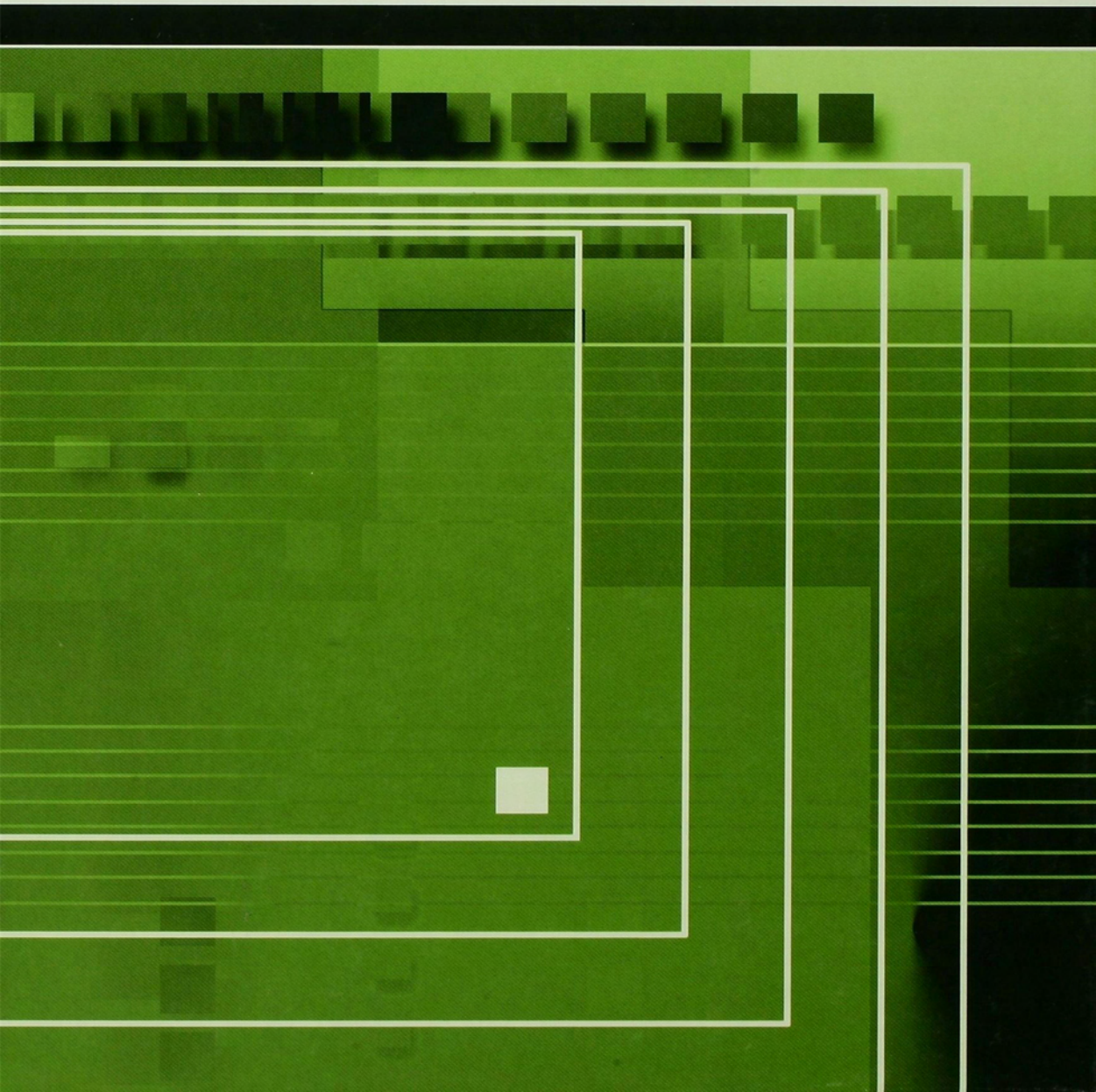
Edited by
Carol Anne Dwyer

Measurement and Research in the Accountability Era



MEASUREMENT AND RESEARCH IN THE ACCOUNTABILITY ERA

Edited by

Carol Anne Dwyer
Educational Testing Service

LEA
2005 LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Mahwah, New Jersey London

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

Measurement and research in the accountability era / edited by Carol Anne Dwyer.
p. cm.
Includes bibliographical references and index.
"The chapters here were originally commissioned as part of the ETS 2003 Invitational Conference . . . held in New York City in October 2003. The authors were invited to expand on their conference presentations for publication"—Introd.
ISBN 0-8058-5330-8 (alk. paper)
1. Educational accountability—United States—Congresses. 2. Educational tests and measurements—United States—Congresses. I. Dwyer, Carol Anne. II. ETS Invitational Conference (2003 : New York, N.Y.)

LB2806.22.M43 2005
379.1'58—dc22 2004059125
CIP

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability.

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Contents

Preface

The 2003 Educational Testing Service Invitational Conference provided an opportunity to convene leading scholars and practitioners to deliberate on the important topic of measurement and accountability. This conference continues a longstanding ETS tradition of seeking to advance the science of measurement, to illuminate important research issues, and to inform policy. Since 1936, the education community has responded to our invitation to address the most pressing technical and policy issues associated with the evolving science of measurement. The 2003 conference was a sterling example of this tradition.

With the theme of "Measurement and Research Issues in the Accountability Era," our hope was to provide the occasion to look at where measurement and research have been and to identify the challenges to our field presented by standards-based reform and accountability policies. Presenters and participants were focused and candid in their analysis of the capacity to use valid data in the service of student learning. This balance and passion are quite apparent in these chapters, and serve to illustrate the professional standards and the sense of civic responsibility required in this age of high-stakes decisions based on measurement products.

As the tradition of the ETS Invitational Conference continues, so does our determination to serve learners. We are most grateful to the authors of these chapters for their support in this regard. We are confident that readers will find invaluable guidance in their efforts to improve learning outcomes for all students.

—Sharon P. Robinson

Introduction

The subject of accountability generates as much heat as light in the context of today's school reform efforts. Little can be accomplished, however, by rhetoric alone. The current focus on accountability creates an opportunity for unprecedented attention to the critical elements of an effective accountability system. These elements include careful specification of what students should learn, creation of realistic opportunities for all students to learn what is required, finding reliable and valid evidence of what learning has taken place, and creating appropriate incentives to improve the system.

This volume attempts to bring to bear the best thinking of leading scholars and experienced practitioners on measurement and research issues in the development and implementation of scientifically rigorous and educationally sound accountability systems. The chapters here were originally commissioned as part of ETS (Educational Testing Service) 2003 Invitational Conference, *Measurement and Research Issues in a New Accountability Era,* held in New York City in October 2003. The authors were invited to expand on their conference presentations for publication.

