CHAPTER 3

BRIDGING ACROSS THE MILE-WIDE AND MILE-DEEP CHASM

Living and Coping with Standards-Based Reform in Science Education

Junlei Li
University of Pittsburgh

INTRODUCTION

Nearly ten years ago, the phrase "a mile wide and an inch deep" entered the public and academic discourse on science education in the U.S. It was the stinging indictment from the Third International Mathematics and Science Study (TIMSS) of the expansive, unfocused, and purportedly less rigorous science and mathematics curriculum in the U.S. as compared to that of other nations (Schmidt, Mcknight, & Raizen, 1997). The TIMSS report, titled, "A Splintered Vision," was wedged between the publications of several important educational policy documents, including the Professional Standards for Mathematics Education (National Council of Teachers of Mathematics Standards, 1991), Benchmarks for Science Literacy (American
Association for the Advancement of Science [AAAS], 1990, 1993), and National Science Education Standards [NSES] (National Research Council [NRC], 1996). Jointly, these documents ushered in the latest era of standards-based educational reform across U.S. Over-quoted and under-debated, the letter and the spirit of these documents engendered countless versions of copycat standards, how-to guides, and political decrees in the areas of science and mathematics, at state, district, even school and grade levels.

With ten years gone by and the proliferation of “standards” (if not the spirit, at least the form), what has happened to the “mile wide” and “inch deep” problem? In this chapter, I argue that, instead of making the difficult but necessary choices to prioritize the divergent objectives of science education, the present standards-based science education system sidesteps the fundamental depth versus breadth dilemma and simplistically demands universal excellence in content areas miles wide and miles deep (thus the title for this chapter). The resulting chasm between what is demanded and what can be achieved, by schools, teachers, and students alike, threatens to engulf the idealistic intent of the standards-based reform movement and obscure the ever pressing needs to engage underrepresented student populations and achieve educational equity in science.

Yet the situation is not without hope or remedy. Through this writing, I hope to not only build upon, but also move beyond the existing theoretical and ideological criticism of standards, ably presented by a minority of scholars amidst the chorus advocating for science standards and accountability. I suggest practical coping strategies grounded by the acceptance of, rather than the resistance to, the presence of standards and accountability in our science education system. While academic scholars have a duty to continue challenging the status quo as ideals, theories, and evidence so compel them, applied researchers need to recognize what cannot be easily changed (at least in near term) and do their best to help students, teachers, and school to live with the system not as it should be, but as it is.

**RESEARCH CONTEXT**

Before proceeding, let me briefly describe the research context that gave rise to the ideas presented in this chapter. I did not begin as a critic of the standards-based science education reform. Quite the opposite, as a research psychologist specialized in cognition and instruction, I set out to develop and apply research-inspired and evidence-based practices for the express purpose to improving urban school science achievement. Funded through a three-year grant from the Institute of Education Sciences, U.S. Department of Education, I believed, or rather, assumed that standards-based reform was the way to lift urban schools out of the achievement gap.
I believed that inquiry learning as defined by various science standards was both the means and the ends of such an endeavor. The standards' focus on inquiry matched the interest of a typical research psychologist. The bulk of the experimental and psychological research in the area of children's scientific knowledge focus on the development of reasoning and conceptual learning, rather than the amassing a vast base of factual knowledge. I believe my initial assumptions are both representative of traditional academic researchers with applied interests (minus in-person field experience), particularly in the context of research funding shifts to engage bench scientists in educational research.

Despite such naivety, I fortunately had the sense to design the research study using non-traditional methods. Instead of developing "research-based" instructional units and offering teachers "professional development" to implement in classrooms, I adopted the co-teaching methodology (Roth & Tobin, 2002) and placed myself as a co-teacher in the urban classrooms. Both the feasibility of inquiry-based science teaching and the applicability of research knowledge were put to harsh tests as I attempted to teach in urban classrooms.

One small classroom episode may serve to illustrate the gradual process by which my ideologies of standards-based reform and research-based intervention crashed to the ground:

Determined to help a classroom of 7th grade students (all African American, nearly all of whom qualified for free and reduced lunches) to develop deep and conceptual understanding of earth's rotation along its axis and its orbit around the sun, my co-teacher (their regular science teachers) and I had been pushing them for two weeks using a variety of strategies including discussions, physical modeling of earth's movement, and drawing. The class's progress had been stalled on the question of why the length of day and night varied depending on earth's position in orbit. The students' interest in the subject, at times high, was definitely waning after two weeks' push towards "mastery."

In an attempt to re-energize the class, I asked the students to imagine the life of farmers in order to lead the discussion towards the necessity of mastering nature's rhythm in shifting daylight hours. For me, this was a textbook example of situating abstract scientific concepts in concrete and novel contexts. As I excitedly rambled on about how different life would have been for the farmers if they could not predict seasonal change, a female student raised her hand. As politely as she could (though it was clear that her patience was being tested), she voiced her objection, "I know all that you are saying. But we are not farmers. I've never been to a farm. I don't know anybody who lives on a farm. We've been talking about the same thing for two weeks. Can we talk about something else?" Stumped, I stammered to continue on with the lesson, but felt deflated.
A series of bubble bursts like this forced me to re-examine my own assumptions about inquiry, deep understanding, robust learning, mastery, high expectations, and many other presumably desirable and achievable ideas. Only then did I seek out literature that critiqued the standards-based science education reform. Informed by some of the critical perspectives, I found little in ways to help my co-teachers, who live within the system as compared to my daily "guest appearances" in the classrooms, to cope with the concrete demands of the standards-based reform and the renewed push for test-based accountability (Anderson, 2004; Bauer, 1992; Donmoyer, 1995; Shamos, 1995; Wolk, 1999, 2004). Thereafter, I focused the research effort on seeking practical strategies, rather than theoretical or ideological hypotheses, to "cope" with the standards and accountability reform. Such a quest also led to further questioning, both theoretically and empirically, of the assumptions upon which the standards had been advocated.

