Junlei Li, David Klahr, Stephanie Siler

What Lies Beneath the Science Achievement Gap: The Challenges of Aligning Science Instruction With Standards and Tests

Amongst all *instructional* issues facing science education, the one that exerts the most substantial impact on the lasting achievement gap is the "mile-wide, inch-deep" curriculum, which is created by superficial alignments among standards, tests, and instructional materials.

"If science educators had a dime for every time the phrase 'standards-based' or 'aligned with standards' pops up in science textbooks, instructional product brochures, conference programs, or in-service workshop presentations …"

Diminishing "standards" and "alignment" to overused buzzwords or superficial checklists masks the dire need for truly systemic and operational standards-based alignment in science education. In this article, we report the findings of an ongoing collaborative effort between cognitive researchers and urban science teachers to align everyday teaching with standards, tests, and research-based pedagogy. We begin with an analysis of how the width vs. depth dilemma in science teaching manifested itself in yearly test scores and the achievement gap. We review the problematic issues of alignment among standards, instruction, and assessment. We argue that simply matching standards with so-called "standards-based" materials creates a false sense of comfort in a superficially aligned curriculum. We advocate for schools, districts, even states to undertake the difficult but necessary planning process to create a framework of performance objectives to serve as the critical hinge linking standards, instruction, and assessment. Such curriculum planning must set as its first priority the goals of effectively cutting down the girth of yearly science content while efficiently managing the handoff of students between grade levels.

Research Context and Data Collection

Our research takes place in three urban parochial schools (> 90% eligible for free and reduced lunch programs and > 95% African American). We use one affluent parochial school as a comparison group (< 10% eligible for free and reduced lunch programs and < 10% African American). Science teachers for 6th through 8th grades in the three urban schools collaborate with the research team in both bi-weekly meetings during the school year and summer workshops. In addition, the researchers learn, observe, and co-teach in the urban classrooms. The comparison school is not directly involved in any intervention efforts.

… simply matching standards with so-called "standards-based" materials creates a false sense of comfort in a superficially aligned curriculum.

All schools use the same district-wide curriculum guidelines, though the instructional materials vary from school to school. All science teachers are certified in elementary education with approximately half also certified to teach science at elementary or secondary levels. This teacher profile

is comparable to nationwide statistics for science teaching in public schools (National Center for Education Statistics, 2002). All schools are annually assessed using the Terra Nova Comprehensive Test of Basic Skills [CTBS] (CTB/McGraw-Hill, 2001), which includes a 40-item multiple choice assessment for science for each grade level. The parochial district evaluates schools based on the annual tests and exerts administrative pressure on principals and teachers to improve performance. There is no "high-stakes" accountability system (e.g., sanction, merit-pay). The achievement data reported here was collected by obtaining students' answer sheets for CTBS tests from both the urban schools and the comparison school. We analyze test performance by test item rather than relying on the gross subject-level data reported by CTB/McGraw-Hill. In addition, we were able to collect, through interviews, surveys, and in-class observations, a detailed record of what each teacher taught in each school. We connected test data and everyday teaching through item analyses that categorized items by topic area and by cognitive demand (Bloom, 1956).

Science Achievement Gap

What follows is a tale of two gaps:

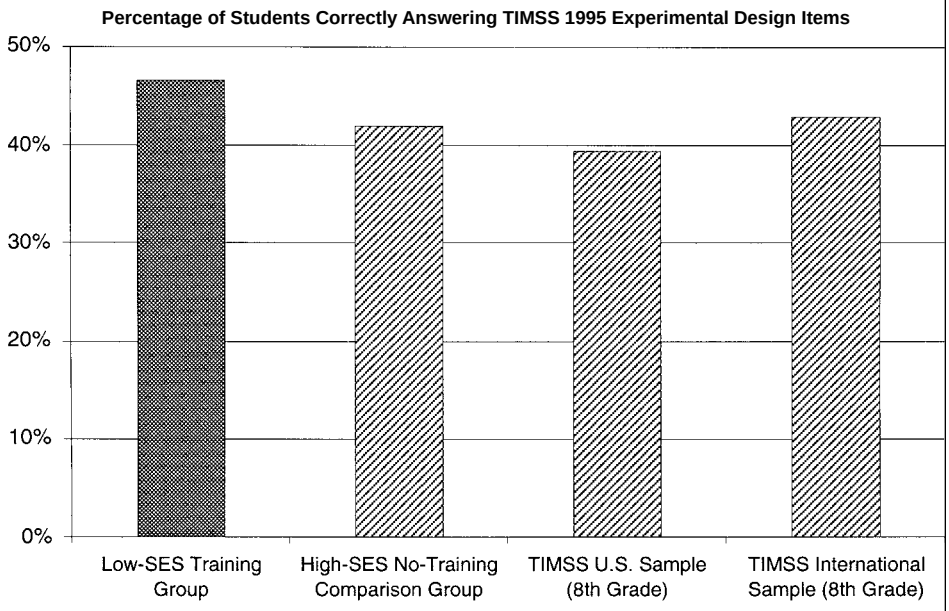
1) The *learning gap* in particular topic areas, which, through teacher and researcher collaboration and intense instructional investment, can be narrowed or even closed;
2) The *test gap* across the entire science curriculum, which, despite teacher and researcher collaboration and intense instructional investment and professional development, remain wide open.

In urban school settings, teaching for mastery requires time and patience.

We began with a set of science instructional strategies first developed in cognitive psychology laboratories and subsequently validated in diverse classroom settings (Chen & Klahr, 1999; Klahr, Chen & Toth, 2001; Klahr & Nigam, 2004; Klahr & Li, 2005; Strand Cary & Klahr, 2005; Toth, Klahr, & Chen, 2000; Triona & Klahr, 2003). For the purpose of our discussion here, the operational details of our instructional method are not particularly important (see Klahr & Li, 2005, for a more detailed discussion). It suffices to say that our proposed methods push for mastery by narrowing our focus on skill or concept domains through a sequence of cognitively-balanced instructional activities, including goal-directed exploration, elicitation of student's justification and explanation, repeated formative and performance assessment, and explicit instruction. The argument we are making here is not that our method is the best way or even that it is better than some other alternative. Instead, we present evidence that our method can close the learning gap while still leaving the test gap wide open.

In urban school settings, teaching for mastery requires time and patience. For example, we had developed instruction to help students achieve high levels of mastery in designing valid scientific experiments. In affluent high-achieving schools, students achieved mastery in two days. In our urban schools, it took one to three weeks depending on classroom and school conditions. But the intense investment of teacher's planning and teaching in urban schools, carried out through iterative lesson studies and in-class teacher-researcher collabora-

Figure 1
Low-SES Training group and high-SES comparison group's performance on select TIMSS 8th Grade science items pertaining to controls and variables, compared with U.S. and international benchmarks.