Accountability systems often appear simple on the surface. Unfortunately, however, as H. L. Mencken wrote, "For every complex problem there is an answer that is clear, simple, and wrong" (Mencken, 1990), and never was this more true than in the case of school reform and educational accountability systems. As the stakes associated with accountability and school reform have risen to unprecedented heights, the need for better sci-

entific evidence on what works has become more and more apparent. Unfortunately, as this evidence emerges, new problems of interpretation are created: It becomes more and more difficult for policymakers, practitioners, and the general public to be sure what this evidence means to them.

It would be a mistake, however, to assume that adequate public awareness of the most basic accountability issues exists. In their 2002 survey commissioned by ETS, *A National Priority: Americans Speak on Teacher Quality,* the bipartisan pollsters Peter Hart and Robert Teeter document both the American public's general dissatisfaction with the state of public education and their lack of knowledge of important educational events. This dissatisfaction with schools is pervasive and of long duration, and there is little good news about progress over time in improving negative perceptions of our educational system. In Hart and Teeter's 2002 survey, over half of the general public felt that the American system of public education was deeply defective, and almost three quarters of the public are in favor of testing students and teachers, and of holding teachers and school administrators responsible for students' progress. From the point of view of the federal government and many states, great strides are being made to improve education through comprehensive (and expensive) accountability programs. Unfortunately, however, there is a tremendous gap between accountability efforts on the grand scale and even a minimal awareness of these efforts on the part of the public and among educators themselves. Interest in the topic clearly outpaces factual information. For example, Hart and Teeter found that only 12% of the public and 36% of teachers said they were aware that a major national school reform bill with bipartisan support and the approval of both houses of Congress, the No Child Left Behind Act (NCLB, 2002), had been signed into law in 2001. Although it is heartening that 63% of those identified as policymakers indicated awareness that NCLB exists, this figure is still far from the level of awareness that one might expect from educational policymakers concerning such a major piece of legislation as NCLB.

Findings such as those of Hart and Teeter make it clear that extraordinary efforts are needed if we are to fully inform the public and to marshal the political will required to meet the educational needs of all students, including those most at risk of being left behind: students living in poverty, English-language learners, and students with severe disabilities. Without effective intervention, the achievement gaps experienced by these students will continue to grow. Testing students is a necessary but not sufficient step to take to close these gaps; we need clear plans for effectively teaching those who have not learned up to standards, not just labels for these individuals and their schools. In addition to the ethical issue of taking collective responsibility for their education, we should also remember that all of these groups are growing in number, and will probably continue to do so.

The inevitable result will be that the practical and economic issues associated with the quality of their education will be greatly magnified in the future. If the promise of school reform for all children is to be realized at least in part by accountability systems, we must understand better how these systems can be made to work for the benefit of all students.

The authors in this book address the context in which educational reforms are taking place; present policy and technical analyses of the design and implementation of the NCLB and other major accountability systems currently in use; project trends for the future; and address the large framing questions of what works and how to bring all of the many elements of school reform and accountability into effective alignment.

We asked the authors to distill their research and measurement findings to provide guidance to the reader in understanding where we might find ways forward to educational improvements. For example, Andrew Porter's chapter focuses on achievement gaps. He first summarizes the enormous amount of research on current achievement gaps (how big they are, how stable over time, how stable over children's developmental span, and how important they are in practical terms), and then systematically evaluates the prospects for each of the major kinds of reform (preschool, teacher, instructional, standards based) that have been hypothesized to decrease these gaps. Eva Baker's chapter focuses on an equally large question, the alignment of components of educational accountability systems. She provides lucid guidance to understanding the nature of alignment itself, what we can reasonably expect of it, and how we might improve on the current state of widespread gaps in alignment.

Contributions such as these, and those of the other authors in this volume, help educational and measurement scholars, practitioners, policymakers, and others develop a deeper understanding of the data and the logic of accountability systems. In our new era of accountability, the importance of solid facts and empirically informed debate has never been more critical. As Ellen Lagemann reminds us in her chapter, it is a fundamental responsibility of the measurement and research community to provide reliable information that supports improved service to learners. This makes it a moral imperative as well as a technical challenge to improve the quality of measurement and research. Only then will we have the foundation needed to advance the learning of all students.

REFERENCES

Hart, P. D., & Teeter, R. M. (2002). *A national priority: Americans speak on teacher quality.* Princeton, NJ: Educational Testing Service. Retrieved April 7, 2004, from ftp://ftp.ets.org/pub/corp/survey2002.pdf

Mencken, H. L. (1990). *The vintage Mencken* (A. Cooke, Ed. Reissued ed.). New York: Vintage.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425 (2002).

PART

I

THE SCIENCE OF EDUCATION AND SCIENTIFIC EVIDENCE

CHAPTER

1

Toward a More Adequate Science of Education

Ellen Condliffe Lagemann
Harvard Graduate School of Education

I am honored to be at this conference today. The ETS Invitational Conference has a long and proud history. There was a predecessor to the Conference held under the auspices of the American Council on Education and several other testing services beginning in 1936. That was discontinued during World War II, and, then, the Conference literally became the ETS Invitational Conference on Testing when ETS (Educational Testing Service) was founded in 1947 (Woods, 1956). By that time, it had shifted its focus from state educational testing to personality and educational testing of many different kinds. Subsequently in the 1960s and 1970s, it shifted focus again, now paying more attention to making testing useful and understandable to teachers (Anastasi, 1966). Today, ETS is reviving a Conference that has not been held in a number of years and, if the current gathering is consistent with the ever-broadening trend line of the Conference's history, I suspect our conversations will expand again, ranging even beyond matters of testing per se to more general questions concerning accountability in education.

As we all know, we are living today in a new era in education. It is an era of unprecedented accountability at the state, local, district, school, and even classroom level. It is an era when we are expecting more of students, teachers, and school leaders and when we are paying more and more—and perhaps too much—attention to "high stakes tests." We live in an era when, to a degree unprecedented in history, we are asking that education policies be buttressed by rigorous, "scientific" research. These new expectations

are appropriate. We should provide an excellent and equal education to all children. We are the richest society in the history of the world and we owe it to all young people to guarantee that they will come of age ready to live meaningful and productive lives. That said, I also think we need to be realistic about the hurdles we will have to surmount to achieve this goal.

First of all, the economic downturn that is plaguing our country is undermining what progress there has been toward improving student achievement. State and local budget cuts will make it difficult to sustain professional development and tutoring programs. Financial stringency may also preclude the continuation and expansion of the after-school programs, mentoring, and summer internships that often make the difference between a student being able to finish school or dropping out. We continue to lose too many teachers and to see many of the best teachers leave high-poverty urban areas, where they are most needed, for higher paying suburban areas. Clearly, the opportunity side of the accountability equation is under strain, to say the least. That being the case, it is not clear that we will have the capacity to do what we hope in education.

Even without our current economic woes, however, meeting our expectations will be very, very difficult. We are just beginning to understand all that is involved in translating theory into practice in education. We now realize, for example, that, in addition to science, we need what I like to call "usable knowledge." Usable knowledge is knowledge derived from research that is then translated into the toys, texts, tests, and teaching materials that teachers and learners themselves can actually use to promote learning. Even though we are beginning to understand the importance of usable knowledge, there are relatively few researchers able to do this kind of work, there is little infrastructure to support it, and there are few opportunities for training.