In providing this background, I hope readers, whether advocates or critics of the science standards, may find some resonance in the process, even if not the product, of my conversion from a highly specialized laboratory scientist to a field researcher. While I will attempt here to synthesize the evidence gathered in my own research context with the broader literature, the discussion in this chapter is far from being a comprehensive literature review of the standards debate. Having made clear the gaps and biases in my own knowledge, then and now, I present what I had observed as persisting gaps in standards-based science education amongst rhetoric and data, policy and practice, and theories and reality.

**SCIENCE EDUCATION CRISIS—PUBLIC RHETORIC VERSUS DATA**

While the scholarly critique of science standards (Anderson & Helms, 2001; Hewson, Kahle, Scantlebury, & Davies, 2001; Settlage & Meadows, 2002) were a minority amongst science education literature (much of which had been about how and why to implement standards rather than the validity and feasibility of standards themselves), the public discourse of the state of U.S. science education has been strikingly one sided—"Science education is and has been a national crisis!"

To trace the crisis rhetoric to its origin, we can go at least as far back as the Sputnik era, when our fear over the Cold War arms race with the Soviets give birth to the associated anxiety over inferior science education. The still often-quoted "A Nation at Risk" report (National Commission on Excellence in Education, 1983) lamented a "rising tide of mediocrity" and equated the performance of our nation's schools to self-waged "act of war." The rhetoric today continues to be as alarmist and militaristic. In a report titled "Road Map for National Security: Imperative for Change," issued by
the U.S. Commission on National Security/21st Century, science education was listed as among the threats to our national security. In addition to linking K–12 science performance to the broad issue of human capital deficit, the report declared that the failure to manage science education constituted, in part, a threat “second only to a weapon of mass destruction detonating in an American city.” The report was issued more than six months before September 11th, 2001. The commission, headed by well-respected leaders aiming for bipartisan consensus, was not alone in its dire pronouncement. Two months following 9/11, reacting to the release of yet another round of dismal results from the National Assessment of Educational Progress (NAEP), both the U.S. Secretary of Education and the National Science Teachers Association cited the report’s warning and the quote “second only to a WMD” to add weight to the NAEP data, which predictably provided ample evidence for the crisis label.

NAEP, along with TIMSS, have been the standard-bearer for bad news for U.S. science and mathematics education. Whatever woes NAEP reveal is said to be confirmed by TIMSS, and vice versa. The headlines and speeches following each round of data release from NAEP and TIMSS have become routinely grave. Such consistency engenders two popular and robust beliefs:

1. U.S. lagged significantly behind international competitors (e.g., Japan, Korea) in science achievement.

2. An overwhelming majority of U.S. students, K–12, failed to reach even the most basic standard of proficiency.

Together, these beliefs implicate K–12 science performance as a key culprit in our losing competitive edge (jobs, market share, etc.) in the global market. Judging by coverage in public speeches or media reporting, other issues, such as the science achievement gap across class or race, do exist, but pale in comparison with the larger problem of national competitiveness and security. The “crisis” in science education has become such an irrefutable truism that I included the phrase “rising tide of mediocrity” and the conclusions of NAEP and TIMSS all in the opening paragraph of my grant proposal to the U.S. Department of Education. After spending just one year in the urban schools, I symbolically removed these references from the text with much chagrin. What was most embarrassing for me about using NAEP and TIMSS results to jump on the science crisis bandwagon was that I had accepted the popularly held conclusions without ever examining the data.

By examining the “data,” I do not mean to comb through the massive database released by both studies for statistical mining. I refer merely to a less presumptuous read of the data charts printed in the published reports themselves. On the question of whether U.S. severely lags countries like
Japan, the consistent result has been that our fourth graders consistently rank among the top nations tested even while our older students (8th grade and 12th grade) fall statistically below the top band of nations. But the media reporting and public statements citing TIMSS results neglect to go beyond the "averages that hide the true extremes" (Berliner, Op-ed by the same title, Washington Post, January 28, 2001). TIMSS results reveal that there are school districts in the U.S. that compete quite well against the international counterparts, along with many sites with underserved students who perform significantly worse. Berliner argued that the science education problem was not that U.S. could not educate scientifically competitive students, but simply that we were not doing so equally for all of our children. Rebuttals for such criticism invoke the apples vs. oranges analogy and argue that it was not fair to compare the best of U.S. to the average of countries like Japan (e.g., Japan would also have a wide spectrum of high vs. low performing sites) and that our problem as a whole was still dire.

What if we do take into consideration the spread within the relatively homogenous (economically and ethnically) international competitors like Japan or South Korea? Figure 3.1 below compares across these three countries using TIMSS 2003 8th grade science scores, disaggregated by the percentage of students in schools that were considered economically disadvantaged.