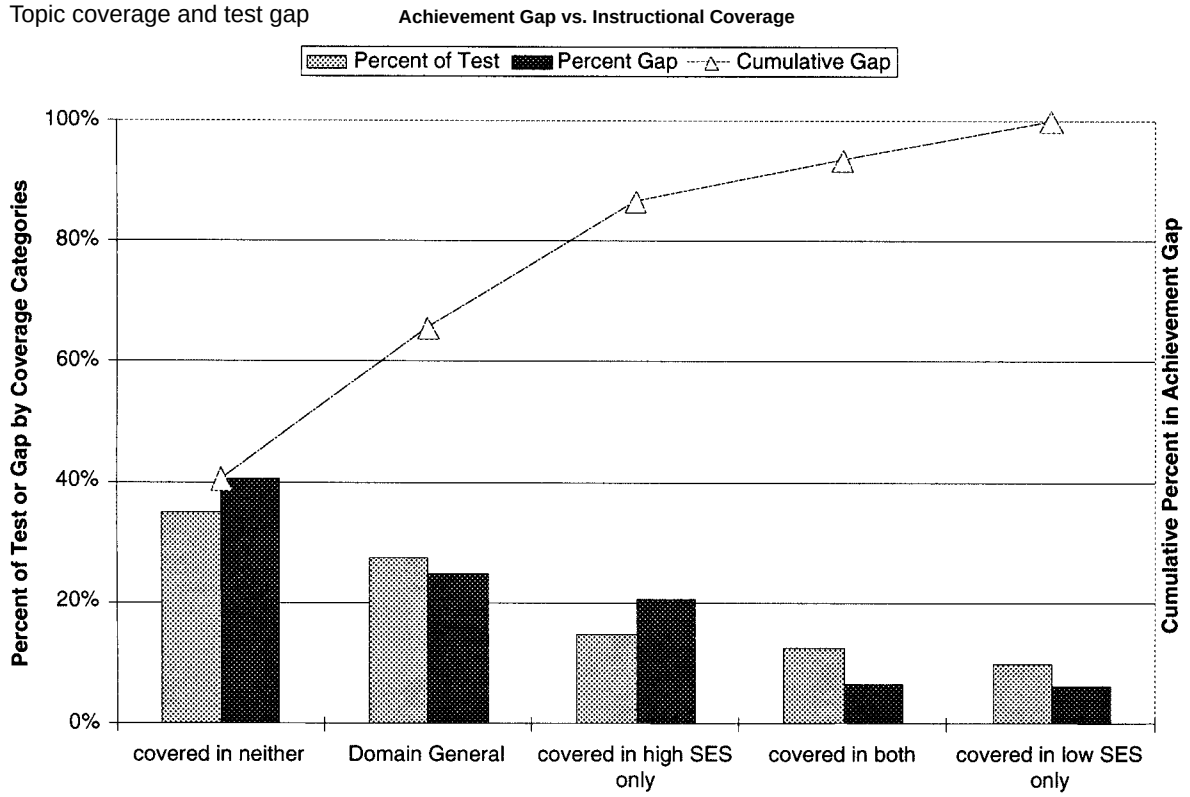
tion, do pay off. In designing scientific experiments, for example, 5th and 6th grade urban students achieved a level of mastery exceeding their same-age counterparts in the affluent school. Their performance also matched or exceeded national and international benchmarks on standardized test items reused from the National Assessment of Educational Progress [NAEP] and Trends in International Math and Science Study [TIMSS] tests (Figure 1). In another example, over a three week period, students in two 6th grade urban classrooms learned to explain day and night and the seasonal change in daylight hours. Their performance on relevant TIMSS 8th grade items not only exceeded that of the U.S. average, but matched that of international leaders like Japan. These results are encouraging indications that, with adequate investment of time, professional development, and research-practice collaboration, we can narrow the learning gap.

One would expect that, with mastery at the topic level, the overall test gap would also narrow. But our efforts did not lead to the narrowing of the gap as measured by yearly standardized tests. The heavy investment in closing the learning gap topic by topic incurs a great cost on the breadth of topic coverage. For each lesson we planned, there were bound to be many that we could not, due to the lack of teacher preparation time. For each topic we taught to mastery, there were bound to be many that we could not, due to the lack of instructional time. In one year, by the time the CTBS test was administered, our three urban schools only managed to cover just over half of the planned curriculum.

The heavy investment in closing the learning gap topic by topic incurs a great cost on the breadth of topic coverage.

The test gap between the urban and comparison schools is illustrated in Figure 2. The 40 items on the test were grouped, based on teacher interviews, surveys, and item analysis, into five categories. The domain-general cat-

Figure 2
Topic coverage and test gap



Note. This parietal chart compares the achievement gap on categories of items based on coverage 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 absolute size of the test gap. The lines show the same information but in an accumulative fashion.

egory includes inquiry or reasoning items that do not rely on any specific content knowledge. The remaining categories include items that, without specific content knowledge, a student cannot answer. Figure 2 shows perhaps the "obvious"—when a test item relates to content topics that were taught in the urban schools or skill areas that required no particular content knowledge (i.e., domain general), the associated test gap is smaller per item than that for test items under topics "not taught". This supports our assertion that intense investment in teaching is beginning to narrow the learning gap in the specific topics or skills taught, but not nearly fast enough or "wide" enough to catch up on yearly tests. The test items that fall under topics "not taught" by urban schools contribute to about 60% of the total test gap (adding together the "covered in neither" and "covered in high-SES only" columns). In other words, 60% of the test gap can be attributed not to the quality of teaching in the urban schools, but merely to the breadth of coverage or opportunity to learn. Furthermore, the single biggest source of the test gap is the "covered in neither" category, suggesting that even when both urban schools and the affluent school were limited in their breadth of coverage, differences in prior knowledge alone could account for 40% of the total test gap. It is tempting to jump to the conclusion that breadth of coverage is what we need. But with breadth, we will lose the depth of mastery per topic. During our intervention, the teachers only taught one third of the topics they had taught in past years, but the overall test scores were no different from years prior.

