Solving the Funding-Achievement Puzzle in America's Public Schools

Schoolhouses, Courthouses, and Statehouses

Eric A. Hanushek and Alfred A. Lindseth

7 Science and School Finance Decision Making

Courts and legislatures alike find themselves looking for scientific research and evaluation to help in making decisions about education policy and appropriations. Unfortunately, while there are plenty of consultants, researchers, academicians, advocates, vendors, and others ready to help answer the key questions, the information they tend to provide is frequently flawed, one-sided, or misleading—pushing decision makers in costly and ineffective directions.

The twin questions policymakers and courts continuously grapple with are these: "What should money be spent on in order to achieve the desired results?" and "How much should an adequate education cost?" If science answers the difficult factual questions about what different programs accomplish and their cost, then courts or legislatures can concentrate on setting appropriate outcome goals and finding revenue sources.[1]

Unfortunately, answering these questions is not so simple. While science is potentially a source of reliable, objective information about programs and their expense, applying scientific methods to complex educational and funding decisions is fraught with problems. Necessary data may be unavailable; or they may be unreliable; or they may be subjected to unreliable analysis that fails to yield the desired information—and judges and legislators may not appreciate the importance of these issues. Particularly when presented with conflicting analyses, they might simply conclude that scientists frequently disagree, and thus they are free to choose whichever evidence matches their preconceptions or meets their purposes. But this conclusion can lead to very bad decisions—ones that are ineffective and expensive.

We begin with a simple summary: much of the information entered into educational debates does not pass muster as reliable and acceptable scientific evidence. We illustrate this by considering standard analyses introduced to answer the "how much" question and then turn to illustrative analyses of "how" to spend resources.

A Simple Decision Model

With a fixed amount of money to spend, devoting more money to one item means that less money is available for another. This simple truth leads to a general rule about how to allocate funds across different resources or policies in order to get the highest achievement: spend additional funds on whatever yields the highest added achievement.

For example, suppose that by spending an additional $1,000 per pupil, a school could attract a high-quality teacher and improve student achievement by five points, or it could spend an extra $1,000 per pupil to reduce class size and improve achievement by one point. Obviously, spending the money to improve teacher quality is the better choice (assuming that these are the only two choices). We can calculate that improving teacher quality costs $200 for each point of increased achievement, while reducing class size costs $1,000 for each point of increased achievement.

What happens if we do not choose the most productive use of funds? If we put the $1,000 a student into class size reductions instead of teacher quality, that spending yields only one point of achievement. But would we still say that a point of achievement "costs" $1,000? Of course not, because we know we can buy an additional point for just $200 if we spend the additional funding on teacher quality instead of reducing class sizes.

Now translate these ideas to the situation at hand. States and local districts must choose among a variety of educational resources or policies. The added dollars should be spent where they have the biggest payoff—i.e., the "biggest bang for the buck." Moreover, the "correct answer" may change as more money becomes available to the district. For example, for a low-spending

district, buying new textbooks might initially yield the largest improvements in achievement, but after everybody has their own new textbook, added spending on textbooks might do nothing for achievement, and something else will yield a more productive use of funds.[2]

What if the added funds were spent on hiring more teachers with master's degrees, a choice that, experience shows, is unlikely to have any significant effect on student achievement? Under this scenario, no matter how much was spent on master's degrees, a five-point gain in achievement would be unlikely.

This discussion illustrates two common errors that have been made over and over again in estimating what an "adequate" education should cost. First, the determination of true "cost" cannot be separated from the efficient use of funds. In estimating costs, we cannot ignore more economical or efficient ways of obtaining the desired achievement gains. Second, determining the funds that would be required to obtain some achievement goals depends upon accurately assessing the impact on achievement that each input or policy might have. This frequently proves difficult, particularly since the available research is often insufficient for projecting the impacts of varying programs on student achievement.[3] With this background in mind, we move to a discussion of various approaches to answering the fundamental question raised inevitably in consideration of school funding: how much will an adequate education cost?

How Much Is Enough?

Both legislatures and courts have sought to ensure that schools have enough money to enable students to reach desired performance levels. In court, plaintiffs present "costing-out" studies as the scientific means of determining appropriate levels of funding; the implication, sometimes baldly stated, is that the process relied on by legislative bodies is both irrational and arbitrary. Costing-out studies have been conducted or are in progress in a majority of the states, and the demand for new ones has continued to rise as

adequacy lawsuits proliferate.[4] Plaintiffs have discovered that there is great value in submitting a specific dollar amount for total required state spending—an amount, they argue, that should be treated as both necessary and sufficient for an "adequate" education.[5] Courts have clearly been influenced by this strategy, as judges have been willing to incorporate the results of such costing-out studies into their remedies or to order that such studies be conducted to guide the legislature in setting appropriation levels.[6] Indeed, during the liability phase of adequacy cases it has become commonplace for plaintiffs to disavow having any views on remedies and instead to ask simply that a costing-out study be commissioned so that everybody can proceed with knowledge of exactly what an adequate education would cost.

Legislatures also commission such studies to guide their appropriations, particularly when they are under pressure from the courts. They also sometimes request such studies on their own, thinking perhaps that outside experts can provide them with objective and scientific recommendations. In some cases, studies are commissioned with the hope that by going through the seemingly rational process of "costing out" an adequate education, the state can protect itself against a threatened adequacy lawsuit. This strategy, however, backfires more often than not. Such studies almost always call for sizeable increases in spending, which politically the legislatures are not ready to approve. However, once the legislature has commissioned its own study, these cost estimates often become exhibit A to the plaintiffs' lawsuit. Moreover, any estimates from these studies gain a higher profile by virtue of having been sponsored and sometimes endorsed by the legislature or state educational authorities themselves. Indeed, in the Kansas adequacy case, the court held that the results of a study by an outside consultant were conclusive against the state, even though the legislature itself had never endorsed them.[7]

Calculating what an adequate education should cost sounds like a straightforward undertaking, but it is anything but. While educators may have an idea of which programs and measures will result in better test scores (although as we have argued throughout much of this book, the evidence is, at best, mixed), there is no consensus

in the scientific community on what it would cost to overcome the adverse effects of poverty and other social ills that now keep many children from achieving at levels comparable to middle-class children. The overarching problem —facing courts, legislatures, and the consultants hired by them—is the nonexistence of empirical evidence on which to base estimates of the costs of attaining desired levels of student proficiency, particularly for the most disadvantaged students—the central focus of most court cases. And there is no evidence because the outcome sought—high achievement for all or most students in a district—has not been achieved before, except in isolated instances.[8]

Calculating the cost of an adequate education would be simple if scholars could consistently show something like the following: an additional expenditure of $1,000 per pupil appropriately spent will translate, on average, into a five-point gain in student proficiency. Unfortunately, decades of research have not been able to show a clear causal relationship between the amount that schools spend and student achievement; certainly our analysis in the previous chapter showed none.[9] After hundreds of studies, it is now generally recognized that *how* money is spent is much more important than *how much* is spent. If schools do not spend additional money effectively, that spending is unlikely to have any effect on achievement. It remains difficult, if not impossible, to infer from current spending patterns what it would really cost to change achievement, even assuming efficient use of funds (or for that matter, continued inefficient use of funds).[10] This finding is particularly important in considering judicially ordered changes in school finances, because such orders seldom require or even suggest that money be spent differently from how it has been spent in the past.

It is, of course, possible to wonder why determining the amount of money needed to offer a sound education is so different from other appropriation decisions. After all, government officials repeatedly estimate how much it will cost, for instance, to build roads. But roads and schools really are different. We not only know the outcome we want—a new highway—but we can also identify with reasonable certainty the resources needed to achieve that outcome—so much asphalt, so many hours of labor, and so on. That

is, we know the technology, and we can fairly accurately specify the recipe for a new road. We also know from experience what those resources will cost.

Little of this reasoning applies to the goal of providing an adequate education—largely because no coherent description of what it would take to meet the goal currently exists. Partly it reflects the fact that the process is so dependent on the skill and actions of the primary actors—teachers, principals, and of course the students.[11] Because it is difficult to measure or specify differences among these primary inputs, it is not feasible to identify just what would be required from schools to reach any level of achievement. Additionally, a multitude of factors outside the school's control affects performance. The child's ability, the education of the child's parents, their involvement in their child's education, the resources in the home, how much the child studies, how much TV the child watches, the child's motivation, the child's health, and a host of other circumstances beyond the control of the school authorities all enter into the equation. In building the road, of course, a variety of outside factors such as terrain, soil type, and weather conditions also affects construction costs, but these factors are not as dependent upon actions of direct participants in the project, and, again, the technology for dealing with them is better understood than with education.

Providing an adequate education, like building a road, also depends on the efficiency with which resources are used. But in public education, this efficiency factor plays a much more central role in thinking about costs than it does either in the private sector or in other governmental sectors.[12] The road contractor not only must competitively bid for the contract, but also must go on to build the road for his contract price if he expects to make a profit; therefore he has a strong motive to make the most efficient use of his money. (Note that when there is just one bidder, we tend to be more skeptical about whether the bid truly reflects the minimum cost of the project.) Public schools face much different constraints. They are a governmental monopoly, and parents, especially the poor, typically have few options except the local public school. Furthermore, both

traditional behavior and union constraints may operate to inhibit efficient hiring and placement. An obvious example we return to later is that teachers are paid on the basis of years of experience and education degree, regardless of their performance or their results in the classroom. A principal desiring to pay teachers based on merit cannot do so in most states and districts. Nor is the pay for many principals based on student outcomes.

While in some districts vigilant parents and taxpayers may demand that school funds be well spent, the "right" answer is not well understood by school personnel, let alone parents. Moreover, school decisions are frequently designed to be insulated from outside pressure, implying that only the most aggressive and persistent parents and taxpayers can hope to have any influence. In many other districts, the concerns of school district constituents may be quite different. For example, in many poor communities, parents may, rightfully or wrongfully, believe they can have little impact on decisions. Further, the school district may be the biggest employer, and the pressure to provide and protect jobs may trump any considerations of efficiency.

The absence of a systematic positive relationship between spending and achievement presents a real challenge to the consultants who purport to describe the spending necessary to achieve adequate levels of student achievement. This dilemma is best exemplified in a candid statement from Augenblick & Myers (2002), a consulting firm that has conducted more costing-out studies than anyone else. In most of their studies, they readily acknowledge that

> resource allocation tends to reflect current practice and there is only an assumption, *with little evidence*, that the provision of money at the designated level will produce the anticipated outcomes.[13] (emphasis added)

Here you have the whole problem in a nutshell. There is "little evidence" that "provision of money at the designated level will produce the anticipated outcomes." Quite obviously, this is not simply a "disadvantage" of the study; it undercuts the entire study and its recommendations for additional spending.

We look now at four approaches used by consultants, each of which attempts to deal with the dilemma just pointed out in a different way. As might be guessed, these approaches all fall far short of standards for scientific validity, although they demonstrate some considerable ingenuity in crafting arguments with surface plausibility. They give the illusion of providing valid, useful, and reliable information, but it remains that—an illusion.

Professional Judgment Approach

Perhaps the most commonly applied approach is the "professional judgment" method.[14] With a few nuances, the approach involves asking a chosen panel of educators—teachers, principals, superintendents, and others—to develop an educational program that would, in their collective opinion, produce certain specified achievement outcomes. Their efforts typically produce "model schools," defined in terms of class size, guidance and support personnel, and various programs that they deem necessary. The consultant in charge of the study then provides the missing elements (for example, costs for central administration or for computers and materials) and employs externally derived cost factors (for example, average teacher or principal salaries) to determine the total funds needed for the model schools.[15] The panel may or may not provide guidance on extra resources needed for disadvantaged children, special education, or the like. The end result is a cost figure based on a "basket of resources" that the panel members believe are necessary to meet the desired goals, which supporters of the methodology like because it is easy to understand and explain to legislators.