What's to be done? Should we abandon our high hopes for education? Should we retreat from our commitment to educational equity? I think not. Rather, to create the knowledge and tools we need to meet our new expectations for education, I believe we need new standards of accountability for the education research community, new infrastructure for research, and new programs of research training. If we can create these things, I believe we will have gone a long way toward creating the conditions necessary to link educational theory and practice in powerful ways. I would like to talk about each of these in turn.

First, in this new era of ever-higher accountability, we need, I think, to ask more of ourselves as scholars of education. By tradition, most researchers have believed that they should be held accountable only for the significance of the questions they have asked and the appropriateness of the methods they have used to answer those questions. Although a number of researchers have focused their efforts on the consequences of testing, and

the late Sam Messick and others have argued that those consequences should be considered an aspect of validity, matters of actual use have generally been seen as beyond a scholar's control or responsibility. Now, I would propose that we should change this and hold ourselves accountable for the applied value of our research. We should commit ourselves to generating new knowledge that will actually have a positive effect on policy and practice. We should commit ourselves not only to illuminating theories, but also to engineering products.

Though some scholars of education will continue to be concerned with what has traditionally been known as "basic research," others should now direct their attention to the actual day-to-day problems of learning and teaching in real-world educational settings. They should engage in what the late Donald Stokes called "use-oriented basic research" (Stokes, 1997). This means that they will do mission-oriented work, directed toward the need to understand practical problems. Some should also move even beyond that to engage in actual design work or engineering. To suggest that we need to supplement more traditional, theoretical work with more novel forms of research and engineering seems to me to be an appropriate new level of accountability for the field as a whole.

How do we achieve this new level of accountability? We do it by working to develop new methods for education research, all the while being careful to match our expectations and promises for research with clear understandings of what particular methods can yield (Shavelson & Towne, 2002).

At least since the 17th century, people have assumed that our understanding and mastery of the world has depended on slowly accumulating knowledge, piece by piece, and adding each new piece to what was already known about some given phenomenon. Physical, chemical, and biological phenomena were dissected in order to isolate, observe, and analyze all of the discrete molecules, gases, or genes of which they were composed. In a sense, the guiding principle of science was reductionism, the assumption being that if one understood all the various parts, one would be able to grasp the overall operation of whatever one was studying. For positivists, it was also important to move sequentially from the positing of assumptions based on theories to the testing of those assumptions with empirical evidence.

Today, this view is still held to be very important in some circles and it is being challenged in others. Much of the work currently being supported by the new Institute of Educational Sciences falls in line with a "hard" science view in which one intends to isolate and define cause-and-effect relationships. One example is the research going on under the auspices of the What Works Clearinghouse, which is generating research syntheses of studies of the same educational phenomena. The Campbell Collaboration located at the University of Pennsylvania is doing similar work.

The goal of efforts such as these is to produce tested generalizations that can inform educational policy and practice. Just as important, these projects seek constantly to refine methods that have been developing over nearly 30 years, which allow one to integrate the findings of different studies. As Harris Cooper and Larry Hedges (1994) put it in their introduction to *The Handbook of Research Synthesis,* synthesizers are analogous to "the bricklayers and hodcarriers of the science world" (p. 4). They are meant to "stack the bricks . . . and apply the mortar" (p. 4) that can hold the edifice together. The problem is that no two bricks—no two studies—are exactly alike. The synthesizer's task, then, is to identify how discrete studies are both similar and different and to account for the differences, especially if those differences pertain to the effects of a treatment.

Even though research synthesis is a technique that was pioneered by scholars of education, it has been less widely used in education than in medicine. This must change because research synthesis carries significant promise of helping to strengthen education research. Cooper and Hedges (1994) insist, for example, that primary research should not be included in a research synthesis unless the study's findings were subjected to some kind of a statistical test. No longer will it be sufficient to say: "I looked at the treatment and control scores and I find the treated group did better" (p. 7). Instead of such vague statements, Cooper and Hedges demand more rigorous, replicable evidence. Assuming one tempers one's expectations for such work with knowledge that "final," "sure" answers are impossible, our real hope lying in achieving better and better estimates, efforts to improve our understanding of causation in education should be important.

Especially in light of the current, often ideological, federal emphasis on methods that some see as providing "final" knowledge about cause-and-effect relationships, it is important to acknowledge the limitations of "what works" research. As Frederick Erickson and Kris Gutierrez (2002) noted in a recent article, "a logically and empirically prior question to 'Did it work?' is 'What was "it"?'—'What was the "treatment" as actually delivered?' " (p. 21). They rightly argue that an overemphasis on matters of causation can oversimplify the complexity of education and the myriad local variations that always creep into actual interventions. They insist, too, that in rushing to determine "what works," we need to be careful not to overlook the side effects that may emerge later on. Respecting calls for more rigor in education research should help us build an ever more reliable body of knowledge in education so long as those calls are tempered with well-informed cautions about what increased rigor can and cannot contribute to knowledge about education.

While people in education research debate the pros and cons of research syntheses, randomized trials, and other forms of "rigorous" research, across the sciences one finds more and more efforts to go beyond

what some see as reductionist approaches to ones that are more complex. These may also be helpful to our thinking about methods in education research.

Some scholars are relying on graphing theory to study networks and discern the principles that organize them. Requiring what Duncan Watts, a Columbia sociologist who studies networks, has described as "the mathematical sophistication of the physicist, the insight of the sociologist, and experience of the entrepreneur," such efforts often involve teams of researchers (Watts, 2003, p. 304). That is also true of activities directed at developing a new science of "chaos." Scientists at the renowned Santa Fe Institute in New Mexico are leaders in this. As the science writer James Gleick (1987) observed in a book about this new field, chaos theory tries to make sense of "the irregular side of nature, the discontinuous and erratic side" (p. 3). It is concerned with the "puzzles," the "monstrosities," and the anomalies that science has traditionally left aside.

Efforts to develop new, more complex approaches to science are also appearing in education. Setting these within the context of more general changes in science should be helpful in matching expectations to methods. As I have argued in *An Elusive Science: The Troubling History of Education Research,* I believe that too often in the past, education research has proceeded in a vacuum, disconnected from developments in the arts and sciences. This is the historic legacy of the gender-related low status in which educational scholarship has been held (Lagemann, 2000). I believe this has weakened our field. If we want to develop the kind of knowledge we need to meet higher expectations for education, then I believe we need to be in conversation with people in other fields, who are on the cutting edge of thinking about problems of causation and scientific explanation.

I would like to give two examples of the kind of new work in education we need to discuss in that context. I believe that both have high potential to advance research in education. The first example has to do with new methods for describing changes in test scores over time. Judith Singer and John Willett (2003), two of my colleagues at the Harvard Graduate School of Education, have been developing these methods along with a number of other statisticians and they have now written a book designed to teach social investigators how to model and analyze change, relying on longitudinal data. Using multilevel statistical models, they have demonstrated both how one can describe within individual changes and analyze interindividual differences in change.

