New Educational Foundations

A Trans-ideological Journal of Criticism, Research, and Review

The Gauntlet: Think Tanks and Federally Funded Centers Misrepresent and Suppress Other Education Research

Richard P. Phelps

The Advanced Placement Program's Impact on Academic Achievement

Russell T. Warne and Braydon Anderson

Schoolhouse Solutions 1.4: What It's All About
Wade A. Carpenter

Leadership: The Philosopher's Stone of the Twenty-first Century
Edward G. Rozycki

BOOK REVIEWS

Raymond E. Callahan's *Education and the Cult of Efficiency: A Study of the Social Forces That Have Shaped the Administration of the Public Schools*

Anya Kamenetz's *The Test: Why Our Schools Are Obsessed with Standardized Testing—but You Don't Have to Be*

SUMMER 2015

New Educational Foundations

A Trans-ideological Journal of Criticism, Research, and Review

New Educational Foundations, a refereed online journal of ideas and dialogue, has been established as a forum for independent-minded educators, adminstrators, and researchers.

We favor no particular ideological bent. We unanimously reject the cultures of complaint and compliance; our audience will be professionals who question conventional thinking and the roar of the crowd.

We encourage you to download and distribute any of the materials in this introductory issue. Potential contributors are encouraged to send a monograph or letter of interest to us as info@newfoundations.com.

From the Editors of NEF, *New Educational Foundations*

New Educational Foundations is off to a great start. Our first three issues have been downloaded nearly 70,000 times altogether.

In an age of prohibitively expensive journals, NEF is available free of charge. We also do not charge authors, as many journals do. We minimize our expenses by paying only for services we cannot donate ourselves.

Your financial support will help us to publish NEF more frequently. If you value NEF's approach to vital issues in education, please consider making a donation of whatever you can afford to protect and sustain this unique platform. Please use our secure link to PayPal:

Thank you,

The editors and staff of *New Educational Foundations*

Contents

COMMENTARY

Schoolhouse Solutions 1.4: What It's All About 2
by Wade A. Carpenter

Leadership: The Philosopher's Stone of the Twenty-first Century 8
by Edward G. Rozycki

RESEARCH AND ANALYSIS

The Gauntlet: Think Tanks and Federally Funded Centers Misrepresent and Suppress Other Education Research 15
by Richard P. Phelps

The Advanced Placement Program's Impact on Academic Achievement 32
by Russell T. Warne and Braydon Anderson

DEPARTMENTS

BOOK REVIEWS

Education and the Cult of Efficiency: A Study of the Social Forces That Have Shaped the Administration of the Public Schools 55
reviewed by Gary K. Clabaugh

The Test: Why Our Schools Are Obsessed with Standardized Testing—but You Don't Have to Be 63
reviewed by Richard P. Phelps

Wade A. Carpenter was a high school teacher in the Charlotte-Mecklenburg (N.C.) Schools for fourteen years, and for the past twenty-seven years he has taught at the college level. He is Associate Professor of Education (Secondary and Foundations) at Berry College in Rome, Georgia. Email Wade at wcarpenter@berry.edu.

Schoolhouse Solutions 1.4: What It's All About

by Wade A. Carpenter

On October 3, 2014, the Berry College Charter Fellows (an alumni organization) presented Dr. Carpenter with its annual award "For Outstanding Service to the Profession of Teaching." He delivered the following remarks upon accepting the award.

When I first started teaching back in the early '70s, it was all about connecting with the kids. Every day it was "How do I reach this kid? How do I get through to that kid?" Now we're more concerned about boundaries. Heaven knows, I understand why—some people made the wrong kind of connections, and every sordid headline hurts our cause. But in our efforts to regain public confidence, we are in danger of losing something. When Socrates first started teaching, it was all about love—in his case, love of Truth. If you've read Plato's *Apology,* you'll recall that nothing offended Socrates more than the allegation that he was a professional teacher. He really went off on Aristophanes for that, insisting that he was an amateur—one who does what he does out of love. Now we rightly insist on professional dispositions, professional behaviors, professional ethics, et cetera, and heaven knows I understand why. But we are, once again, in danger of losing something. When John Dewey first started teaching, it was all about accumulating data, scientifically analyzing it, and applying the conclusions that followed. Just like now. But if we evaluate as objectively as edTPA insists, we may lose something.[1] If we are to be *educators*, teachers need both connections and boundaries, we need to be both professionals and amateurs, and we need to master both science and humanity.

An educational foundations professor is supposed to give an overview of our field, the positive and the negative, to enable young people to make informed career decisions and to help them develop the "critical, normative, and interpretive" perspectives that can help them survive and succeed—to at least have a working understanding of what it's all about. Hence, some find the "gatekeeper course" rough, as any boot camp is supposed to be. With that in mind, I usually end the course with some statement along the lines of "I'm not optimistic about the future of American education, but I am profoundly hopeful—and I'm looking at some of the reasons why I am hopeful."

But for the first time in many years, I don't think I can say that. For the first time in many years, I am *optimistic* about the future of American education as well. Guardedly optimistic perhaps, but optimistic nonetheless. I see some things happening nationwide in professional ethics, teaching, and evaluation that might very well bring about dramatic increases in learning. I'm *guardedly* optimistic because I'm also a bit concerned about how those changes will be implemented.

• • •

Across the country, we now have codes of ethics for teachers. I hope we'll apply them thoughtfully and carefully. *Care*-fully. An example:

Many years ago as a high school teacher, I had a young lady in my required senior-level world history class who just wasn't the sharpest tack on the bulletin board. Nice kid from a loving family, but just not too bright. The effects of poverty, growing up in the projects, and some pretty lame teaching early on had made academics difficult for her. But somehow, hope hadn't been extinguished, and she was willing to work, and willing to ask for help. So I helped her. A lot. We didn't have much in the way of resources to work with in those days, but slowly and painfully, her achievement improved. By the end of the year she was passing—not by much, but she was passing. Quite an accomplishment for both of us.

So her relatives had all come to town to watch their first family member graduate from high school, some coming from hundreds of miles away. But then she crashed the final. Flamed out, a total breakdown under pressure. On top of her long-term issues, she was now working on her third pregnancy, and her father was working on his third suicide attempt—the man was visibly shaky. Her chance for a future was even shakier. So when a terrified little girl came in to see me as I was entering grades, I faced a moral dilemma. I had a reputation for "69.5 passes, 69.4 doesn't" (which, by the way, is a very good reputation to have). But this situation had multiple highly mitigating circumstances. Could I flunk that

poor kid just because she couldn't remember who Charlemagne was from first semester? The risks were too high, and the cost could be too great. So I pointed to my desk and said in my best Mayberry, North Carolina, accent and syntax, "Honey, you see the top of my desk? Neither do I. I'm afraid I done lost your exam paper." It took her a moment to figure out what I was saying, but then she fell all over me, thanking me, crying her eyes out, and just about ruining a perfectly good shirt with her mascara.

Nowadays, under the Georgia Code of Ethics, what I did would be considered unethical. I falsified a test result. It's a good rule, as a number of teachers in Atlanta, Philadelphia, and D.C. are learning to their shame. Things are getting better. But I am a little concerned about the dangers posed by an unbending legalism and an unforgiving moralism. (Moralism is, of course, morality's evil twin.) Maybe nowadays it's unethical, but given the same resources I had then, I'd do the same thing today. Maybe it is now unprofessional, but what I did then was educational. In fact, it was Education, a lesson in mercy and kindness for both of us that will never harm a soul on this earth. You have far more resources and supports available, so if you want to avoid that sort of moral dilemma, develop your street smarts as a resource broker as well as your dispositions as a teacher.

• • •

That brings us to changes taking place in teaching. So let's begin with two foundations-type questions: What is teaching? And what works in teaching? Both are really very easy questions. Yes, teaching is the facilitation of learning, helping kids discover and explore Truth and Beauty, and those occasions in which we can do that are some of the coolest episodes of our career. But make no mistake, it's also *teaching*. If you are living the kind of life good teachers live, and with a Berry education to boot, you will have a great deal of great value to teach them. Don't let reformers silence the teacher's voice—and there are some enthusiasts of one stripe or another who would do just that: high-tech profiteers, self-serving politicians, and corporate curriculum scripters among them. Which brings us to the "what works?" question. Again, that's easy. Darned near anything works—somewhere, sometimes, with some learners. Heck, stark terror works very well in the U.S. Marine Corps. In preparing for war, being more afraid of your sergeant than of the enemy can enhance your survival chances considerably. We, on the other hand, have the great good fortune to be preparing people for something better. Berry and Berry graduates will continue to do that well as long as we remember Martha's "Head, *Heart*, and Hands."[2] And I'm delighted to see

that American education seems finally to be getting past the old "process *versus* product" false dichotomy. The traditional Aristotelian accumulation of content knowledge and the progressive Deweyan emphasis on thinking skills are *both* important, two sides of the same coin. It's kinda hard to think well if you don't have anything to think about, isn't it? Teaching's getting better.

My two favorite courses to teach: in high school, Current Issues; in college, Perennial Questions in Education. Current Issues was a lot of fun—with no resemblance at all to the stereotypical "bring in a newspaper clipping on Friday." It was an advanced scholarship-preparation course in which we studied a huge range of topics, from war in the Middle East to AIDS in Africa to poverty in America to how the stock market really works to how a bill really becomes a law. In other words, what it's all about. As you can imagine, with no set standards, the curriculum changing every year—heck, every *day*—no textbook even possible, and most everything highly controversial, it involved an enormous amount of work year-round for me, and an enormous amount of risk for my principal, a man who was keenly aware of Murphy's Law. I'm grateful for the man's trust. We never had a single parental complaint from that course. And one year every one of the thirty-six kids in that class got an academic scholarship. One of them is now president of a research foundation, another is online editor for a national political magazine, a third is the psychologist who did the court-ordered competency testing in the D.C. sniper case, and a fourth is now a university professor in South Carolina specializing in autism. A fifth kid, the captain of my High Q team and the senior voted most likely to succeed, served five years for embezzling $600,000 from a presidential campaign that had hired him. (Well, nobody wins 'em all!)

My favorite course here at Berry was a Perennial Questions in Education class we had in the honors program for a few years, in which first-semester freshmen started off with Plato's *Republic* (the whole thing), and then went deep. Heavy-duty readings with heavy-duty seminars, good old-fashioned perennialist style, and Truth was discovered, explored, deconstructed, and re-created. The kids loved it and I loved it. One of the students went on to be Berry's valedictorian, several of them decided to become teachers, and one is now a university professor in Texas specializing, curiously enough, in autism.

My concern: with the improved but very top-down, and perhaps even scripted, curricula of the future, and the high-activity teaching methods that are *rightly* favored by education reformers, would those two courses even be possible?

• • •

And that brings us to evaluation. Accountability is what it's all about nowadays, and in general, I think that's a good thing. I'm a taxpayer, a parent, and a citizen. I don't want to see my money going to waste, anybody's children getting a botched childhood, and uninformed idiots voting. Evaluation is a good thing. As a wise old educator who taught here at Berry, Jesse Laseter, was fond of saying, "You get what you expect, and you get what you *in*spect." But our growing obsession with assessments and data points and standards and scores could have some troubling outcomes as well. In fact, it appears to me that the current iteration of the accountability movement is based on a level of distrust and even disrespect for teachers and students that may coerce better teaching and learning but is incompatible with anything I'd care to call education. Paint by Number may make art more accessible to more people, but it won't hold on to artists.

An example of the sort of examination madness I'm worried about happened to my son, Daniel, when he was in high school. He had been dating a cute-as-a-button girl. His first love—you remember, don't you? Then one day she was killed in an automobile accident. Her family invited Daniel to sit with them at the funeral. Unfortunately, the funeral coincided with his history final exam, and school-system policy was that immediate family members would be excused and allowed a retake, but otherwise, it was the teacher's decision. Daniel's teacher said no. The policy itself was a good one—it's not difficult to imagine the consequences if they had had a softer policy. But that teacher . . . well, let's be nice and just say she wasn't a Berry graduate. So the son of an old history teacher walked away from the disrespect with which he had been treated, went to the funeral anyway, and failed history. And I'm proud of him. In the bad old days, at least where I taught, that teacher's response would somehow have been accidentally shredded somewhere between the secretary's desk and the principal's office. But not now. Oh, and by a mortifying coincidence, the assistant principal who had to back up the teacher had been one of my all-time best student teachers some years previously. And he was right to back her up. You just can't get by with playing that loose with policy and documentation nowadays, and I reckon that too is an improvement. But to the extent that schools lose their humanity, they lose their value. And it's a pretty safe bet they'd also lose the public confidence that all this accountability was designed to regain.

So in conclusion, four unconventional (even downright odd) lessons I've picked up over the years that may have great value for you and your kids:

One, ethics: Don't let justice kill kindness.

Two, teaching: Don't let good teaching get in the way of great education.

Three, evaluation: We get what we expect, we get what we inspect, but let's not forget to respect.

The fourth and final lesson is the one that enables me to make sense of it all: redemption. Obviously, in the sense of spiritual redemption, I'll leave that to God. He's better at it than I am, and public schools have a notoriously difficult time handling spiritual questions satisfactorily. But redemption also means redeeming kids from social evils like poverty and violence and despair and failure. It means redeeming all of us from three intellectual vices that threaten our nation every day:

- One, ignorance (when you don't know much).
- Two, stupidity (when you only know what somebody else has told you).
- Three, silliness (when you only know what you want to know).

This life of redemption is, if you choose to live it, the greatest and most joyous part of our calling. It's what provides the hope that makes optimism possible, and it's in living it that we meet and join with God, even in public schools. And I am grateful to all of them and to all of you who have helped me live it. The prayer for education written for my church (and Martha's) is what it's all about:

> *Almighty God, fountain of all wisdom, enlighten by thy Holy Spirit those who teach and those who learn, that, rejoicing in the knowledge of thy Truth, they may worship thee and serve thee from generation to generation, in thy Name and for our sakes we pray. Amen.*

Notes

1. The prevalent "one best system" for preparing teachers has been trademarked "edTPA" by the American Association of Colleges for Teacher Education (AACTE).
2. Martha Berry (1866–1942) was the founder of Berry College.

Edward G. Rozycki, Ed.D, served seventeen years as an associate professor of education at Widener University, Widener, Pennsylvania. He is the webmaster and co-sponsor of the article banks at www.newfoundations.com.

Leadership: The Philosopher's Stone of the Twenty-first Century

by Edward G. Rozycki

> [T]he effectiveness of . . . symbolic action is enhanced by the confusion of all involved between substantive and symbolic results.
>
> —Jeffrey Pfeffer, "Management as Symbolic Action"

> The appearance of moral authority and even a sacred aura at the top of the hierarchy is essential to sustain the privileges of leadership.
>
> —Jeffrey S. Nielsen, *The Myth of Leadership*

What Are We After?

For thousands of years, would-be power holders have searched for some magic that would enhance their lives. Reputedly, the philosopher's stone (Arabic, *al-iksir*; Buddhist or Hindi, *Cintamani*) enabled its possessor to transmute metals of lower value into gold. Other rumored powers included changing common crystals into precious stones; healing illnesses; lengthening life; and creating *homunculi*.

One can find thousands or even millions of characterizations, in all types of media, of the terms *leader* and *leadership*. Hardly less common is the hope of finding or creating "true leaders" who can transmute any group, even a corporation or a society, into something wondrous. Any logical and practical prior determination of the targets, i.e., of what most would find wondrous, is generally passed over by those impatient to effect magical transformations.

Looking for the "Essentials" of Leadership

> *I have plenty of clever generals, but just give me a lucky one.*
>
> —Napoleon (anecdotal)

> *Luck is not something you can mention in the presence of self-made men.*
>
> —E. B. White, *One Man's Meat*

What is interesting about the volumes written on leadership is that many of them focus largely on the personal characteristics of what they term "leadership," ignoring not only the influence of luck but also the constraints of organizational structure, tasks, and goals. Thus we find books and blogs and newspaper articles on leadership—as abundant today as horoscope columns for the lovelorn have ever been—focusing on the behavior of the reader as a monologic actor in a group. Generally overlooked is the often-stultifying influence of the context of action.

That context has long been studied by industrial analysts, but it is generally ignored by public school reformers, who are no doubt well aware that their markets contain many individuals who seek to acquire "generalship" and esteem as "self-made" persons. The irony here is particularly striking, since many current public school critics envision reformed schools that will provide the economy with more-productive graduates "in the twenty-first century." However, Joan Woodward's *Industrial Organization: Theory and Practice* (a sociological study unfamiliar, in my experience, to educational administrators) provides an analysis that relates personnel relationships to organizational type, inputs, and outputs.