Professional judgment panels generally are instructed not to consider where the revenues will come from to pay for their model schools, or any other possible constraints on spending. They are allowed to "dream big," unfettered by any sense of reality or thought of trade-offs. Typically, they will be instructed as follows:

> You should not be concerned about where revenues will come from to pay for the program you design. Don't worry about federal or

> state requirements that may be associated with some kinds of funding. You should not think about whatever revenues might be available in the school or district in which you work or about any of the revenue constraints that might exist on those revenues.[16]

With no incentive to be mindful of costs in coming up with their model school, panel members tend to go on a shopping spree and order everything their hearts desire, not the minimums actually needed to provide an adequate education.

This feature of costing-out studies is very attractive to those inclined to pursue additional funding through the courts, which are also single-issue-oriented and often ignore revenue constraints.[17] A 2003 study by Augenblick Palaich & Associates illustrates not only the extent to which professional judgment panels are urged to forget real-life spending constraints but also the kind of interplay that occurs between consultants and professional judgment panels (this time in North Dakota). In their final report, the consultants state: "We worked hard to push people to identify resources they thought were needed to help students meet state and federal standards in spite of their natural tendency to exclude items because local voters might not approve of them or schools could 'get by' without them."[18]

In fact, these admonitions to professional judgment panels to "dream big" amount to a fundamental redefinition of the term *cost*. The term *cost* is usually understood to mean the *minimum* expenditure necessary to achieve a given outcome, whether one is purchasing a car or seeking to raise student performance.[19] The idea is to establish the desired level of quality and determine the least amount of money required to obtain it. But that is not the way a professional judgment panel works. The panel is first told to put aside cost considerations in imagining an outcome—a high-quality education—and then asked what that high-quality education "requires," ignoring all other less costly options.[20]

Consider a father wanting to buy his teenage son a car. He asks his son to find out the cost of a good, safe, dependable car that he can drive to and from school. A different father tells his son, "I want you to find a car to drive to school, and money is no object."

In the first case, the son might price out a Honda Civic, an economical but very serviceable car. In the other, he might bring home a Hummer, an outrageously expensive model with all kinds of unnecessary bells and whistles. Both cars could do the job, but one is many times more expensive than the other. Professional judgment panels are like the second son; they operate without constraints of any kind on spending—except that the panels cannot even be assured that their Hummer will do the job.

Another major problem with this method is the inherent conflict of interest for panel members who may stand to personally benefit from the higher spending recommended by them. If the state wanted to determine what it should spend to build a two-mile stretch of highway, it would not convene a panel of road contractors likely to be given the contract, tell them not to worry about where the money to pay for the project is going to come from, and ask them to come up with a price. Yet this is exactly the way a professional judgment panel costing-out study works. Educators who benefit financially and otherwise from higher spending on education are told that money is no object, and then asked to tell the legislature how much they think should be spent on education in the state. Their evident self-interest in the outcome makes such studies little more than "wish lists," as one court described them.[21]

Some consultants try to ameliorate the conflict of interest by using educators from outside the state as panel members, since they will not themselves financially benefit from the results of the study. However, this tactic is commonly criticized by other consultants, who contend that out-of-state teachers and principals do not know enough about the particular state to make the necessary judgments. In a more outrageous example of bias and self-interest, the interest groups who commission the studies select the panelists themselves. In Missouri, the plaintiff not only sponsored and paid for the costing-out study it relied on in its adequacy suit; along with its supporters, including the state teachers' union, it handpicked the panel members.[22] In Massachusetts, where four plaintiff school districts were involved, teachers and principals from those districts, who had a direct financial interest in the outcome of the study and lawsuit, were allowed to sit on the panels. Amazingly, the mother

of the named plaintiff in the original adequacy suit filed in Massachusetts was a member of one of the panels.[23] No wonder that even the special master who otherwise sided with plaintiffs in that case did not give any credence to the study.[24]

Professional judgment panels are effectively encouraged to maximize expenditures in the hope that the resulting amount will be enough to produce proficient students. A 2004 New York study conducted by a consortium of researchers from two groups—the American Institutes for Research and Management Analysis and Planning, Inc. (AIR/MAP)—even used a two-stage process in which subpanels first estimated the desirability of various educational components, and a super-panel then aggregated the results, input by input, from each of the subpanels. This design effectively maximized expenditure estimates by ensuring that any trade-offs between programs and resources made by the individual subpanels were ignored.

Courts relying on professional judgment studies to mandate spending levels assume that the panelists' model school will produce the desired results just because that was the panel's objective. But none of the reports ever test that assumption or evaluate the possibility of such achievement levels being reached. In fact, just the opposite holds. Reports generally include a disclaimer, citing other possible reasons why students may not actually achieve at these levels, despite the provision of the recommended additional funding. Take, for example, the statement in the New York City study sponsored by plaintiff CFE. After recommending an increase in annual spending for the New York City public schools of $5.63 billion, it added:

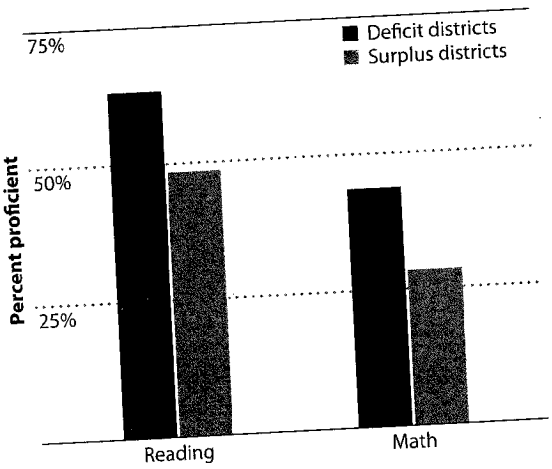
> It must be recognized that the success of schools also depends on other individuals and institutions to provide the health, intellectual stimulus, and family support upon which public school systems can build. Schools cannot and do not perform their role in a vacuum, and this is an important qualification of conclusions reached in any study of adequacy in education. Also, success of schools depends on effective allocation of resources and implementation of programs in school districts.[25]

Incredibly, the very experts who tell courts how much money will produce the desired student success admit in the same report that this money may do nothing of the kind.

Such studies do not rely on any empirical evidence. Even worse, they often ignore empirical evidence that directly contradicts the results of the studies. Consider a 2003 study by Augenblick Palaich & Associates, which used a professional judgment analysis to determine the spending level necessary for each of North Dakota's school districts to attain desired outcome goals in 2002.[26] It also specified the additional amount of aid each district would need to attain these outcomes or, in about twenty-five school districts, the amount by which spending already exceeded its recommended levels. Because information is available on the actual performance of North Dakota students for 2002, we were able to calculate the relationship between their performance and the fiscal deficits and surpluses determined by the Augenblick study. (Here, spending less than the study found necessary is termed a "PJ [professional judgment] deficit"; spending more is termed a "PJ surplus.") We would expect that student performance in districts with PJ surpluses would exceed, or at least meet, the panel's achievement goals, and that districts with larger PJ deficits would be further from achieving their goals than those with smaller PJ fiscal deficits. These expectations are appropriate, since the methodology already takes into account differing needs that arise from variation in school size, the concentration within a district of a disadvantaged population, and the like.

Yet the results of the PJ study were exactly the opposite of what one might reasonably expect. On the average, student achievement in districts with funding greater than the panel recommended, or a PJ surplus, was significantly worse than that found in districts with lower than recommended spending, or a PJ deficit (see figure 7.1).

Using a more sophisticated analysis, a regression of reading and math proficiency percentages of North Dakota districts, indicates a *positive* relationship between a PJ deficit and student achievement. In other words, the larger the PJ deficit, the higher the students' performance. This positive relationship between deficits and



7.1 Student Achievement vs. Adequacy of Funding, North Dakota

achievement levels holds true even after eliminating all surpluses and deficits greater than $2,000 to ensure that the analysis is not distorted by outliers. In short, the information provided by the PJ study was worse than no information. Its results would have dictated less, not more, spending for districts whose students were struggling the most to meet the state's academic standards, even though spending overall would substantially increase.

Such obviously silly results are far from unique. In another costing-out study by Augenblick's firm in Missouri, the results were similar. That study concluded that an average increase in funding of $4,874 per student was needed in Missouri's top twenty-five performing school districts, compared to only $2,551 in its twenty-five lowest-performing districts—again, the opposite of what one would expect.[27] These anomalous results pervade the professional judgment approach, which never tries to calibrate results to known achievement patterns.

In sum, it should surprise nobody that the professional judgment approach yields biased and unreliable recommendations, albeit ones that tend to be very useful to people interested in expanding educational spending. Panel members generally lack expertise in designing programs to meet objectives that are outside of their experience. While they may have experience making trade-offs within current school or school district budgets, they rarely have the re-

search skills or personal experience to know how resource needs will change if they design a program for higher student outcomes or for different student body compositions. We have already pointed out the palpable conflicts of interest endemic to such panels. But their most important flaw is the failure to even consider other reforms, such as vouchers or charter schools or stronger accountability measures. As stated in the consultant's report on the New York study:

> PJPs [professional judgment panels] were not asked to reform other, often quite important, components of New York's education system. School district consolidation, charter schools, devolution of authority in large districts, school board structural reform, and a long list of other possible changes might well be in order. However, they were not the focus of PJP deliberations.[28]

The panels were given only one choice: spend money. At the hearing before the judicial referees in the *CFE* case, when one of the parties asked the referees to include charter schools in their recommendations to the legislature, the response denying the request was simple and to the point: "This is about money."[29]

State-of-the-Art or Evidence-Based Approach

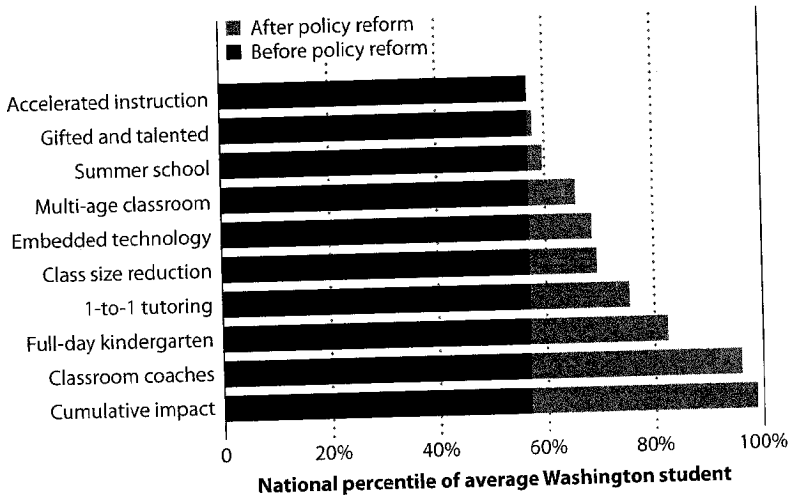
An alternative to the professional judgment approach is one that relies on the judgments of the consultants themselves. This approach has been immodestly called "state-of-the-art" by the major consulting firms using it.[30] Seeking to give their studies enhanced scientific cachet, they also refer to it in more recent applications as the "evidence-based" method. Generally, the method works as follows: The consultants review available research and select specific studies that relate to the typical programs and services considered necessary for a model school, such as small class sizes, and so on. They then develop a set of model schools based on the programs and services they select and price out their cost. They then use the costs of these model schools to determine statewide costs.

The most recent versions of the model are very precise: they quantify the effectiveness of *each* program or service included in

their model schools by actually showing the impact on achievement that would result from implementing it for a single year (grade) for each student.[31] Perhaps they have supplied too much information. Their results, whether evaluated for each separate component or for the combined package, are simply not credible.