I am not a statistician, and I am not going to go into the details of Singer and Willett's (2003) work—for that, I would refer you to their new book, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* In any case, the more important point for my argument is that Singer and Willett, among others, have developed a dynamic model that will enable

scholars to describe and analyze multiple layers of change. This is crucial if one wants to understand human development, learning, or growth, not as single events, but as the multivariable, longitudinal processes they actually are. By collecting multiple waves of data over time, change becomes a discernible process that can be studied in systematic ways. Because understanding change is so essential to studying education, this represents more than incremental progress in the science of statistics. It represents a fundamental increase in our capacity to do education research and should be recognized as such.

If these new statistical methods provide one example of more dynamic, multifaceted approaches to education, the work Deborah Loewenberg Ball, David Cohen, Stephen Raudenbush, and Brian Rowan have been doing at the University of Michigan provides another. In a very large study of the relationship between resources and student outcomes, which was written up by Cohen, Raudenbush, and Ball (2002), the Michigan group found that to discern a relationship between resources and student outcomes researchers must, first, disaggregate the term *resources* to include "teachers' knowledge, skills, and strategic actions ... [as well as] students' experiences, knowledge, norms, and approaches to learning" (p. 85).[1] Then, they must examine how resources are used in instruction, which, in turn, depends on a wide range of variables—everything from parental attitudes toward the curriculum to school leadership.

The question the Michigan group asked has, of course, been contested among education researchers since even before publication, in 1966, of James Coleman's *Equality of Educational Opportunity.* And, now, using a much more complex and dynamic approach to its analysis, the Michigan researchers have addressed that question by posing a fundamental challenge to scholars of education. Finding that the value of resources for increased student outcomes depends on their use in instruction, they have observed that researchers will need to manage everything that is involved in shaping instruction. As the Michigan group itself did, researchers will have to step over the traditional question, which has been "how do the available resources affect learning," and instead ask: "What instructional approach, aimed at what instructional goals, is sufficient to ensure that students achieve those goals?" (Cohen et al., 2002, p. 109). Asking that question, they will, then, need deliberately to design and manage different programs of instruction. This will be necessary because understanding that the use of resources in instruction determines their value in terms of student achievement forces one to recognize that one must manage incentives for teachers to teach in particular ways and for students to learn. One also needs to

[1]Although he was not a coauthor of this particular essay, Brian Rowan is the study director for the Michigan group.

manage the educational environment, that is, the links between a school and surrounding institutions and between people working within a school.

Put most simply, if one accepts the findings of the Michigan group, one must acknowledge, as they do, that we need new, active approaches to the study of instruction. We need to develop specific and varied instructional programs—in their words "instructional regimes"—and then we need to experiment with different levels of resources to discern what the impact of these differing levels of resources will be on instruction (Cohen et al., 2002, p. 110). Again, using the language the Michigan group borrowed from medicine, scholars of education will have to run both efficacy trials, in which resources will be plentiful, and effectiveness trials, in which resources are constrained.

In the Cohen et al. (2002) write-up of the Michigan project, they acknowledge that there will continue to be important roles for other forms of research, ethnographies, surveys, and the like, that do not require large-scale active experimentation. Those kinds of research can offer important descriptive data. That said, the new approach they have modeled and described represents the boldest challenge of which I am aware to standard linear science in education. As they observe, taking up this challenge will be enormously expensive and especially difficult in the United States, owing to our highly diffuse systems of educational accountability and governance.

As part of our effort to develop new methods and to match those to realistic estimates of what they can yield, we need to concern ourselves with the norms of the education research community. Indeed, I believe we should face something that is widely known, but little discussed in public. Education is a very large field, with a great many different subspecialties, and a good deal of important research going on. Despite that, it seems clear that we do not have a strong research community for the field as a whole. To be sure, subfields have strong research communities, but these are not well linked to one another. That is the case because we do not have common norms and standards across the field to differentiate good from bad research and we do not have common standards for research training. There is not a body of knowledge that everyone engaged in education research needs to master and there are no skills that all need commonly to hold. In consequence, we have a cacophony of theories and methods, but lacking common norms, we cannot produce authoritative, warranted knowledge to offer to policymakers and practitioners.

More a community in name than as a result of powerful interdependencies, the education research community mirrors the diffuse systems of accountability and governance for public education that have grown up in the United States. These derive from long-established traditions of local control. Leaving room for local traditions, cultures, and values to play a role in the shaping of educational policies and practices is very important. Indeed,

some would have it that most practical questions in education should be left to local decision making. The linguist David Olsen has recently argued, for example, that research should provide local educators with "an elaborated set of options," that they can then "combine with the accumulation of local experiences as to what works well and what works less well with their students and staff, in their school, in their community" (Olsen, 2003, p. 23). Even acknowledging the value of local influences, it is nonetheless clear that, to build a stronger research community in education, there will need to be concerted efforts to develop common standards and to gain support for them. This will be especially important as design and development work become increasingly significant alongside discipline-based studies. Disciplines are built around common standards. Newer styles of more applied research will need to develop them.

Design experiments are an example of this. According to Allan Collins, a leader in pioneering this work, design experiments are intended to introduce an innovation into a "real" educational setting and, then, through careful observation and quantitative comparison with control settings, to refine it and refine it again, all the while working at a practice site (Collins, Joseph, & Bielaczyc, 2002). In contrast to laboratory work, design work is located in "messy situations." It involves multiple rather than single dependent variables. One cannot hold variables constant, follow a fixed design, or test a hypothesis. One cannot even control the experiment. One is merely a coparticipant with others—notably, teachers, students, parents, and so on (Collins, 1999).

Design experiments represent very new, really emergent methods for simultaneously improving and studying education. Although they have gained credibility in the learning sciences community—one of the subcommunities of our field—they have not yet gained sufficient credibility in the wider world of educational scholarship. For that to happen, standards will need to be developed that, as Allan Collins, Diana Joseph, and Katerine Bielaczyc have argued, "make design experiments recognizable and accessible to other researchers" (Collins et al., electronic communication, Summer, 2003). The fact that proponents of design work are taking up this challenge bodes well for the emergence of explicit, shared norms for education research.

In combination, I believe that the continuing development of new methods coupled to realistic expectations and strengthened norms for research will slowly, over time, enable us to meet the heightened standards of accountability that, I believe, we, as education researchers, should now feel obligated to meet. In this way, we will provide the knowledge and tools we need to educate all children well. However, as I said at the start, I believe we also need new infrastructure for research and new models for research training.

The infrastructure that already exists at universities and other research centers is sufficient to support discipline-based research. It is not sufficient, however, to support more applied design and development work and explicit programs of research such as the one carried out by the researchers from the University of Michigan. To ensure that such work continues to advance, we will need to build new structures to facilitate research in multiple sites, to promote collaboration between and among researchers working on related problems, and to share data and works-in-progress even before they are ready for formal publication.