Figure 3.1. Comparing U.S., Japan, and South Korea TIMSS Score, 8th Grade, 2003 (by % of economically disadvantaged students in schools).
Clearly, even when we compare apples to apples (i.e., students in similar economic situations), U.S. 8th graders hold their own internationally so long as they are privileged enough to attend schools with less than 25% economically disadvantaged students. The pattern changes direction and worsens significantly for the U.S. as the percentage of economically disadvantaged students in schools increase. This comparative relationship is similar to other sites such as Taipei, Hong Kong, and Singapore. But because these sites are on the geographical scale of cities, they cannot usefully included in comparisons of nations.

Conspicuously missing from the above chart is the data point for Japan in the last column (schools with >50% low-SES students). The official TIMSS explanation was "(sampling in Japan for that category) did not satisfy guidelines for sample participation rates." Given that TIMSS always strive to sample representative populations in each country, one can safely conclude that it had been difficult to identify schools that fit such description in Japan. If we can trust TIMSS sampling as representative of the nations’ student population, Figure 3.2 compares the percentage of students sampled in each of the above economic categories revealed why reporting the average TIMSS scores hide the extremes (at least within the U.S.).

Compared with Japan and South Korea, the schools (and thereby the students) are much more segregated by economic status in the U.S. The achievement gap among these economically segregated schools within U.S.

---

![Figure 3.2. Comparing U.S., Japan, and South Korea Participant Samples, 8th Grade, 2003](image-url)
is reflected in the overall lower average score in the international context, even while our relatively affluent students are performing just as well as their counterparts internationally. Similar pattern exists if the average score is disaggregated by race within the U.S. For example, of the same score in question (TIMSS 2003 8th Grade), the average score for U.S. Whites is exactly the same as that of Japan while the average scores for African and Hispanic/Latino Americans lag farther behind with countries like Iran and Bulgaria. Note that the comparison here refers to Whites across all social economic status sampled. It is a reasonable comparison given that Japanese students, relatively to U.S. standards, can be treated as an ethnically homogenous population much as U.S. Whites are.

In short, the average science education provided by the U.S. to its White students or its moderately well-off students regardless of race (schools with <25% low-SES students are hardly elite prep schools) are as good as that provided to students in Japan or South Korea.

But how does one explain that NAEP, dubbed “The Nation’s Report Card,” consistently fails the majority of U.S. K–12 students in science, regardless of race or economic status? From 1995 to 2005, across three separate NAEP science tests, the highest percentage of proficiency achieved by any one ethnic group was at 40%, with White and Asian/Pacific Islanders leading the other ethnic groups. Among those who are not eligible to receive free school lunches, the highest percentage of proficiency also happened to be 40%. These domestic results certainly sound alarming independent of how one might interpret the TIMSS results.

But one cannot interpret NAEP results “independently” from TIMSS results. TIMSS, by using international benchmarks, provides perhaps the most reliable relative standard of what is “good enough” or “world-class” (i.e., if a nation is near the top of the world). NAEP, with its rather complex and subjective process of defining proficiency, has an absolute “good enough” bar set by more by value judgments than benchmarking data, albeit through a consensus-building process by committees of experts.

Within my limited knowledge, I am unaware of any objective benchmarking done for NAEP tests to set achievement criteria (i.e., basic, proficient, advanced). How do we know if NAEP had set its proficiency bar too high, or too low, or just right? Ideally, we would need to take the NAEP tests to countries like Japan and South Korea. If the vast majority (much more than 40%) of Japanese or Korean students tested proficient in science by NAEP standards, then we would have reason to worry.

Of course, such benchmarking had not been done. However, within the data provided by NAEP and TIMSS, one could estimate ballpark figures. Given that moderately affluent students (those attending schools with <25% low-SES students) in U.S., Japan, and South Korea all performed comparably on TIMSS, we can make the assumptions that Japa-
Chinese and South Korean students in this category would also do comparably on NAEP and those below this category would do correspondingly worse. Given that the bulk of Japanese and South Korean students fall within the <25% economical category, we can use the U.S. NAEP performance in the <25% category as a ceiling for these Japanese and South Korean students might do had NAEP been administered there. In the NAEP dataset, the definition of a school’s low-SES enrollment was percent eligible for free and reduced lunch. Based on the 2005 NAEP data, only 39% of the students coming from schools with less than 25% low-SES enrollment were considered proficient in science. Thus, the estimated-NAEP-proficiency rate for both Japan and Korea would also have to be less than 39%. If these assumptions are true even within the ballpark, then headlines for NAEP science results might as well read, “Most Japanese/Korea Students found to be below proficiency in science, by U.S. tests.” That would have been an attention grabber.

Given the vastness of NAEP and TIMSS dataset, one may criticize the above analyses as data mining along the old adage “you can use data to prove anything.” The analyses is in fact selective, but for the purpose to illustrate a central concern—the reporting of NAEP and TIMSS data, particularly in political statements or media, has been so one-sided that it implies that the data is unequivocally in support of declaring our science education a crisis in terms of international competitiveness or national security. The analyses presented here at least raise some data-backed counter arguments. What is unequivocal across TIMSS and NAEP is not in terms of global markets or WMD, but of equity among economic classes and races in science education within U.S.