Let us take their claims at face value by looking at a recently conducted cost study using the evidence-based approach in Washington State conducted by Allan Odden and Lawrence Picus, the consultants who are the chief proponents of this methodology.[32] As is typical, the consultants designed a school around a series of programs that have surface plausibility: smaller class size, full-day kindergarten, expanded summer school, more professional development for teachers, and the like. They then reported what they believed to be the best evidence about how much improvement each component would bring about, and recommended including all of the components, regardless of their respective cost or effect. Figure 7.2 summarizes the consultants' evaluation of the achievement impacts of each component, garnered from the best research they could find, if Washington accepted their recommendations.[33] Thus, for example, they report that, according to their review of the best scientific literature, class size reduction would move the average Washington student (who starts at the fifty-seventh percentile of the national achievement distribution) up to the seventieth percentile.[34]

The consultants do not typically show the combined effect of the components they specify, but it is easy to calculate their projected combined impact. We have added to the chart our computation of the combined impact of implementing all nine components, which is found by simply summing the impacts of the separate programs. Such a calculation illustrates, perhaps better than anything, the unreliability and bias of the studies upon which they rely. As figure 7.2 shows, under their interpretation of the research, implementation of their recommendations would turn the *average* student in Washington, now performing at the fifty-seventh percentile, into the best-performing student in the country, scoring above the ninety-ninth percentile.[35] Indeed, by their reckoning, introducing only one of the nine components—classroom coaches—would in a



7.2 Picus-Odden Estimates of One-Year Impact of Specific Reforms on Average Student Performance in Washington State

single year skyrocket the average student in the state of Washington from the fifty-seventh percentile in the nation to the ninety-sixth percentile. Such results are simply not credible.

Past experience makes it is obvious that the consultants' programs—which are nothing more than repackaged versions of various existing programs—will not have any such results. Either the program evaluations are deeply flawed or the consultants have selected a particularly biased set of program evaluations.

This methodology has other serious problems. It specifically eschews attempts to calculate the minimum costs of achieving any level of achievement. In fact, it appears that the consultants seek instead to maximize expenditures. The programs they repeatedly recommend in state after state vary widely in their predicted effectiveness and cost, according to the consultants' themselves. Yet instead of concentrating on programs that yield high achievement per dollar invested, the consultants typically recommend doing everything, even though some of their programs would purportedly produce ten times the achievement of others for each dollar spent. For example, if instituting classroom coaches will raise achievement to the ninety-sixth percentile, does it really make economic

sense to implement eight other programs, thereby doubling or tripling the cost, to move to the ninety-ninth percentile?[36] Rational government decision making would never make programmatic decisions in this manner (unless cost efficiency were really deemed irrelevant).

The only empirical bases for state-of-the-art analyses come from a small number of selected research studies that do not necessarily apply to particular states. And, most important, because those studies have been selected from the research base to suit the consultants' own purposes, there is no reason to believe that they provide an unbiased estimate of the more general empirical reality. Indeed, studies of the same programs that show little or no impact on student achievement are routinely disregarded in favor of those showing the highest impact. The usual response to such objections is simply to say that "while the evidence may not be perfect, it is the best we have." If it is that bad, however, it should not be used for policymaking.

While the concerns just cited are serious, another problem renders the entire process meaningless, even if these other concerns did not exist. For the evidence-based method to work (to actually improve achievement), the states and school districts must necessarily spend their money on the approaches they recommend and on which their cost estimates are based. Yet in Wyoming, where the legislature has relied upon cost figures generated by Picus and Odden using the "evidence-based" approach, the consultants' own studies show that nothing of the sort is taking place.[37] The consultants recommended smaller core classes, more professional development, and extending the school day. But Wyoming school authorities continue to spend money much as they did in the past, while ignoring the consultants' recommendations. They paid significantly higher salaries to existing teachers, dramatically increased the use of aides, and introduced more elective classes rather than concentrating attention on core classes.[38] Thus, although Odden and Picus have set forth a theoretical cost for an adequate education, the cost figures are unrelated to any potential outcomes, because the specific programs underlying them are being wholly ignored. Whether Wyoming's actual spending choices lead to the

level of achievement posited by the two consultants will be totally happenstance, having nothing whatsoever to do with their "scientific" study and its calculation of the cost of an adequate education. Put another way, there is an infinite number of potential programs that could be costed out and subsequently ignored in practice. It is easy to invent imaginary schools that are either more expensive or less expensive than their choice. What makes the choices of these consultants any better than others in the set of imaginary schools?[39]

Successful Schools Approach

Historically, the other method most commonly used to calculate the cost of an adequate education has been the "successful schools" approach. It begins by identifying schools—or districts—in a state that have been most effective at meeting some set of educational goals. Efforts to identify successful schools typically concentrate on student achievement, normally with no adjustments based on student background.[40] Spending on special programs, e.g., remedial education or special education, is stripped out of budgets in order to obtain a "base cost" figure for educating regular students in each successful district. Exceptionally high- or low-spending schools are often excluded, and the base costs for the remaining schools are averaged to arrive at a level of base spending that can reasonably be expected to yield high performance. (The high-spending districts are presumably eliminated because they are spending inefficiently, i.e., spending more than the minimum to achieve the observed outcomes. The rationale for eliminating low-spending districts is less clear).

The assumption underlying this method, and articulated by Ron Edmunds, the founder of the Effective Schools Movement, is that, if some schools in the state meet the required performance standard and spend in the range of, say, $8,000 per pupil, similarly situated school districts should also be able to meet the standard on similar amounts.[41] Thus far, the methodology makes a lot of sense. It is based on empirical evidence about spending by those schools that have been most successful.[42]

However, this is not the end of the process. It is necessary somehow to generalize from these selected schools in two ways. The selected schools may differ from others in the state in terms of the backgrounds of students. And, building on the accountability and adequacy issues surrounding the analysis, it is often necessary to consider the implications for performing at a higher level than *any* of the schools currently observed in the state.

The method typically selects the highest-performing schools in the state, defined by student test scores and other educational outcomes, as the gauge for spending levels. These schools are almost always predominantly white and middle-class, meaning that the base cost excludes many of the nonschool factors that affect student performance in many less successful schools, such as family background, peer relationships, and previous schooling experiences. There is no reliable evidence that similar funding will yield similar achievement in two schools with student bodies from very different backgrounds; indeed, there is powerful evidence to the contrary. Therefore, the base costs have to be adjusted to take into account the requirements of special needs and other at-risk children. In most studies, the consultants simply add a weighting that increases funding for special needs students. (For example, a student qualifying for the free and reduced-price lunch program may be treated as 1.5 students, and a student qualifying for special education services may be treated as 2 students.) The consultant is then able to calculate the cost of being "successful" for schools with varying mixes of at-risk student populations. From those figures, they calculate total statewide costs of providing an "adequate" education, i.e., one akin to that offered in the successful schools selected by them.

As previously mentioned, the problem is that no one knows what the proper weightings should be. Does it cost 1.2 times or 1.6 times the cost of an average regular pupil to educate an at-risk pupil, typically defined as one eligible for the free and reduced-price lunch program? Does spending 20 percent more, rather than 60 percent more, on such children significantly increase the odds that they will perform better? Past spending programs, such as the federal Title 1 program for disadvantaged students, have a very poor perfor-

mance record, implying again that there is not sufficient experience with success to predict how spending on disadvantaged students relates to outcomes. Plaintiffs and cost consultants routinely rely on very high weightings, but in truth there is no consensus in the education or research community about what the weightings should be, or whether increasing them leads to higher achievement. In 2004, over half of the states had no multiplier for poor children. For those states that did, the weightings in 2002 varied from 1.02 to 1.59.[43] Other programs—those for English learners and for special education—also vary widely in form and in amount across the states, and also have little empirical basis. Ultimately, the successful schools approach seems to depend on guesswork as much as the professional judgment method.[44]

The adjustments for different types of students can also lead to anomalous results. For example, Augenblick & Myers performed a successful schools analysis on behalf of the Massachusetts teachers' union and other supporters of the plaintiffs in connection with that state's renewed adequacy case in 2004. The court found that two-thirds of the seventy-five schools that they had identified as "successful" under the current funding levels did not have as much money as his methodology declared necessary for success. According to the results of his study, spending would have to be increased in those districts in order for them to be "successful" or, by definition, adequate, even though under his definition of adequacy, they were in fact already meeting his standards of "success."[45]

Quite apart from such considerations, the successful schools approach faces the second basic methodological problem: there are limited (or no) data available that enable making accurate predictions about the cost of improved student performance. Consultants project the levels of student proficiency that would occur in the future if spending were increased, basing their projections on a school's current operations and current performance levels. They can say something about meeting the performance goals established under NCLB *only if* some subset of schools is currently achieving at the level that NCLB requires. However, essentially no district has yet reached the NCLB standards. Because the approach relies on observations about a set of schools with a given level of

success, it has no way to project those observations to any higher performance level. For example, if 70 to 80 percent of students perform at the proficiency level in the schools identified as successful, there is no reliable way to extrapolate those results to a 95 percent level.

The inability to extrapolate to different performance levels seriously compromises the ability of this technique to consider alternative policy goals. Moreover, because it never identifies any set of policies, personnel decisions, or the like that contribute to the successful schools (but are absent from unsuccessful ones), the approach provides no real policy guidance to schools that want to do better.

Cost Function Approach

The "cost function" approach, sometimes referred to as the "econometric" approach, relies on current spending and achievement patterns across all schools in a state to predict optimum spending solutions.[46] The most popular forms use statistical methods to describe how spending across a given state varies with achievement and with characteristics of the districts and their students.[47] They then use the results of their statistical analysis to derive appropriate spending levels for each district depending on its particular pupil mix and other measured characteristics.

Cost functions have entered into the adequacy debates largely as an alternative to the prior methods. The prior three approaches to judging adequacy have the appeal of being easily understood and of having surface plausibility. In contrast, early participants in developing costing-out methods, James Guthrie and Richard Rothstein, commented a decade ago, "At the most practical level, because of its technical complexity, there is little chance that statistical modeling can be proposed in any state as the primary means of calculating the cost of an adequate education or as the primary way of estimating how the costs of education may vary from place to place or from student to student."[48] But, as the flaws of the prior methods have become obvious, their usefulness and

impact have diminished in both courts and legislatures. The econometric approaches have now gained almost the opposite appeal of the other methods: they are complicated and difficult to understand, so they must provide the scientific foundation for reliable costing-out of adequacy.

The technical complexity of these studies presents a challenge to us also, because understanding the nature of the studies, and why they fail to provide reliable answers to the "how much" question, requires more technical discussion than needed for the others. Nevertheless, given their growing use, it is important to lay out the fundamental issues, even if it requires dipping into statistical modeling issues.

For all their scientific pretensions, cost function studies also cannot come to grips with the lack of any significant causal relationship between student performance and spending.[49] As we have noted, there is a large body of statistical research examining how various levels of funding in schools influence student achievement, after taking into account differences in a range of background characteristics. This research, often called "production function" analysis and reviewed previously in chapter 3, has found little consistent relationship between spending and student outcomes. Even those studies showing that more money does improve student performance typically find only a very small effect of spending on student outcomes.[50] Given these small, if any, spending effects, these analyses imply that, absent other reforms to make the education system more efficient or responsive, it would take extremely large and unrealistic spending increases to make any meaningful progress in raising achievement.