Industrial organizations, writes Woodward, fall into three classes depending upon their goals, the kinds of products they make, and the kinds of technology they use to produce them. For those of us who contemplate school change, Woodward's crucial finding is that

the goals and technologies of the most successful organizations she analyzes profoundly affect both the productive and the social relationships between workers. In other words, what we try to do and how we go about doing it affect the way we work together, our productivity, and our politics.

Despite a not-quite-comfortable fit, large-batch and mass-production industries have provided Americans with a factory image of the modern school. According to Woodward, the goal of such industries is to produce uniform items for a pre-existent mass market. Their technologies, although complex, can be made piecemeal. Causal connections are generally clear. Uniform inputs produce uniform outputs, a process that diminishes the need for research and development. Management separates itself from low-skilled workers even as it controls them through a variety of highly elaborate sanctions. Communication occurs only to exchange information of interest with management. The technical rationality of the workplace tends to fragment social relationships that might undermine efficiency.

American schooling reflects the industrial model in its attempts to standardize curriculum, testing, and promotion standards. Nonetheless, the model is ignored by special educators and by those who try to meet the individual needs of every student. It is also undermined by opening the public schools to all comers, using age as the only prerequisite for acceptance rather than standardizing admissions criteria in any productively, e.g., pedagogically, relevant way.

Process industries such as oil refineries, chemical plants, and pharmaceutical companies are technology rich. They produce specialized products for hard-to-identify specialty markets. Complex though well-defined causal processes are built into their plant equipment to minimize the need for workers. The few workers needed tend to be highly skilled technicians who can maintain and troubleshoot production. Management control is of little concern since both the equipment and the technical orientation of the workers help ensure success. As in the mass-production industries, communication is necessary only for exchange of information. The technical rationality of the process does not support—though it need not undermine—the social relationships of organization members.

That model is beloved of all technically adept educators, whose main failing is often little more than the assumption that their successes rest solely on their own pedagogical skills rather than on classroom or student characteristics beyond their control. Even so, the model is actively promoted by teacher-accrediting organizations,

eager to convince would-be teachers that all children can learn and that even future adult behavior can be determined through early school interventions. Unfortunately, little consensus exists that those propositions have scientific support.

Unit and small-batch industries, in Woodward's typology, produce custom-designed specialty items, such as locomotive engines and custom cars. Specialty demands provide the impetus for researching and developing processes and methods that take the very specific characteristics of inputted material and, with skilled worker attention, transform them into relatively unique outputs. Management-worker relationships tend to be nonhierarchical, and communication occurs operationally as the process dictates. Since teaming and mutual support are often necessary, social relationships are as important as technical ones.

That is what private (and otherwise small-school) education is about, although limited budgets may hinder the recruitment of highly skilled teachers or administrators. But all parents (especially of smaller children) like to believe that their offspring will receive special treatment. For each public school student admitted to special education, that wish is addressed by an IEP (individualized educational placement).

Chart 1 below summarizes the relationships between inputs and outputs for the different organizational types.

SYSTEM Dimension		Unit & Small Batch	Mass & Large Batch	Process
Kind of Production Control		Very difficult Reliance on skilled practitioner	Very elaborate Highly developed system of sanctions	Built into process Of little concern
Dominant Personnel		Engineers	Production	Marketing
Manufacturing Cycle		Marketing	Development	Development
	MOST CRITICAL	Development	Production	Marketing
		Production	Marketing	Production
Relationship between Task Functions		Day-to-day operational relationship	Normally, information exchange only	Normally, information exchange only
Relationship of *Technical* to *Social* Functions		***T,S*** equally important; teaming necessary	***T*** conflicts with ***S***; fragments social relations	***S*** less important since planning controls ***T***

Chart 1. Characteristics of Production Systems (adapted from Joan Woodward, *Industrial Organization Theory and Practice* (London: Oxford University Press, 1966)

American schooling ideologies that try to characterize schools as production systems render those models problematic. Both progressive and "scientific" ideologies tend to view the relationship of technical to social functions of the school as one of unit and small-batch systems, but the production control of most schools, given their size, defaults to that of large batch and mass systems. *(For more on this point, see "Productivity, Politics and Hypocrisy in American Public Education": http://www.newfoundations.com/EGR/ProductivityWEB.html.)*

Two Classes of Leadership: Role-Leaders versus Performance-Leaders

> *Management is doing things right; leadership is doing the right things.*
>
> —Peter Drucker

Some leaders are so designated because of positions attained in organizations. Job titles and roles are ranked on organization charts, the higher considered superior to the lower. Such positions are bestowed for various reasons, e.g., tradition, experience, nepotism, political connection, friendship, seniority, and (hallelujah!) possession of needed skills.

For organizations that rank these role positions, "leadership"—so the expectation (the hope?) goes—is a function of rank. A given job holder possesses a leadership position over everyone in line of command below him or her on the chart. A job title, however, may be a poor indicator of competence: some high-ranking people may lack the experience, abilities, or the attitude to be good leaders. They are called "leaders" nonetheless, gaining unearned much of the approval, compensation, or deference the term implies.

By contrast, respondents to informal surveys I have conducted over the years discuss the desired characteristics of leaders or leadership primarily in terms of performance leadership. They more or less agree with Peter Drucker about what it is to be a "manager."

Why, then, is role-based leadership tolerated, even celebrated? Particularly in many long-established organizations, it is because top-ranking leaders are the most powerful, controlling organizational resources, particularly through ownership. The top-rankers needn't possess the productive skills necessary to hire and evaluate lower-ranking members. Princes need tend no gardens.

In *The Myth of Leadership: Creating Leaderless Organizations,* Jeffrey S. Nielsen treats all leadership as role-leadership oriented. (The term he uses is "rank-leadership.") By "leaderless" he means "without rank-defined leaders." What he does propose is that organizations gravitate relentlessly toward the performance-leadership model.

Two Domains of Any Kind of Leadership: Social versus Technical

Strangely enough, workers in organizations where rank is important, funding is sufficient, and competitive pressures are minimal may still yearn for the kind of performance leadership that powerful top-rankers would likely not tolerate. However, performance leadership can be tolerated, even in organizations where role-leadership is dominant, by restricting its range of action.

We can distinguish persons exercising either role- or performance-leadership as occupying one of two different domains of functioning: the social or the technical. To protect the organization or its subparts against internal or external threats, social functioning involves behavior aimed at maintaining various forms of consensus among organizational members.

Typical social-functioning examples can involve such activities as participating in advertising, representing the firm at public functions, and commenting on public events. The language of social functioning tends to be cordial, collegial, celebratory, noncommittal, and vague. Replete with slogans, truisms, compliments, and happy (or angry) ambiguities, it requires little technical training beyond that of an undergraduate liberal arts major.

Technical functioning ordinarily requires planning and skillful strategizing to maximize the efficiency of social functioning, but—an important note—it presumes a context of consensus already established through social functioning. Technical leadership, whether role-based or performance-based, is evaluated by its success at achieving restricted aims. The language of technical leadership is about cause-and-effect, cost-and-benefit, and efficiency.

A leader may appear cordial and celebratory, but that behavior, as a stratagem to secure a desired aim, constitutes technical leadership: honey set out to attract more flies than would vinegar.

Leadership for a Common Humanity: A Philosopher's Stone?

> *A genuine leader is not a searcher for consensus but a molder of consensus.*
>
> —Martin Luther King, Jr.

Martin Luther King, Jr.'s characterization of a "genuine leader" is intriguing. Not needing to search for a community of consensus assumes either that a) you are already fortuitously embedded in such a community or that b) you have an irresistible technique for consensus building. So powerful is this community-building technique, evidently, that those who earlier did not agree with you will suddenly drop their resistance and acknowledge that your consensus-molding efforts have successfully recruited them.

What underlying beliefs support King's characterization of a "genuine" leader? I can think of two that King, an obvious performance-leader himself, expressed on many occasions:

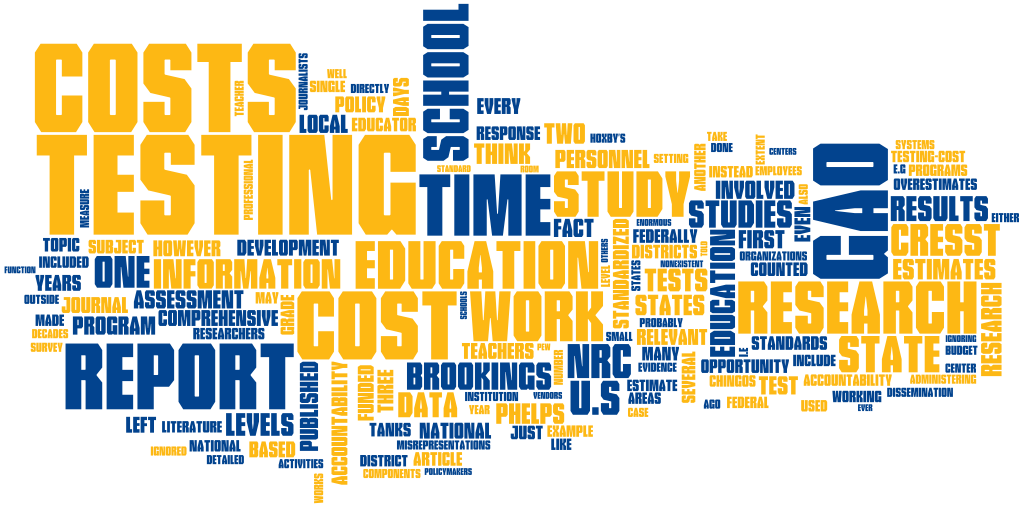
1. We are all children of the same God—i.e., we belong to a universal community—and
2. Non-violent confrontation is the method that will awaken that sense of community and moral consensus in those who right now don't feel it.

King's first belief acts upon the social domain, enabling a logical foundation for a consensus. His second belief supports his choice of a technical approach, non-violence, to the goal of racial equality. Whether we share Dr. King's beliefs or not, we can agree that they provide a coherent basis for the actions he took to achieve the goals he wanted. That logic stands, even though the evils he faced—much like those we ourselves face in all arenas of our lives—have proved far more resistant to his efforts than we might wish.

(To examine these issues in further detail, see "Controlling the School: Institutionalization," http://www.newfoundations.com/OrgTheory/Institutionalization.html.)

Notes

Display epigraphs. Jeffrey Pfeffer, "Management as Symbolic Action," *Research in Organizational Behavior* 35 (1981); Jeffrey S. Nielsen, *The Myth of Leadership: Creating Leaderless Organizations* (Palo Alto, Calif.: Davies-Black, 2004).



The Gauntlet: Think Tanks and Federally Funded Centers Misrepresent and Suppress Other Education Research

by Richard P. Phelps

I BECAME INVOLVED IN the education-testing debate purely by chance. I did not begin as an advocate for standardized testing. And, truth be told, I am still not motivated primarily by a fondness for standardized testing, despite the fact that I have come to deeply appreciate its benefits and strengths. I am strongly motivated, however, to battle deliberate misrepresentation, censorship, and information suppression.

THE GENERAL ACCOUNTING OFFICE

More than two decades ago, while working at the U.S. General Accounting Office (GAO, now called the Government Accountability

Office), I completed a study that measured the extent and cost of standardized testing in the United States (U.S. GAO). The first President Bush, George H. W., had proposed a national assessment system that would test U.S. students in five core subject areas at three grade levels. You probably have not heard of the proposal because it died a natural death after President Bush lost his re-election bid in 1992. Part of my job at the GAO was to estimate the proposed new testing system's overlap with current testing—the time and cost it would add. In the process, I would also build a highly detailed database of state and local district assessment practices based on the GAO data collection.

We did an exceptionally thorough job. We developed surveys carefully, reviewed and pretested them, and through enormous persistence, achieved very high response rates. We collected budgets from most states and many school districts to use in benchmarking the survey results. A "Who's Who" of notables in the evaluation, statistical, and psychometric worlds (e.g., William Kruskal, Lee Sechrest, Mark Lipsey) reviewed various aspects of the study. Nothing like it in quality or scale had ever been done before, or has been done since.

The many peer reviews from both inside and outside the GAO were rigorous, as one would expect for an investigation into a key aspect of a major presidential proposal. On all GAO quality measures (e.g., survey response rates, fact-checking) the study exceeded GAO norms.

The study results, however, were surprising, at least to me. I had been led to believe by the most accessible education-policy literature that education testing was exceptionally costly and time-consuming. The evidence showed that it wasn't, even when one accounted for all the opportunity costs in personnel time at all levels—national, state, school district, school, and classroom. In 1990–1991, systemwide (i.e., external) testing and test-related activity made up on average about seven hours per year of a student's time and about fifteen dollars in purchase costs and staff time.

The results surprised others as well. One outside review provided my first taste of a type of reaction, one more emotional than substantive, that would later become very familiar. My results could not possibly be correct, went the argument: I must have left something out. Tests cost more and take up more educator time than I had found, this reviewer was certain: additional calculations were needed, which I made, but my critic judged them unsatisfactory as well.

For those unfamiliar with such research, judgments of its quality and the trustworthiness of the results are typically benchmarked by two aspects: the size and representativeness of the sample of relevant units—public education administrative units in this case—and the scope of the measures (i.e., accounting for all relevant components of cost and time). I made every effort to ensure that not a single relevant cost or time component was neglected and conversely that no extraneous cost or time components were included.

Since then, as far as I can tell, no study of the extent or cost of testing in the United States has come anywhere close to matching the scale and coverage of the GAO study. Forty-eight states that used testing programs in 1990–1991 as well as more than six hundred school districts—a robust, nationally representative sample—had delivered complete survey responses.

Most studies undertaken since then have reported partial information: for the state level only, from a few to several school districts only, or for the purchase costs of tests and test-contractor services only (not the opportunity costs of education personnel time).[1]

The GAO, however, has a single client—the U.S. Congress. Once a report has been presented to Congress, no further effort at dissemination is made.

TREATMENT OF THE GAO REPORT

Case One: The Center for Research on Educational Standards and Student Testing (CRESST)

I left the GAO before the report was actually released in January 1993; pressure to suppress the report and its findings—essentially, that standardized testing is not excessively burdensome or expensive—apparently began even before its release.[2] Over the ensuing months, I learned of additional efforts to suppress or misrepresent the report's findings. Conference panels, to which I was not invited to participate, criticized the report. Reports written by the federally funded Center for Research on Evaluation, Standards and Student Testing (CRESST) and elsewhere lambasted the report and suggested that better studies were needed.[3] The critics claimed that the GAO report omitted information that in fact was not, and that it included information that in fact was not. But reasonable people who heard CRESST et al.'s version of the story believed it, so the GAO report, along with probably the most thorough and detailed database on

state and local testing practices ever developed, began fading into obscurity.

In place of the GAO study, other conference presentations and journal articles in mainstream education journals purported to show that standardized tests cost an enormous amount and overwhelm school schedules with their volume. Other 1990s-era studies were based on tiny samples: a single field trial in a few schools, a few telephone calls, one state, or in some cases, facts that were just invented. The cost studies among them that actually used some data for evidence tended to heap all sorts of non-test activities into the basket and label them costs of tests.

The two testing-cost studies that CRESST promoted in three successive annual conferences were based on a tiny sample (from a New Standards Project field trial) and a single state (Kentucky; Picus and Tralli). In the latter, survey responses were apparently accepted *as is* without review: for example, they included a response claiming that salaries of school personnel for the entire school year should be considered test preparation and added to the cost of tests. Both studies were widely praised and disseminated. The first of the two studies was summarized and published as the lead article in a 1995 issue of the *Journal of Education Finance* (Monk, 1995), along with misrepresentations of the GAO report.

Giving such work the benefit of the doubt, those authors may have merely misread the GAO report's specifications of the opportunity costs of personnel time. The opportunity costs of testing, however, are noted starting on page 1 and on most pages thereafter. They are noted in the introduction; the conclusion; and every chapter in between. They are included in many of the figures and tables.

I wrote dozens of letters and made dozens of telephone calls to the researchers of the testing-cost studies mentioned above; to those responsible at the organizations promoting their work; and to the U.S. Education Department (US ED), which funded (and continues to fund) CRESST. At one researcher's request I furnished him with technical documents and instruments from the GAO project work. In most cases, I was simply ignored. In a few cases, I received assurances, first, that the matter would be looked into—it was not—and second, that an *erratum* would be published in the CRESST newsletter; it never was. Articles I submitted based on the GAO study were rejected by mainstream education journals for outlandish and

picayune reasons, or because "everyone knows" that the GAO report was flawed.