Some would argue that we have two urgent problems and both need attention. But judging from the NAEP and TIMSS data, our first priority ought to be closing the science gap within our borders rather than fretting about the purported gap with students overseas. Return again to the 8th grade scores in TIMSS: the score gaps within U.S. amongst races or economical classes are three times the size of the score gaps between U.S. and leading countries like Japan and South Korea. As much as we fret about our lagging behind Asian competitors, should we not proportionally be much more concerned (three times, to be exact) about the achievement gap within? Such proportional reasoning is not reflected in reactions to each round of NAEP and TIMSS results. The bulk of the emphasis is placed upon minor year-to-year score fluctuations to make an urgent case for crisis, even when such fluctuations are miniscule in comparison to the persisting achievement gap in every assessment.

From a political perspective, an argument can be made that so long as science education is getting public attention (and therefore increased funding), it does not matter which particular issue (gap overseas or gap
between races and SES classes) generated the attention, ultimately it would benefit science in general and all students who wish to learn science. Such argument may underlie why NSTA, an advocacy group of teachers, would join the chorus of federal officials to describe our science education problem to be “second only to a WMD.” But as I will argue throughout the rest of this chapter, misplacing focus from a real gap to a purported gap has far reaching and negative consequences beyond simply getting attention and funding for science education.

SCIENCE STANDARDS—THEORY VERSUS IMPLEMENTATION

It is within the above-described climate of public discourse that science standards were born. Both AAAS and NRC developed the early standards, with National Science Education Standards (NSES) (NRC, 1996) being more visible and cited version in the long run. Created as a political product through a complex process of consensus building (Collins, 1998), NSES advocated for scientific inquiry as both the means and ends of education while narrowing (at least as its intent) the vast science content to a set of core content standards. Though NSES suggested the use of other complementary methods along with inquiry, it made clear that inquiry was the spirit and emphasis of the NSES documents. This message has been carried through the many replicating (or duplicating) versions of state and school district standards and guidelines in the years following NSES, though not to the level of implementation.

NSES, by its very nature as a document of political consensus, allowed great flexibility in interpretation and enactment. This characteristic made the document very marketable and citable but at the cost of endangering the fidelity of implementation of the inquiry spirit of NSES. The state standards, modeling themselves onto the NSES framework, largely inherited this characteristic. Reviews of science standards across the states, regardless of ideological orientation, all consistently pointed to the lack of specificity and usability of the documents (e.g., Thomas B. Fordham Foundation [TBF], 2005). Consequently, vagueness permeated even as the phrase “standards-based” proliferated in the marketing of instructional materials.

It had been difficult to meander through the critiques of the standards document. On every issue, there are critics advocating for both sides. These disagreements are based on ideological, theoretical, and practical grounds. Here is just a brief list of the contentious issues, with representative sources. Some were concerned that the needed emphasis on inquiry and process had been diluted through the equal emphasis on content standards (e.g., Shamos, 1995), while others felt that the emphasis on inquiry was neither
justified by research evidence (e.g., Shiland, 1998) or by practical efficacy (e.g., TBF, 2005). Many were concerned that diversity issues were rendered "invisible" through the "one size fit all" science standards (e.g., Barton, 1998; Lee, 1999; Rodriguez, 1997; Vesilind & Jones, 1998), while others contend that too much emphasis has been given to issues of ethnicity, gender, and culture in a subject matter that was meant to objective and rigorous (e.g., TBF, 2005). Perhaps the most remarkable aspect of the ongoing debate was how impervious the standards are to criticisms. For example, despite the highly visible pronouncements by reports like "State of State Science Standards" (TBF, 2005), few states reacted constructively to the bad grades TBF gave to their science standards document. Compare the two press releases issued by Texas Education Agency (TEA) in response to two reports, five years apart, from the one and same Thomas B. Fordham Foundation. In 2000, Texas was one of only five states to be listed among TBF's "honor roll" for "solid academic standards" (TBF, 2000). The TEA press release (January 6, 2000) read, "This latest report from the Fordham Foundation is yet another sign that Texas public schools are on the right track." In 2005, TBF downgraded Texas' science standards to a dismal "F" grade, finding fault with the document's legalistic, broad, and breadth over depth approach. TEA press release (December 7, 2005) read, "A report issued today by the Thomas B. Fordham Foundation that purports to rate each state's science curriculum is more science fiction than science." Granted, the TBF reports had their own biases and limitations and many an involved educator may disagree with their findings. This episode illustrates that the resiliency of the standards is fostered through the blanket acceptance of praise and the blanket rejection of criticism by the very same governing body that both created and is now responsible for revising the science standards. Such defensive ownership of standards by states is by no means limited to the state of Texas. Tests performances are used to hold schools, teachers, and students accountable for standards, not the bodies that create and uphold the standards.

Amidst the conflicting critics of standards, some consensus does emerge even if the proposed solutions vary. One important consensus is concerning the breadth of topics included in state science standards. There is a growing recognition that the standards as they are implemented at state and district levels have greatly deviated from the intent of the movement, which was to push for depth of understanding by reducing needless breadth of coverage (Kesidou & Roseman, 2002). Instead, the standards documents across the country had become a kind of "anything goes" catch-all. To illustrate, from the Texas Assessment of Knowledge and Skills (TAKS) information booklet (TEA, 2004), there appeared an explanation regarding the distinction of the phrases "such as" and "including" as used in the science standards. For instance, "The student is expected to identify
patterns of change *such as* in weather, metamorphosis, and objects in the sky” or “The student is expected to classify matter based on its physical properties *including* magnetism, physical state, and the ability to conduct or insulate heat, electricity, and sound.” Here is the “difference” between the two phrases, as according to TAKS:

The term *such as* is used when the specific examples that follow it function only as representative illustrations that help define the expectation for teachers…. Teachers may choose to use them when teaching… but there is not requirement to do so. Other examples can be used in addition to those listed or as replacements for those listed.