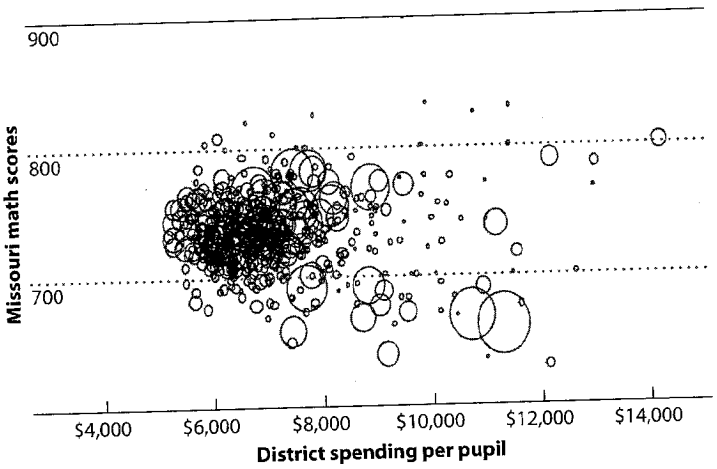
Consultants doing "cost function" analysis take a slightly different approach and purport to have found a way to predict how spending must vary in order to obtain a given level of achievement. Instead of asking how much achievement would rise if a school were given an extra $1,000 per student (as the prior work has asked), they consider how much spending would increase if achievement went up by five points.[51] To a nonspecialist, these questions sound almost the same, and the nonspecialist would be right. Both methods rely on exactly the same data, and one would

expect similar answers regardless of the method used. However, for mathematical reasons having to do with the drawing of the regression line in each approach, the results not only differ, but differ so much that it renders the results downright silly.

Let us look at what happened from parallel studies in California. As part of a broad set of studies of California's public schools, a number of costing-out studies were conducted to determine how much would have to be spent to allow the state's students to meet certain academic standards.[52] Jennifer Imazeki, a well-known expert in the area, applied both statistical approaches to the same data, but got such strikingly different results that it was obvious that the results of both studies should be disregarded. Using cost function analysis, she estimated that annual education spending in California would need to be increased by $1.7 billion a year to be adequate (on a base of approximately $50 billion).[53] Using production function analysis, based on the same data, she estimated spending would need to be raised by $1.5 *trillion per year—almost a thousand times greater.* Put into a slightly different perspective, this latter figure is thirty times the total amount California currently spends on its schools and three times larger than the spending on K–12 education for the entire United States.

The use of the same techniques in other states has also resulted in widely disparate results. For example, in recent studies conducted in Missouri, the use of one method indicated an increase of $367 per student would be required to meet academic targets, while the other estimated a $22,000 increase per pupil would be needed. Similar studies introduced in the Texas adequacy case also reached widely differing results.[54]

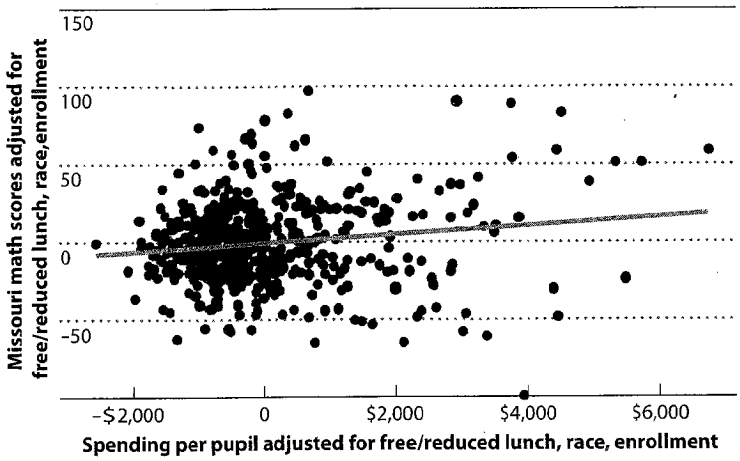
Importantly, the higher numbers obtained by the consultants come from the approach most accepted in the research literature (understanding how much achievement would change with additional resources), and their magnitude reflects the common finding that spending variations are not significantly associated with student achievement. Thus, it takes vast amounts of money to get any achievement change if schools stick to the current way of doing things. As these studies show, and as Imazeki acknowledges, the relationship between dollars and student achievement is so uncer-



7.3 Missouri District Average Eighth Grade Math Scores vs. District Spending

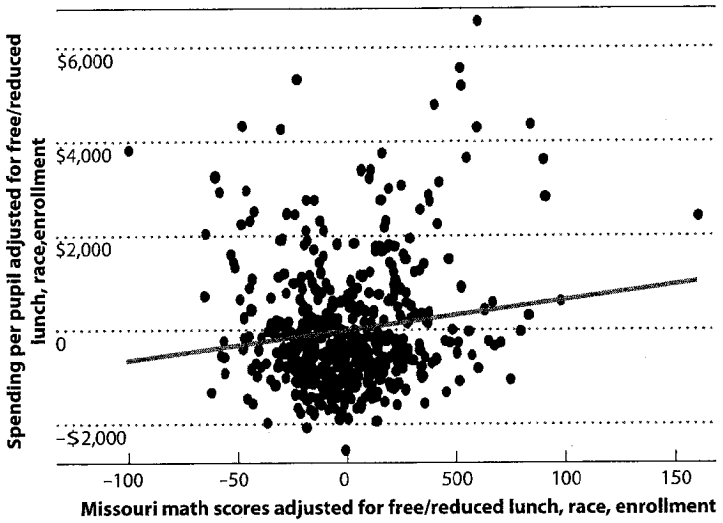
tain that such results should not be used to gauge the potential effect of resources on student outcomes.[55] One can readily see this lack of any meaningful relationship by plotting achievement scores against per-pupil spending for the different school districts in a state. For example, figure 7.3 portrays the 2006 scores for each Missouri school district on the Missouri Assessment of Performance for eighth grade math.[56] The cloud of points, each representing the average test score of a Missouri school district, shows vividly the lack of any relationship between spending and achievement.

It is possible that the lower-scoring districts have disadvantaged student populations that are more difficult to educate and that therefore figure 7.3 does not accurately reflect whether higher spending is having a beneficial effect. However, using regression analysis, we can adjust for that possibility. We have done exactly that in figures 7.4 and 7.5, allowing for differences among districts in the racial composition of their student bodies, the percentage of students in poverty, as indicated by their participation in the federal government's free and reduced-price lunch program, and total enrollment. Figure 7.4 shows the relationship between achievement and spending, after controlling for these differences using the standard achievement-spending or production function approach,



7.4 Achievement–Spending Relationship: Production Function Results, after Controlling for Poverty, Race, and Enrollment

while figure 7.5 reflects the results using cost function analysis. Again, the central feature of both graphs is the "cloud" of points, each representing a Missouri school district, indicating little relationship between spending and achievement. The regression lines on each figure slope slightly upward, indicating that for lots of money some slight achievement gains might be realized. However, as pointed out earlier, the amounts are so large as to render the methods useless in trying to determine the amount of spending it would actually take to reach such achievement levels. For example, in the cost study presented by plaintiffs in the Missouri court case, the target for performance in Missouri called for increasing achievement in mathematics by sixty-seven points over their 2006 level. The slope of the line in figure 7.4 indicates that this would require additional spending per pupil of over $22,000, i.e., the regression line would have to be extended a very long way to the right before it was high enough on the *y* axis to reach sixty-seven added points. Put in context, this would require a quadrupling of K–12 education spending in the state, clearly something that for political and practical reasons is not going to occur. Indeed, any public official, including judges, who even suggested such an increase would be laughed out of office.



7.5 Achievement–Spending Relationship: Cost Function Results, after Controlling for Poverty, Race, and Enrollment

The key part of the cost function approach, found in figure 7.5, is that the "no relationship" clearly seen in the figure takes on a different meaning. Now it appears that one can travel a long way in student achievement (on the horizontal axis) while only rising a little bit in terms of spending (the vertical axis). In sum, the cost function approach simply tries to capitalize on a different way of expressing the lack of relationship of spending and performance, a different way that appears more plausible than the outlandish interpretation of no relationship in figure 7.4.

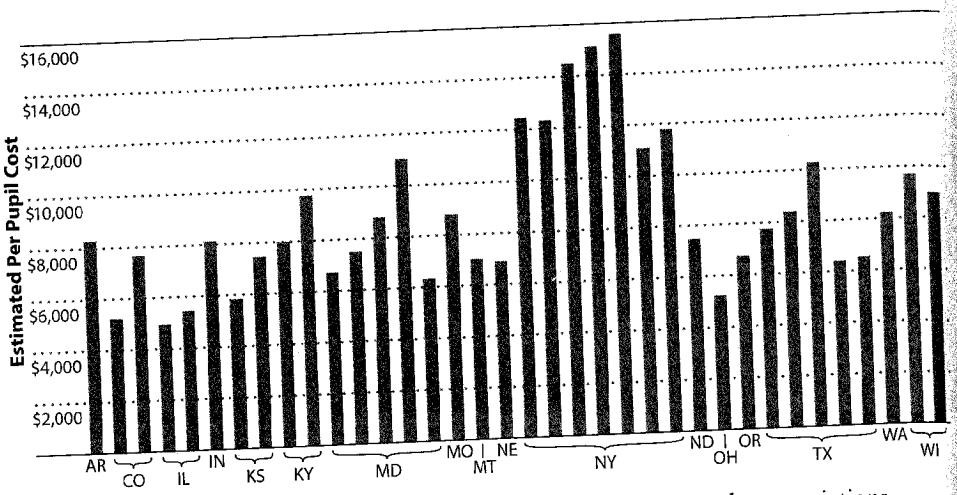
These average spending relationships—whether estimated in the way of the "production functions" or in the way of "cost functions"—signify nothing more than the pattern of spending in a state. In order to interpret either of these relationships as describing the "cost" of changing achievement, one would have to presume that any inefficient spending is eliminated from the analysis. The trouble is that there is no budget item labeled inefficiency that can be factored into the statistical analysis. Inefficiency differences across districts are impossible to remove from the observed spending–achievement relationship in a state, resulting in the patterns we have observed: many high-achieving districts with low spending and vice versa.

Inefficiency can be buried deeply in the operations of districts in a thousand different ways: as salaries for teachers that are unrelated to the teacher's performance or quality, as programs that have no effect on student outcomes, or as the purchase of very costly inputs such as class size reduction instead of more cost-effective means of raising student achievement. Cost and production function studies do not consider any of these factors. Instead they typically include ad hoc measures of district characteristics in the statistical analyses and claim that they are related to efficiency. For example, in an attempt to provide some orderly pattern, the cost consultants in Missouri included characteristics of the district such as the amount of state aid per pupil, the percentage of the population aged sixty-five or older, and voter turnout in elections,[57] though how these factors are associated with more or less efficiency in a district is hard to fathom. The statistical analyses themselves show that these measures are only slightly related to spending differences across districts and could not possibly explain the wide disparities in spending for districts achieving at the same levels—i.e., the cloud of points in either figure 7.4 or 7.5.

Given the fundamental problems discussed, it adds little to go into detail about how current observations must be extrapolated. Obviously, if one projects to higher achievement—such as 100 percent proficiency in 2014 according to NCLB—it is necessary to go outside of the current observations.[58] Additionally, since much of the legal and policy focus is on low achievers, it is generally necessary to assume that high socioeconomic status (SES) districts directly show what low SES districts can do in terms of achievement.[59] If the other concerns about cost functions were appropriately addressed, these issues alone would be showstoppers. Nonetheless, one seldom gets to this point with cost functions.

Comparing Approaches and Studies

But it is not just the two different cost function studies that produce wildly different results; the results of the four different approaches consultants use to estimate how much an adequate education should cost vary considerably as well— and fundamentally call into



7.6 Costing-Out Results in Eighteen States (adjusted for regional cost variations and expressed in 2004 dollars)

question their reliability. Even though the standard of adequacy may differ somewhat from state to state, it should presumably cost more or less the same to provide an "adequate" education in New York as in Illinois, assuming the cost numbers are adjusted to reflect different costs of living. However, the truth is that the cost study results not only vary from state to state, they are not even close.