The response from the relevant U.S. Education Department program officer was particularly revealing. CRESST has operated for three decades under repeatedly renewed federal grants. Consequently, no other federally funded research center has focused on testing policy. Those many millions of federal dollars have granted CRESST directors and affiliated scholars enormous power to decide which and whose research becomes known and which and whose does not. I complained to the U.S. ED grant program officer that CRESST had misrepresented the GAO report at three successive annual conferences, denied my request to attend, and ignored my requests to add errata in their publications. CRESST, I was told, was responsible for any "editorial" matters.

The trend continued even when I was finally allowed to present the results of the GAO study at an education-research conference (Phelps, 1998). During the question-and-answer session following my presentation, one individual standing at the back of the room suggested that the study's failure to address opportunity costs deprived it of any value. I asked my questioner to identify which costs were left out, but he did not respond and soon left the room. The damage had been done—the misrepresentation of the GAO study had once again been reinforced.

Finally, I decided to send the *Journal of Education Finance* a commentary rebutting such misrepresentations as a response to a lead article the journal had published in 1995, but my initial approaches were rebuffed. I then contacted the chief editor of the journal directly. She approved the manuscript for publication and provided space for her board member to respond (Monk, 2006; Phelps, 2006). In my space in the school finance journal, I criticized the disparagement of the GAO report as censorial and its misrepresentations as tending to discredit it. The response? My criticism of the disparagement was itself censorial.

The critics continued their assault after publication of the commentary-response. Two years later my other critic from CRESST published another report, with the misrepresentations intact (Picus and Tralli). I managed to get one offending paragraph excised, but several others remained. Ultimately, I wrote an article summarizing the methods and results of the GAO report, which won two national prizes.[4] Later, in 1999, I updated the GAO study results with data

from 1998–1999 and inflation-adjusted cost figures, detailed the combined results in an article with up-to-date estimates of the extent and cost of testing in the United States, and submitted it to the same journal whose article a few years earlier had precipitated the rebuttal-response episode recounted above. That journal published it in its back pages (Phelps, 2000).

Case Two: The National Bureau of Economic Research

My journal article was published just prior to the 2000 U.S. presidential election campaign, the first in which standardized testing was a key issue. After the new administration took office, President George W. Bush proposed a national testing program in the accountability provisions of the No Child Left Behind (NCLB) Act. The program was modeled on one he had promoted in Texas.

As a result, the current extent and cost of testing, and any possible increase due to the president's proposal, again became national issues. Studies were conducted on some aspects of the topic, for example by Ted Rebarber of Accountability Works and the Pew Center's Stateline.org. (See Accountability Works, 2004, and Danitz.)

The most widely publicized testing-cost report from the early 2000s, however, came from Carolyn Hoxby (2002), a faculty member at Harvard, then Stanford, universities and the long-time director of the education program at the National Bureau of Economic Research (NBER). Her work is the best-known on the topic because of her affiliation with organizations, such as the NBER, Harvard's Program on Education Policy and Governance, and the Brookings and Hoover Institutions, that invest a great deal of money in publicity and dissemination.

I first became interested in Hoxby's work after noticing that several reports published by NBER on education topics claimed to be the first ever to study a topic or declared that no prior research on a topic existed (Phelps, 2012a). Normally, that might not seem interesting, but in each case many previous studies had been conducted.

Hoxby's own study of testing costs doesn't refer to earlier work at all. Her work is hardly noteworthy, either. She examined budgetary expenditures for testing programs from fewer than half the U.S. states. Even had she obtained them from all states, such data are problematic: some costs induced by testing end up in other categories in accounting spreadsheets, and vice versa. Moreover, Hoxby's study

took no cognizance of local school and school district costs, which can dwarf state costs.

Case Three: The National Research Council

CRESST re-entered the testing-cost debate with a report from the Board on Testing and Assessment (BOTA) at the National Research Council (NRC), a group that CRESST captured in the late 1980s and has held as its own since (Phelps 2008/2009, 2012b). The 2008 BOTA-NRC report, *Common Standards for K–12 Education?*, asserts, again, that the GAO report left something out and so underestimated the cost of testing (Beatty).[5] And again, the assertion is false. This time, the NRC accused the GAO study of neglecting to consider the cost of "standard setting" during test development; in fact, this cost was fully accrued in the GAO calculations.[6]

Claiming a void in others' calculations can be used as an excuse to bulk up testing critics' own cost estimates massively. Here are just a few ways that the NRC report, *Common Standards for K–12 Education?*, overestimates the cost of testing:

- One-time-only start-up costs—e.g., standard (passing-score) setting—are counted as annual recurring costs.
- Educator travel and lodging expenses for serving on standard-setting and other test-development panels are counted twice, both as direct educator expenses and in the budget of the state education agency (which, in fact, reimburses the educators for these expenses).
- The full duration of all testing activities at a school—said to be 3–5 days—is allotted to each and every educator participating. For example, take the time of a fifth-grade teacher who administers a one-hour math exam on Tuesday of testing week and who otherwise teaches regular class that week. That time is counted as if s/he were involved in administering each and every exam in every subject area and at every grade level throughout the entire 3–5 days. Moreover, the time of each teacher in the school is counted as if the teacher is present in each testing room for all subject areas and grade levels. By this method, the NRC overestimates the educator time spent directly administering tests about twentyfold.
- Another way of looking at the problem is to ignore the fact that a school administers a series of one-hour tests across the

tested subject areas and grade levels over the span of 3–5 days and instead assume that all classes in all subject areas and grade levels spend 3–5 days doing nothing but take day-long exams—which, in fact, is not what happens.

- The NRC calculates the number of teachers involved by using a federally estimated average pupil-teacher ratio, rather than an average-class-size estimate. Pupil-teacher ratios underestimate class sizes because they include the time of teachers when they are not teaching. By this method, the NRC overestimates the number of teachers involved in directly administering tests by another 50 percent.
- The NRC counts all teachers in a school, even though only those in certain grade levels and subject areas are involved in testing—usually amounting to fewer than half a school's teachers. By this method, the NRC overestimates the number of teachers involved in directly administering tests by another 50 percent or more.
- In calculating "data administration costs" of processing test data in school districts and states, the NRC classifies all who work in those offices as "management, business, and financial" professionals who earn $90,000 per year. Anyone who has worked in a state or local government data-processing department realizes that this classification grossly overestimates the real wages of the majority of employees, who essentially work as clerical and often contingent staff.
- The NRC is told by one school district that its average teacher spends twenty hours every year in professional development related to assessment and accountability. Despite how preposterous the number should sound, the NRC has used that one piece of hearsay to estimate the amount of time that teachers everywhere, whether involved in testing or not, annually spend in related professional development.
- Moreover, professional development related to testing and accountability is assumed to be unrelated to regular instruction, so it is counted as a completely separate, added-on (i.e., marginal) cost.
- The NRC counts educator time working on standard-setting and other test-development panels as "two or three days," which anyone who has worked in test development knows is a high estimate. One to two days is more realistic.

Finally, the NRC studied testing and accountability in only several school districts, in only three states. Nonetheless, according to the NRC, the GAO report—which as we have seen analyzed more detail from all forty-eight states with testing programs and more than six hundred school districts—is the study that left stuff out. In the end, the NRC's estimates for testing and accountability costs are, in the council's own words, "about six times higher" than previous estimates.

For several years afterward, each of the two most recognizable sides in U.S. education policy debates had its own testing costs research champion. Education reformers, think tankers, and Republican Party advocates had Carolyn Hoxby's numbers, which hugely underestimate the cost of testing programs. The education schools, educator professional associations, and Democratic Party advocates had the CRESST-NRC numbers, which greatly exaggerate the cost of testing programs. Anything in between was either ignored or misrepresented.[7]

Case Four: The Brookings Institution

These days, the education policy topic *du jour* is the Common Core Standards, and standardized testing is a key component of the planned program. Naturally, one would expect a think tank to weigh in on the matter of the possible costs, and the Brookings Institution has done so with the work of yet another Harvard University Ph.D. in economics or political science—in this case Matt Chingos, a political scientist.

Several months ago, Brookings began promoting its own report, which begins by clearing the field.

> Unfortunately, there is little comprehensive up-to-date information on the costs of assessment systems currently in place throughout the country. This report seeks to fill this void by providing the most current, comprehensive evidence on state-level cost of assessment systems, based on new data gathered from state contracts with testing vendors. (Chingos, p. 1)
>
> [Other] Estimates of these costs are based primarily on assumptions and guesswork. . . . The most comprehensive nationwide data were collected about a decade ago, in separate investigations by Caroline Hoxby and the Pew Center for the States. (p. 4)

The latter criticism—estimates "based primarily on assumptions and guesswork"—was directed at two other studies that Chingos presumably also considers not as "comprehensive" as his, cited in the accompanying footnote. A detailed reading of the Brookings report, however, reveals its own abundance of assumptions and guesswork.

Like Chingos's own work, the Hoxby and Pew Center studies he cites examined only the direct costs of testing at the state level, not the more consequential data at the local level or any data at all on personnel time (outside the easiest-to-locate line items in state budgets). Because Chingos's study did not examine those cost components—an absolutely necessary step for a complete cost estimate—perhaps he did not wish to draw attention to other studies that included them (e.g., Accountability Works, 2004; and Phelps, 2000).

As for those other cost components, Chingos pleads that they are too difficult to measure. Take for example the time spent by state employees in "selecting contractors and overseeing the vendors":

> But such costs are difficult to track consistently across states, and usually represent a small fraction of the testing budget. (p. 7)

That may fairly be termed disingenuous. State employees typically do far more than just "oversee" the vendors, and such costs are not "small," though they may be a small fraction of the *testing budget*. The costs are absorbed in other parts of the budget—in the regular salaries for staff positions that probably would not exist if there were no testing program. Collectively, they can represent a large portion of the cost of a testing program.

> The roles played by school and district officials who aid in test administration and scoring are important as well, but the cost of this work is challenging to measure. Calculating such costs requires information on which employees have these responsibilities, their compensation levels, how much time they devote to test-related activities, . . . (p. 7)

Yes, it is challenging to measure. Yes, it does require information on responsibilities, compensation levels, and time devoted to test-related activities. So did the Brookings Institution meet those challenges and gather that difficult-to-gather information? (Note: the GAO study did both.) No, the Brookings report claimed that it was too hard.

Brookings dismisses the BOTA-NRC cost estimates of 2008 as irrelevant because "these costs are data collected from only three states and reflect the costs of standards and accountability systems in addition to the assessment costs" (Chingos, p. 27, note 10). In fact, however, the BOTA-NRC estimates did not reflect the costs of standards and accountability systems in addition to the assessment costs. Those estimates had simply double counted the cost of "standard setting" (i.e., "passing score" setting) sessions. Like the National Research Council report, the Brookings report ignores how tests are actually developed.

Other excuses for not being comprehensive, even while repeatedly boasting about being *the most* comprehensive:

> Time spent preparing for end-of-year tests may also be considered a "cost," but it is one that is nearly impossible to measure given the difficulty of separating instructional time that is geared specifically towards preparation for the test as compared to for some other purpose. (p. 38, note 36)

> For these contracts, we either ignore the development costs (instead focusing on the contract costs during operational test years) or divide the development costs equally over the operational years. (p. 8)

The Brookings estimates of testing costs are suspect because they are far from comprehensive. They do not include, or even attempt to include, personnel costs at either the state or the local levels. Neither do they include any local costs. Ironically, for a work that repeatedly touts its comprehensiveness, the report's single greatest lack is comprehensiveness. (For an interesting contrast, see Accountability Works, 2012, or Nelson.)

After the truncated, skewed testing-cost estimates, all that is left of value in the Brookings report is the revelation about saving money on testing through state consortia, an idea that could have been lifted right out of the GAO report.

CRONY RESEARCH DISSEMINATION

The GAO project work was not just unfairly slighted by education's vested interests: it was repudiated. All that effort, all that expense—funded by U.S. taxpayers—was so thoroughly and effectively discredited by its opponents that barely a trace remains in the

collective working memory of education policymakers, or anywhere else outside my own cranium and computer hard drive.

To discredit my GAO report, education's vested interests falsely accused my work of ignoring the costs of personnel time. Ironically, the think tankers' own work has comprehensively ignored the opportunity costs of personnel time and has apparently felt no obligation to include it, yet still claim comprehensiveness.

It would seem that even substandard education research from the think tanks or federally funded centers is deemed praiseworthy, while the highest-quality work from those of the vast research working classes is flicked away like a stinkbug.

This latest report from the Brookings Institution continues a twenty-first-century tradition of information suppression, misinformation, and self-promotion in education policy research from our country's best-known and best-funded think tanks. But censorship isn't the only problem: the process fosters a nonmonetary form of corruption. The currency of scholars is attention, providing the "richest" among them a confluence of honors, awards, and remuneration streams.

Both the NRC and the think tank reports mentioned above may be used to proselytize and mislead. More emphatically, they are expropriated to showcase the careers of those involved: their authors declare the work of other researchers inferior or nonexistent, while at the same time they liberally cite their own work and that of like-minded colleagues and package the combination as if it were all that mattered. The stated mandates of these organizations are to objectively review all the research available; instead they promote their own work and declare most of the rest nonexistent. They are mandated to serve the public interest; instead they serve their own.

As a result, journalists assume that the easily accessible work of think tanks and federally funded centers represents the research literature as a whole and that the dissemination of education research is objective and fair. They couldn't be more wrong.

Some journalists step further into an ethical abyss—they help promote dismissive reviews. No journalist has the time to validate such claims; it can take years to learn a research literature. When journalists mention a "paucity of research on this topic" or the like, they are probably taking one quite self-interested person's word for it. When they write "[So-and-so's] study is the first of its kind" without

further investigation, they are complicit in the corruption. Analysis and debate on education are adversely impacted at all levels—local, state, and federal.

The National Research Council's BOTA was captured decades ago by CRESST-affiliated researchers. A clique of faculty members from a handful of elite universities has monopolized the education-policy function at the country's most prominent think tanks. (Similarly, many argue that the education research function at the National Science Foundation has been captured by radical constructivists who fund the type of research they like and pretend the rest of the research literature does not exist.)

The disastrous results illustrate how federal and foundation money can concentrate power to achieve results exactly the opposite from those intended. Once small, cohesive groups control the larger organizations, they can focus their efforts on restricting entry into policy arenas to those in their own circles. The careers of those inside these groups have flourished. Meanwhile, the amount of objective information available to policymakers and the public—our collective working memory—has shrunk.

Another ramification is that too few people acquire too much influence over those who control the purse strings of education research. And those who control the purse strings wield excessive influence over policy decisions. Until the folks at the Bill and Melinda Gates Foundation and the U.S. Education Department—to mention just a couple of consistent funders of education-policy debacles—broaden their networks, expand their reading lists, and open their minds to more intellectual diversity, they will continue to produce education policy failure.

The problems of American schools can hardly be ameliorated by ignoring sound, relevant information. It would help if funds were available to a wider pool of legitimate education researchers, evidence, and information. In recent years, grantors have instead encouraged the converse—funding a saturating dissemination of a narrow pool of information—thereby contributing to U.S. education policy's number-one problem: pervasive misinformation.

SO WHAT?

Not only are these badly behaved researchers the only sources that most journalists and policymakers consult, but the effects of their bad behavior are also spreading overseas. The education-testing

research function at the World Bank, for example, has been handed down over the past few decades from one scholar affiliated with Boston College's School of Education to another. True to form, they cite the research they like, some of which is their own, most of the rest from CRESST, and imply that the vast majority of relevant research is nonexistent.[8]

Recently, the Organisation for Economic Co-operation and Development (OECD) published a study on educational assessment that followed the template of ignoring most relevant research literature and highlighting work conducted at a certain U.S. federal research center and several U.S. think tanks (Phelps 2013, 2014).

Their skewed recommendations are now the world's.

Recommended Citation: Phelps, R. P. (2015). The gauntlet: Think tanks and federally funded centers misrepresent and suppress other research: *New Educational Foundations* 4, http://www.newfoundations.com/NEFpubs/NEFv4Announcement.html

References

Accountability Works, (2004, January). NCLB under a microscope: A cost analysis of the fiscal impact of the No Child Left Behind Act of 2001 on states and local education agencies

Accountability Works, (2012, February). National Cost of Aligning States and Localities to the Common Core Standards, Boston, Mass.: Pioneer Institute.