Can we expect this trend to improve over time? Would multiple years of intervention narrow the gap? Though we would like to believe that, based on our success in closing some topic-level learning gaps, our further analysis reveals a more pessimistic answer. Recall that our instruction is focused on mastery and deep understanding. By mastery, we mean that students not only could recognize and reproduce factual information, but could apply their learning robustly in an inquiry context—a goal aligned with the spirit of the standards movements. To what extent is our instructional focus on knowledge application aligned with the assessment instrument? Figure 3 shows the break-down of the 40 test items by cognitive objectives (Bloom, 1956). Over 80% of the achievement gap is contained within the most basic level of Bloom's hierarchy of cognitive objectives, involving mostly terminologies and facts. If we follow the "getting the biggest bang for your buck" principle, we would be tempted to suggest that the quickest path to closing the test gap on the CTBS tests is to target instruction towards the lowest levels of cognitive objectives. This suggests that our instructional focus for understanding and mastery is aligned with the standards but misaligned with the emphasis of the tests.

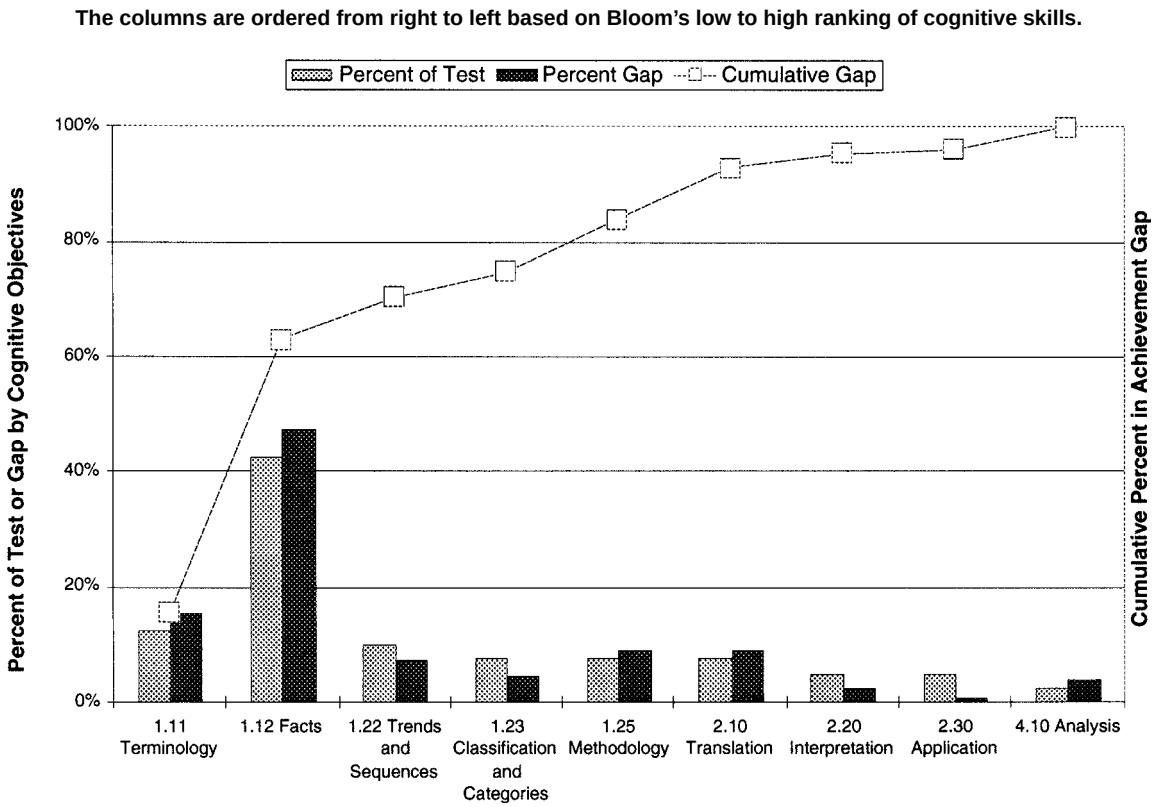
Standards and tests are here to stay and nearly every state has adopted science content standards.