Look at the results found by Bruce Baker, a supporter of costing-out studies and a consultant working in various school finance court cases, when he compared the results of thirty-six "costing-out studies" from seventeen different states. The results were striking, in several ways. As shown in figure 7.6, the "base" costs, with no monies added for special student needs, varied all over the map, from $5,210 in Illinois to $17,647 in one of the New York studies, when expressed in 2004 dollars.[60] When adjusted for regional cost of living differences, the wide disparities still remained, with some concluding it would cost about $5,000 per student and others concluding that over $15,000 per student would be required. Nor were those numbers outliers. The cost figures were arrayed all along the spectrum, as figure 7.6 demonstrates.[61]

One might try to account for such differences by arguing that different states have different quality requirements for an adequate education. But in Illinois, where the costing-out study concluded

that $5,000 per student was sufficient, the state constitution contains the highest standard of the fifty states, requiring the state to provide its students with a "high quality" education. In contrast, New York, where costing-out studies concluded that more than three times as much money was necessary to satisfy the constitution, has one of the lowest constitutional standards of the fifty states. It simply requires "free common schools" and does not specify their quality.

Even within the same states, the cost studies result in vastly different cost figures depending on which of the four methods consultants used. On average, studies using the successful schools method conclude that the cost of providing an adequate education is about $2,000 a student less than the studies relying on the professional judgment model, even when such studies are of the same state education funding system. The Missouri costing-out study by Augenblick & Myers on behalf of the members of the Missouri Education Association used both these models and got two dramatically different results. The professional judgment model indicated that it would cost $7,832 per student in Missouri to provide an adequate education; under the successful schools model, the cost was $5,664, which was not only roughly the amount the schools were already spending, but also over $2,000 less per pupil than indicated by their professional judgment study.[62] The results of the successful schools study indicated no additional money was needed; the results of the professional judgment model indicated that almost a billion dollars more was required. It is difficult to take seriously the results of any of these studies, or consider any of the four methods generally reliable, when the same consultant studying the same schools can produce two such different results.

An analysis of costing-out studies by Augenblick & Myers in four separate states, undertaken by *Education Week* in 2005, found that the successful schools method resulted in significantly lower costs, varying from $735 per pupil lower in Maryland to $2,461 lower in Missouri.[63] This fact has not been lost on plaintiffs or state government officials who want higher spending on education, and consequently they have insisted that the professional judgment approach be used in most studies conducted in recent years.[64]

In summary, there are many analytical problems with costing-out studies, and there is little reason to believe that they can accurately determine the amount of funding necessary to provide an "adequate" education. Even some of the school finance consultants who have been paid hundreds of thousands of dollars to conduct these studies are no longer willing to vouch for them. James Guthrie, who is generally given credit for pioneering the professional judgment model in Wyoming and whose company MAP, in conjunction with AIR, conducted the costing-out study relied upon by CFE and the trial court in the *CFE* case in New York, has become thoroughly disillusioned with the manner in which such studies are being conducted and used, concluding that the claims made in them are "unsubstantiated" and "unreasonable."[65]

It is impossible even in retrospect to assess the accuracy of such studies, that is, to see after several years if they have accomplished what their purveyors claim. The amounts recommended in the studies conducted thus far have been so unrealistically high that it is doubtful that any state legislature will ever fund them at the recommended levels.[66] Therefore, if performance goals are not reached in two, three, or ten years, study advocates will always be able to argue underfunding as the reason for failure. Even if such levels of funding were actually provided and children still failed to meet the state standards, advocates of such studies could point to their disclaimers that money alone cannot ensure success[67] or plead failures in implementing the recommended programs.[68]

Most damaging, however, is the implication of such costing-out studies that our country's education problems, particularly those related to the low achievement of at-risk students, can be solved by increased spending alone. By their very nature, such studies ignore other more effective steps a state or school district might take to increase achievement, and focus solely on increasing spending.

How Should the Money Be Spent?

Costing-out studies are not the only "scientific" evidence appealed to in courtrooms and legislative chambers. Research on the

effectiveness of particular programs and educational strategies also plays a role in discussions of educational policy because it helps to answer the all-important question: how should education funds be spent?

It makes some sense, as a matter of principle, to justify particular policy decisions—regulatory frameworks or the use of state funding for particular programs—by citing scientific evidence that the program works. If science says that the program will raise student achievement, then the state may want to insist that all districts follow it. Legislatures can write such a policy into law, or state departments of education can require districts to pursue it, or funding can be structured to supply money if it is used. Or courts seeking remedies in adequacy suits may order that districts implement the program.

Whether the program, once implemented, bears the fruit suggested by the research depends on a variety of factors. First and foremost, the underlying scientific research clearly must be highly reliable and widely applicable. This proves to be a tough test, because many plausible-sounding policies linked to research prove ineffectual in application. Second, the implementation must be faithful to the underlying design if the research is to be a good predictor of the policy's impact. Policies that prove effective in some circumstances do not always travel well; small differences in implementation or local variations can produce quite different results. Third, policy initiatives aimed at a specific program can lead to unanticipated and sometimes harmful consequences as local decision makers react in unexpected and unintended ways to the incentives that are created. Finally, of course, the policy has to be a good one in the economic sense; that is, its educational returns must be sufficient to justify the expense of the program, as well as any related impacts on other programs.

It would, of course, be impossible to consider here all the existing scientific research on education, and how well scientifically based programs have fared once implemented by specific schools and districts. But we can illustrate some of the problems that have arisen in applying scientific research by looking at two popular policies: class size reduction and preschool education.

Debates over these two policy ideas have become central to many legislative and judicial deliberations in the last decade.[69] In other years, different programs might have made the top of the list. For example, in the 1990s, the central focus of many educational policy discussions was whole school reform.[70] This idea even produced a public–private partnership called the New American Schools with a design competition and federal start-up money.[71] Yet, after much hoopla and high hopes, implementation problems and the lack of obvious achievement gains led to its gradual demise.[72] Prior to that, the Office of Economic Opportunity experimented with "performance contracting"—hiring private firms to provide education and paying them based on results. But here the inexperience in writing contracts doomed the effort.[73] Other efforts in other years were likewise trumpeted, and likewise faded.

As the following discussion will make clear, the application of science to difficult questions of education policy has not been very effective. We believe that science could potentially be enormously helpful in finding solutions to education problems, but only if it is based on reliable and appropriate conclusions. Certainly, some research was unsuccessful because of flaws in the research itself, but more often, we believe, the problem has been the lack of appropriate data and the failure to properly evaluate policies before putting them in place. In the last section of this chapter, we address how better use could be made of science in education decision making, and why features of performance-based funding are critical to achieving those benefits.

Class Size Reduction

The impact of class size on student outcomes is perhaps the most studied policy issue in education. By the early 1990s, most studies looking at this issue found only limited evidence that reductions in class size were associated with improved student achievement. While teacher–pupil ratios fell steadily after the 1960s, these reductions were seldom a primary policy choice.[74] In 1996 this situation changed when Governor Pete Wilson of California, looking for a

way to put more money into schools without devoting all of it to higher teacher salaries, hit upon the idea of reducing class sizes across the state. Wilson's political popularity consequently shot up, inducing a majority of other governors and the federal government to very quickly announce their own programs to reduce class sizes.

The policy proposal is aided considerably by a surface plausibility. Smaller classes permit a teacher to individualize instruction and to spend more time with each student. As a result, they must by popular consensus lead to improved achievement.[75] Unfortunately, the issue is vastly more complicated.

The result of the rush to smaller classes was a search for evidence supporting the initiative, and the proponents of the policy—including parents, teachers' unions, and school personnel—seized upon a random assignment experiment in Tennessee that previously had been largely overlooked. Project STAR was a legislatively mandated evaluation of an experimental reduction in class size from the then standard of twenty-two to twenty-five students per class to thirteen to seventeen students per class.[76] Beginning with a set of seventy-nine schools that volunteered for the experiment and that had at least three sections of classes in each grade from K to 3, kindergarten students in 1985 were randomly placed in either small or large classes. Students were to stay in small or large classes for the entire four years of the experiment (K–3), and each spring the students would be tested in reading and math to gauge achievement growth in the different-sized classes.

The random assignment experiment is widely regarded as one of the most valid ways of separating the causal effect of a policy from other factors that might be influential.[77] Other analytical designs, including those most frequently applied to study class sizes before the STAR experiment, are more prone to questions about the influences of other factors, even though efforts are made to control statistically for such factors through regression analysis.

In the STAR experiment, student performance in the small classes averaged about one-quarter of a standard deviation higher than that in the larger-sized classes. An impact of this magnitude implies that a student at the center of the distribution would move up to the sixtieth percentile. Based on these estimates, findings from

this experimental program were immediately cited as support for reductions (even as all other evidence was disregarded). Across the country, teacher–pupil ratios declined rapidly. Indeed, the two costing-out methods described earlier that rely on professional or analyst judgments always include a significant class size reduction component based largely on Project STAR. Similarly, most trial presentations in adequacy cases include evidence from Project STAR in order to justify increased funding of schools.

By the standards of random assignment experiments, Project STAR was not a very high-quality experiment.[78] Over half of the initial students dropped out of the experiment and were replaced before the conclusion of the four-year evaluation; a sizeable portion of the students transferred from the large to the small classes between kindergarten and the third grade; and a large percentage of students (larger in the small class groups) did not take all of the examinations. Perhaps most importantly, because the inferences from the analysis depend upon no systematic assignment of teachers to classes of different sizes, there is no evidence about the allocation of teachers to the different treatment groups in the experiment—raising questions about the validity of the experiment.

Strikingly, although the experiment has received enormous attention and its design is very straightforward, no efforts have been made to replicate it in the two decades since its completion. Thus, any questions about its validity and its relationship to the vast amount of other evidence about class size remain subject to continuing, often heated, debate.[79]

A central issue, something often disregarded in the policy discussions, is what generalizations can be made from such evidence. Ignoring the shortcomings in the study just summarized, it remains the case that the experimental evidence refers just to a particular—and particularly large—reduction, from twenty-four to fifteen students per class. The experiment does not provide information about the achievement effects of reducing from, say, twenty-six to twenty-one students or twenty-nine to twenty-five. Nor does it provide any information about grades other than K to 3. Further, since *all* of the impact of moving to smaller classes actually was observed in the first year, the experiment itself says nothing about the effects

of smaller changes in class sizes, class size reductions in later grades, or what have you. The evidence, it appears, cannot be the basis for broad generalizations.

In this case, we can observe the effects of trying to replicate the results of a small, controlled experiment with the blind implementation of such a program on a statewide basis. The problems, some unforeseen, that can arise during implementation are huge, and are best illustrated by what happened in California's well-publicized (and often-praised) program. That state's program was designed to provide extra funding to school districts to enable them to reduce class sizes in grades K–3 to twenty students or less.[80] However, the sudden reduction in class sizes across the state resulted in a commensurately sudden and sharp increase in the demand for teachers. In response, many school districts set out to attract and hire existing teachers from other districts. In the end, districts serving disadvantaged populations disproportionately lost their most qualified teachers, potentially harming the most important target group. Whatever positive effect reducing class size had in those districts was offset by the effect of losing many of their best teachers, and the net result was little or no improvement in achievement in such districts.[81]

Another unintended consequence of class size reduction programs in California has been to actually increase some class sizes. Since the additional funding in California is for reductions in grades K–3, school districts under financial pressure may seek to cut costs by hiring fewer teachers—and thus increasing the class sizes—in later grades. Therefore, classes of twenty students in K–3 are frequently followed by much larger classes in grades 4–6. The average class size in the state for grades 1–3 in 2006 was less than twenty, while that for grades 4–5 was twenty-nine.[82] No evidence suggests that this is an optimal strategy.