Beatty, A. (2008). Common Standards for K–12 Education?: Considering the Evidence: Summary of a Workshop Series. Committee on State Standards in Education, Washington, D.C.: National Research Council.

Chingos, M. (2012, November). Strength in numbers: State spending on K–12 assessment systems. Washington, D.C.: Brookings Institution. Retrieved March 12, 2014, from http://www.brookings.edu/~/media/research/files/reports/2012/11/29%20cost%20of%20assessment%20chingos/11_assessment_chingos_final.pdf

Clarke, M. [moderator] (2013). What does the research tell us about how to assess learning? Panel discussion for World Bank Symposium: Assessment for Global Learning, November 7–8, 2013, Washington, D.C.

Danitz, T. (2001, February 27). Special report: States pay $400 million for tests in 2001. Stateline.org. Pew Center for the States.

Harris, D. N., & Taylor, L. L. (2008, March 10). The Resource Costs of Standards, Assessments, and Accountability: A Final Report to the National Research Council.

Hoxby, C. M. (2002). The cost of accountability, in W. M Evers and H.J. Walberg (Eds.), *School Accountability*, Stanford, Calif.: Hoover Institution Press.

Koretz, D. (2013, November 7). Learning from research on test based accountability? Paper presented at World Bank Symposium: Assessment for Global Learning, November 7–8, 2013, Washington, D.C.

Monk, D. H. (1995, Spring). The costs of pupil performance assessment: A summary report, *Journal of Education Finance*, *20*(4), 363–371.

Monk, D. H. (1996, Spring). The importance of balance in the study of educational costs, *Journal of Education Finance*, *21*(4), 590–591.

Nelson, H. (2013, July). Testing More, Teaching Less: What America's Obsession with Student Testing Costs in Money and Lost Instructional Time, Washington, D.C.: American Federation of Teachers.

Phelps, R. P. (1996, Spring). Mis-conceptualizing the costs of large-scale assessment, *Journal of Education Finance*, *21*(4), 581–589.

Phelps, R. P. (1998). Benefit-cost analysis of systemwide student testing, Paper presented at the annual meeting of the American Education Finance Association, Mobile, Ala.

Phelps, R. P. (2000, Winter). Estimating the cost of systemwide student testing in the United States. *Journal of Education Finance*, *25*(3), 343–380.

Phelps, R. P. (2005, September). A review of Greene (2002), High school graduation rates in the United States, Practical Assessment, Research, and Evaluation, *10*(15). http://pareonline.net/pdf/v10n15.pdf

Phelps, R. P. (2008/2009). The National Research Council's Testing Expertise, Appendix D in R. P. Phelps (Ed.), Correcting fallacies about educational and psychological testing, Washington, D.C.: American Psychological Association. Available at: http://supp.apa.org/books/Correcting-Fallacies/appendix-d.pdf

Phelps, R. P. (2012a). Dismissive reviews: Academe's Memory Hole. Academic Questions, Summer. http://www.nas.org/articles/dismissive_reviews_academes_memory_hole

Phelps, R. P. (2012b). The rot festers: Another National Research Council report on testing. New Educational Foundations, *1*, 30–52. http://www.newfoundations.com/NEFpubs/NEFv1n1.pdf

Phelps, R. P. (2013). The rot spreads worldwide: The OECD: Taken in and taking sides. New Educational Foundations, 2. http://www.newfoundations.com/NEFpubs/NEFv20f0513.pdf

Phelps, R. P. (2014). A review of Synergies for Better Learning: An International Perspective on Evaluation and Assessment. *Assessment in Education: Principles, Policy, and Practice, 21*(4), 481–493.

Picus, L. O., and Tralli, A. (1998, February). Alternative assessment programs: What are the true costs? CSE Technical Report 441, Los Angeles: CRESST.

Shepard, L. (2013, November 7). How can classroom assessment inform learning? Keynote Presentation presented at World Bank Symposium: Assessment for Global Learning, November 7–8, 2013, Washington, D.C.

U.S. GAO. (1993, January). Student testing: Current extent and expenditures, with cost estimates for a national examination. GAO/PEMD-93-8. Washington, D.C.: U.S. General Accounting Office.

Notes

1. Some have argued that an opportunity cost of student time "lost" to testing should also be included. That assumes, however, that students learn nothing when taking a test and that they would be learning something if the time were not used for testing. As it turns out, a massive research literature affirms that students are more likely to learn when taking a test (see, for example, Phelps, 2012). Hence, if the opportunity cost of student time in testing were to be considered for inclusion, it should be subtracted from the cost calculations.

2. For reasons never explained to me, the working title that I gave the study, and that had passed through all internal and external reviews—"Student Testing: Current Extent and Cost, with Estimates for National Examination"—was changed to "Current Extent and Expenditures." This, despite the fact that we used line-item budget data—expenditure data—only to validate the survey data from state and local testing directors, which could be quite different. Line-item expenditures may or may not categorize relevant expenditures neatly; usually they do not. As it turned out, this change substantially aided the censorial efforts the leading critiques of the GAO report, which claimed that it ignored the opportunity costs of personnel time. In fact, the majority of costs in the GAO calculations were of personnel time.

3. For example: 1993 CRESST Conference (Assessment Questions: Equity Answers: What Will Performance Assessment Cost?), Monday, September 13; 1994, CRESST Conference (Getting Assessment Right: Practical and Cost Issues in Implementing Performance Assessment), Tuesday, September 13; 1995, CRESST Conference (Assessment at the Crossroads: What are the Costs of Performance Assessment?), Tuesday, September 12. CRESST report #441 still contains mostly erroneous

claims related to the GAO report, on pages 5 and 64–66, and mostly erroneous claims about CRESST's work on the issue, in the first seventeen pages.

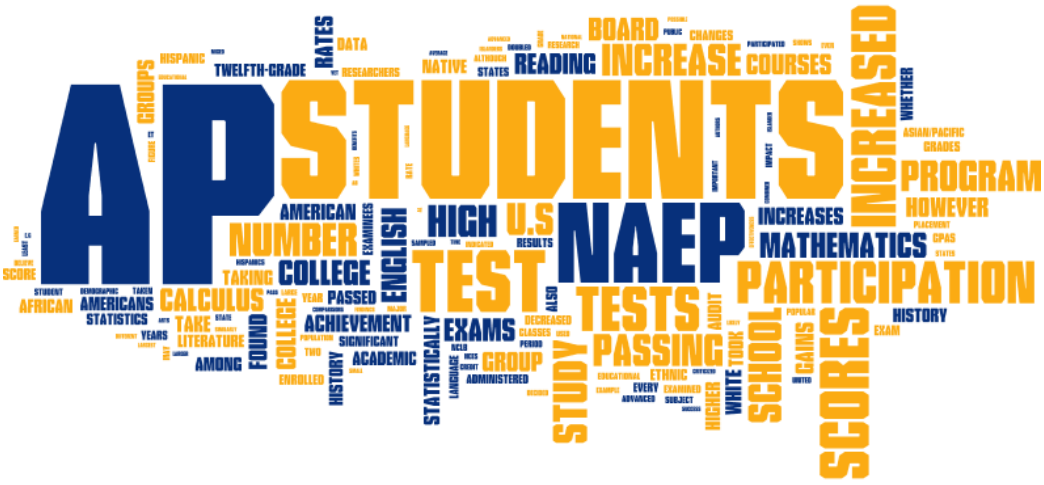
4. The Doctoral Scholar Award of the National Center for Education Statistics (NCES) and the New Scholar Award of the Association for Education Finance and Policy (AEFP), both in 1997.

5. On pages 8–9 of the background paper "The Resource Costs of Standards, Assessments, and Accountability" (Harris and Taylor, 2008) one reads, "On the other hand, neither Phelps nor the GAO study ascribes any costs to standard setting. . . ."

6. Test developers often confusingly use the phrase "standard setting" to identify two entirely different phases of test development. There is the writing of academic content standards and expected performance levels that takes place before the development of a standardized test even starts. Then, much later in the test-development process, after some test forms have already been administered, groups of educators, experts, and public officials gather to decide how to score the new test. Often, but not always, the "standard" being set at these meetings is the passing score for the new test, and the meetings are sometimes called "passing-score setting" meetings. But the traditional, albeit confusing, label of "standard setting" is still widely used. The GAO study included all costs for the latter type of standard setting—passing score setting—contrary to the claims in the NRC report.

7. This is hardly the only issue where education establishment and think tankers present opposing assertions as facts, with both being wrong, misleading, or exaggerated. Until the mid-2000s, for example, education establishment folk favored the use of a "graduation rate" that grossly overestimated the actual proportion of students who begin high school and later graduate. Since then, think tankers have managed to institute a different measure that grossly underestimates that proportion (e.g., by counting those who take more than four years to graduate or transfer schools as dropouts). (See Phelps, 2005.)

8. See Clarke (2013), Koretz (2013), and Shepard (2013). Long a junior partner in CRESST's censorial efforts, the even more radically constructivist and (anti-) reliable, high-stakes testing-policy group at Boston College has somehow maintained control of the educational testing function at the World Bank for decades (viz. various works of Kelleghan, Greaney, and Clarke). Leadership succession in this office of the World Bank is not meritocratic; it is filial.



The Advanced Placement Program's Impact on Academic Achievement

by Russell T. Warne and Braydon Anderson

Abstract

The number of high school students who have taken and passed Advanced Placement (AP) exams has more than doubled since 2000. In this article, we examined whether this increased participation in the AP program has impacted twelfth-grade students' scores on the National Assessment of Educational Progress (NAEP) in mathematics, reading, and U.S. history for all students and for five major ethnic/racial groups: White, Black, Hispanic, Asian American, and Native American students. We found that the drastic increase in AP tests taken has coincided with improved NAEP scores in mathematics, but not in reading or U.S. history. We explored possible explanations for this situation, such as the AP program's possible ineffectiveness in raising overall academic achievement, the small proportion of students who actually take AP tests, and more. We conclude by providing suggestions for future research on the AP program.

Keywords: Advanced Placement program, standardized tests, academic achievement, high school achievement, NAEP

The College Board's Advanced Placement (AP) program enables high-achieving high school students to take college-level classes taught by high school teachers. To demonstrate mastery of the course content and to earn college credit, the students take a standardized AP test at the end of the year (Jeong, 2009). Recently, AP tests have grown more popular in the United States. Between 2000 and 2010, the number of students taking AP tests has more than doubled, and tests taken has increased by a factor of 2.53 times (College Board, 2010a) while the population of fifteen- to nineteen-year-olds in the United States increased by only 9.9% (U.S. Census Bureau, 2011a). The numbers from the College Board are confirmed by the U.S. Department of Education's high school transcript study, which found that the average high school student in the United States was enrolled in 0.58 AP courses in 2000; by 2009, the number had increased to 1.08 (National Center for Educational Statistics, 2011a).

Several factors have contributed to the increased popularity of the AP program. First, economic barriers to participation have been reduced. For example, the College Board, which sponsors the AP program, offers fee waivers for students from low-income families. The federal government has also recently made grants available to states to subsidize AP fees (Klopfenstein & Thomas, 2009). Forty-eight states provide financial assistance to students from low-income families so that they can meet the costs of the AP testing fees (Dounay, 2007). In addition to reducing the financial burden of taking AP tests, at least five states have mandated that every public high school offer AP courses (Dounay, 2006). Those incentives and other techniques devised by states and districts have not only increased the number of students who take AP tests but also increased the number and

proportion of traditionally underserved students (Hispanic, African American, Native American, and low-income students) who are participating in the AP program (College Board, 2010b).

Success in the AP program has been linked with positive outcomes, many of which have been researched by personnel working for or in association with the College Board. One College Board study found that college students in nine different academic majors earned higher college grade-point averages (GPAs) if they had passed AP exams for introductory courses in their majors. Moreover, the number of major-related AP exams was also a statistically significant predictor of college GPAs in engineering, the social sciences, and natural sciences (Patterson, Packman, & Kobrin, 2011). Those findings were supported even after controlling for high school characteristics, college variables, student demographics, and student academic ability.

Research by the scientists affiliated with College Board (Mattern, Shaw, & Xiong, 2009) indicated that students who scored a 3, 4, or 5 on English Language, Biology, Calculus AB, and U.S. History exams achieved higher first-year GPAs and higher second-year retention rates than students who did not pass the same AP test. However, students who took the AP exams but scored only a 1 or a 2 did not earn statistically significantly higher first-year GPAs than students who took no AP exams, regardless of the AP test examined. Mattern and her colleagues concluded that "the results of this study do provide support for the role of participation in the AP Exam in subsequent college performance and success" (2009, p. 12).

Similarly, other College Board researchers found that at a large, elite public university, students who had earned credit through AP examinations outperformed their classmates in subsequent courses in the same major, whether the comparison group consisted of students who did not pass the AP exam, did not take the AP exam, or earned credit through concurrent enrollment programs (Keng & Dodd, 2008). The researchers observed the same pattern of achievement across ten different AP exams. However, the authors could not determine whether any of the students who did not take AP exams were enrolled in AP classes. The authors also could not determine whether success in the AP program *caused* higher achievement in later courses or whether more-successful students were simply more likely to pass the AP test and earn higher grades in more-advanced college courses (Keng & Dodd, 2008).

The previously mentioned research may be considered suspect because of the College Board's financial incentive to sustain and propagate the AP program. Therefore, studies conducted on the AP program by scientists not affiliated with the College Board should have particular value. One such study found that Texas students who had taken and passed AP tests in high school were more likely to graduate from college than students who did not pass AP tests or who did not enroll in AP classes (Dougherty, Mellor, & Jian, 2006). Another study found that among college students who enrolled in introductory science courses, those who had passed the corresponding AP tests in high school received the highest grades. The study also scrutinized students who enrolled in the corresponding AP classes in high school but chose not to take the AP tests. Both groups of AP students surpassed their high school classmates who had enrolled in only honors or regular-level science courses (Sadler & Tai, 2007). Similarly, success in AP courses has been linked to higher college-admissions test scores (Warne, Larsen, Anderson, & Johnson, in press) and an increased likelihood of obtaining an advanced degree (Bleske-Rechek, Lubinski, & Benbow, 2004).

In another study researchers concluded: "[F]or students with similar high school rank or SAT scores, those with advanced placement credit significantly outperformed their peers with no advanced placement credit. Performance of AP students was higher, regardless of gender or ethnicity" (Scott, Tolson, & Lee, 2010, p. 30). The results of the study also indicated that first-semester college students who had taken AP courses or exams in high school earned higher GPAs than students who had not. However, only first-semester college outcomes were examined, and the study left many questions about the long-term benefits of the AP program.

A study conducted by Thompson and Rust tested whether success on AP tests resulted in higher college GPAs in natural sciences and English when compared to the performance of other high-achieving college students who did not take the AP courses or exams in high school. Although the researchers found no difference in college GPAs among AP and non-AP students (possibly because of a restriction in the range of GPAs), the authors found that students who took AP courses thought that the AP program benefited them more than the general high school curriculum (Thompson & Rust, 2007).

Despite such positive findings, questions about the effectiveness of the AP program have been raised as it becomes more popular (e.g.,

Tai, 2008). Hallett and Venegas (2011), for example, found that the AP programs available to inner-city students are of subpar quality and are often taught by unqualified teachers. That finding is bolstered by a study in which AP scores and AP course grades had a low correlation of $r = +.336$ (Sadler & Tai, 2007, p. 8), indicating that many high school AP teachers' grading systems do not accurately reflect AP exam grades. Students in low-quality AP programs have been found to be more likely to fail AP tests and to develop a distaste for the AP program (Jeong, 2009). Lichten (2000) criticized the College Board and the AP program for labeling an examinee's score of 3 as the minimum passing score. Lichten provided compelling evidence that even moderately selective universities require a score of at least 4 to consider a student qualified. He uses this fact as evidence that what the College Board considers "qualified" differs from what many universities consider a "qualified" student.

The authors of a highly cited study critical of the AP program found that the number of AP courses in which a student was enrolled had no relationship to first- and second-year college grades. However, AP *test scores* were found to be the second-most-powerful predictors of college grades, with only high school grade-point averages having a stronger relationship with future grades (Geiser & Santelices, 2004). However, the participants in that study were students at an elite public university, the University of California at Berkeley, which may make the results nongeneralizable to the population of AP students. Moreover, the Geiser and Santelices study has been vigorously criticized by scientists at the College Board on methodological grounds (Camara & Michaelides, 2005).