The term *including* is used when the specific examples that follow it must be taught. However, other examples may also be used in conjunction with those listed.

While the intent of such distinction is to offer teacher’s flexibility and to clarify what is required, it is important to note that no example, as according to this distinction, can ever be *excluded* from the science curricula. That, in turn, implies that instruction and materials can indefinitely expand in size so long it covered the required components within its bulk. From the publishers’ point of view, there is every incentive to expand the size of coverage to meet the varying needs and specifications of state markets and very little constraint in trimming down. When Project 2061 from AAAS reviewed various existing instructional materials against its own criteria of scientific rigor, the results were best summarized by the title of one such review, “Heavy Books, Light on Learning” (AAAS, 1999; see also Stern & Ahlgren, 2002; Stern & Roseman, 2004).

There are deeper and systemic reasons as to why the standards movement had failed to slim down. The process by which standards are made is a political one. Science educators and scientists are well represented in the committees that craft standards. Within such diverse advocate groups for science education and science at large, there is no practical rule or theoretical principle by which one may determine “this is one topic too many” while there are plenty of rational arguments to be made for “here is one more.” The original AAAS and NRC standards, though well intended, have started this trend in the first place. Both organizations represent research scientists, who, when it comes to providing input to the standards, fall more easily into the advocacy category rather than the gatekeeper category. In subsequent iterations at the state or district level, the effort to define “science literacy” as science for all bordered making *all* of science for all. This is understandable given the broader political climate. Who can soundly argue to remove even a fraction of rigorous science content from existing standards when our country’s science education had
been in a crisis mode, threatening our economical competitiveness and national security?

The intent of the science standards, particularly with regards to inquiry, was to engage students in personally, socially, and scientifically meaningful investigations. The mile-wide, mile-deep standards threaten such intent by taking away the time that can be devoted to meaningful inquiry. The political discourse overly focused on workforce and competition simply downgrades such intent by substituting individual motivation with economic needs. It exacerbates the mile-wide and mile-deep problem by creating a climate in which slimming down was less and less of a politically defensible option. Even assuming that the economic crisis was substantiated, there is scant data to support the idea that a broad-based K–12 science education is required. Much of the standard-setting committees rely on experts and key stakeholders’ subjective evaluations of what was needed rather than actual research of what productive adults in the U.S. (or anywhere in the world) actually use in their professional or personal lives. Using mostly engineering and science students from Carnegie Mellon University (considered among the top engineering institutions in the U.S. and across the world), we conducted a small-scale benchmark test by giving them elementary and middle school NAEP and TIMSS items. If these students are representative of the future scientific literate adults, then their performance on NAEP and TIMSS items certainly fail to demonstrate convincingly. While they excelled on content-lean inquiry items that required mostly reasoning and analyses, their performance varied greatly on content-heavy test items and are far from perfect. For example, on a multiple choice question from NAEP asking for the two most common elements in earth’s crust, these elite college students did no better than chance (i.e., only 25% correct when given four choices). If somehow this performance is regarded as yet another example of U.S. lagging in science, then the question has to be asked—what percentage of adult population, in the U.S. or anywhere in the world, can be reasonably expected to meet the science literacy requirements which we have asked of our K–12 students? The data that could be used to answer this question was pitiful. The best available may be the rudimentary surveys (with less than 15 items) conducted by NSF across U.S. and Europe (which, by the way, found no significant difference in adult scientific literacy across the Atlantic). Any studies of literacy within U.S. without international comparison lack objective criteria to determine adequate versus inadequate scientific literacy.

What is the consequence of having a bloated, albeit high, expectation for our students’ scientific literacy? With limited time, financial sources, and teacher expertise and education, the expansive content standards directly conflicted with the other important goal of the standards movement, which was to help students develop deep and lasting understanding
of science rather than to promote the cursive coverage of topics (NCES, 2002). This was a zero sum game. If a teacher had to spend more time covering a breadth of topics, less time would be spent on pushing students to master each topic in a meaningful and deep way; vice versa. This problem is not unanticipated when NSES was first issued. In one section, NSES specifically addressed the question, “How can teachers cover everything in the curriculum if they use inquiry-oriented materials and teaching methods?” (NRC, 1996)

...the Standards do not suggest that all science should be learned through inquiry. However, investigations are important ways to promote deep understanding of science content and the only way to help students practice inquiry abilities. So there is still the issue of coverage vs. learning strategy to address.

Analysis of data collected in the Third International Mathematics and Science Study (TIMSS) reveals that the typical U.S. eighth-grade science textbook includes about 65 topics. A similarly large number of science topics appears yearly in state and local science standards and curriculum guides. Teachers, understandably, feel obligated to teach all of the topics called for in their local science curriculum. The result can be the "mile wide and an inch deep" curriculum often decried in U.S. education. Furthermore, research shows that this "cover everything" approach provides few opportunities for students to acquire anything but surface knowledge on any topic (Schmidt et al., 1997).

There are several steps that teachers and administrators can take to deal with this problem. They can renegotiate the expectations embodied in the curriculum. They can carefully select a few areas to emphasize, spending more time teaching those areas though inquiry. They can carefully analyze the curriculum expectations and combine several learning outcomes in lessons and units. They can work with other grade-level teachers to eliminate the redundancies that.