Our point is not that small classes have no positive attributes. It is undoubtedly the case that small classes can be very beneficial in certain circumstances—with certain teachers, with certain students, in certain subjects, and in certain grades. But the existing research does little to pinpoint those circumstances, even as it demonstrates that across-the-board class size reductions are, at best, a

mixed bag. One important difficulty with such research is that in applying it, little effort is made to tailor the program to the specific circumstances of a school or district. It is this inflexibility, finally, that casts most doubt on the appropriateness of the policy. Class size reduction is the most expensive broad policy that is commonly contemplated by either courts or legislatures. Because of the increased need for teachers, even a small reduction from twenty-six to twenty-three students raises operating expenses by 10 percent,[83] to say nothing of the sometimes huge capital cost of constructing additional classrooms.

Debates frequently fail to ask whether spending an equal amount on some other policy, such as higher teacher quality, would not produce even greater achievement gains than those anticipated from class size reductions.[84] But because class size reductions garner large increases in funding for schools while involving essentially no changes in schools' structure or system of incentives, parents and teachers alike support them. For these reasons, programs to reduce class size have been very popular, whether for general policy purposes or for achieving adequacy.

As the projected benefits of class size reduction have failed to accrue, the policy push behind the initiative has diminished. Yet many states are left with expensive class size reduction programs in place—California currently pours $2 billion per year into its K–3 program—that are very difficult to modify because of continuing public and school support. Because the program is so widespread, it is impossible to assess precisely what its real value for student achievement may be. In the end, the mixed scientific evidence on the merits of general class size reduction programs and the experience in California and elsewhere make clear the risks of relying on limited, selective studies to approve hugely expensive programs.[85]

Preschool Education

A second focus of current policy discussions with large ramifications for school finance is preschool education. In whatever version (e.g., universal or means tested), preschool education is fre-

quently mentioned as the next "obvious" fix for the current schooling problems.

Support for preschool education is based on three different and plausible arguments.

First, preschool would address the problems of disadvantaged students who come to school far behind their middle-class peers. If language and other deficits, which have lasting effects on student outcomes, were lessened for these children by preschool attendance, the result might be smaller achievement gaps in the future.

A second argument for preschool rests on a variety of conceptual arguments for early investments in human capital—most notably articulated by Nobel laureate James Heckman. Heckman has argued that early investments in education are critical, since, in his words, "learning begets learning."[86] Investments made early in life enhance learning later in school and even into careers, making such investments economically attractive.

Third, key studies with strong research designs based on random assignment of students to programs have supported the efficacy of preschool education for aiding the school readiness of disadvantaged youth. The most well-known is the Perry Preschool Program, but others, such as the Abecedarian Program and the Early Training Program, also provide important evidence in favor of early childhood education.[87] A set of benefit-cost analyses of the Perry Preschool Program shows that this program appears to have been effective, conferring social benefits in excess of expenditures.[88]

Support for expanding preschool programs is currently very strong; courts in South Carolina and New Jersey, for instance, have found preschool education to be an essential element of an adequate education.[89] Yet serious questions remain about the reliability of the studies affirming the benefits of preschool, and about how generally applicable their results are. The evaluation of the Perry Preschool, Abecedarian, and Early Training programs relied upon a random assignment methodology that followed subjects over extended periods of time. But the numbers of children taking part in the experiments were relatively small. The Perry Preschool treatment group consisted of just fifty-eight children, the Abecedarian Program just fifty-seven children,[90] and the Early Training Program

just forty-four children. Quite clearly, samples of this size raise the concern about whether the evaluation results can be generalized to much larger programs, especially when, upon reanalysis, many of the originally reported findings have turned out to be fragile.[91]

The experimental evidence has been supplemented by observational studies in other locations. Perhaps the most commonly cited study is the Chicago Child-Parent Center program, a program currently operating in the Chicago public schools.[92] This program is lower in cost than Perry or Abecedarian, although the benefits are also considerably less certain. More recently, studies of Tulsa—meant to provide evidence on a universal program in Oklahoma—have provided another interesting, albeit limited, evaluation of the results of broader programs.[93]

The beneficial results that have been identified are quite varied. First, as Michael Anderson demonstrates, girls experienced virtually all of the programs' benefits; boys were likely to experience no benefit or worse.[94] Second, a substantial part of the benefit fell outside of schools and the development of cognitive skills; the benefits found for girls related to reduced criminal behavior.[95] Thus, even if the programs are valuable for society, they do not appear to be a panacea for school achievement problems. Third, the results for varying preteen, teen, and adult outcomes differed across programs, so that it is not enough to simply recommend "preschool," but is rather necessary to identify the precise kind of preschool.

Probably most important, these programs are not your typical community or school-based program found in most states. The Perry Preschool Program, estimated to cost over $15,000 per child (in 2000 dollars), involved intensive treatment by teachers with master's degrees in child development, student–teacher ratios of 6:1, and regular home visits—but they ran just from October to May.[96] The Abecedarian Program is full day, five days per week, fifty weeks per year for five years beginning at birth, and includes medical care and home visits.[97] Over the five years of program services, it is estimated to cost $75,000 per child (in 2002 dollars).[98] Clearly, these programs are not within the range that would be considered on a broad scale in most states, but the experiments provide no information about which components are most valu-

able or what a more modest version (with significant benefits) might look like.

All this evidence indicates that there may well be benefits for society to instituting expanded preschool programs for disadvantaged students, but there are also potentially huge costs associated with doing it right—if in fact we even understand what "doing it right" would be. The rationale for preschool is that it is easier to remediate earlier rather than later, and that school will supplement what children experience at home to foster stronger educational development in the future.[99] At the same time, the educational benefits of existing programs that have been evaluated, except perhaps the most intensive and expensive, have been small and short-lived. The limited number of models that has been evaluated provides uncertain guidance about design of effective programs, particularly those that reach boys.

We believe the existing evidence supports a more extensive set of experiments and evaluations into the efficacy of different approaches to providing early education. In the words of preschool proponent William Gormley, "Preliminary results from a growing body of pre-K programs are encouraging, but not entirely convincing."[100] This is just the situation where strategic experimentation offers the most promise. But the evidence does not support instituting broad, full-scale programs—particularly when doing so would make evaluating and learning from the experiences very difficult. Implementing preschool programs on a statewide basis would be much like introducing the class size reduction programs in California, where universal coverage eliminated any appropriate comparison group and thus made any reliable evaluation impossible.

One other aspect of program design is also important. Any proposals of governmental support for preschool must consider which groups should receive programmatic help, how the programs should be organized, and how they should be financed.[101] The existing evidence on preschools is limited largely to their impact on disadvantaged students. There is no evidence about positive impacts for middle- and upper-income students.[102] Many who support expanded preschool appear to assume that the programs will be universal and will simply extend the current public schools back

to earlier grades. But the scientific evidence for preschool provides support, albeit limited, for programs targeted only to disadvantaged students.

Finally, we add this important reminder. None of the rigorously evaluated experiments just discussed has been a public school program, and none of them suggests that the public schools would be any more effective at providing preschool programs than they have been at the K–12 level. [103]

Indeed, the disappointing results of Head Start indicate otherwise. Many people tend to forget that we in fact have a large public preschool program, introduced with the programs on the War on Poverty in 1965. Over nine hundred thousand three- and four-year-olds from families in poverty are currently enrolled in Head Start programs around the country. The federal Head Start program is considerably different from the Perry and Abecedarian programs. In 2005, just 35 percent of its teachers had a bachelor's degree, and the programs varied considerably in length and intensity.[104] The cost of Head Start is usually reported as slightly over $7,000 per pupil per year (in 2003–4 dollars), derived by dividing total program costs by the number of participants. As welfare specialist Douglas Besharov and his colleagues point out, however, this mixes together a variety of different programs; if run on a full-time, full-year basis, the program costs would be over $20,000 per year.[105] Against these expenditures, there is considerable uncertainty about the benefits. Many studies find no lasting impact, while others find a modest initial impact but one that tended to fade over time.[106]

Some states are moving to universal pre-K programs, including Oklahoma, Georgia, and Florida. The evaluation of these programs is difficult, because of their universal nature. And there is disagreement about the impacts on achievement. Those studies that show improvements in short-run achievement are matched by those more skeptical.[107] As with the previously described preschool programs, some reason for optimism about the potential for disadvantaged students survives, but the uncertainties also point to a need for gathering more systematic evidence before starting full-scale universal programs.

Finally, we bring the evidence back to the education policy context. The existing evidence suggests that the achievement gains from current and past preschool programs are relatively small and may fade out altogether as the student progresses in school. Thus, even the most optimistic view of the evidence does not suggest that preschool is likely to close the existing achievement gaps by itself.[108]

Using Science More Effectively

Our position on science and the development of public policy should not be misunderstood. It is emphatically not meant to indicate that scientific evidence is inappropriate or that science should be avoided in developing educational policies—quite the contrary. We believe that better use of empirical evidence is the *only* way that policies will improve. A key element of our performance-based funding plan, described in the next chapter, is a scientific improvement process.

What we object to is the misapplication of scientific methods and the misinterpretation of scientific conclusions. Costing-out studies are simply bad science—political approaches masquerading as studies following scientific principles. On the other hand, other studies are pushed beyond their limits. For example, the science is just too uncertain on the impacts of class size reduction and preschool education to be applied as some would wish and without the appropriate caveats about the unreliability of available evidence.

Our concerns focus on the potential for misuse of the available evidence. Generalizing from small, specific studies to universal application of a legislated policy is, at the very least, a risky strategy and most likely an expensive and ineffective approach to improving schooling. Even where programmatic evidence is reliable and valid, applying the same program to heterogeneous districts of a state is unlikely to be successful. Consider what state educational systems look like. Nine states have over five hundred separate school districts.[109] California, the most populous state, has more than a thousand districts, including Los Angeles Unified with over 700,000 students and twelve others having enrollment exceeding

49,000.[110] At the same time, over half of the state's districts are K–8 districts. An effective common program that is legislated from Sacramento to cover the variety of activities in the state is, as the data have shown, impossible.[111] Yet, most discussions of class size reduction, preschool, and the like tend to assume that a specific program could be faithfully and universally implemented across an entire state.

There is actually little evidence that *any* policy extracted from the scientific evaluation literature has ever been successfully implemented across the heterogeneous districts of states. The available science does not support effective regulatory and implementation policies that reliably replicate evaluation results on a broad basis. The overall lack of relationship between spending, which is frequently related to programs that are justified by reference to scientific evidence and outcomes, provides evidence for this conclusion. One interpretation of the oft-mentioned difficulties of "going to scale" with a program is that it is rarely faithfully implemented across districts.[112] But the implementation issues remain, nonetheless, a central part of school policy, because different districts with different needs and capacities will tend to modify virtually any policy mandated (or voluntarily selected). Indeed, in places where funds have increased in accord with evidence-based costing-out studies (e.g., Arkansas or Wyoming), school districts have not used the extra funds in anything like the ways the costing-out studies prescribed.[113]

We object also to the tendency for advocates on different sides of an issue to rely on science selectively—that is, to choose just the studies that support a particular point of view.[114] A variety of commentators see the political use of research to be a very natural state of affairs.[115] Natural or not, it obviously compromises the value of scientific evidence, since choosing studies on the basis of results, rather than the quality of the study, is both an invalid scientific approach and one likely to lead to unsatisfactory outcomes. Certainly, science itself is not immune to bias, but the continuing dialog within disciplines, the scientific peer review system, and the mores of science work militate against such problems; whereas the

political, or, for that matter, the judicial process, lacks such scientific checks and balances.