Standards-based Reform in Science Education

We do not infer from the above analysis that schools should do away with standards or tests. Standards and tests are here to stay and nearly every state has adopted science content standards. Beginning in the 2007-2008 school year, all states must also measure science achievement with assessments that align with state standards (No Child Left Behind Act, Public Law 110-107). All of these reform efforts have the intention of narrowing the achievement gap—albeit the gap has mostly been defined as the test performance differences between rich and poor or predominantly white or minority schools.

Before we can operationalize a systemic alignment in science education, we need to first understand the relationship between the achievement gap, standards, tests, and everyday instruction. During the last decade, large scale investigations have focused on international comparisons of test performance, curriculum, and teaching. Two prominent reports—*A Splintered Vision* (Schmidt, McKnight, & Raizen, 1997) and *The Teaching Gap* (Stigler & Hiebert, 1999)—argue that U.S. science and mathematics education are "splintered" by "mile wide, inch deep" curriculum aims and textbooks, and that U.S. teachers have neither the supporting resources nor the ongoing collaborative professional practice to iteratively plan, evaluate, and revise their lessons. The outcry against bloated science curricula and advocacy for professionalizing science teaching are among the core issues that inspired the standards movement (American Association for the Advancement of Science [AAAS], 1990, 1993; National Research Council [NRC], 1999). But is science education less "stuffed" and more "nourished" today than it was more than a decade ago? The debates persist as the standards reform movement, in the course of state-by-state implementation, triggered unintended consequences such as ballooning the

Figure 3
Cognitive objectives and test gap



Note. This parietal chart compares the achievement gap on categories of items based on cognitive objectives (using Bloom's 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 absolute size of the test gap. The lines show the same information but in an accumulative fashion.

scope of science content, limiting the choice of instructional strategies, and imposing one-size-fits-all goals and solutions for diverse student populations (Anderson, 2004; Anderson & Helms, 2001; Barton, 1998; Bauer, 1992; Donmoyer, 1995; Hewson, Kahle, Scantlebury, & Davies, 2001; Li & Klahr, 2005; Rodriguez, 1997; Settlage & Meadows, 2002; Shamos, 1995; Shiland, 1998; Thomas B. Fordham Institute, 2000, 2005; Vesilind & Jones, 1998; Wolk, 1999, 2004).

Academic and policy debates seem somewhat remote to practitioners on the frontline of science education. Science teachers, department heads, and instructional specialists need to survive and thrive in a teaching environment increasingly driven by standards and measured by accountability tests. They are the ones who must, here and now, find solutions to the pressing problems of standards-based reform, including (but not limited to) the following three interrelated claims (Anderson, 2004, p.1):

- The reform agenda is more ambitious than our current resources and infrastructure will support.
- The standards advocate strategies that may not reduce achievement gaps among different groups of students.
- There are too many standards, more than students can learn with understanding in the time we have to teach science.

The analysis of the width vs. depth dilemma in our local context supports each of these three claims. The state of science education, we argue, requires the application of a basic economic principle: *scarcity necessitates choice*. The scarcity of instructional and planning time, physical materials and resources, and teacher preparedness necessitates

difficult choices that teachers have to make about what and how to teach on a daily basis. These choices converge on the issue of *alignment*—the streamlining of instructional goals and strategies within policy constraints that maximally utilize the available human, time, and physical resources towards closing achievement gaps.

The outcry against bloated science curricula and advocacy for professionalizing science teaching are among the core issues that inspired the standards movement.

Alignment

There is a broad consensus among practitioners, policy makers, and researchers that "alignment" is a prerequisite for educational improvements in today's high standards and high accountability system (see Olson, 2003 for a succinct summary and advocacy). Everyday instruction in the science classroom should align with standards, be informed by formative and summative assessments that also align with standards, incorporate instructional products that are standards-based, and apply pedagogical strategies that are also standards-based.