The steps proposed by NSES were for teachers and administrators to make curricular decisions on the presumption that such choices would in fact given to teachers and administrators. At the time of publications, authors of NSES may have reasons to believe in such an empowering system. But with the increasingly focus on accountability and the inclusion of science tests into No Child Left Behind’s testing requirement slated to begin in 2007–2008, such choices are not made in the classrooms or schools, but from state level down. What the standards movement in science education lacked, more than research theories, was a theory of action or implementation. What are the practical and feasible mechanisms, given our existing system of science education, which can deliver on the promises of both depth and breadth? Ten years after NSES, we not only have not answered the ques-
tion, but also have in fact lost the few plausible answers originally proposed. The standards movement raised the bar on the depth of understanding students should achieve, but failed to reduce the breadth of topics across which such achievement is demanded. The mile wide, inch deep problem has since become the mile wide and mile deep dilemma. It forces a false choice: either hurry students through a vast array of topics in which they have little natural interest and in ways which would unlikely generate such interest, or to focus on a smaller set of topics and teach it in ways that engage the students (which takes time) and face the consequences of not testing well.

FROM STANDARDS TO ACHIEVEMENT GAP

Based on our in-class observations and interviews with 14 science teachers across 6 schools within the parochial district, we hear one unanimous message from all teachers: "I don’t have enough time to teach everything. I start slow but then have to rush things through and try to get as much done as I can.” This is the aforementioned false choice forced upon the teachers through the existing system of curriculum and tests. Using middle-school science as an example, here is how the purported “alignment” among standards, textbooks, and tests contributes to this impossible situation.

Despite the intent to narrow down content areas, the AAAS benchmarks, NSES, and subsequently developed state/district standards still encompass most areas of modern science. Table 3.1 summarizes the broad topic areas that we synthesized from AAAS, NRC, and Pennsylvania standards (Pennsylvania Department of Education, 2002).

Using the above-defined “content clusters,” one can map out the alignment between a typical textbook and a popular standardized test (Terra Nova Comprehensive Test of Basic Skills, or CTBS). For example, Table 3.2 shows that the majority of the test items on a particular year fall within the topics covered by the textbook for the same year. Both the textbook and the test seemed to have passed the muster of “standards-based” alignment (Olson, 2003). Yet such alignment is achieved by “casting the net wide.” Note how the textbook also covered nearly as many topics that were not part of the test. Teachers, rightfully ignorant of the actual item selection on the test, must then cover a total of 24 (out of 30) standards-based topics in one single school year in order to make sure that the students have at least been taught something in the area where they might be tested.

At what level of depth would such teaching necessarily be? Using the particular 6th grade textbook as an example, it is divided into 59 lesson units, which, if divided by the available school days in a year, require on average 2.5 class periods each. The scope of a lesson unit in heredity contains traits, DNA, gene, Watson & Crick, DNA base types, DNA structure,
### TABLE 3.1
Topics Identified across Three Content Standards, NSES, AAAS, and PA

<table>
<thead>
<tr>
<th>Earth Science</th>
<th>Life Science</th>
<th>Physical Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster I: Earth Composition, Plate Tectonics &amp; Related Processes</td>
<td>Structure &amp; Function of Cells</td>
<td>Physical Properties &amp; Phases of Substances</td>
</tr>
<tr>
<td>Cluster II: Erosion &amp; Deposition</td>
<td>Levels of Organization &amp; Development</td>
<td>Chemical Changes &amp; Reactions</td>
</tr>
<tr>
<td>Cluster III: Rock Cycle &amp; Soil Formation</td>
<td>Human Body Systems</td>
<td>Elements &amp; Compounds</td>
</tr>
<tr>
<td>Cluster IV: Natural Resources &amp; Environment</td>
<td>Disease</td>
<td>Motions &amp; Forces</td>
</tr>
<tr>
<td>Cluster V: The Atmosphere</td>
<td>Reproduction</td>
<td>Forms &amp; Transfer of Energy</td>
</tr>
<tr>
<td>Cluster VI: Water</td>
<td>Heredity</td>
<td>Sound Energy</td>
</tr>
<tr>
<td>Cluster VIII: Planetary Characteristics &amp; Composition</td>
<td>Populations &amp; Ecosystems</td>
<td>Electricity &amp; Magnetism</td>
</tr>
<tr>
<td>Cluster IX: The Universe</td>
<td>Energy use in Ecosystems</td>
<td></td>
</tr>
<tr>
<td>Cluster X: Gravity &amp; Movement in the Solar System</td>
<td>Classification of Organisms</td>
<td></td>
</tr>
<tr>
<td>Cluster XI: Seasons</td>
<td>Extinction &amp; Fossil History</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 3.2
The Alignment Among Textbook, Test, and Content Standards in 6th Grade