This concern about the selective (i.e., result-dependent) use of studies pervades the policy process, but it is most acute in the courtroom and legal setting. The adversarial system encourages testimony that is biased toward one or the other positions and that discourages nuance. Differing rules of evidence in the courtroom and in science compound the problem. Hearsay rules, while lessened with expert witnesses, work against broad development of scientific evidence in the courtroom. Additionally, the complexity of statistical analyses and the difficulty that nonspecialists have in penetrating complicated methodologies and the testing of hypotheses mean that scientific opinions and evidence are evaluated only in the crudest way in courtrooms.

The answer is to recognize the role that science can and cannot play in understanding and formulating educational policy.[116] Scientific evaluation, properly applied, can assess whether an already implemented policy is likely to be effective and is worth the expenditure. This is extraordinarily important.

In order to facilitate scientific comparisons of alternative programs, the Institute of Education Sciences (the research branch of the U.S. Department of Education) has developed a fledgling program to compile evidence on programmatic effectiveness, to evaluate the scientific reliability of the evidence, and to compare costs and effectiveness of alternatives where possible. This activity, the What Works Clearinghouse (WWC), currently addresses a limited number of areas of study, and each has a limited number of studies evaluated as being high quality—reflecting the recentness of emphasis in education research on scientific validity.[117] Nonetheless, an example is instructive. In March 2008, the WWC could provide comparative evaluations of ten separate programs to aid reading comprehension of English language learners. From this, one could readily discover that one program (Instructional Conversations and Literature Logs) had "potentially positive effects," while another (Read Naturally) showed "no discernible effects."[118] The interested person can quickly get a summary but can also dig deeper into the evidence and the potential for generalizing the results. Top-

ics that are covered are largely driven by the availability of suitable scientific evidence, but the potential can be seen from available reviews ranging from elementary math programs to dropout prevention and character education interventions.

On the other hand, existing evaluations seldom provide much assurance that a mandated policy will have any given outcome. Indeed, under current operations, mandated programs are seldom evaluated at all. And the available evidence on educational programs cannot supplant the legislative process by providing a scientific answer to the question "How much should we spend on education?"

In the next chapter, we discuss in more detail the components of a performance-based funding system, which we believe can make science a more important and productive part of educational policymaking. A key component of these ideas is recognition that a single best solution that can be applied universally simply does not exist. There are some underlying general policies that might be developed through improved research and evaluation. But their applicability and specific use are likely to vary considerably across districts with different student demands and different district capacities. Both states and districts have to think in terms of a continuous improvement process—one that modifies basic policies over time to retain the good parts and drop the bad parts. The process clearly should rely heavily on scientific evaluation methods to disentangle the causal effect of policies on student outcomes, i.e., to go beyond mere associations that are observed across schools and districts and to pinpoint what can be expected if a given policy is instituted.

Unfortunately, scientific research and evaluation analyses got off to a slow start in education. Too little effort went into identifying the true effect of various policies and resources on student achievement, mainly because of the lack of appropriate data. Until recently, it was difficult to trace the impacts of teachers, schools, and policies on student achievement. This shortcoming has gone a long way toward being remedied with the development of accountability systems in all states (even if a number of states have resisted

assembling these data into useful databases that permit such evaluation). And the tide appears to be turning. The No Child Left Behind Act mentions "research" over two hundred times as it calls for a closer linkage of evidence and policy.[119]

But the policy process has itself hindered the ability to judge the effectiveness of policies. All scientific evaluations are based upon having appropriate control groups to indicate what happens without any intervention. Simply put, it is very difficult to understand the impact of a policy without having a good idea of what would have happened without the policy. Yet programs are often implemented universally across districts or states—perhaps because of ignorance of the uncertainty surrounding a program, perhaps because of a sense that the problem is urgent and must be dealt with immediately, perhaps because legislators and governors do not want to admit that they are proposing policies that are uncertain, or perhaps because they are reluctant to deprive any children of the assumed benefit of the policy. But when applied universally, programs can seldom be evaluated with any reliability owing to the lack of any possibility to observe outcomes in situations where the policy was not implemented.

We propose making scientific research and evaluation a more fundamental part of the educational policy process. Doing so will require a variety of changes in the normal way of doing educational business, including the following:

- More comprehensive and relevant databases and assessment techniques, including value-added measures
- Evaluation components as part of the implementation of any new education policies or strategies
- Dedicated funding for independent evaluations
- Pilot programs to road test policies before they are implemented on a broad scale
- Enhanced roles for state departments of education in evaluating and identifying effective strategies and in providing information on them to local districts
- Increased flexibility for local districts so they can adopt and implement promising programs that fill their particular needs

Such changes will ensure that scientists have the necessary data and funding to conduct focused research on what works and what does not work under a variety of situations, that programs and strategies are appropriately and periodically evaluated after implementation, that the research is more credible, and that large-scale programs are not implemented until they have been tested in controlled pilot programs. They will strengthen the science behind education policy and ultimately strengthen our schools. They will not, however, eliminate the need for policymakers to make important judgments in evaluating and relying upon scientific evidence.

after school finance judgments offset increases partially or completely." Berry (2007), pp. 214, 223.

Chapter 7
Science and School Finance Decision Making

1. John Myers, partner in Augenblich & Myers consulting firm, stated: "Historically adequacy was determined politically using input measures and available resources. . . . Now adequacy is technically determined and output oriented" (Dunn and Derthick [2007b]). Indeed, consultants do not limit themselves to determining "needed" inputs. They also decide upon the desired outcomes in most such studies. For example, in the New York study sponsored by CFE, the outcomes on which the study was based were specified by the consultants after obtaining input from public forums organized by CFE, and bore little, if any, relationship to the constitutional standard enunciated by the court.

2. The same rules apply to making decisions in other policy areas, although the actual application is generally beyond current knowledge. For example, one could conceptually judge whether spending more on education is better than spending a dollar more on health care by comparing the gain in society's well-being from a dollar spent on education and health care. But in reality we have poor measures of society's "well-being" and how well-being is affected across areas. Thus, we ask legislatures to make their best judgments about how limited public money should be allocated between, among other things, education and health care.

3. See, for example, Slavin (2007).

4. A review of past costing-out studies can be found in *Education Week*'s annual report for 2005, *Education Week* (2005). See also the ACCESS Project website (www.schoolfunding.info), a project of the Campaign for Fiscal Equity (CFE), the plaintiffs in the New York City adequacy case.

5. This explains why the websites for advocacy organizations give top billing to costing-out studies. For an example, see the ACCESS Project website.

6. See, e.g., *CFE II* (2003); *CFE III* (2006); *Montoy* (2005A).

7. Augenblick & Myers, who have conducted costing-out studies in more than twenty states, were hired by the Kansas legislature, and concluded that an "adequate" education in Kansas would cost approximately $850 million over what the state was spending at the time. The study and its results were introduced into evidence at trial, and the court held they were binding against the state, ruling that they showed the state was underfunding education to the tune of about $850 million. The Kansas Supreme Court also held the state to the study's conclusions and affirmed Judge Bullock's order. *Montoy* (2005). Subsequently, the amount required was reduced to approximately $755 million based on the results of another costing-out study. *Montoy* (2006).

8. See, for example, the discussion and analysis of poverty and California school performance in Sonstelie (2007) and Loeb, Bryk, and Hanushek (2008).

9. Hanushek (2003).

10. As noted, the research on spending and achievement does not imply, as some critics say, that "money never matters" or that "money cannot matter."

11. An early and insightful discussion of the importance of skill differences is found in Murnane and Nelson (1984).

12. Health care, on the other hand, has many parallels to education. Not surprisingly, the best way to provide health care services is subject to many of the same controversies.

13. An example of this is Augenblick Palaich & Associates (2003), p. II-3.

14. Examples of this approach include the following costing-out study reports: Augenblick & Myers (2002b); Augenblick, Myers, Silverstein, and Barkis (2002); Augenblick Palaich & Associates (2003); AIR/MAP (2004); Picus, Odden, and Fermanich (2003); Verstegen and Associates (2003). A majority of these reports are available from the websites of the relevant states.

15. Often the consultants will enhance the teacher salaries if they believe the prevailing salary schedules are lower than the surrounding states, e.g., Augenblick Palaich & Associates (2003). Such an approach is obviously arbitrary (but one-sided since it always involves raising salaries). As noted, variations in teacher salaries are unrelated to effectiveness in the classroom, and little research suggests that changing the average salary will affect student outcomes. See Hanushek and Rivkin (2004, 2006); Hoxby and Leigh (2004).

16. For example, Augenblick & Myers (2002a), appendix D-1 (instructions to professional Judgment School Site Panel members).

17. *Campbell County* (1995), p. 1279.

18. Augenblick Palaich & Associates (2003), p. IV-15.

19. The terms *cost*, *cost efficiency*, and *efficiency* are often used interchangeably to indicate the minimum spending required to achieve a given outcome. For example, in comparing different programs that are designed to achieve the same outcome, the cost-efficient one would be that requiring the least spending. It is important, however, to understand that the cost-efficient choice applies only to alternative ways of achieving the same objective. If one wished to, say, compare programs that yielded different outcomes, the efficient one would not necessarily be the one with the smallest expenditure.

20. See notes 16 and 18, this chapter.

21. *Hancock* (2004), p. 283.

22. Augenblick and Myers (2003), p. IV-5.

23. Interview with Deidre Roney, counsel for state during Hancock trial.

24. *Hancock* (2004), pp. 283–84.

25. AIR/MAP (2004), p. 3, note 12. This "warning label" contrasts sharply with the claim in the November 2002 AIR/MAP proposal that their study would answer the question, "What does it actually cost to provide the resources that

each school needs to allow its students to meet the achievement levels specified in the Regents' Learning Standards?" See Hanushek (2006c).

26. Augenblick Palaich & Associates (2003).

27. *Committee for Educational Equality* (2004), State Exhibits 761, 766.

28. AIR/MAP Study, p. 7.

29. Williams (2005), p. 28; Williams (2007), p. 202.

30. See Odden, Fermanich, and Picus (2003).

31. See Odden, Picus, Goetz, and Fermanich (2006). This analysis is similar to an earlier study conducted by the same consultants for the state of Wyoming. Odden et al. (2005).

32. For more detail on the problems with these analyses, see Hanushek (2007a).

33. Odden, Picus, Goetz, and Fermanich (2006) provide information about the "effect size" of each program they select. The effect size indicates how large of a change in student outcomes could be expected from the program, where change is measured in standard deviations of the outcome. The figure simply translates these changes into movements in the overall performance distribution. For example, a change of one standard deviation would move a person at the middle of the distribution to the 84th percentile of the distribution.

34. Odden, Picus, Goetz, and Fermanich (2006) report that class size reduction would improve student achievement by 0.34 standard deviations. We simply report the implied change in achievement. This magnitude of change is equivalent to moving up 13 percentiles in the distribution—the result shown in the figure.

35. The technical basis for this conclusion comes from their assessment of the "effect sizes," or the predicted standard deviations of improvement in achievement. Their model school is reported to have a total effect size of 3.0 to 6.0 standard deviations, a completely implausible outcome that would place the average beyond the 99.9 percentile of the prior distribution.

36. The consultants often justify their inclusion of all the recommended programs as being required to achieve "adequacy." However, the notion that the *average* student in a state must score at, for example, the 99th percentile, in order for the education to be "adequate" is ludicrous.

37. Odden et al. (2008) reports on the results of a study for the Wyoming legislature about district spending after acceptance of their prior study on an evidence-based funding for Wyoming schools (Odden et al. 2005). The follow-up study is available at: http://legisweb.state.wy.us/2008/interim/schoolfinance/schoolfinance.htm (accessed July 24, 2008). A similar analysis for Arkansas supports the same general findings.