Because little previous research had disaggregated AP participation from other measures of high school curricular rigor, recent researchers studying the AP program have begun to control for curricular rigor in their studies. After taking the degree of curricular rigor into consideration, one study of Texas public university freshmen found that AP participation provided few unique benefits (Klopfenstein & Thomas, 2009).

Evaluating the Effectiveness of the AP Program

Several methods of evaluating the AP program are possible. For this article, we have decided to compare the increases in AP participation with scores on the National Assessment of Educational Progress (NAEP). In this context, NAEP will be used to examine the

impact of the increase in AP participation. NAEP is frequently used to determine whether gains on a different test are unique or whether the learning gains generalize to other instruments (Brennan, 2001). For example, several studies show that NAEP scores are higher in states implementing high school exit exams than in states that permit their students to graduate from high school without passing a standardized test; the finding demonstrates that the preparation for such high school tests is useful beyond just the state exam and may indicate real learning (Bishop 2005). Similarly, Haney (2008) and Loveless (2008) used NAEP to examine the effectiveness of the educational reforms implemented by the No Child Left Behind (NCLB) Act for different groups of students. When used to confirm findings from other tests, NAEP functions as an *audit test*. Additional examples of using NAEP as an audit test are plentiful (e.g., Haney, 2009; Linn, Graue, & Sanders, 1990).

NAEP is so commonly used as an audit test because its sampling procedures make it the only test given that permits *group* comparisons across state lines (Lane et al., 2009). In fact, NCLB mandates NAEP comparisons across state lines for accountability purposes (Koretz, 2003). Moreover, NCLB has codified into federal law NAEP's status as an audit test for examining states' educational progress (Koretz, 2003). For those reasons, and to maintain a connection to the larger body of K–12 educational research, we decided to use NAEP as an audit test.

Methods

Data for this study were drawn from two principal sources: the College Board and the National Center for Educational Statistics (NCES), which houses the data from NAEP online (NCES, 2011a). NAEP scores for twelfth-grade students from 2001 to 2010 were collected for the total sample of each year and for the major racial/ethnic subcategories that NAEP reports: Whites, African Americans, Hispanics, Asian Americans, and Native Americans. We decided to download NAEP data for reading, mathematics, and U.S. history because those subjects correspond to the most-popular AP exams. In 2011, the U.S. history test was the single-most-popular AP exam—administered to 406,086 students (College Board, 2011). The three mathematics exams (AP Calculus AB, AP Calculus BC, and AP Statistics) were administered to a total of 483,461 students; the two English exams (AP English Literature and Composition and AP English Language and Composition) were administered to 780,428

students in 2011 (College Board, 2011). It is important to note when reading this study that NAEP does not test every subject every year (Lane et al., 2009). Therefore, we downloaded only AP data from the College Board (2001, 2002, 2005, 2006, 2009, 2010c) for years that corresponded to the years the same subject was tested by NAEP.

Although AP participation increased throughout the 1990s (College Board, 2010a), we decided to examine trends for the 2000s only, because changes in NAEP make score comparisons before 2001 difficult. Moreover, we thought not only that the educational changes ushered in at the national level by NCLB were important and drastic enough to represent a convenient break with the past, but also that they made comparisons with the pre-NCLB era less useful.

Data Analysis

The statistical analysis of the publicly available data from the College Board and NCES is actually quite simple. For each subject, we calculated the percentage of increase in AP tests administered compared to the earliest year in our study. Those percentages were calculated within each subject for the entire population of AP examinees and for each ethnic/racial group. The values were then compared graphically to NAEP scores in the same subject for the corresponding years.

It should be noted that there are three AP tests in mathematics (Calculus AB, Calculus BC, and Statistics) and two in language arts (English Language and English Literature). For data analysis we always combined the data from the two AP Calculus courses. For language arts AP tests, we analyzed data from each AP test separately and combined the results.

Results

Mathematics

Table 1. Number of AP Calculus and Statistics Tests Administered, Passed, and Not Passed during Years Examined in This Study

Group	AP Calculus (AB and BC combined)			
	Year	Tests Passed	Tests Not Passed	Total
White	2005	100,994	53,180	154,174
African American	2005	2,862	6,105	8,967
Hispanic	2005	6,505	9,823	16,328
Asian/Pacific Islander	2005	28,393	12,438	40,831
Native American	2005	407	478	885
All Students	**2005**	**187,006**	**105,616**	**292,622**
White	2009	122,528	60,039	182,567
African American	2009	4,058	8,968	13,026
Hispanic	2009	10,703	15,210	25,913
Asian/Pacific Islander	2009	39,385	15,257	54,642
Native American	2009	567	524	1,091
All Students	**2009**	**147,217**	**86,801**	**234,018**
Group	AP Statistics			
	Year	Tests Passed	Tests Not Passed	Total
White	2005	32,677	17,839	50,516
African American	2005	841	2,442	3,283
Hispanic	2005	1,793	3,337	5,130
Asian/Pacific Islander	2005	7,949	4,235	12,184
Native American	2005	125	142	267
All Students	**2005**	**45,830**	**29,838**	**75,668**
White	2009	45,818	26,931	72,749
African American	2009	1,520	4,648	6,168
Hispanic	2009	3,385	6,444	9,829
Asian/Pacific Islander	2009	12,667	6,334	19,001
Native American	2009	184	268	452
All Students	**2009**	**67,006**	**47,492**	**114,498**

Note. Ethnicity-group totals do not add up to the total number of students within a year because ethnicity was reported as "other" or "unknown" for some students.

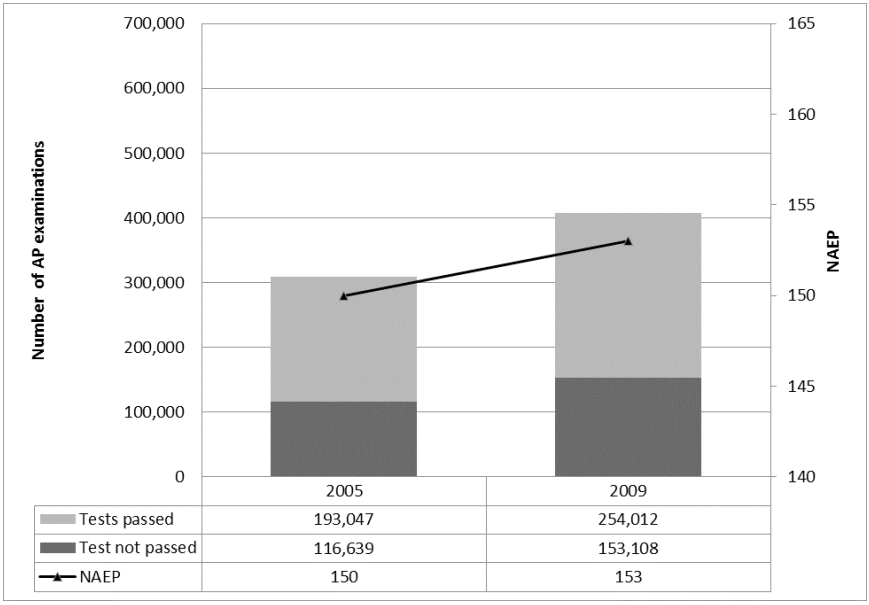


Figure 1. Increases in the number of students taking and passing the AP Calculus and Statistics examinations, compared to NAEP mathematics score trends.

AP Calculus. Table 1 shows the number of students who took AP tests in mathematics during the same years that NAEP was administered. Figure 1 shows the changes in NAEP scores and the changes in AP participation in the same years. The NAEP mathematics test was administered twice during the period we examined in this study. During 2005, 234,018 AP Calculus (AB and BC) tests were administered, with 147,217 tests being passed. By 2009, the number of AP Calculus tests had increased to 292,622; 187,006 of those tests were passed. This means that in 2009 25% more students were taking the AP Calculus test than in 2005 and that the number of students passing the tests increased 27% in that time frame, figures that also correspond to a slight increase in passing rates (from 62.9% in 2005 to 63.9% in 2009).

The number of students taking and passing the AP Calculus exams increased in all major racial or ethnic groups examined by the College Board. The greatest increases occurred among Hispanic students (a 58.7% increase in AP Calculus exams taken and a 64.5% increase in the number of examinations passed). Whites were the ethnic group with the smallest proportional increase in AP Calculus

participation: an 18.4% increase in the number of AP Calculus examinations between 2005 and 2009 and a 21.3% increase in the number of passing AP Calculus exams. However, White students recorded the largest numerical increases in examinees and passing scores because of the group's larger size.

Although AP Calculus participation increased among all ethnic groups in the United States, passing rates did not. We found increases in the passing rates of AP Calculus exams for White, Asian/Pacific Islander, Native American, and Hispanic students. However, the passing rates of African American students decreased slightly—from 31.9% to 31.1%. The greatest increases in passing rates from 2005 to 2009 were observed among Native Americans (from 46.0% to 52.0%) and Hispanics (from 39.8% to 41.3%).

AP Statistics. The results of the AP Statistics test bear strong similarities to the findings on the AP Calculus tests, though with some minor differences. As in the calculus exams, each racial group increased its participation in the AP Statistics test. The increase in White participation was the greatest (22,233 students from 2005 to 2009); the rate of participation increased the most among Hispanics (91.6% from 2005 to 2009).

Passing rates for the entire population of AP Statistics examinees decreased, however, from 60.6% to 58.5%. Only Asian/Pacific Islanders increased their passing rates (from 65.2% to 66.7%). The largest decrease in the passing rate was among the Native American group, whose 46.8% passing rate in 2005 declined to 40.7% in 2009. However, that demographic group was so small—in both years fewer than 500 Native Americans took the AP Statistics test—that the change is statistically insignificant.

NAEP. As is apparent in Figure 1, national scores from the twelfth-grade NAEP mathematics assessment have demonstrated positive, statistically significant changes from 2005 to 2009 in every ethnic group and for all ethnic groups combined. Overall, NAEP scores increased from 150 to 153, with individual ethnic groups' increases ranging from 4 to 12 points.

Table 2. Number of AP English Literature and English Language Tests Administered, Passed, and Not Passed during Years Examined in This Study

Group	AP English Literature			
	Year	Tests Passed	Tests Not Passed	Total
White	2002	106,830	42,457	149,287
African American	2002	3,669	8,039	11,708
Hispanic	2002	6,348	9,703	16,051
Asian/Pacific Islander	2002	13,722	7,103	20,825
Native American	2002	457	502	959
All Students	**2002**	**139,375**	**71,799**	**211,174**
White	2005	119,003	54,139	173,142
African American	2005	4,322	12,179	16,501
Hispanic	2005	8,555	15,029	23,584
Asian/Pacific Islander	2005	15,937	9,271	25,208
Native American	2005	574	765	1,339
All Students	**2005**	**158,243**	**97,464**	**255,707**
White	2009	137,496	66,719	204,215
African American	2009	6,626	20,713	27,339
Hispanic	2009	13,460	25,831	39,291
Asian/Pacific Islander	2009	20,665	12,490	33,155
Native American	2009	802	1,072	1,874
All Students	**2009**	**190,518**	**135,210**	**325,728**
Group	**AP English Language**			
	Year	**Tests Passed**	**Tests Not Passed**	**Total**
White	2002	70,271	32,331	102,602
African American	2002	2,590	5,989	8,579
Hispanic	2002	5,450	10,828	16,278
Asian/Pacific Islander	2002	9,963	5,975	15,938
Native American	2002	357	390	747
All Students	**2002**	**94,573**	**59,193**	**153,766**
White	2005	93,035	51,918	144,953
African American	2005	3,681	10,672	14,353
Hispanic	2005	8,224	18,907	27,131
Asian/Pacific Islander	2005	14,322	10,268	24,590
Native American	2005	554	738	1,292
All Students	**2005**	**128,057**	**98,830**	**226,887**
White	2009	136,814	61,398	198,212
African American	2009	7,737	19,737	27,474
Hispanic	2009	16,332	30,749	47,081
Asian/Pacific Islander	2009	23,799	12,414	36,213
Native American	2009	968	1,113	2,081
All Students	**2009**	**198,089**	**134,390**	**332,479**

Note. Ethnicity-group totals do not add up to the total number of students within a year because ethnicity was reported as "other" or "unknown" for some students.

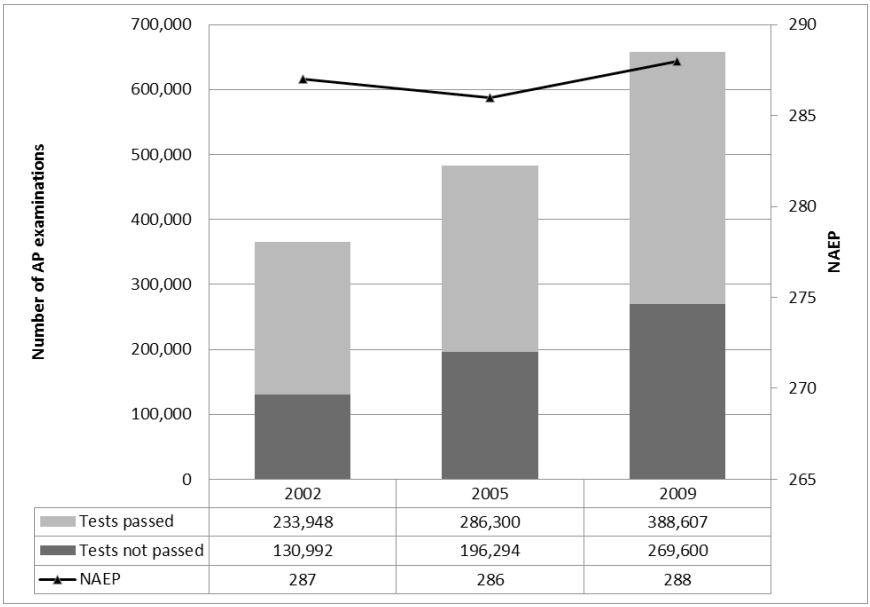


Figure 2. Increases in the number of students taking and passing the AP English Literature and English Composition examinations, compared to NAEP reading score trends.

Reading

AP English Literature. Measures of reading achievement were administered to twelfth-grade students for NAEP in 2002, 2005, and 2009. During that time, the number of examinees taking the AP English Literature test increased from 211,174 to 325,728 (shown in Table 2 and Figure 2)—a 54.2% increase. During the same period, the number of students passing the AP English Literature test increased from 139,375 to 190,518, a figure that represents a 36.7% increase. Because the number of examinees increased faster than the number of students who passed the test, the passing rate for the AP English Literature test decreased from 66.0% in 2002 to 58.5% in 2009.

For all groups the number of examinees and of students passing the test increased, but the passing rates decreased. African American students' passing rates decreased the most: from 31.3% to 24.2% between 2002 and 2009. During the same period, the number of African American students who took the AP English Literature test more than doubled, from 11,708 to 27,339. Although a large increase, it was exceeded by the increased Hispanic participation in the AP

English Literature test. From 2002 to 2009 the number of Hispanic students who took the test increased from 16,051 to 39,291—an increase of 144.8%. The Hispanic passing rate decreased from 39.5% to 34.3%. The decreases in passing rates for other demographic groups ranged from 3.6% (for Asian/Pacific Islanders) to 4.8% (for Native Americans).

AP English Language. As with the AP mathematics and English Literature tests, the results from the AP English Language test indicated that it became much more popular in recent years. Table 2 shows that from 2002 to 2009, the number of students taking the test grew from 153,766 to 322,479—an increase of 109.7%. However, increased participation produced a more mixed impact on passing rates than that observed in the AP English Literature test. Overall, the passing rate decreased modestly, from 61.5% to 59.6%. However, the passing rates of White, Asian/Pacific Islander, and Hispanic groups increased (ranging from 0.5% to 3.2%), while those of African American and Native American examinees decreased (2.0% and 1.3%, respectively).

Participation in the AP English Language testing program increased more dramatically than for any other test we examined. The number of African American students taking the AP English Language test more than tripled, from 8,579 to 27,474 (an increase of 220.2%), and the number of African American students passing increased proportionally almost as much, from 2,590 to 7,737 (an increase of 198.7%). In fact, for every ethnic group except Whites, the number of participants in the AP English Language test more than doubled from 2002 to 2009, and for White students the increase was 93.2%.

