and when he did go to observe it, what he saw justified his ambivalence: the class was chaotic, noisy, and did not present a productive learning environment. Further, the teacher of the program appeared stressed.

When it was time to build next year's budget, the superintendent was fully expecting to cut the math program. However, examining data on program effectiveness was an important principle of the district's budget process, so it was important to honor this principle and examine the data for these programs. Regardless, the superintendent reasoned, the math program would be cut because the data would show the program's presumably poor results, which would help build support among the rest of district's management for cutting the program. Then he got a surprise: the scores of the students in the math program greatly exceeded expectations. On average, students made 18 months' progress in a year. Meanwhile, there was virtually no detectable improvement in the abilities of the participants in the English writing program, considering both grades and the quality of writing samples. It turned out the noise that the superintendent observed in the math program was actually the natural byproduct of middle schoolers getting excited about something (in this case, math) and the chaos was partially a result of students actually sneaking into the class because they had heard that this was the place where they'd finally conquer math. Conversely, the apparent order in the English class was actually because, as attendance data showed, about half the students were cutting the class (so they weren't there to cause disorder) and the concentrated work between students and teachers turned out to be not much more than a glorified study hall where students would get tutoring on their regular classwork, rather than systematic instruction on how to improve their lagging writing abilities. Perhaps less surprisingly, the English program cost more — almost four times as much!

The decision, then, was clear — the English program was cancelled and a new English program was modeled on the successful math program.

Lessons Learned

Establish your principles. The superintendent's first inclination was to cancel the program based on his gut-level assessment of how the program was performing. However, because the district's management had agreed to a principle of using data to inform their decision making, he was required to look at the data first. Without this principle in place, it is likely that a program that was very good for kids (math) would have been cut and one that was ineffective (reading) would have been kept.

Beware of cognitive biases. Although we can't know exactly which cognitive biases may have affected the superintendent, here are some hypotheticals:

- *Overconfidence bias.* Because the English problem aligned with the superintendent's philosophy on how learning should happen, he overestimated the potential of the English program.
- *Confirmation bias.* When the superintendent visited the classrooms he interpreted what he saw in a way that supported the conclusion he wanted.
- *Familiarity effect.* Repeated visits to the classroom further increased his affinity for the English program.

"We're blind to our blindness. We have very little idea of how little we know. We're not designed to know how little we know."

—Daniel Kahneman,
Nobel Prize-winning
psychologist

Use multiple types of data. Data are an abstraction of reality. As such, looking at just one kind of data gives us an incomplete perspective on reality. For example, looking at both grades and writing samples provided a much better assessment of the reading program than just one or the other. Attendance data also provided a third perspective: the level of student engagement with the class.

Bring in the cost. The academic impact of the two programs was striking, but when cost differential was also taken into account the decision that needed to be made was clear. Imagine that the costs of the two programs had not been presented along with the student achievement data. It is plausible that the district might have allowed the English department to keep its program and make adjustments in an effort to increase student learning next year. However, when cost differences were part of the decision, it is almost inconceivable that the program would have been maintained.

Appendix 3 –

Excerpt from WCPSS Program Inventory

Central Program Name	Grades Included	Response to Instruction (RTI) Service Tier	Description
Academically/ Intellectually Gifted (AIG)	K-12	Tier I	Provides an appropriately challenging educational program for identified AG students. Provides professional development for regular classroom teachers to increase the cognitive challenge of learning opportunities for all students.
Academy of Math	6-12	Tier II	This dynamic program uses systematic instruction that begins with simple concepts and moves to more complex skills. Ongoing assessment and progress monitoring provide robust data to inform instruction and show students' progress.
Academy of Reading	6-12	Tier II	Provides software and teacher support to develop foundational reading skills in students to the level of automaticity.
Accelerated Learning Centers	9-12	Tier II	Students who fail a regular course can retake the course in an online environment.
Child Find (Special Education)	2.3-5	Tier III	Communicates and coordinates with outside agencies who notify WCPSS of children suspected of having disabilities; Meets with parents to determine the appropriate course of action for their child; If eligible, coordinates with service delivery to ensure appropriate setting to meet the child's needs to implement the IEP.
College Prep Success (CPS)	7-9		Analyzes difficult or confusing concepts from core classes, organizes, synthesizes, and evaluates course content. Utilizes high level questioning, discussion and reflection, and collaboration. Explores skills, interests, and personal preferences for career-planning, and ways to stay focused on learning, goal-setting, inquiry, and mathematical reasoning.
Communities in Schools (CIS) Wake	K-12	Tier II	Connects community resources with students in order to provide special enrichment and academic opportunities for students. Works to prevent school failure through mentoring, tutoring, credit recovery, and support for athletic eligibility.
CTE Career Development	6-12		The Career Development Coordinator (CDC) coordinates career development service for all.