What resources are available to classroom teachers, school principals, and district leaders to create such a system of alignment? We are bombarded with documents titled, "content standards", "benchmarks", "teaching standards", and "curriculum frameworks", many of which overlap and restate each other. Educational product brochures are strewn with variants of the "alignment checklist"—generically and superficially claiming how each lesson unit or module is aligned with a host of inquiry and content standards. The overuse and misuse of the term "alignment" belies the genuine alignment process—*to be in or come into precise adjustment or correct relative position* (Webster's Dictionary)—that demands a system in which everyday teaching, standards, and tests can be brought into "correct relative position" through "precise adjustment".

We argue that the lack of an operational process of alignment is not due to the lack of trying, but a dearth of *specificity* and *transparency* in the reform infrastructure. In order to ward off excessive width or depth in teaching, a teacher needs to know *specifically* what content should be taught at what grade level, to what level of mastery, and measured by what set of performance objectives. For example, standards statements like "students should develop general abilities, such as … identifying and controlling variables" (NRC, 1996) and "design controlled experiments, recognize variables, and manipulate variables" (Pennsylvania Department of Education, 2002) could easily have been used to describe goals in undergraduate or graduate level research methods classes. These statements do not offer a usable specification of the level of mastery expected of students in grades 5 through 8. The alternative is to build grade-level performance objective based on the standards, such as:

> "In 5th grade, students should be able to design a controlled experiment when the key variables are already given, in simple topic areas such as, 'Does water make a plant grow faster or slower?' or 'Does sugar dissolve faster in warm water?' In addition, students should be able to discriminate a controlled experiment from an uncontrolled experiment when they are given the variables and the procedures. Also, students should be able to identify the important variable to contrast when the research question has been specified, such as "water" or "temperature of water" in the two topic examples above."

In order to align day-to-day teaching and formative assessment with yearly accountability assessments, a teacher also needs a *transparent* roadmap that leads from topic-specific performance objectives to the skills and knowledge demanded by accountability tests. This roadmap should make it clear and unambiguous what the mandated test expects of the student within a specific topic area at a specific grade level, not some general descriptions of "proficiency". Most states and test publishers release teacher's guides and assessment handbooks in the hopes of providing such a roadmap. But guideline statements often are just as vague and generic as those in the standards, for example:

> "Students must have an understanding of the concepts and terms included in the standards through grade 7. This understanding should go beyond simple knowledge recall (Bloom's Level One). Students should be able to translate and apply the terms to new situations when answering an item." (Pennsylvania Department of Education, 2002)

Using our example earlier, how would a teacher know, from this gen-

eral assessment guideline, what level of performance is expected from the students when it comes to controlling variables and designing experiments? The alternative is to provide a topic-level roadmap so that the teacher can clearly see the linkage (i.e., transparency) between the standards, the performance objective, and the test requirements. It may look something like this for our example topic:

> "At a 'recall' level, students can define the words 'variable' and 'control'. At a 'basic use' level, students can identify the target variable from a question statement, such as, 'Does water make the plant grows faster?' At an 'application' level, students can design an experimental procedure based on the variables they can identify from a question statement. At a 'generalization' level, students can examine a given description of an experimental procedure and critique whether the procedure has met the requirements of a controlled experiment. For each of these levels, example assessment items are included. At the 4th grade level, assessment items will emphasize recall and basic use. At the 7th grade level, assessment items will emphasize application and generalization."

The state of science education, we argue, requires the application of a basic economic principle: *scarcity necessitates choice.*

These two aspects of alignment, specificity and transparency, cannot be implemented independently. Without specificity of content and performance standards, there is no framework to which the tests or teaching could align. Without transparency in the tests, the outcome measures can only produce information of a coarse grain size, unusable to inform and improve everyday teaching. We believe that, as a prerequisite for improving achievement, we must have a system of alignment that can reduce the burden of the "mile-wide" content and enable meaningful and mastery-focused teaching. This is easier said than done. Using our local context, we review the challenges of using traditional approaches and existing resources to attempt this daunting task.