NAEP. The scores on the twelfth-grade NAEP reading assessment, however, were somewhat mixed, as indicated in Figure 2. The overall decrease in the reading scores from 2002 to 2005 was statistically significant, but the decline ended in 2009. The White students' NAEP score of 296 in 2009 was statistically higher in significance than were the scores of both 2005 and 2002. The most drastic gain in NAEP reading scores, from 286 in 2002 to 298 in 2009, took place among Asian/Pacific Islanders. The other groups had NAEP reading-score increases ranging from one to four points, none statistically significant.

Table 3. Number of AP U.S. History Tests Administered, Passed, and Not Passed during Years Examined in This Study

Group	Year	Tests Passed	Tests Not Passed	Total
White	2001	78,148	65,054	143,202
African American	2001	2,609	7,786	10,395
Hispanic	2001	4,084	10,735	14,819
Asian/Pacific Islander	2001	12,299	10,682	22,981
Native American	2001	336	527	863
All Students	**2001**	**104,625**	**100,215**	**204,840**
White	2006	116,712	82,129	198,841
African American	2006	4,621	13,937	18,558
Hispanic	2006	8,676	21,601	30,277
Asian/Pacific Islander	2006	20,093	14,846	34,939
Native American	2006	568	983	1,551
All Students	**2006**	**163,790**	**144,767**	**308,557**
White	2010	137,052	94,359	231,411
African American	2010	7,104	21,636	28,740
Hispanic	2010	15,304	35,150	50,454
Asian/Pacific Islander	2010	27,929	16,816	44,745
Native American	2010	762	1,312	2,074
All Students	**2010**	**201,994**	**182,572**	**384,566**

Note. Ethnicity-group totals do not add up to the total number of students within a year because ethnicity was reported as "other" or "unknown" for some students.

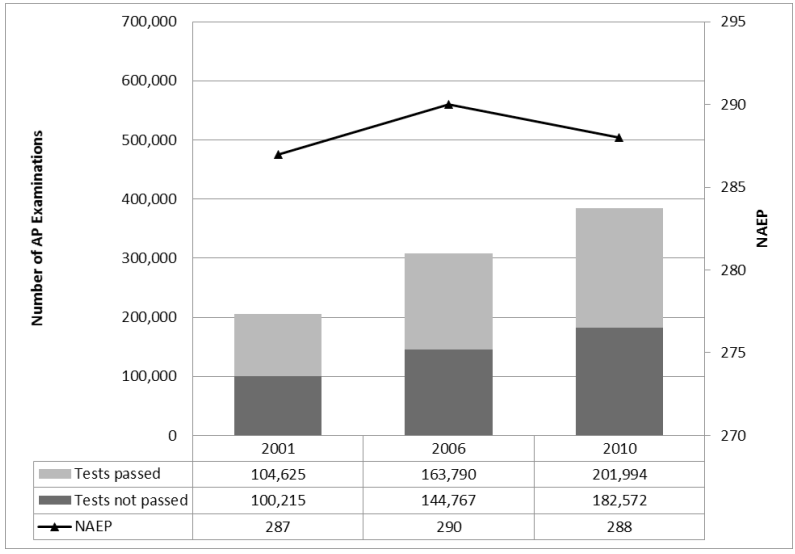


Figure 3. Increases in the number of students taking and passing the AP U.S. History examination, compared to NAEP U.S. history score trends.

U.S. History

AP U.S. History. Table 3 and Figure 3 show the trends in AP participation and NAEP scores for U.S. History. Like the other AP tests discussed earlier, the AP U.S. History test has become extremely popular in recent years, with 87.7% more students taking the exam in 2010 than in 2001 (an increase of 179,726 students). Additionally, similar to what we observed in the other AP tests, each ethnic group increased its participation: the increases ranged from 61.6% (for White students) to 240.5% (for Hispanic students). Again, because of the sheer size of the group, White students accounted for the largest share of the increase in test takers (88,209 students). However, the number of students participating in the AP U.S. History exam and those passing the exam more than doubled for all three groups traditionally underrepresented in AP classes—Hispanics, African Americans, and Native Americans. Among Hispanics, 240.5% more students took the test in 2010 than in 2001, and 274.7% more passed; 176.5% more African Americans participated, and 172.3% more passed; 140.3% more Native Americans participated, and 126.8% more passed. Overall, changes in passing rates for the AP U.S. History exam were minimal, with no increase larger than 4.7% (for White students) and no decrease larger than 2.1% (for Native Americans).

NAEP. Figure 3 shows that despite the increase in AP participation, for the nation as a whole twelfth-grade NAEP scores in U.S. history remained statistically equal from 2001 to 2010. The only ethnic group with improved U.S. history NAEP scores during that period was White students, whose numbers increased from 292 to 297. Every other ethnic group score remained statistically constant from 2001 to 2010.

Discussion

This study utilized the period from 2001 to 2010 to compare the changes in AP testing-program participation with changes in NAEP scores in reading, mathematics, and U.S. history. Those subjects were chosen because they represented the most popular AP tests and because two—mathematics and reading—are generally considered the most important subjects in the core curriculum. We found that AP participation skyrocketed from 2001 to 2010, yet only NAEP mathematics scores showed consistent gains during the same period. However, White and Asian/Pacific Islander students demonstrated statistically significant gains in NAEP reading scores, as did White students in U.S. history.

When we started this study, we expected that the rapid increase in AP participation would lead to higher academic achievement, as measured by an audit test like NAEP. We were disappointed, for example, to find that the increase of almost 80,000 students per year passing the AP U.S. History test did not lead to statistically significant rises in NAEP scores. However, at the same time that the number of students taking the AP Calculus tests increased by 58,604 and those taking the AP Statistics test increased by 38,830, NAEP mathematics scores increased for every major ethnic/racial group in the United States. It is also important to consider a mixed result: all groups' participation in the AP English Literature and Composition tests increased by at least 59.8%, yet only Whites and Asian/Pacific Islanders demonstrated statistically significant gains in NAEP reading scores during the same time.

Overall, we believe that those results are contradictory and their meaning is unclear. We think it reasonable to expect increases in participation in an academically rigorous program—like the Advanced Placement program—to lead to score increases on an audit test that assesses more basic content. The fact that gains in NAEP scores and in AP participation coincide only in mathematics is troubling, and it raises questions about the effectiveness of the AP program. In the following section we explore potential explanations for our observations.

First, it is possible that the AP program is not an effective method of raising academic achievement, at least in language arts and U.S. history. Another possibility is that the number of students in AP courses is not large enough to raise the average NAEP score in reading or U.S. history. During 2009, 4.21 million students between the ages of fifteen and nineteen were enrolled in the twelfth grade (U.S. Census Bureau, 2011b). Consequently, only about 9.1% of students took even the most-popular AP test during that school year—U.S. History. Similarly, about 17.0% of students would take the two English tests combined. It is questionable whether such a small minority of advanced students could impact the average score on NAEP. That is especially true for groups underrepresented in AP classrooms. For example, 651,000 African American students between the ages of fifteen and nineteen were enrolled in the twelfth grade in 2009. That same school year, a paltry 4.4% took the AP U.S. History test, and only 1.1% passed it. Considering the relatively small numbers of high-achieving African Americans, it is unlikely that even the large proportional increases in AP participation would impact NAEP scores significantly.

Nevertheless, that possibility does not explain why NAEP scores among every demographic group increased for mathematics. Only some 7.0% of U.S. students take the two AP Calculus tests each year, yet the NAEP mathematics scores show the strongest gains. While the present data are not clear, the gains in NAEP scores may be due to educational reforms enacted during the students' earlier education. Indeed, previous research has suggested that NCLB reforms have been most beneficial to diverse students and to students who were struggling the most in their schooling (e.g., Haney, 2008; Loveless, 2008). While that proposition suffices to explain why mathematics scores have increased strongly from 2005 to 2009, it does not explain why scores in reading—another area in which most students have made gains on NAEP during grades 4 and 8—are flat for twelfth-graders.

The idea that the non-AP students' performance is masking the gains in AP participation is supported by the latest NAEP report on mathematics and reading (NCES, 2011b), which indicates that reading students at the 75th and 90th percentiles—the students most likely to take AP classes—have increased their scores by 3 points since 2001 (for reading) and 2005 (for mathematics). Readers should note, though, that the increased achievement at the 90th percentile in mathematics is not statistically significant. Nevertheless, those results strengthen the argument that AP courses are translating into achievement gains, even if those gains are not raising the average NAEP scores of the entire population of high school seniors. However, the increase in participation in the AP U.S. History program has not led to similar gains in U.S. history achievement on the NAEP scores of twelfth-grade students at the 75th and 90th percentiles (NCES, 2011c).

Limitations

As with all research, our study has its limitations and shortcomings. One strong assumption in our calculations is that students taking AP tests are high school seniors. Although such students make up the largest portion of AP examinees (College Board, 2011), they are not a majority. However, we believe that twelfth-grade students who took AP tests earlier in their high school careers would still manifest gains on NAEP in twelfth grade.

Another limitation to this study is that NAEP participation rates in twelfth grade are quite low, which may impact the validity of academic gains at the national level. Schools, states, and students participate in NAEP at very high levels in grades 4 and 8 because NCLB mandates

participation if schools are to receive federal funds. However, participation in twelfth-grade assessments remains optional under current law (Noell & Ginsburg, 2009). Indeed, in the 2005 mathematics assessment, only 57% of sampled twelfth-graders participated in NAEP, versus 90% of sampled eighth-graders and 93% of sampled fourth-graders. In reading, only 55% of sampled twelfth-grade students participated in NAEP assessments, while 88% of eighth-graders and 90% of fourth graders who were sampled participated (Chromy, 2005, pp. 3–4). Until twelfth-grade participation in NAEP increases, the validity of NAEP scores for that grade level will be questionable, and NAEP's usefulness as an audit test will be impaired.

A further limitation of this study is that the only AP students examined were those who actually took AP tests. Students commonly enroll in AP classes and receive exposure to the more-advanced curriculum but decline to take the tests. Lichten (2000) estimated that one-third of students enrolled in AP courses forgo taking the AP test. Data from Dougherty et al.'s (2006) study indicated that 46.6% of AP enrollees decided not to take AP exams, while Geiser and Santiclices (2004) found 55% to 60% of AP students making that decision. No matter the exact proportion of AP students who decide not to take AP tests, it is a considerable portion of high school students, and their data were not included in this study. Whether mere enrollment in an AP course leads to academic benefits is an open question, although Dougherty et al. (2006) did find that enrolling in AP courses eventually benefits high school students in college.

Finally, NAEP has been criticized as an audit test for a variety of technical reasons. Yet NAEP is the only audit test established as such by federal law. Although NAEP may not be perfect, it is widely accepted as an audit test among educational researchers, and we believe that no better option is available for monitoring group academic-achievement gains nationwide. Even if NAEP were the ideal audit test, this study is merely correlational in nature, and—based on these data—we cannot state whether AP participation actually *causes* increases in test scores on an academic achievement test. At best, we could merely say that the increase in AP participation coincided with changes in NAEP scores. However, we believe that this study nevertheless presents an important review of the effectiveness of the AP program, which has recently been criticized in education-research circles (e.g., Lichen, 2000; Tai, 2008).

Conclusion

We believe that the impact of the AP program's increased popularity on overall academic achievement is mixed or negligible. The only subject that shows a simultaneous increase in both AP participation and NAEP scores is mathematics. In language arts, only White and Asian/Pacific Islander students improved their NAEP scores, despite dramatic increases in AP test participation rates among all demographic groups. In U.S. history, moreover, only Whites experienced a statistically significant increase in NAEP scores, yet the number of AP U.S. History examinees increased by at least 61.6% among every demographic group.

We urge further study to determine which AP programs raise academic achievement most effectively. We also suggest that future researchers use state databases—which are often much more detailed than our data—and methodologies that permit stronger inferences (such as random assignment or propensity score modeling) to study whether AP participation causes other increases in academic achievement (see Warne et al., in press). Surely researchers and the public at large could only benefit from more independent scrutiny of this highly popular program.

References

Bishop, J. (2005). High school exit examinations: When do learning effects generalize? *Yearbook of the National Society for the Study of Education, 104*(2), 260–288. doi:10.1111/j.1744-7984.2005.00033.x

Bleske-Rechek, A., Lubinski, D., & Benbow, C. P. (2004). Meeting the educational needs of special populations: Advanced Placement's role in developing exceptional human capital. *Psychological Science, 15*, 217–224. doi:10.1111/j.0956-7976.2004.00655.x

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*(4), 6–18. doi:10.1111/j.1745-3992.2001.tb00071.x

Camara, W., & Michaelides, M. (2005). *AP use in admissions: A response to Geiser and Santelices*. Retrieved from College Board website: http://research.collegeboard.org/publications/content/2012/05/ap-use-admissions-response-geiser-and-santelices

Chromy, J. R. (2005). *Participation standards for 12th grade NAEP*. Retrieved from National Assessment Governing Board website: http://www.nagb.org/publications/chromy_paper_revised.doc

College Board. (2001). *National summary report* [Data file]. Retrieved from http://www.collegeboard.com/prod_downloads/student/testing/ap/sumrpts/2001/xls/national_2001.xls

College Board. (2002). *National summary report* [Data file]. Retrieved from http://www.collegeboard.com/prod_downloads/student/testing/ap/sumrpts/2002/xls/national_2002.xls

College Board. (2005). *National summary report* [Data file]. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/programsummaryreport_47494.xls

College Board. (2006). *National summary report* [Data file]. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/ap06_prog_summary_rptx.xls

College Board. (2009). *National summary report* [Data file]. Retrieved from http://www.collegeboard.com/prod_downloads/student/testing/ap/sumrpts/2009/xls/NATIONAL_Summary.xls

College Board. (2010a). *Annual AP program participation 1956–2010*. Retrieved from College Board website: http://professionals.collegeboard.com/profdownload/AP-Annual-Participation-2010.pdf

College Board. (2010b). *The 6th annual AP report to the nation*. Retrieved from College Board website: http://professionals.collegeboard.com/profdownload/6th-annual-ap-report-to-the-nation-2010.pdf

College Board. (2010c). *National summary report* [Data file]. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/AP-Program-Summary-Report-2010.xls

College Board. (2011). *Program summary report*. Retrieved from College Board website: http://professionals.collegeboard.com/profdownload/AP-Program-Summary-Report.pdf

Dougherty, C., Mellor, L., & Jian, S. (2006). *The relationship between Advanced Placement and college graduation* (2005 AP Study Series, Report 1). Retrieved from National Center for Educational Accountability website: http://www.nc4ea.org/files/relationship_between_ap_and_college_graduation_02-09-06.pdf

Dounay, J. (2006). *Advanced Placement: State mandates AP course offerings*. Retrieved from Education Commission of the States website: http://mb2.ecs.org/reports/Report.aspx?id=996

Dounay, J. (2007). *Advanced Placement: Subsidies for testing fees*. Retrieved from Education Commission of the States website: http://mb2.ecs.org/reports/Report.aspx?id=1003

Geiser, S., & Santelices, V. (2004). *The role of Advanced Placement and honors courses in college admissions* (Research & Occasional Paper Series

CSHE.4.04). Retrieved from Center for Studies in Higher Education website: http://cshe.berkeley.edu/publications/docs/ROP.Geiser.4.04.pdf

Hallett, R. E., & Venegas, K. M. (2011). Is increased access enough? Advanced Placement courses, quality, and success in low-income urban schools. *Journal for the Education of the Gifted, 34*, 468–487. doi:10.1177/016235321103400305

Haney, W. M. (2008). Evidence on education under NCLB (and how Florida boosted NAEP scores and reduced the race gap). In G. L. Sunderman (Ed.), *Holding NCLB accountable: Achieving accountability, equity, and school reform* (pp. 91–101). Thousand Oaks, Calif.: Corwin Press.

Jeong, D. W. (2009). Student participation and performance on Advanced Placement exams: Do state-sponsored incentives make a difference? *Educational Evaluation and Policy Analysis, 31*, 346–366. doi:10.3102/0162373709342466

Keng, L., & Dodd, B. G. (2008). *A comparison of college performances of AP and non-AP student groups in 10 subject areas* (College Board Research Report No. 2008-7). Retrieved from College Board website: http://research.collegeboard.org/publications/content/2012/05/comparison-college-performances-ap-and-non-ap-student-groups-10-subject

Klopfenstein, K., & Thomas, M. K. (2009). The link between Advanced Placement and early college success. *Southern Economic Journal, 75*, 873–891.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice, 22*(2), 18–26. doi:10.1111/j.1745-3992.2003.tb00124.x

Lane, S., Zumbo, B. D., Abedi, J., Benson, J., Dossey, J., Elliott, S. N., . . . Willhoft, J. (2009). Prologue: An introduction to the evaluation of NAEP. *Applied Measurement in Education, 22*, 309–316. doi:10.1080/08957340903221436

Lichten, W. (2000). Whither Advanced Placement? *Education Policy Analysis Archives, 8*(29). Retrieved from http://epaa.asu.edu/ojs/article/view/420/543

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5–14. doi:10.1111/j.1745-3992.1990.tb00372.x

Loveless, T. (2008). An analysis of NAEP data. In *High-achieving students in the era of NCLB* (pp. 13–48). Washington, D.C.: Thomas B. Fordham Institute.