Endnotes

1 “Recognition for Due-Diligence Well Earned.” *Traverse City Record Eagle*. May 26, 2016.

2 Phillip E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton University Press: Hoboken, New Jersey). 2005.

3 Louis Menand. “Everybody’s an Expert.” *The New Yorker*. December 5, 2005.

4 Don Hovey and Ulrich Boser. “The New Education CFO: From Scorekeeper to Strategic Leader.” Center for American Progress and Education Resource Strategies. June 2014.

5 Definition paraphrased from: Richard E. Nisbett. *Mindware: Tools for Smart Thinking*. MacMillan. 2015.

6 This statement is based on a study described in: Richard P. Larrick, James N. Morgan, and Richard E. Nisbett. “Teaching Cost-Benefit Reasoning in Everyday Life.” *Psychological Science*. Vol 1, No 6. 1990. Note that in their study, Larrick, et al only addressed the principles of cost-benefit analysis, which we have stayed very close to in this paper. The principles of evidence-based decision-making were derived through interviews with experts in the field. (See acknowledgments section of this report.)

7 Tali Sharot. *The Optimism Bias: A Tour of the Irrationally Positive Brain*. Pantheon. 2011.

8 Robert B. Zajonc. “The Attitudinal Effects of Mere Exposure.” *Journal of Personality and Social Psychology* 9 (1968): 1-27.)

9 Richard E. Nisbett. *Mindware: Tools for Smart Thinking*. MacMillan. 2015.

10 Heuristic originally described in: Daniel Kahneman. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011. Descriptions that appear in this paper are adapted from: Richard E. Nisbett. *Mindware: Tools for Smart Thinking*. MacMillan. 2015.

11 This concept is suggested in: Chip Heath and Dan Heath. Decisive: *How to Make Better Choices in Life and Work*. Crown Business. 2013.

12 Survey data comes from: The Conference Board, “Ready to Innovate: Are Educators and Executives Aligned on the Creative Readiness of the U.S. Workforce?” Research Report R-1424-08-RR (October 2008). Survey and conclusions referenced in: Daniel H. Pink. *To Sell Is Human: The Surprising Truth About Moving Others* (p. 244). Penguin Publishing Group. 2012.

13 The name of the district has been left out intentionally. The example is from the consulting experiences of one of the authors.

[14] Coalition for Evidence-Based Policy. *Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects.* July 2013.

[15] Analogy is from: Richard H. Thaler. *Misbehaving: The Making of Behavioral Economics.* WW Norton and Company. 2015.

[16] This concept is suggested in: Chip Heath and Dan Heath. *Decisive: How to Make Better Choices in Life and Work*. Crown Business. 2013.

[17] Hamre, B. K., and R. C. Pianta. "Can Instructional and Emotional Support in the First-Grade Classroom Make a Difference for Children at Risk of School Failure?" *Child Development* 76 (2005): 949-67.

[18] Richard E. Nisbett. *Mindware: Tools for Smart Thinking.* Farrar, Straus and Giroux. Kindle Edition (p. 297).

[19] See for example: Caroline M. Hoxby. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115 (2000): 1239-85.

[20] See for example: Caroline M. Hoxby. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115 (2000): 1239-85.

[21] In-Soo Shin and Jae Young Chung. "Class Size and Student Achievement in the United States: A Meta-Analysis." *Korean Educational Institute Journal of Educational Policy*. 6 (2009): 3-19.

[22] John Hattie. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* Routledge; 1 edition (December 26, 2008).

[23] Joseph A. Durlak and Emily P. DuPre, "Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation." *American Journal of Community Psychology* 41 (2008): 327–50. This work was cited in: "Implementation Oversight for Evidence-Based Programs" An issue brief from the Pew-MacArthur Results First Initiative. May 2016.