Difficulty of Alignment In a Local Context

In the same year as our project began, our parochial district unveiled its newly revised curriculum guidelines based on the adoption of the state science content standards. Because the CTBS tests used by the district proclaim to be aligned with science standards at the national level, we evaluated whether the Pennsylvania state standards align with National Science Education Standards [NSES] (NRC, 1996) and the Benchmarks for Scientific Literacy (AAAS, 1993). We assembled all of the standards pertaining to the middle grade levels (5th through 8th). From this collection of content standard statements from three separate guidelines, we group similar topics together into "clusters"—each containing

Table 1
An example of content standard topics used in the alignment analysis

	NSES 5-8	AAAS 6-8	PA 7th
Light & Solar Energy	• Light interacts with matter by transmission (including refraction), absorption, or scattering (including reflection). To see an object, **light from that object--emitted by or scattered from it--must enter the eye.** p155 • The **sun is a major source of energy** for changes on the earth's surface. The sun loses energy by emitting light. A tiny fraction of that light reaches the earth, transferring energy from the sun to the earth. The **sun's energy arrives as light with a range of wavelengths, consisting of visible light, infrared, and ultraviolet radiation**. p155	• Something can be "seen" **when light waves emitted or reflected by it enter the eye**... 4F p90 • Human eyes respond to only a narrow range of **wavelengths of electromagnetic radiation—visible light**. Differences **wavelength** of within that range are perceived as differences in color. 4F p90 • Light from **the sun is made up of a mixture of many different colors of light**, even though to the eye the light looks almost white. Other things that give off or reflect light have a different mix of colors. 4F p90	• Explain how…**light travels** in waves of differing speeds, sizes and frequencies 3.4.7C • Explain how convex and concave mirrors and lenses change light images. 3.4.7C • Know that the **sun is a major source of energy that emits wavelengths of visible light, infrared and ultraviolet radiation**. 3.4.7B

Note: AAAS states that students should "learn about the electromagnetic spectrum, including the assertion that it consists of wavelike radiations. Wavelength should be the property receiving the most attention but only minimal calculation." (p 90)

Table 2
Topics identified across three content standards, NSES, AAAS, and PA

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

and comparing relevant statements from all three standards. Table 1 shows one example of such clusters and Table 2 shows the total of 30 clusters identified in the three main branches of middle school science across three sets of standards. We have not included the inquiry standards in our analysis with the understanding that students should be engaged in inquiry across all science content areas.

The three sets of standards, for the most part, ask for a similar core body of content. At least on a content level, we seemed to have found alignment among district, state, and national standards. But in practice, a curriculum plan requires a level of specificity that the standards fall far short of providing. We discuss two significant challenges in curriculum planning using standards: sequence and selective emphasis.

Unlike a curriculum plan, standards provide neither transition between topics nor progression within topics for each block of grades (e.g., 5th through 8th). Using our example topic Light and Solar Energy (Table 1), where does it fit sequentially within the whole spectrum of science content (Table 2)? In addition, which aspects of this topic should be taught in 5th grade vs. 6th grade vs. 7th grade vs. 8th grade? The content standards offer no such specification, leaving this enormously complex task to practitioners. Also unlike a curriculum plan, standards tell what topics should be taught, but not the appropriate emphasis or weight one should place upon different topics at different grade levels.

Without specificity of content and performance standards, there is no framework to which the tests or teaching could align.

Obviously, content standards leave much to do for teachers and science instruction specialists. But on what research, knowledge, and practical grounds should such complex decisions be made?

In the absence of a specific curriculum framework, the teachers in our local context rely heavily on existing instructional materials—including textbooks, lab kits, and miscellaneous activities they have attempted in past years. Much of the materials published after the release of the national standards proclaim their alignment with content and inquiry standards. If materials do indeed align with standards, then why not just follow their predefined sequence and emphases? We could keep our fingers crossed that what one teaches based on standards-based materials matches what one's students would be measured on by the standards-based tests.