Recognizing this, the National Research Council has recently published the final report of the Committee on a Strategic Education Research Partnership (SERP). SERP is at once *"a program of use-inspired research and development," an organization* that will provide national and local infrastructure for research, and *a partnership* among researchers, practitioners, state policymakers, foundation officials, and other corporate, government, and nonprofit leaders. It is intended to provide opportunities for long-term, sustained research and development. Based on a careful analysis of current problems in the development and application of fundamental knowledge to educational problems, SERP is an expensive and ambitious plan for a major innovation in the way education research is mounted in the United States (Donovan, Wigdor, & Snow, 2003).

SERP proposes a research center, or "hub," with spokelike relationships between the center and networks of local researchers. Whether this structure will prove too cumbersome remains to be seen. It may turn out that we need instead to build confederations of smaller communities of practice (Wenger, 1998). Such communities would share conceptions of significance and method and offer opportunities for discussion and criticism. They could provide settings for concentrated work over extended periods of time, thereby helping to cumulate knowledge. They would be known for their expertise on the questions their members think most important and study. They would be linked to one another through collaborations born of necessity when the expertise of one community of practice could help another community with its ongoing research (Wenger, 1998).

Communities of practice already exist within some schools of education, within some departments within schools of education, and, in a virtual distributed sense, among colleagues at different locations, who share works-in-progress on a regular basis. At times, I am sure they have also existed within research institutes and think tanks. Given their importance, however, especially for researchers in training, I think schools of education should now make more deliberate efforts to build communities of practice.

Currently, faculty members in schools of education—as well as in other parts of research universities—carry out research in public schools, afterschool settings, Head Start centers, and the like. Following principles of aca-

demic freedom and faculty autonomy, these research projects tend to be located wherever individual faculty members have personal connections. If faculty were instead willing to place some of their projects in locations where their university had a research/practice site, this would help aggregate interventions and perhaps increase their effects. In this way, particular school districts or neighborhoods could become laboratories for the design and evaluation of comprehensive educational services and the scholars, practitioners, and policymakers involved could develop into a community of practice with shared norms and standards. This should not in any way suggest that *all* research should be carried in local practice sites. There will, of course, be times when scholars need national data or data from several different locations. My suggestion is rather that we supplement existing research with studies that proceed within a local research/practice site.

Developing the infrastructure required to supplement basic work in disciplines with programs of more applied studies will be very expensive. It is estimated, for example, that the start-up costs of SERP will be approximately $500 million over the first 7 years. Clearly, therefore, the recognition that we must engage in this kind of work, if theory is to connect with practice in education, represents a challenge to funders of education research, both public and private. If funders believe that it is important to guarantee all children opportunities to learn to high levels, then, individually or in partnerships, they must dedicate the resources needed to build a strong infrastructure for education research. After publication of the Carnegie Foundation's famed *Flexner Report on Medical Education* in 1910, this was done in medicine. Now that must be done in education. Solo-practitioner research projects are fine and may yield very important new knowledge. Traditional research carried out in offices and libraries will always be necessary. But funders now need to go beyond the support of such projects to build the educational equivalent of teaching hospitals, which can invent and experiment with innovative solutions to educational problems.

That brings me to my third and last point, our need for new patterns of research training. I am convinced that there are three essential elements in the preparation of researchers who can work effectively in education. The first is a core curriculum that will promote students' capacity to be articulate about education. Students need to be able to articulate what purposes education can, does, and should serve. They need to understand and be able to describe what education is as a process; how learning is related to neuroscience, cognitive science, human development, and culture; why teaching is both an art and a science; and why educational problems should always be viewed through multiple lenses.

The second essential element of research training in education should, in my view, involve disciplinary study in a faculty of arts and science. Disciplines provide characteristic ways of asking questions, analyzing data, and

presenting findings. They, quite literally, discipline one's thinking and, by doing so, deepen one's capacity to understand problems, albeit from a particular, disciplinary, perspective. Historians think about change over time. Time—when something happened—is always critical to their analyses. Anthropologists, by contrast, think about patterns in cultures via the analyses of language, gesture, and kinship relations, among other things. Time is not a dimension that has as much importance to them as it does to historians. In order to offer doctoral students the disciplinary work they need, faculties of education would do well, I believe, to partner with faculties of arts and sciences to create joint degree programs. In addition to strengthening the doctoral preparation of education researchers, this should help diminish the isolation that has traditionally kept schools of education at the margins of universities.

Finally, research training in education must involve practicum experiences in which students work as apprentices on a research team. Preferably, this experience will expose students to the complexity and importance of applied design and development work. It will help them realize that, although disciplinary thinking is important, interdisciplinary teams most effectively address educational problems. Recognizing that such experiences might prolong graduate study longer than is desirable, it might be that schools of education should develop significantly increased opportunities for postdoctoral fellowships that would enable young researchers to gain experience working in the field. As they do this, they will also have to negotiate with the powers that be in research universities concerning standards for tenure. Because tenure standards have traditionally placed a much higher value on original contributions to knowledge than on the applied value of one's research, they have often served as a disincentive, tending to prevent young scholars from engaging in design and development work. Perhaps following precedents for judging applied work in schools of engineering or architecture, faculties of education can begin to address this issue.

Of course, to sketch a general program of research preparation is much easier than designing curricula in detail. Designing actual curricula will be enormously difficult. And that is the point I would like to make in conclusion. To do all that I have suggested needs doing will present challenges to every person and every institution involved in education research. It will require that researchers rethink the ways they conceive and do research. It will necessitate even more deliberate efforts than exist today to articulate common norms and standards. It will demand that funders operate differently. It will make it necessary for schools of education to revamp their doctoral programs.

As a historian of education, I have sometimes toyed with the idea of writing a history of failed efforts to mobilize a Flexner-like revolution in educa-

tion. These began in 1920 when the Carnegie Foundation published a bulletin entitled *The Professional Preparation of Teachers for American Public Schools.* Attempts to raise standards of training in education continued with J. B. Conant's *The Education of American Teachers,* which appeared in 1963. They were carried forward again by a study Charles E. Silberman did for the Carnegie Corporation of New York about "the education of educators," which was eventually published, in 1970, as *Crisis in the Classroom: The Remaking of American Education.* Representing only a very few of the high points of a continuing refrain, none of these works have had the desired effect. What, then, will it take to develop the research methods, norms, and standards, the infrastructure, and the research training we need to deliver instruction with sufficient power to help all children master the knowledge and skills they need to be productive workers, effective citizens, and satisfied human beings? I suspect that rather than a single dramatic report, it will take steady, determined work on the part of people like all of us in this room. We face huge challenges in education, but none so huge that they cannot be surmounted.

ACKNOWLEDGMENTS

I am grateful to Jen DeForest for help with the research on which this chapter is based. David Cohen, Michael Feuer, Daniel Koretz, Fritz Mosher, and Judith Singer provided very helpful comments.

REFERENCES

Anastasi, A. (1966). *Testing problems in perspective.* Washington, DC: American Council on Education.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2002). Resources, instruction, and research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 80–119). Washington, DC: Brookings Institution.