[24] Description of inventory and example adapted from the July 2014 issue of Data Trends, a newsletter produced by the WCPSS Data and Accountability Department.

[25] The name of the district has been left out intentionally. The example is from the consulting experiences of one of the authors.

[26] Example taken from: "Key Items to Get Right When Conducting Randomized Controlled Trials of Social Program." A research report by the Laura and John Arnold Foundation. February 2016.

[27] Dana Weiner. "Making Policy in the Age of Immediacy." A presentation at the Chicago Humanities Festival. November 6, 2016.

[28] See: Researcher-Practitioner Partnerships in Education Research CFDA 84.305H https://ies.ed.gov/funding/ncer_rfas/partnerships.asp

[29] Coalition for Evidence-Based Policy. *Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects.* July 2013.

[30] The name of the district has been left out intentionally. The example is from the consulting experiences of one of the authors.

[31] Christie Aschwanden. "You Can't Trust What You Read About Nutrition." Jan 6, 2016. http://fivethirtyeight.com.

[32] According to Wikipedia, this phrase was popularized in United States by Mark Twain, who attributed it to the British Prime Minister Benjamin Disraeli.

[33] "False-Positives, p-Hacking, Statistical Power, and Evidential Value". A presentation by Leif D. Nelson, PhD, for the June 2014 Summer Institute of the University of California at Berkley's Berkley Initiative for Transparency in the Social Sciences.

[34] Method originated by the Center for Priority Based Budgeting. Description of the method adapted from: Chris Fabian, Jon Johnson, and Shayne Kavanagh. "The Challenges and Promise of Program Budgeting." *Government Finance Review.* October 2015.

[35] The most appropriate treatment of these costs will depend on the particular of a situation. In some cases, it might be wise to amortize the costs of a multi-year period. In other cases, if the costs are already sunk, it might not be wise to include them in the cost.

[36] H. M. Levin. "Cost-Effectiveness and Educational Policy." *Educational Evaluation and Policy Analysis*, 10(1): 51-69. 1988.

[37] The "value of information" concept and the graphical representation of it adapted from the work of Doug Hubbard. See Exhibit 7.7 in: Douglas W. Hubbard. *How to Measure Anything: Finding the Value of Intangibles in Business.* Wiley. 2014.

[38] "Flaw of Averages" concept originated by Sam Savage in: Sam Savage. *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty.* Wiley. 2012.

[39] Roderick M. Kramer, "Rethinking Trust," *Harvard Business Review*. June 2009.

[40] Ulrich Boser. *The Leap: The Science of Trust and Why It Matters.* (Amazon Publishing: New York, New York). 2014.

[41] The reader can learn more about these statistics and how to calculate them in: Mark W. Lipsey, Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick. "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." Institute of Educational Sciences. November 2012.

[42] Mark W. Lipsey, Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick. "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." Institute of Educational Sciences. November 2012.

[43] Heath and Heath. *Made to Stick*. 2007.

[44] The other "enemies" that Chip Heath and Dan Heath Describe are confirmation bias, short-term emotion, and overconfidence bias, all of which are addressed in this paper. For Chip and Dan Heath's work, see: Chip Heath and Dan Heath. *Decisive: How to Make Better Choices in Life and Work.* Crown Business. 2013.

[45] The concept of attaining distance before decided was originated by Chip Heath and Dan Heath. The science behind why perspective-taking works also was described by Chip Heath and Dan Heath. Chip Heath and Dan Heath. *Decisive: How to Make Better Choices in Life and Work.* Crown Business. 2013.

About GFOA

Government Finance Officers Association (GFOA), exists to promote excellence in state and local government financial management. GFOA views itself as a resource, educator, facilitator, and advocate for its more than 19,000 members and the governments they represent. GFOA provides best practice guidance, leadership, professional development, networking opportunities, award programs and advisory services, concentrated in the following areas:

- Accounting, auditing, and financial reporting
- Budgeting
- Capital planning
- Debt management
- Financial leadership
- Pension and benefit administration
- Treasury and investment management

Government Finance Officers Association

203 North LaSalle Street, Suite 2700
Chicago, IL 60601-1210
312-977-9700 | 312-977-4806 FAX | www.gfoa.org