<table>
<thead>
<tr>
<th>Total of 30 Content Clusters (Grades 5–8, see Table 3.1)</th>
<th>Covered in textbook</th>
<th>Not covered in textbook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tested in CTBS</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Not tested in CTBS</td>
<td>11</td>
<td>3</td>
</tr>
</tbody>
</table>

copies, and ladder, the Human Genome Project, and the use of DNA in police work. Who amongst us can teach a group of 6th grade students (urban or otherwise) all of these ideas, meaningfully, interestingly, and lastingly, in 2.5 class periods?
A suburban class of students who have had strong prior knowledge (from home), prior education (from earlier school experiences), who could read textbook efficiently may well manage to retain a portion of the factual content at least till test time. The challenge is nearly insurmountable with urban students who have had little exposure to formal science outside schools, had major gaps in their prior science education (e.g., due to an emphatic focus on reading and math in early grades, or simply lacking qualified teachers), and lacked the skill and resources to read and comprehend the textbooks as written. The resulted achievement gap is predictable and apparent (Figure 3.3). The chart illustrates the distribution of achievement gap between two neighboring schools. The higher performing school has less than 10% African American students and less than 5% who qualify for free and reduced lunch (i.e., in the previous discussion of TMISS and NAEP results, this school would be one whose performance match its counterparts in Japan and South Korea). The lower performing school has over 95% African American students and over 90% who qualify for free and reduced lunch. Using one single science test (CTBS) in a 6th

![Figure 3.3. Cognitive objectives and test gap.](image)

Note. This parietal chart compares the achievement gap on categories of items based on cognitive objectives (using Bloom’s [1956] Taxonomy) in low- and high-SES schools. The columns show, respectively, the weight of a particular category of items on the test and the extent to the category contributes to the overall test gap between the schools. The columns are ordered from right to left in terms of the rising scales of Bloom’s Taxonomy. The lines show the same information but in an accumulative fashion.
grade year, each of the 40 test items was coded by Bloom’s Taxonomy into cognitive levels. The lowest level (1.*) in Bloom refers mostly to factual/recall type of knowledge. The higher levels correspond more with the “inquiry” as described in the standards. The figure illustrates two important findings:

1. The test is weighted heavily towards the breadth of factual content;
2. Expectedly, the achievement gap is weighted heavily towards the breadth of factual content as well.

In fact, over 80% of the achievement gap (as measured by % correct per test item) can be attributed to the lowest level of cognitive objective as defined by Bloom’s Taxonomy. Inquiry, both as a means and an end, seem hardly relevant to the problem of achievement gap in this context.

COPING WITH AND RISING ABOVE THE SYSTEM

How do you carry out the spirit of standards (inquiry, equity, literacy) while living within the current system of science standards and upcoming test-based accountability? Particularly, how do you narrow the achievement gap across racial and economic lines within the U.S.? I propose two ideas: one with a practical goal of raising test scores and the other with the more idealistic goal of raising students’ motivation.

Raising Test Score? Teach to the Tests

The strategy we have proposed in our classrooms and carried out to varying degrees of success was to somewhat de-couple content and inquiry in science. Particularly in middle school grades, the content portion of the science, so far as satisfying testing requirements are concerned, can largely be taught with strategies that enhanced recall and recognition, rather than the ones that aim for in-depth understanding. We reserve meaningful inquiry to be conducted separately from the content-heavy coverage. Such strategy is not our innovation. In fact, commercial companies like Kaplan has long made efficient test preparation their business and has now become a big provider for schools in their scramble to meet accountability goals, first in reading and math, and now in science.

The challenge, of course, is to do the test preparation piece efficiently so that there would be enough time left over for students to conduct their own inquiries that emphasize meaning and depth rather than coverage. The traditional methods of cognitive task analysis whereby one breaks
down a problem into its knowledge chunks and required strategies can be useful in dissecting the test items on standardized tests.

Using a broad sampling of released items from international (TIMSS), national (NAEP), and state tests, one can derive the "most commonly tested" knowledge chunks and the ways in which they were most commonly assessed. Incorporating such knowledge into testing preparation reduces a great deal of aimless wandering through the thick textbook. Figure 3.4 shows an example of such analysis in a popular middle-school science topic.

What the chart illustrates is a phenomenon (of which we were surprised) that science tests, as diverse in origin as NAEP, TIMSS, state tests, commercial tests, show great convergence in terms of the specific knowledge they require in any particular topic area. As the number of sampled tests increase, the number of requisite and distinct knowledge chunks does not. This is not simply a case by which all tests, based on standards, eventually can be reverse-engineered to reproduce the standards. What we term "chunks" here are far more concrete and specific than most statements in standards. In this particular topic area, the most heavily "hit" chunks are, "Mammals begin their life in the womb as oppose to eggs or larve" and "Insects can be identified by three body segments and six legs." Merely knowing that as a factoid is sufficient to answer most, if all, of the related questions.

This, of course, falls precisely into the broad category of practices termed "teaching to the test." Such practices work in so far as raising test scores (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000). Public schools from Texas to Pennsylvania had spent millions contracting with test preparation companies like Kaplan to monitor and ensure achievement score gains. We do not advocate for such broad approaches of teaching to the test. Within our local context, teaching to the test is only justified if the teacher intends to use it to free up time for the purpose to engaging students in richer and more authentic inquiry experiences, which otherwise would have been squeezed out by the mile-wide demands of mandated standards and accountability tests. This is our more "idealistic" goal for science education.

Raising Motivation: Teaching to and for the Students

In the recent bestseller, "The World is Flat," New York Times columnist Thomas Friedman gave much attention to the issue of science education. The author lamented, among other causes, the "ambition gap" between U.S. students and those in underdeveloped countries. The issue apparently is that students in rising economies like China and India are soon to (or have already) overtake us in engineering and science expertise. This
Figure 3.4. Decoding tests into knowledge chunks (Topic: Classification of organisms)
matches the commonly accepted crisis in the relative decline in our engineering and science graduates from universities. The data for the latter crisis, like the early crisis of K–12 science scores, seemed more debatable than the frequent parroting would suggest. A recent study (Gereffi & Wadhwa, 2005) suggest that, when compared using comparable criteria, U.S. produces more engineering and science graduates per capita than either China or India.