Mattern, K. D., Shaw, E. J., & Xiong, X. (2009). *The Relationship between AP exam performance and college outcomes* (College Board Research Report

No. 2009-4). Retrieved from College Board website: http://research.collegeboard.org/publications/content/2012/05/relationship-between-ap-exam-performance-and-college-outcomes

National Center for Educational Statistics. (2011a). *NAEP data explorer* (Online data aggregator). Retrieved from http://nces.ed.gov/nationsreportcard/naepdata/

National Center for Educational Statistics. (2011b). *Grade 12 reading and mathematics 2009 national and pilot state results* (NCES Report 2011-455). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2009/2011455.pdf

National Center for Educational Statistics. (2011c). *U.S. history 2010: National Assessment of Educational Progress at grades 4, 8, and 12* (NCES Report 2011-468). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2010/2011468.pdf

Noell, J., & Ginsburg, A. (2009). Evaluation of the National Assessment of Educational Progress: Next steps. *Applied Measurement in Education, 22*, 409–414. doi:10.1080/08957340903221691

Patterson, B. F., Packman, S., & Kobrin, J. L. (2011). *Advanced Placement exam-taking and performance: Relationships with first-year subject area college grades* (College Board Research Report No. 2011-4). New York, N.Y.: College Board. Retrieved from College Board website: http://research.collegeboard.org/publications/content/2012/05/advanced-placement-exam-taking-and-performance-relationships-first-year

Sadler, P. M., & Tai, R. H. (2007). Weighting for recognition: Accounting for Advanced Placement and honors courses when calculating high school grade point average. *NASSP Bulletin, 91*, 5–32. doi:10.1177/0192636506298726

Scott, T. P., Tolson, H., & Lee, Y. (2010). Assessment of Advanced Placement participation and university academic success in the first semester: Controlling for selected high school academic abilities. *Journal of College Admission, 208*, 26–30.

Tai, R. H. (2008). Posing tougher questions about the Advanced Placement program. *Liberal Education, 94*(3), 38–43.

Thompson, T., & Rust, J. O. (2007). Follow-up of Advanced Placement students in college. *College Student Journal, 41*, 416–422.

U.S. Census Bureau. (2011a). *Age and sex composition: 2010* (Census Report C2010BR-03). Retrieved from http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf

U.S. Census Bureau. (2011b). *Single grade of enrollment and high school graduation status for people 3 years old and over, by sex, age (single years for 3 to*

24 years], race, and Hispanic origin: October 2009 [Data files]. Retrieved from http://www.census.gov/hhes/school/data/cps/2009/tab02-01.xls

Warne, R. T., Larsen, R., Anderson, B., & Johnson, A. O. (in press). The impact of participation in the Advanced Placement program on students' college admissions test scores. *The Journal of Educational Research*.

Russell T. Warne is Assistant Professor of Psychology in the Department of Behavioral Science, Utah Valley University.

Braydon Anderson, a former research assistant in the UVU Department of Behavioral Science, is now affiliated with Qualtrics, Provo, Utah.

The authors appreciate the feedback that Rosalma Arcelay provided for this manuscript.

Correspondence concerning this article should be addressed to Russell T. Warne, Department of Behavioral Science, Utah Valley University, 800 W. University Parkway MS 115, Orem, UT 84058. Email: rwarne@uvu.edu.

BOOK REVIEW

Raymond E. Callahan

Education and the Cult of Efficiency

A Study of the Social Forces That Have Shaped the Administration of the Public Schools

University of Chicago Press, 1962

reviewed by Gary K. Clabaugh

A Particularly Relevant Classic

TODAY'S WOULD-BE SCHOOL REFORMERS misuse high-stakes testing, manifest an soulless preoccupation with purely vocational objectives, scapegoat educators for the academic consequences of chronic social and economic injustice, and bully teachers when respectful consultation and cooperation are required. Many wonder how such a downright foolish approach ever came to dominate.[1] Raymond Callahan's 1962 classic *Education and the Cult of Efficiency* explains how it all began.

Scientific Management and the School as Factory

Callahan focuses on 1900 to 1930, that critical period when socioeconomic circumstances pushed public schooling into its present industrial mode. Until Congress restricted immigration in the

years after World War I, an unprecedented number of newcomers had been flooding into America. Combined with the simultaneous mass migration of Americans from farm to city and persistent inflation, that produced unprecedented difficulties for public schooling.

Just keeping up with urban enrollment increases proved extremely challenging. Between 1906 and 1917, for example, the School District of Philadelphia had to build forty-four new elementary and six new high schools.[2] Similar explosive growth occurred in city after city.

Urban school administrators were forced to focus sharply on per-pupil costs.[3] Unfortunately, budget constraints often morphed into poorly conceived, industrialized school management that fit the pro-business times but badly shortchanged both students and teachers.

Callahan explains that the public schools' organization made them especially susceptible to the era's pro-business *zeitgeist.*[4] The urgency of the situation made a focus on efficiency all but inevitable. Moreover, admired "experts" were assuring the public that managing schools scientifically, via cost accounting and cost management, would solve the public school funding crises.*5*

Many of the individuals guiding this supposed scientific revolution were professors in newly created departments of educational administration. (Previously, school managers had not been formally trained.) In addition to preparing business-minded school administrators, these education revolutionaries were busy writing influential school-administration texts, acting as consultants to major city school systems, and advising industry.[6]

Was the self-appointed experts' perspective really scientific? Callahan offers evidence that it was not. The school-management methods they advocated were often based on scientifically primitive studies of heavy industry—in one laughable instance, the production of pig iron.

The Evangelists

Callahan provides particularly interesting descriptions of scientific management's major evangelists. Among the most famous is Columbia University's John Franklin Bobbitt. His lectures and publications unapologetically reduced public education to a business model, transformed schools into factories, and sternly advised school administrators to make efficiency their master.

"Scientific" management experts conveniently claimed that large class sizes (thirty-five to fifty students) made no difference in educational outcomes.[7] And they characteristically promised that large schools were superior to smaller ones. Callahan cites the example of the Chicago schools' cost-conscious superintendent, William McAndrew. He "proved" large schools' superiority by citing the solicited opinion of his subordinates and providing tables, compiled by his finance department, that showed a 9.5 percent per student saving at schools of 4,000 students compared to those of 2,500.[8]

One might think that by employing a business model to industrialize public schooling, Bobbitt and his fellow scientific-management evangelists outdid modern-day reformers. None of them, though, advocated management of public schools by privately operated businesses. Nor did they even imagine publicly funded, for-profit, charter school chains that employ scripted lessons written for semi-skilled workers—the very epitome of the school as factory.

Few early twentieth-century educators openly challenged the ascendancy of the school as factory. One who did was Thomas J. McCormick, a high school principal from LaSalle, Illinois. He acerbically informed the National Education Association's factory-school-oriented Department of Secondary Education that the "inordinate zeal to practicalize and popularize education" ignored its real purpose: to "make men and women as well as engineers and rope stretchers."[9]

Such criticism notwithstanding, business-oriented school boards, elected by cost-conscious voters, quickly began to hire the industrial-style school managers that the education-administration programs were turning out.

High-Stakes Tests

Predictably, the cult of efficiency depended on high-stakes tests. Predictably, in a no-tenure era, few educators spoke out against them. William E. Maxwell, the superintendent of the New York City schools, became sufficiently frustrated to say:

> After shedding lakes of ink and using up untold reams of paper and consuming the time of un-numbered teachers in administering and scoring the Courtis standard tests . . . , the learned director reached the conclusion that "29% of the pupils in the eighth grade could exchange places with a like number of students in the fourth grade," I am inclined to exclaim:

> My dear sir, what do you expect? That all the children in a grade would show equal ability in adding, subtracting, multiplying and dividing? Any teacher of experience could have told you that they would not. You should have known it yourself. One flash of Horace Mann's insight would be worth a thousand miles of your statistics.[10]

Current efforts to "reform" public schooling likewise rely on high stakes tests.[11] Nothing is inherently wrong with such tests, per se; in fact, they can be very helpful. The harm arises when they are misused, for example, as a way to prod already-stressed educators.

Test-making services warn that their tests aren't intended to evaluate educators, though those cautions seem suspiciously similar to beer commercials that remind boozers to "drink responsibly." In any case, many contemporary critics charge that the tests' frequent misuse is fundamentally misdirecting public education.

A combination of governmental, media, and public pressure accentuates the misdirection. A headline in a suburban Philadelphia newspaper offers an example of the pressures at work: "Board Addresses Decrease in Test Scores." The story begins: "A number of lower scores on standardized tests left officials in the Wissahickon School District with a lot of explaining to do."[12]

That kind of public pressure is precisely what causes educators to teach to the test. Administrators and teachers alike, feeling under assault, develop test-focused tunnel vision. Some educators, especially those wrestling with the abysmal test scores associated with deep poverty, conclude that cheating is the only way to survive: hence the scandals in such cities as Atlanta, Philadelphia, and Washington, D.C.

Callahan describes how, in the early 1900s, muckraking journalists generated similar public pressure on educators. The muckrakers, who had begun detailing business abuses, also targeted public schooling. Alarmed readers were regaled with stories of waste and mismanagement and told that school reform, usually in the form of scientific management, was urgently needed.[13]

Teacher Accountability

The cult of efficiency also included teacher accountability. In 1913 the *American School Board Journal* reported that administrators in large cities were "almost without exception" working out "elaborate

plans for rating the work of instructors."[14] In many cases the evaluations were extended to include all school personnel, even janitors.[15] Callahan, though, reports that the difficulties of including the full range of relevant social, economic, and educational factors often led to actually assessing teachers on general impressions.[16]

One particularly troubling development was that of rating teachers by the percentage of children promoted. The idea was to save money by encouraging teachers to pass students who would otherwise repeat the grade.[17] Here, as in so many other cases, cost accounting trumped instruction.

Nor is there much evidence that the "experts" designing the ratings ever considered the additional expectations that underpaid, harried teachers faced: not only to successfully instruct, but also to comfort the afflicted; inspire the defeated; suppress bullies; correct disruptive behavior; observe the children for signs of abuse or neglect; instill a love of learning, patriotism, good citizenship, sportsmanship, and fair play; check heads for lice; teach students manners; and cope with kids (and parents) who spoke little or no English. What's more, educators were to do all that with nothing more than some chalk, a blackboard, a bulletin board, and a few books.

Teacher ratings, based largely on student test scores, are also a key feature of the Obama administration's $5 billion Race to the Top. This time, however, the power and financial resources of the federal government are being employed to make sure it happens, despite strong warnings from the American Statistical Association about using student test scores to measure teacher quality.[18]

School Quality Surveys

"Scientific management" also required school-quality surveys similar in intent to the school ratings required by both No Child Left Behind and Race to the Top. Focusing narrowly on price and product while uncritically embracing business-style management, the surveys provoked prestigious opposition.

John Dewey, for instance, strongly opposed applying business procedures and industrial values to schooling. He recognized the power and place of genuine science in education but repeatedly criticized the era's "scientific" management as oversimplified and un-scientific.[19]

Dewey wryly observed that most "scientific" initiatives were really the same old education masquerading as science. Dewey also

charged that testing, although potentially valuable, was being put to exactly the wrong purpose. Instead of being used to gain a better understanding of children, it was being misused to classify and standardize them.[20] Dewey's concerns, along with those of other prominent critics, were largely ignored.

The Book's Strengths and Weaknesses

A salient strength of *Education and the Cult of Efficiency* is its contemporary relevance. Time and again we see connections to contemporary events. Another strength is the care and industry that went into its writing. It was painstakingly researched, albeit somewhat one-dimensionally.

One weakness of the book is its failure to grant full consideration to the inevitability of spending limits. On the other hand, the management methods Callahan describes were so focused on cost reduction that they often proved demoralizing, heartless, and harmful.

The book also neglects to consider how the dominantly female composition of the teaching force influenced the happenings described. During the book's 1900–1930 time frame, female teachers outnumbered men some 5 to 1. That disproportion, combined with the overriding sexism of the age, surely encouraged the dismissive view of teachers described by Callahan. (Since 76 percent of the current public school teaching force is female, perhaps male chauvinist reformers still shrug off teacher knowledge for the same reason.)

Conclusions

Since the publication of *A Nation at Risk* in 1983, school reformers have re-embraced the notion of the school as factory—not the well-run modern factory prescribed by the famed management expert W. Edwards Demming, but an old-fashioned, top-down, condescending despotism that is inefficient at everything except alienating those actually doing the work.[21]

Absent from the contemporary reform agenda are concerns about tradition, pride of work, personal happiness, life fulfillment, depth of character, abiding values, group membership, and "proper" behavior—all manifestations of Callahan's largely ignored ideal of the school as temple.

Similarly absent is a focus on democratic decision-making, individual differences, concern for others, civility, and willingness to

compromise—all key elements of the school as town meeting (and of a functioning democracy).

The ideal schooling types of temple and town meeting are left to elite, government-test-exempt private schools—the schools that would-be reformers' loved ones often attend. Only the children of less financially fortunate people end up in the school as factory.

Today's school reformers are still wasting educators' invaluable time and energy—not on the cost accounting of the early twentieth century but on over-emphasized (and misused) high stakes test scores. That misguided emphasis distracts, disheartens, and demoralizes educators, misdirects scarce resources, and shortchanges students. Present-day reformers, like their predecessors, proceed with imperious disregard for teacher knowledge and experience. Today, however, they also have the power and financial resources of the federal government behind them.

In the end it is hard to grant the policies of contemporary school reformers much respect. If they really want to improve public education rather than posture and play politics, they would stop bullying educators and start working with them. Few worthwhile school reforms will take place without teacher trust and cooperation. Most public school teachers already receive far less credit than they deserve. Continued prodding will only result in further demoralization and a more-rapid reduction in the already-diminished desirability of teaching as a career.

Real reformers would also address the ills that are destroying educational opportunity at its very roots. They include but are not limited to: indefensibly unequal school funding; lax teacher preparation; the cancerous growth of concentrated poverty; the fact that 2.7 million American children have an incarcerated parent and the accelerating erosion of America's middle class, whose children provide the core of public school successes.[22]

This essay is based on a longer review that will be published in Joseph L. DeVitis, ed., Popular Educational Classics: A Reader *(New York: Peter Lang Publishing, 2016).*

Notes

1. Private schooling has generally escaped these testing requirements. For instance, in Pennsylvania private school students are exempted from the recently imposed Keystone tests that students must pass to qualify for a diploma.

2. Peter Williams, *Philadelphia: The World War I Years* (Charleston, S.C.: Arcadia Publishing, 2013), 27.
3. Raymond Callahan, *Education and the Cult of Efficiency* (Chicago and London: University of Chicago Press, 1962), 14–15.
4. President Coolidge summed up the spirit of the times this way: "The chief business of the American people is business."
5. Meanwhile, elite private schools avoided industrialization, which they still do today.
6. Formal training of school administrators was just getting under way.
7. Callahan, *Education*, 235.
8. Ibid., 237.
9. Quoted in ibid., 11.
10. Maxwell, quoted in ibid., 122.
11. This contemporary development had its origin in the 1980s with the publication of *A Nation at Risk* (Washington, D.C.: National Commission on Excellence in Education, 1983).
12. Eric Devlin, "Board Addresses Decrease in Test Scores," *Ambler Gazette*, Sunday, December 14, 2014, 1.
13. Callahan, *Education*, 245.
14. Ibid., 104.
15. Ibid., 108.
16. Ibid., 105.
17. Ibid., 169.
18. ASA Statement on Using Value-Added Models for Educational Assessment, American Statistical Association, April 8, 2014, http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf
19. Callahan, *Education*, 124–25.
20. Ibid.
21. Here, from Brainy Quote (http://www.brainyquote.com/quotes/authors/w/w_edwards_deming.html) are several Deming observations that contrast vividly with recent school reforms: "Hold everyone accountable? Ridiculous!" "All anyone asks is to work with pride." "Whenever there is fear you will get wrong figures."
22. Katie Riley, "Sesame Street Reaches Out to 2.7 Million American Children with an Incarcerated Parent," Factank: New In The Numbers, Pew Research Center, June 21, 2013, http://www.pewresearch.org/fact-tank/2013/06/21/sesame-street-reaches-out-to-2-7-million-american-children-with-an-incarcerated-parent/

BOOK REVIEW

Anya Kamenetz

The Test

Why Our Schools Are Obsessed with Standardized Testing—but You Don't Have to Be

Public Affairs, 2015

reviewed by Richard P. Phelps

PERHAPS IT IS BECAUSE I avoid most tabloid journalism that I found journalist Anya Kamenetz's loose-cannon introduction to *The Test: Why Our Schools Are Obsessed with Standardized Testing—but You Don't Have to Be* so jarring. In the space of seven pages, she employs the pejoratives "test obsession," "test score obsession," "testing obsession," "insidious . . . test creep," "testing mania," "endless measurement," "testing arms race," "high-stakes madness," "obsession with metrics," and "test-obsessed culture."