Collins, A. (1999). The changing infrastructure of education. In E. C. Lagemann & L. S. Shulman (Eds.), *Issues in education research* (pp. 289–298). San Francisco: Jossey-Bass.

Collins, A., Joseph, D., & Bielaczyc, K. (2002). *Design research: Theoretical and methodological issues.* Retrieved from http://www.extension.harvard.edu/2002-03/programs/cte/ext02drt.pdf

Cooper, H., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 4–14). New York: Russell Sage Foundation

Donovan, M. S., Wigdor, A. K, & Snow, C. E. (Eds.). (2003). *Strategic education research partnership.* Washington, DC: National Academies Press.

Erickson, F., & Gutierrez, K. (2002). Culture, rigor, and science in educational research. *Educational Researcher, 31*(8), 21–24.

Gleick, J. (1987). *Chaos: Making a new science.* London: Penguin.

Lagemann, E. C. (2000). *An elusive science: The troubling history of education research.* Chicago: University of Chicago Press.

Olsen, D. R. (2003). *Psychological theory and educational reform: How school remakes mind and society.* Cambridge, England: Cambridge University Press.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academies Press.

Silberman, C. E. (1970). *Crisis in the classroom: The remaking of American education.* New York: Random House.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation.* Washington, DC: Brookings Institution Press.

Watts, D. J. (2003). *Six degrees: The science of a connected age.* New York: Norton.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* Cambridge, England: Cambridge University Press.

Woods, B. D. (1956). *Testing—Then and now: Invitational conference on testing problems.* Princeton: Educational Testing Service.

CHAPTER

2

Scientific Evidence and Inference in Educational Policy and Practice: Implications for Evaluating Adequate Yearly Progress

Robert L. Linn
University of Colorado at Boulder
National Center for Research on Evaluation, Standards, and Student Testing

The No Child Left Behind Act of 2001 (NCLB, 2002), which amends the Elementary and Secondary Act of 1965, is a law that provides billions of dollars in federal aid for a wide range of educational programs. As was noted by Feuer, Towne, and Shavelson (2002), NCLB "exalts scientific evidence as the key driver of education policy and practice" (p. 4). Indeed, "scientifically based research" is one of the dominant themes in the law. Provisions throughout the law require states and districts to demonstrate that funds obtained under the law will be spent on programs that are supported by scientifically based research. In the realm of education, this emphasis on scientific evidence is unprecedented.

Accountability also has a prominent role in NCLB: "The passage of the NCLB Act marked a significant shift in Federal educational policy from an emphasis on standards and assessments to an emphasis on accountability—school, district, and state accountability for students' academic achievement such that **all** students reach, at least a minimum, proficiency on the States academic achievement standards and the State academic assessments by 2013–14" (Marion et al., 2002, p. 5, emphasis in the original). The demonstration of adequate yearly progress (AYP) by schools and school districts is a key component of the accountability requirements in NCLB for schools and districts receiving Title I funds. States are required to define AYP for the state, school districts, and schools in a way that enables all students to meet the state's student achievement standards by 2013–2014.

In keeping with the stress on scientifically based research, NCLB requires states to develop accountability systems that are "valid and reliable." The AYP definitions must apply "the same high standards of academic achievement to all public elementary school and secondary school students in the State"; must be "statistically valid and reliable"; and must result in "continuous and substantial academic improvement for all students" (NCLB, § 1111(b)(2)(C)(ii–iii)). Furthermore:

> [The AYP definitions] must include separate annual measurable objectives for continuous and substantial improvement in both mathematics and reading/language arts for all students considered as a whole as well as for each of the following specific subgroups of students: students who are economically disadvantaged, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency. (Department of Education, 2002)

VALID AND RELIABLE

The terms *valid* and *reliable* are almost redundant in everyday usage and are interpreted to mean trustworthy or accurate. Distinctions are made in scientific uses of the two terms by measurement specialists, however. Reliability refers to the consistency or replicability of measurements. Validity, the more important of the two concepts, refers to the degree to which inferences and interpretations of measurement results are justified by supporting evidence. Thus to say that AYP definitions must be reliable implies that they should result in consistent classifications of schools as meeting or failing to meet AYP targets. Consistency of classification is highly dependent on the number of students that enter into the determination of AYP for a school. The reliability of AYP will be lower for small schools than for large schools (Hill & DePascale, 2003; Linn & Haug, 2002). Consequently, small schools will tend to show uneven results from year to year, so that a small school that meets its AYP target in one year may fail to do so the next, not because instruction has become less effective but because of the low reliability of the results due to the chance variation in the achievement level for small groups of students in a given grade from one year to the next.

The reliability will also be lower for schools where AYP must be reported for multiple subgroups of students than for schools with the same total number of students, but without subgroups for which results must be separately reported (Hill & DePascale, 2003; Kane, Staiger, & Geppert, 2002). Thus, the reliability of AYP results will be considerably lower for an integrated school with, say, 200 students in Grades 3 through 5 comprising roughly equal numbers of African American, Hispanic, and White students than for a school with the same total number of students that belong to a

single racial/ethnic group. This is so because of the requirement that a school meet AYP targets not only for the total group of students, but for each of the subgroups for which separate reporting is required.

Unreliable determinations of the AYP status of a school certainly undermine the validity of the inferences that are based on AYP results. The inference that there was a decline in instructional quality in a small school that exceeded its AYP target by a comfortable margin in 2003, but failed to meet its goal in 2004 is not justified, and therefore invalid, if the result can be attributed to low reliability. In other words, reliability is a necessary condition for validity. A high level of reliability does not guarantee an adequate degree of validity, however. The reliability may be sufficiently high, for example, to be quite certain that the School A met its AYP target whereas School B did not. That is, the classification of the two schools as meeting and failing to meet AYP targets is dependable and would likely be replicated with another cohort of students in another year. Inferences about the two schools, however, may or may not be valid. Evidence may or may not support the inference that the instructional program in School A is good and that students are making good gains in achievement. Similarly, the inferences that the instructional program in School B needs improvement or that only small gains in achievement are being made by students in School B may or may not be justified. Some specific examples may help clarify the fact that reliable determination of the AYP status of schools is insufficient for assuring valid inferences about schools and student achievement.

DEFINITIONS OF AYP

Some additional details about the definition of AYP and AYP targets are needed in order to illustrate some of the features of AYP that affect the validity of inferences about schools based on AYP results. In order to track their AYP toward the goal of 100% proficient or above by 2013–2014, states have to define percentage proficient or above starting points. The starting point for each subject (reading/language arts and mathematics) is defined to be equal to the higher of the following two values: (a) the percentage of students in the lowest scoring subgroup who achieve at the proficient level or above; and (b) "the school at the 20th percentile in the State, based on enrollment, among all schools ranked by the percentage of students at the proficient level" (NCLB, § 1111 (b)(2)(E)(ii)). In most cases the latter value will be the higher one and will define the starting point.