Statistics aside, motivation ultimately is a subjective and personal attribute. Instead of a purported ambition gap overseas, I observe more readily a motivation gap within our borders. In my local context, I found a gap of aspirations between my inner-city school students and the motivations for political consensus (e.g., the American Competitiveness Initiative announced by the President’s 2006 State of the Union address and echoed by a bipartisan chorus of national leaders). Politically, science education seems a rare unifying agenda. The argument goes:

1. America is losing its competitive edge to China and India;
2. Our children are flunking science and unfit for high-paying jobs;
3. Science education is the engine of America’s competitiveness.

Yet despite the fact that most of my students are poor enough to quality for free and reduced lunches, I do not believe I can convince them to study science in order to fight for a job against some low-wage worker in Asia or to contribute to shareholder values of America’s corporations. My students had a very different agenda. When I gave them the freedom of deciding the science fair projects they would do, here is a partial list of the ideas that mattered to them:

- To find better and more efficient ways to recycle;
- To find out what causes asthma because most of his/her family suffers from it;
- To find out how to cure diabetes because his/her grandmother has it;
- To make a book that flips itself so disabled people can read it;
- To make a Braille book so she can teach a blind child to read;
- To make food safer to eat and take out all the bad chemicals in it;
- To make lightening, to make a laser;
- To know what his/her peers think about sex, drugs, and alcohol.

Perhaps both the children and our national leaders are motivated by necessity. But there is a fundamental difference between what is considered worthy goals by children and what our leaders are presenting as the urgent case for science education. My students want to study science to understand their world and themselves, to serve their family, community, and mankind, and to “make gentle the life of this world.” Surely, most kids
do not know enough science yet to do their projects properly. But they want to and that is a start. For all my coaxing and coercion, nothing I ever did brought forth the kind of initiative and innovation my students voluntarily put into their own projects, even when the science involved was far more difficult than anything on my tests. My seventh graders even agreed to put weeks of cooperative effort into test preparation on my promise that, once the annual standardized science test is over, we will together study science ideas that will help them to do their projects.

Albert Einstein once wrote, “The ideas which have lighted my way, and time after time have given me new courage to face life cheerfully, have been Kindness, Beauty, and Truth.” When the Smithsonian magazine honored 35 great innovators of our time, each of the honored scientists, writers, and artists embody one or more of these attributes. I cannot imagine a higher calling for science or a nobler purpose for science education. Our children get it. Much of political discourse on science education is missing it. It was during the Great Depression that FDR consoled our nation with “the only thing we have to fear is fear itself.” It was following Sputnik launch during the cold war that JFK energized a generation of youth with “ask what you can do for your country [and] what we together can do for the freedom of men.” Today, with U.S. as the only remaining superpower in the world, is “preparing our nation to compete in the world” the best we have to inspire our children?

The political and social motivations that bloated the science standards are neither grounded in data nor likely to inspire the next generation of students to pursue or at least engage in science. The most promising science teaching practices (or any teaching practice for that matter) for socially disadvantaged students had often been ones that aim to engage the students in a deep, meaningful, rigorous attempt to connect good science to their lives. Relying on such practices exclusively is unlikely to produce higher test scores, particularly in the context of pushing for short-term, rapid gains (e.g., annual yearly progress, or AYP). The aforementioned mentioned test preparation strategies may deliver both the time and the political cover for teachers to continue to pursue, preserve, and protect these practices for students.

**SUMMARY**

This separation of strategies that serve content knowledge and inquiry skills is most definitely unintended by the standards-based movement. The intent of NSES, in contrast, was to integrate what was traditionally called “content” and “process.” But the ideal of teaching content through the inquiry process was simply unfeasible given the instructional time severely
limited by the breadth of required topics. Much of the accountability tests (commercial or state-made) continue to reflect this division (see Figure 3.3) despite declarations of doing the opposite. The status quo, resulting from the conflicting visions and implementations, lowers science learning for all students. It penalizes underserved students in particular by forcing upon them what they are least prepared or motivated to do—the breadth of factual content while squeezing out time for what they are most likely to engage—meaningful, relevant, and in-depth scientific inquiry.

In proposing a coping strategy that contradicts the vision of the standards-based reform but meets the existing reality in urban schools, I am not vouchering that such an approach would ultimately lead to better scientists, more competitive workforce, or simply more scientifically literate citizens. But through our field research, I know to a reasonable certainty that this approach is better than the status quo. It is more practical than ideological. It accepts the reality despite disagreeing with the rhetoric. It copes with the constraints imposed by standards and accountability on science education. It ultimately falls far short of the ideals of the standards, but for the moment, it is something one can live with while making gradual and steady, albeit small steps towards narrowing the achievement gap.

ACKNOWLEDGMENTS

The research described in this article is funded through the Cognition and Student Learning Program at the Institute for Education Sciences, U.S. Department of Education. We thank all the science teachers in six Pittsburgh parochial schools who opened their classrooms for our research. We thank them particularly for their patience in searching for a workable solution amidst the most challenging instructional environment.

REFERENCES