38. Odden et al. (2008).

39. The same question, of course, arises with the professional judgment model, where there is little reason to believe that schools actually array themselves in the ways chosen by the panels. One difference, however, is that the professional judgment panels seldom get into any programmatic detail but instead stop at ge-

neric descriptions of numbers of people of specific kinds and perhaps numbers of computers and the like.

40. See, for example, Augenblick & Myers (1997); Myers and Silverstein (2005); Standard & Poor's School Evaluation Service (2004).

41. In some instances, there may be a considerable difference between spending levels in the "successful" districts, and the issue arises as to whether the high spenders, or a portion of them, should be eliminated because they are not efficiently using the funds available to them. This was a major issue in the *CFE* case. While the trial court rejected the use of an "efficiency screen," the New York Court of Appeals reversed and approved the state's cost study, which eliminated the top half of the spenders in calculating the spending levels needed to be a "successful" district. *CFE III* (2006), pp. 18–20.

42. Another potentially important issue is whether the analysis has identified schools that are truly successful. The test score measures that are used to pinpoint those schools doing well are often fraught with error. As pointed out in terms of school accountability, this can lead to some schools being identified as good during one year and bad during another. See Kane and Staiger (2002) and Peterson and West (2003).

43. Thompson and Silvernail (2001); Carey (2002).

44. Bruce Baker, who reviewed weights for the evaluation of school finance systems by *Education Week*, stated: "Rarely is there any empirical evidence to influence weighting. . . . [T]here are many layers of arbitrary and political decisions applied in each weight" (Park 2005). Even Michael Rebell, a strong proponent of costing-out studies, acknowledges that such weightings have little empirical or scientific basis. Rebell (2006b), pp. 56–57.

45. *Hancock* (2004), p. 280.

46. In economics and other quantitative sciences, one variable is said to be a function of another if its level is shown to vary, whether positively or negatively, in response to changes in another variable. For example, when the price of gas increases, demand for gas goes down; demand for gas is therefore a function of price. The cost function label reflects the assumption made in these studies that the level of required spending in a district varies predictably along with various observable characteristics of its students and the desired achievement level.

47. Cost function analyses generally employ standard regression models but at times use alternative methods that weight certain districts, those presumed to be more efficient in their spending, more heavily. See, for example, Gronberg, Jansen, Taylor, and Booker (2004). These alternatives also suffer from a series of problems, but they are not explicitly discussed here. For a critique of cost function studies, see Costrell, Hanushek, and Loeb (2008).

48. Guthrie and Rothstein (1999), p. 223.

49. Duncombe (2006) argues that one way to assess these models is measurement of "predictive validity." Thus, if the predicted spending by districts in years other than those on which the estimation is based is close to actual spending levels,

he would say the models are validated. However, this only shows a continuation of the school spending patterns of the past, and not that such funding was needed to reach any particular achievement levels. The validity of cost functions should be based not on their ability to predict future expenditures but on the ability to predict the outcome improvements that would result from different levels of spending.

50. Hanushek (2003). These studies are generally referred to as "production function" analysis. They are statistical analyses (using many of the same approaches as the "cost function" analyses discussed here), except they generally try to explain variations in achievement across students, schools, and districts. The cost function studies try to explain variations in spending.

51. A more complete description of the problems of the cost function estimation and the comparison with the production function approach is found in Costrell, Hanushek, and Loeb (2008).

52. The studies into the finance and governance of California schools were part of an overall evaluation of how to improve the state's schools. See Loeb, Bryk, and Hanushek (2008).

53. Imazeki (2007, 2008).

54. Guthrie and Springer (2007a), p. 106.

55. Imazeki (2007, 2008) and Loeb, Bryk, and Hanushek (2008).

56. There is nothing special about the subject, the grade, or the state, and using a different grade or test would still yield a picture qualitatively similar to figure 7.3.

57. Baker (2006) included measures for state aid/pupil, income/pupil, the resident tax ratio, and the population aged sixty-five or greater in order to eliminate differences in the efficiency of spending across districts. Duncombe (2007) included measures of the fiscal capacity of the district and voter turnout.

58. Both methods have to deal with the fact that typically there are no districts that achieve at the performance levels defined as adequate. In such cases, the consultants typically assume that the relationship between spending and achievement they identify remains the same regardless of achievement level. That is, if they observe proficiency levels to be increasing by ten percentage points for every additional $1,000 per pupil spent in a set of districts with a maximum proficiency rate of 60 percent, they assume that the relationship remains unchanged as districts near the target of 100 percent proficiency. There is, of course, no way to know whether that is true.

59. Cost function analyses make assumptions about the way in which various factors based on student characteristics, such as the percentage of low-income students in a district, affect required costs. However, it is unclear, for example, whether the evidence from Westchester County is at all informative about how to improve student achievement in the Bronx or about precisely what adjustments would have to be made to account for the many differences in the two locations. Yet that is exactly the kind of analytic leap of faith that a cost function study

conducted in New York State was forced to make, leading one study to conclude, apparently seriously, that New York City needed to spend 3.5 times as much per student to obtain the same level of achievement as other districts in New York State. Duncombe and Yinger (1998). Their subsequent estimates of adequacy for New York City are lower but still far above those of the other costing-out studies for New York. See, for example, Duncombe and Yinger (1998), pp. 129–53.

60. The studies were done at different times. To ensure that he was not comparing apples with oranges, he adjusted the results of each study to reflect the dollar amounts deemed necessary to achieve adequacy in 2004 dollars.

61. *Education Week* (2005), p. 39.

62. Augenblick & Myers (2003), p. ES-3.

63. *Education Week* (2005), p. 36.

64. E.g., *Williston*, Deposition of Tom Decker, August 17, 2005, p. 312.

65. Guthrie and Springer (2007b).

66. The state of Maryland actually used such a study as the basis for its school finance system; however, whether the system will ever be fully funded remains to be seen. Wyoming may be an exception where the court, until its recent change of direction, ordered the legislature to fully fund the consultants' cost function results. See chapter 6.

67. See the discussion in the text of the AIR/MAP (2004) study and note 25, this chapter.

68. See note 37, this chapter.

69. For example, in 2002, two constitutional amendments were passed in Florida, one requiring class size reductions and the other requiring the implementation of preschool programs. Fla. Constitution, art. IX, §1.

70. For a history of comprehensive school reform, see Borman, Hewes, Overman, and Brown (2003). The Abbott school districts in New Jersey were, for example, required to adopt an approved whole school reform model.

71. For a history of the New American Schools, see Mirel (2001).

72. Evaluations of the implementation and success of these schools can be found in American Institutes for Research (1999) and a series of RAND reports including Berends, Kirby, Naftel, and McKelvey (1996) and Glennan (1998). In 1994 its demise was complete when it was absorbed into the American Institutes for Research; see Olson (2004).

73. Gramlich and Koshel (1975).

74. The analytical portions of this section are elaborated on in Hanushek (1999a). A broad independent assessment of class size policies can be found in Ehrenberg, Brewer, Gamoran, and Willms (2001a, 2001b).

75. That is, smaller classes provide these gains if teachers actually change how they teach and how they interact with the class as class size is reduced. It also assumes that the necessary hiring of additional teachers that goes along with a general class size reduction policy brings in new teachers who are equally effective as the existing teachers. These important matters are discussed later.

76. A description of the program and the results can be found in Word et al. (1990). The original design also investigated regular classrooms with teacher aides, but this portion was abandoned after initial investigation showed that the aides added nothing to student performance.

77. See Mosteller (1995) on the advantages of this design for studying class size reduction.

78. An analysis of the quality of the experiment is found in Hanushek (1999b).

79. For a discussion of the range of issues in the experiment and policy debate, see Mishel and Rothstein (2002).

80. The exact class sizes required are specified in detail. For example, the classes had to average near twenty between September and April, even if larger than twenty students in some month. Schools could implement it in some but not all grades K–3, but they had to start with all first and second grades. None of these details materially affect the discussion, however.

81. See the evaluation of the California program in Stecher and Bohrnstedt (1999).

82. Data from the state website: http://www.ed-data.k12.ca.us/welcome.asp (accessed December 31, 2007).

83. Calculating the cost of class size reduction is actually a complicated process that depends on the range of changes that take place with it. See Brewer, Krop, Gill, and Reichardt (1999).

84. As noted, the costing-out studies such as Odden, Picus, Goetz, and Fermanich (2006) do not attempt to address these issues (see Hanushek 2007a). There are questions about the cost of increasing teacher quality, but most analyses suggest that it would be less than the cost of the large-scale class size reductions of Tennessee. See, for example, Hanushek, Kain, O'Brien, and Rivkin (2005); Hanushek (2004).

85. The pattern of legislative changes across the states can be found in Zinth (2005).

86. Heckman (2006); Heckman and Masterov (2007); Carneiro and Heckman (2003).

87. Schweinhart et al. (2005) and Witte (2007). A comprehensive review of different pre-K programs and their evaluations can be found in Besharov, Germanis, Higney, and Call (2008).

88. Gramlich (1986); Barnett (1992); Galinsky (2006); Belfield, Nores, Barnett, and Schweinhart (2006).

89. *Abbeville County* (2000) (on appeal); *Abbott* (1998), pp. 473–74.

90. Campbell and Ramey (1995) and Campbell et al. (2001).

91. An extensive reanalysis of the data from these programs has been conducted by Anderson (2007). He attempts to correct for attrition and multiple outcome evaluations along with statistical innovations.

92. Reynolds, Temple, Robertson, and Mann (2002). The evaluation of the program relies on matching participants with comparable students in similar

schools. The validity of this approach requires that the students similar on a few measured characteristics provide a good comparison group for what might happen if the program students were not in the program—an uncertain assumption.

93. Gormley, Gayer, Phillips, and Dawson (2005).

94. Anderson (2007).

95. The impact of differences in criminal activity are particularly important in the case of the benefit-cost analyses; see Gramlich (1986). The females did, nonetheless, generally have positive school completion results; Anderson (2007).

96. Cost estimates and programmatic comparisons are found in Witte (2007). Children actually participated one or two years, and the cost figure represents an average for the actual program. For a new program, costs would presumably be larger per child if all students participated two years.

97. Campbell and Ramey (1995).

98. Costs reflect per-child costs over preschool years. See Belfield and Schwartz (2007).

99. Early action is estimated to be very important in saving future costs of remediation and special education. For example, Lyon and Fletcher (2001) conclude that addressing early reading problems can substantially reduce subsequent special education costs.

100. Gormley et al. (2005), p. 873.

101. For a sense of the disagreements about programs and purposes, see Kirp (2007), Fuller (2007), and Finn (2009).

102. Fuller (2007), chapter 6.

103. The Chicago Parent-Child Centers is a public program, and thus differs from the other programs discussed, but its research design is notably weaker than the randomized experiments. Thus, the evidence marshaled to support broader adoption of preschool programs generally focuses on the experimental but non-public programs.

104. Head Start Bureau (2005).

105. Besharov, Myers, and Morrow (2007).

106. Relevant studies include Fryer and Levitt (2004); Currie and Thomas (1995, 2000); Garces, Thomas, and Currie (2002); Administration for Children and Families (2005).

107. Gormley, Phillips, and Gayer (2008) consider the impact of preschool for a special sample of students in Tulsa, Oklahoma, and conclude that the Tulsa program had positive impacts, ones that were larger than for Head Start. Lisa Snell, policy analyst at the Reason Foundation, observes nonetheless that students in Oklahoma have actually lost ground in fourth grade reading on the NAEP tests since universal preschool was introduced (Snell 2008).

108. This is not to say that the benefits—particularly in areas outside of cognitive achievement—are insufficient to potentially justify some types of public programs.