Once a state has established an AYP starting point, it must then set intermediate goals for AYP that will ensure that all students meet the state's definition of proficient achievement by 2013–2014. The intermediate goals must "increase in equal increments over the period covered by the State's time-

line, . . . provide for the first increase to occur in not more than 2 years, . . . [and] provide for each following increase to occur in not more than 3 years" (NCLB, § 1111 (b)(2)(H)). The equal-increments provision has been interpreted by the U.S. Department of Education (2002) in a way that allows states to vary the number of years between constant increments in the percentage of students at the proficient level or above. Thus, two states that have the same starting points and 2013–2014 goals may set different intermediate goals, as is illustrated in Fig. 2.1 for States X and Y. The straight-line definition of intermediate goals shown for State X is the pattern that was presented by the U.S. Department of Education to illustrate the setting of intermediate goals. It is more common, however, for a state to elect to adopt a pattern of AYP goals that is similar to that shown for State Y. That is, AYP growth functions are specified that have increments that occur every 3 years until 2010, after which increments are required every year, thereby postponing until later years gains that have to be realized every year.

There is considerable evidence that gains in student performance on the tests tend to be greatest in the first few years after they have been introduced as part of an accountability system and then taper off in later years. That is, the pattern of improvement in percentage of students scoring at the proficient level or above is a trend line that shows a decelerating rate of improvement over years rather than the accelerating curves that a number of states have adopted for the AYP intermediate goals between 2002 and 2014. Thus, it can be anticipated that the AYP goals, which are likely to be hard to meet in the early years, will become increasingly difficult to meet in the out years of the program.

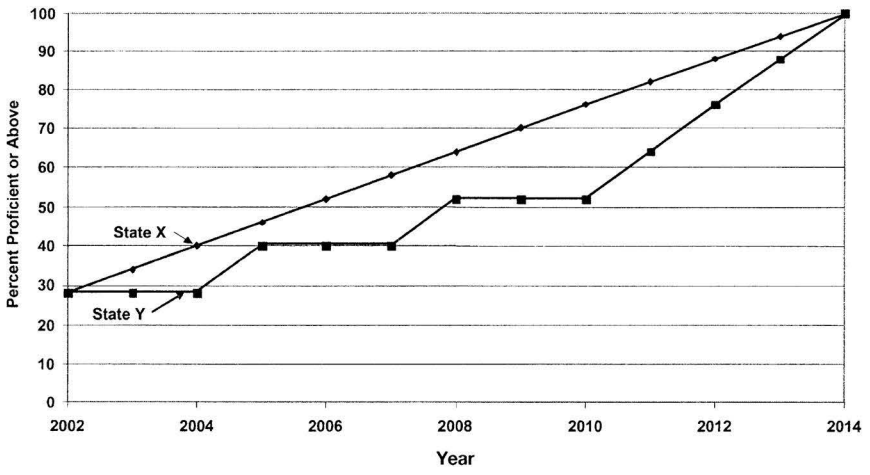


FIG. 2.1. Intermediate AYP percentage-proficient goals for two states with the same AYP starting points.

STATE-TO-STATE VARIABILITY IN AYP GOALS

Because states have their own assessments and establish their own definitions of proficient achievement, the starting points for states are radically different from state to state. Although some states have yet to define their starting points because they are in the process of introducing new assessments, starting points expressed as the percentage of students at the proficient level or above are available on state department of education Web sites for more than half the states. Starting points for Grade 4 mathematics assessments that are specified on the Web sites of 34 states range from a low (most stringent standard) of 8.3% proficient or above in Missouri to a high (most lenient standard) of 79.5% in Colorado. The corresponding range for mathematics assessments for the 34 states at Grade 8 is from 7% in Arizona to 74.6% in North Carolina (Linn, 2003b; Olson, 2003).

I doubt that anyone would believe that mathematics achievement is that much better in Colorado than in Missouri at Grade 4. Nor is the difference in starting point percentage proficient or above at Grade 8 a reflection of vastly better mathematics achievement in North Carolina than in Arizona. Instead, the huge differences in starting points reflect radically different definitions of proficient achievement in the different states. It is clear that valid inferences cannot be made about the relative proficiency of students in different states based on comparisons of percentage proficient statistics from state assessments employing such widely discrepant definitions of proficient achievement. Furthermore, the goals of 100% proficient by 2013–2014 lack comparability across states due to the different definitions of proficient student achievement.

The starting points and intermediate year percentage proficient or above AYP goals for the Grade 4 mathematics assessments for Colorado and Missouri are displayed in Fig. 2.2. A similar display is provided in Fig. 2.3 for the Grade 8 mathematics assessments for Arizona and North Carolina. As can be seen in the figures, all four states adopted a pattern of intermediate goals that follow a pattern similar to that shown in Fig. 2.1 for State Y. Because of the low percentage proficient or above starting points for Missouri and Arizona, the increments required in years with the big changes are necessarily quite large in comparison to the increments required for Colorado and North Carolina where the percentage proficient starting points are quite high.

From a comparison of the intermediate AYP goal lines for Colorado and Missouri in Fig. 2.2, and of Arizona and North Carolina in Fig. 2.3, it is evident that meeting AYP goals in any given year will mean quite different things for the two states involved in each comparison. Admittedly, the four states represent the extremes in terms of high and low percentage proficient AYP starting points at each grade level. Nonetheless, it seems clear that valid compari-

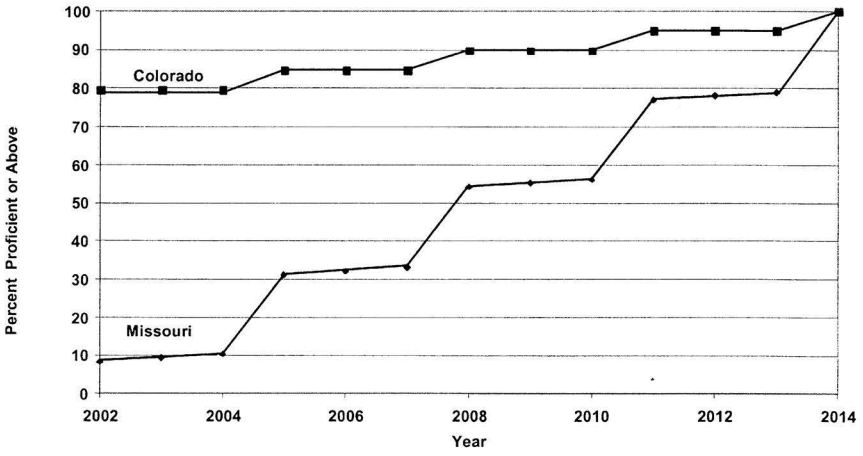


FIG. 2.2. Grade 4 mathematics percentage proficient or above AYP goals by year for Colorado and Missouri.