Those unmeasured words fit tightly alongside assertions that education testing, standardized testing, or high-stakes testing is responsible for numerous harms, ranging from stomachaches, stunted spirits, family stress, "undermined" schools, demoralized teachers, and paralyzed public debate, to the Great Recession (pp. 1, 6, 7), which was initially sparked by problems with mortgage-backed financial securities (and parents choosing home locations in part based on school

average test scores). Oh, and tests are "gutting our country's future competitiveness," too (p. 1).

Kamenetz tells us, "[T]here's lots of evidence that these tests are doing harm, and very little in their favor" (p. 13), but she has made almost no effort to search for counter-evidence.[1] Her sources for information in the relevant research literature include some of the country's most prolific proponents of her claim.[2] Ergo, why bother to look for it?

Had a journalist covering the legendary Hatfield-McCoy feud talked only to Hatfields, one might expect a surplus of reportage favoring the Hatfields over the McCoys and a deficit of reportage favoring the McCoys over the Hatfields.

Looking at tests from any angle, Kamenetz sees only evil. Tests are bad because they were used to enforce Jim Crow discrimination (p. 63). Tests are bad because some of the first scientists to use intelligence tests were racists (pp. 40–43).

Tests are bad because they employ the statistical tools of latent trait theory and factor analysis (the same tools, incidentally, currently used by tens of thousands of social scientists worldwide), but the "eminent paleontologist" Stephen J. Gould doesn't like them (pp. 46–48). (Gould argued that if you cannot measure something directly, it doesn't really exist.) And by the way, did you know that some of the early-twentieth-century scientists of intelligence testing were racists (pp. 48–57)?

Tests are bad because of Campbell's Law: "When a measure becomes a target, it ceases to be a good measure" (p. 5). Such a criticism, if valid, could be used to condemn any measure used to evaluate anything in society's realm. Forget health and medical studies, sports statistics, Department of Agriculture food-monitoring protocols, *Consumer Reports* ratings, Angie's List, the Food and Drug Administration. None of them are "good measures" because they are all targets.

Tests are bad because they are "controlled by a handful of companies" (pp. 5, 81); testing companies "determine . . . the quality of teachers' performance" (p. 20); and "tests shift control and authority into the hands of the unregulated testing industry" (p. 75). Criticisms such as Kamenetz's, if valid, could justify nationalizing all businesses in industries with high-scale economies (e.g., there are only four big national wireless telephone companies, so perhaps the federal

government should take over) and outlaw all government contracting. Most of our country's roads and bridges, for example, are built by private construction firms under contract to local, state, and national government agencies to the latters' specifications, just like most standardized tests: but who believes that those firms control our roads?

Kamenetz swallows anti-testing dogma whole. She claims that multiple-choice items can test only recall and basic skills (p. 35), that students learn nothing while they are taking tests (p. 15), and that U.S. students are tested more than any others (pp. 15–17, 75). That's true if you make calculations the way her information sources do—counting at minimum an entire class period for each test administration, even a one-minute DIBELS test; counting all students in all a school's grades as taking a test whenever any students in any grade are taking a test; counting all subtests in the United States independently (e.g., making each ACT count as five because it has five subtests) but only the whole tests in other countries; etc.

Standardized testing absorbs way too much money and time, according to Kamenetz. Later in the book, however, she recommends an alternative education universe of fuzzy assessments that, if enacted, would absorb far more time and money.

What are the author's solutions to the insidious, obsessive mania of testing? She engages in some Rousseauean fantasizing: all schools should be like her daughter's happy pre-school, where each student learns at his or her own pace (pp. 3–4) and the school's job is "customizing learning to each student" (p. 8).

Some of the book's latter half addresses "innovative" (of course) solutions that are not quite as innovative as National Public Radio's "lead digital education reporter" seems to believe. True, some interesting recent technologies suffuse Kamenetz's recommendations. But even jazzing up the context, format, and delivery mechanisms with the latest whiz-bang gizmos will not eliminate the problems inherent in her old-new solutions: performance testing, simulations, demonstrations, portfolios, and the like. Like so many Common Core Standards boosters advocating the same "innovations," she seems unaware that they have been tried in the past, with disastrous results.[3]

Lacking personal acquaintance with Ms. Kamenetz, I must assume the sincerity of her beliefs and her decisions about what to write. Nonetheless, if she had naively allowed herself to be wholly misled by those with a vested interest in education-establishment doctrine, the result would have been no different.

The book is basically a slapped-together rant, unworthy of an established journalist. Ironically, however, I agree with Kamenetz on many issues. Like her, I do not much like the assessment components of the old No Child Left Behind Act or the new Common Core Standards. (My solution would be to repeal both programs, not eliminate standardized testing.) Like her, I oppose the U.S. practice of relying on single proficiency standard for all students (pp. 5, 36). (My solution would be to employ multiple targets, as most countries do. Kamenetz would dump the tests.)

Again like Kamenetz, I believe it unproductive to devote more than a smidgen of time (at most half a day) to test preparation, with test forms and item formats, that is separate from subject-matter learning. And like her (p. 194), I am convinced that most test prep does more harm than good. Kamenetz, however, blames the tests and the testing companies for the abomination; in fact, the testing companies prominently and frequently discourage the practice. The advocates of test prep are actually the same testing opponents Kamenetz has chosen to trust. Trying to establish the legitimacy of non-subject-matter-related test preparation serves the argument of testing opponents because, if true, it would expose tests as invalid measurement instruments that can be gamed with tricks.

Like Kamenetz, I oppose firing teachers based on student test scores, as current value-added measurement (VAM) systems do, while the students suffer no consequences. I believe the VAM systems wrong because they rely on too-few data points and because student effort in such conditions is unreliable, varying by age, gender, socioeconomic level, and more. I would eliminate VAM programs, or drastically revise them; Kamenetz, by contrast, would eliminate the tests.

Like Kamenetz, I believe that educators' cheating on tests is unacceptable, far more common than is publicly known, and should be stopped. I say, stop the cheating. She says, dump the tests.

It defies common sense to have teachers administering high-stakes tests in their own classrooms. Rotating test-administration assignments so that teachers do not proctor their own students is not particularly difficult, nor is rotating assignments further so that every testing room is proctored by at least two adults. So why aren't these and other remarkably simple fixes for test-security problems implemented? (Note that the education professionals responsible for managing test administrations are often the same individuals who complain that testing is impossibly unfair.)

The sensible solution is to take test-administration management out of the hands of those who may welcome test-administration fiascos and to hire independent professionals with no conflict of interest. Like many education insiders, though, Kamenetz would ban the testing and thereby reward those who have mismanaged test administrations, sometimes deliberately, by giving them a vacation from reliable external evaluation.

If Kamenetz were correct on all these issues—that the testing is the problem in each case—shouldn't we also eliminate examinations for doctors, lawyers, nurses, and pharmacists (many of which rely on the multiple-choice format, by the way)?

Our country has a problem. More than in most other countries, our public education system is independent, self-contained, and self-renewing. The education professionals who staff school districts make the hiring, purchasing, and school catchment-area boundary-line decisions. School district boundaries often differ from those of other governmental jurisdictions, confusing the electorate. In many jurisdictions, school officials set the dates for votes on bond issues or school board elections and can do so to their advantage. Those school officials are trained, and socialized, in graduate schools of education.

A half century ago, many faculty members in graduate schools of education may have received their own professional training in such core disciplines as psychology, sociology, or business management. Today, by contrast, most members of education school faculties are themselves education school graduates, socialized in the prevailing culture. The dominant expertise in schools of education can maintain that dominance with faculties that support the conventional wisdom and deny tenure to those who stray. The dominant expertise in education journals can control education knowledge when article submissions with agreeable results are accepted and those without are rejected.

Even doctoral training programs in testing and measurement now reside mainly in schools of education, inside the same cultural cocoon.

Standardized testing is one of the few remaining independent tools American society has for holding education professionals accountable to the public interest, rather than their own. Without valid, reliable, objective external measurement, education professionals can do largely what they please inside our schools, with our

children and our money. When educators are the only arbiters of the quality of their own work, they tend to rate it consistently well.

A substantial portion of *The Test*'s girth is filled with complaints that tests fail to measure most of what students are supposed to or should learn: "It's math and reading skills, history and science facts that kids are tested and graded on. Emotional, social, moral, spiritual, creative, and physical development all become marginal. . . ." (p. 4). Kamenetz quotes Daniel Koretz: "These tests can measure only a subset of the goals of education" (p. 14). Several other testing critics are cited making similar claims.

Yet standards-based tests are developed through a multi-year process that enlists scores of legislators, parents, teachers, and administrators to serve on a variety of decision-making committees. The citizens of a jurisdiction and their representatives choose the content of standards-based tests. They could choose content that Kamenetz and the critics she cites prefer, but they don't.

If the critics are unhappy with test content, they should take their case to the appropriate decision-makers, voice their complaints at tedious standards commission hearings, and contribute their time to the rather monotonous work of test-framework review committees. I sense that such patient effort holds little interest for them; they would instead prefer to wield all decision-making power *ex cathedra*, to do as they think best for us.

Moreover, I find some of the testing critics' assertions about what *should be* studied and tested fraught with dangers. Public schools should teach our children emotions, morals, and spirituality?

Likely that prospect would concern most parents, too. But many parents' first reaction to a proposal allowing schools to teach children *everything* might instead be something like: first show us that you can teach our children to read, write, and compute: *then* we can discuss further responsibilities.

So long as education insiders insist that we must hand over our money and children and leave them alone to determine—and evaluate—what they do with both, calls for "imploding" the public education system will only grow louder, as they should.

It is bad enough that so many education professors write propaganda, call it research, and deliberately mislead journalists by declaring the absence of countervailing research and researchers. Researchers confident in their arguments and evidence should be

unafraid to face opponents and opposing ideas. The researchers Kamenetz trusts do all they can to deny dissenters a hearing.

In addition to testing, another potential independent tool for holding education professionals accountable could be an active, skeptical, and inquiring press knowledgeable about education issues and conflicts of interests. Other countries have it. Why are so many U.S. education reporters gullible sycophants?

Notes

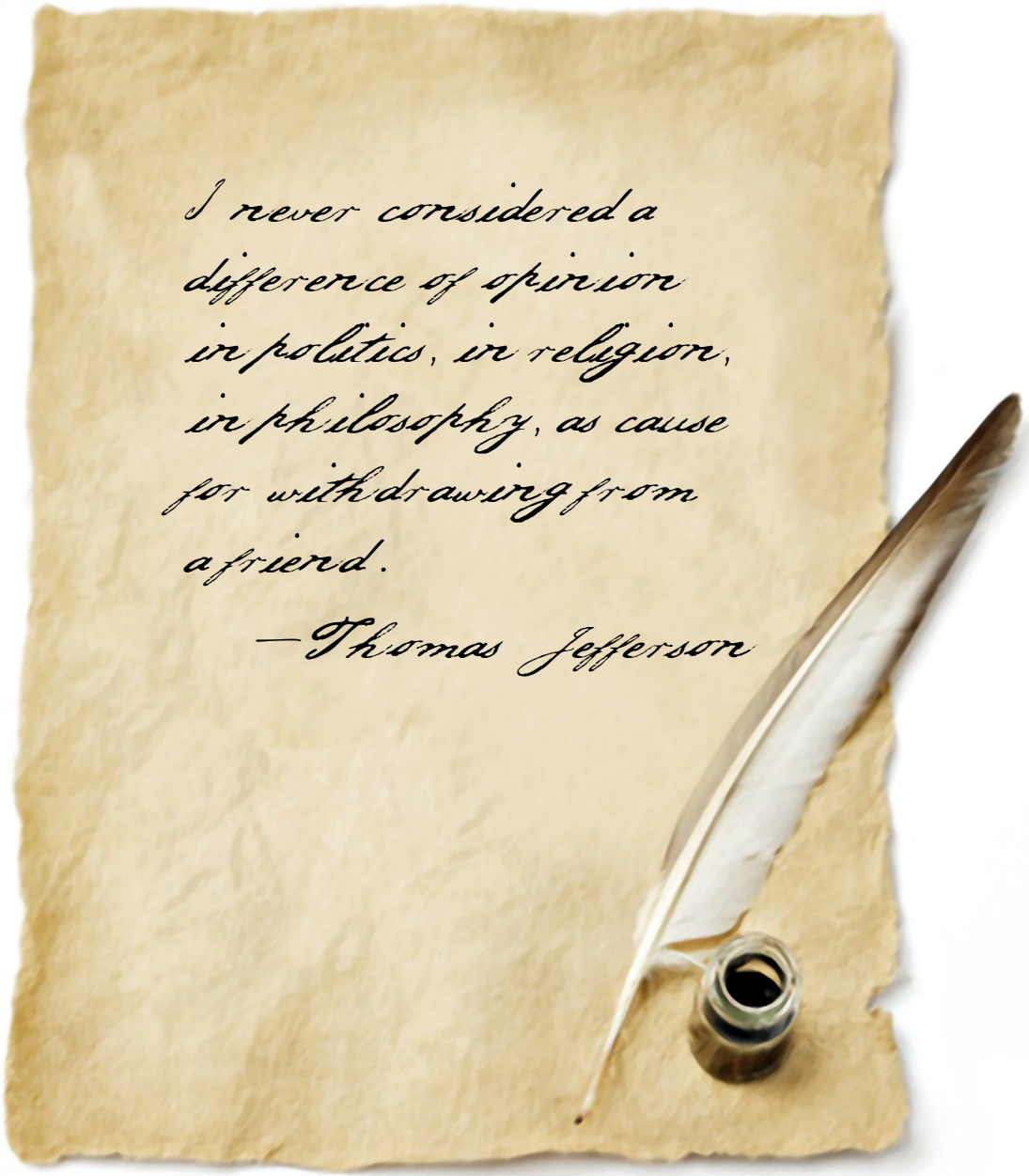
1. Kamenetz did speak with Samuel Casey Carter, the author of *No Excuses: Lessons from 21 High-Performing High-Poverty Schools* (2000) (pp. 81–84), but she chides him for recommending frequent testing without "framing . . . the racist origins of standardized testing." Kamenetz suggests that test scores are almost completely determined by household wealth and dismisses Carter's explanations as a "mishmash of anecdotal evidence and conservative faith."
2. Those sources are Daniel Koretz, Brian Jacob, and the "FairTest" crew. In fact, an enormous research literature revealing large benefits from standardized, high-stakes, and frequent education testing spans a century (Brown, Roediger, and McDaniel, 2014; Larsen and Butler, 2013; Phelps, 2012).
3. The 1990s witnessed the chaos of the New Standards Project, MSPAP (Maryland), CLAS (California) and KIRIS (Kentucky), dysfunctional programs that, when implemented, were overwhelmingly rejected by citizens, politicians, and measurement professionals alike. (Incidentally, some of the same masterminds behind those projects have resurfaced as lead writers for the Common Core Standards.)

References

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Cambridge, Mass.: Belknap Press.

Larsen, D. P., & Butler, A. C. (2013). Test-enhanced learning. In K. Walsh (Ed.), *Oxford textbook of medical education* (pp. 443–452). Oxford: Oxford University Press. http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2008.03124.x/full

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. International Journal of Testing, 12(1), 21–43. http://www.tandfonline.com/doi/abs/10.1080/15305058.2011.602920


I never considered a
difference of opinion
in politics, in religion,
in philosophy, as cause
for withdrawing from
a friend.
—Thomas Jefferson