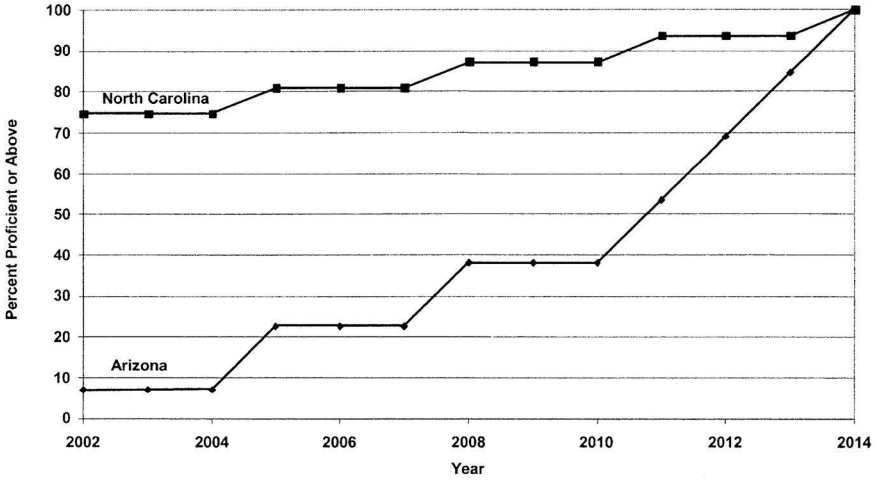


FIG. 2.3. Grade 8 mathematics percentage proficient or above AYP goals for Arizona and North Carolina.

sons across states will not be possible from the state assessment and percent proficient results reported under NCLB by each state.

NCLB does require states to participate in state administrations of the National Assessment of Educational Progress (NAEP) in reading and mathematics at Grades 4 and 8 in every other year starting in 2003. The law does not say what use will be made of the NAEP results, but there is some sense that they will be used to provide some kind of benchmark against which the

state results can be compared. In other words, NAEP will provide an external check on the validity of the reported percentage proficient or above and on the changes in those percentages over the next few years.

Because Arizona and North Carolina both participated in the 2000 NAEP mathematics assessments, it is possible to get some foreshadowing of the use of NAEP results as a benchmark. At Grade 8 the percentage of students scoring at the proficient level or above on the 2000 NAEP mathematics assessment was somewhat higher in North Carolina (30%) than in Arizona (21%) (Braswell et al., 2001). The difference on NAEP is much smaller, however, than the difference in the Grade 8 mathematics starting points for NCLB for these two states. Arizona's starting point of 7% proficient or above is 14 points lower than the Arizona Grade 8 percentage proficient or above in mathematics on NAEP in 2000, whereas North Carolina's Grade 8 mathematics starting point of 74.6% is almost 45% higher than the percentage of students who were proficient or above on the 2000 Grade 8 NAEP mathematics assessment.

Colorado did not participate in the 2000 NAEP mathematics assessment so comparisons similar to those for Arizona and North Carolina are not possible for 2000. Both Colorado and Missouri did, however, participate in the Grade 4 NAEP mathematics in 1996. In 1996 the percentage of students scoring at the proficient level or higher in mathematics at Grade 4 was 22% in Colorado and 20% in Missouri—a difference of only 2%, which is tiny in comparison to the difference of 71% in their Grade 4 mathematics starting points for NCLB (Shaughnessy, Nelson, & Norris, 1997). It is also worth noting that the improvements in the percentage of students performing at the proficient level or above provides another sharp contrast with the increases that have been realized on NAEP. The percentage proficient or above AYP goal for Grade 4 mathematics increases from the 8.3% starting point in 2002 to an intermediate goal of 31.1% in 2006—an increase of 22.8 percentage points in 4 years. Missouri participated in the Grade 4 NAEP mathematics assessment in 2000 as well as in 1996. From 1996 to 2000 the percentage of Grade 4 students in Missouri who were at the proficient level or higher increased by 3 percentage points (from 20% to 23%) (Braswell et al., 2001). Rapid acceleration of the gains in percentage of students performing at the proficient level or above will clearly be needed for the goals to be met not only in Missouri, but in other states as well (Linn, 2003a). Changes in the percentage of students at the proficient level or above on NAEP will provide a check on the validity of the increases reported by states.

USING AYP RESULTS TO COMPARE SCHOOLS

It seems clear that it will be difficult to make valid comparisons of states based on their AYP results. But what about the validity of within-state comparisons of schools or school districts? Are valid inferences likely to be

made about the quality of the instructional programs in schools? These questions are of fundamental importance to schools and districts that will face sanctions if they fail to meet AYP goals. Schools that fail to meet AYP goals for two consecutive years are placed in school improvement programs and the district must offer public school choice within the district. If the school fails to meet AYP goals the year after it is placed in school improvement, the district must provide supplemental services and technical assistance, which scientifically based research has shown to be effective. If the school still fails to meet AYP goals for the fourth consecutive year, it is subject to corrective action, which may include the replacement of staff. The school would be restructured in the following year if the school still did not meet AYP goals for a fifth consecutive year.

The solid line in Fig. 2.4 shows the percentage proficient or above AYP goal line that is similar to a fairly typical state. Also shown by the dashed lines are the percentage proficient or above results for three hypothetical schools. School A is a school where students have very low achievement with only 5% proficient or above in 2002, but where there is a steady and substantial increase in percentage of students who are proficient or above. Because School A has such a low starting point, it never reaches the AYP goals set by the state and would be subject to sanctions starting in 2004. Indeed, it would have to be reconstituted in 2007 despite the steady increase in the achievement of its students. School B just barely exceeds the state starting point in 2002 and has a steady, but more modest increase than School A in the percentage of students performing at the proficient level or

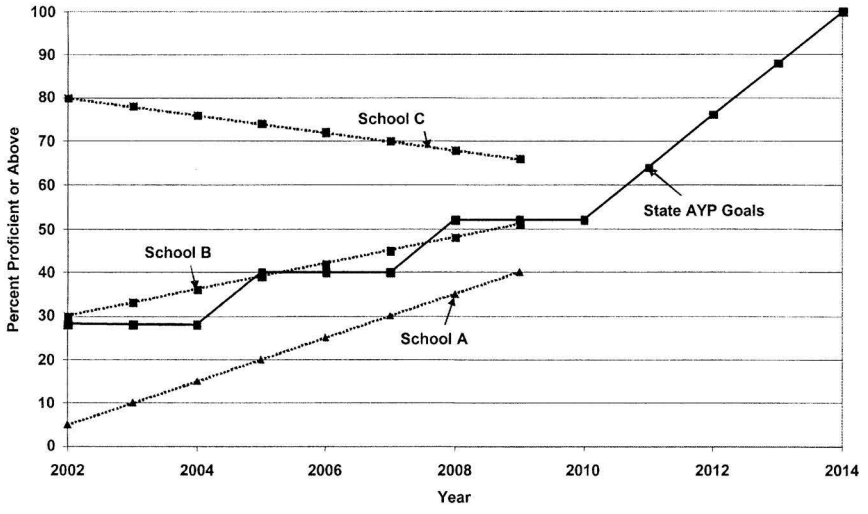


FIG. 2.4. Trends for three schools in comparison to state percentage proficient or above AYP goals.