The alignment between popular instructional materials and science standards has been extensively studied, particularly in middle school science (Kesidou & Roseman, 2002; Stern & Ahlgren, 2002; Stern & Roseman, 2004). Across the board, popular instructional materials fail to convey the "big ideas" intended by the standards and to provide meaningful assessments appropriate to the knowledge level demanded by the standards. We do not re-investigate these issues, but rather, focus on three commonsense practical issues. First, do the textbooks "cover" the topics in the standards? Using 6th grade as an example, we find that the textbook covers or touches upon 24 of the 30 total clusters (Table 2). The textbook 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. How much content is

included in one single lesson unit that is to be taught in 2.5 class periods? Using heredity as an example, the textbook lesson unit contains the following concepts—traits, DNA, gene, Watson & Crick, DNA base types, DNA structure, copies, and ladder, the Human Genome Project, and the use of DNA in police work. Though standards should in theory help us narrow down the coverage, the lack of specificity in the language of the standards invariably favors *inclusion* rather than exclusion of topics. One can quite easily make a case that all of the concepts listed above fall under the relevant state standards, which include statements such as, "know that every organism has a set of genetic instructions", "identify and explain inheritable characteristics", "describe how traits are inherited", "recognize that mutations can alter a gene", and "describe how … genetic technologies can change genetic makeup". Lest these topic-level statements not be inclusive enough, there are always some "catch-all" topic-general standards under broad headings such as, "Science, Technology, and Human Endeavors", with inclusive statements like, "explain how human ingenuity and technology resources satisfy specific human needs and improve the quality of life." The lack of specificity in standards all but ensured that the textbooks will always "cover" standards-based topics.

Second, do the test items align with the topics in the standards? From the CTBS tests, we identified all test items that demanded specific content knowledge and matched them with appropriate topics (inter-rater reliability 85%, disagreements resolved by consensus). All of the content-based items in the 6th grade science test in CTBS fall within 16 of the 30 total clusters. This alignment between test and standards is expected given the general "inclusiveness" of the standards language. For example, the 6th grade test included two test items in the general topic area of "gravity and movements in the solar system". One item asks about the causes of tide and the other compares all nine planets' orbiting times. Easily, the topic and level of these two items align with the standards. The problem is, so would many other possible test items. What about the causes of day and night, summer and winter, sunrise and sunset, changes in length of daylight, or the comparisons of gravitational force on each planet and the moon? How does a teacher know which of these many topics need to be taught deeply when there is no time to teach all of them equally in-depth? None of these ideas are trivial, by any means. The famous "Private Universe" video shows how Harvard graduates and faculties stumble on these supposedly "middle-school" science questions.

As our nation's science education crosses the threshold of accountability testing, it is imperative to build, at whatever level feasible—by state, by district, by school, or by science department if need be—a coherent and operational system of alignment among everyday teaching, content standards, and assessments.

This leads to our last question—does the instructional material used in a particular year cover what is needed to perform on the test items used for the same year? This would seem highly unlikely considering that the textbook is published before the test was ever made and by a different publisher. But like magic, the majority of the test items fall within the topics covered by the textbook (Table 3). Both the textbook and the test seemed to have passed the muster of "standards-based" alignment. Can we simply follow the instructional materials and be confident that, if we teach by these materials, our students would achieve on these aligned tests?

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." Referring back to Table 3, it is easy to see why this would happen. The 30 content topics are meant for all four grade levels from 5th through 8th grade. They are not designed to be taught in a single grade level. The 6th grade textbook, for examples, covers 24 of the 30 clusters. This is the amount of coverage of all general science textbook we surveyed, regardless of grade levels. So teachers are repeating many topics year after year, yet each time could not afford to spend more than a few class periods on each lesson unit. This echoes the depictions of the "mile wide, inch deep" curriculum in the TIMSS report on U.S. Science Education (Schmidt, McKnight, & Raizen, 1997). Nearly a decade after NSES and seven years after "A Splintered Vision", we see ghosts of pre-standards days materialize in our schools, or perhaps, they never left. In this system, we may be able to speak of

Table 3
The alignment among textbook, test, and content standards in 6th Grade

Total of 30 Content Clusters (Grades 5-8, see Table 2)	Covered in textbook	Not covered in textbook
Tested in CTBS	13	3
Not tested in CTBS	11	3

Table 4
CTBS test coverage in life science across four grade levels

	LIFE SCIENCE	GRADE 5	GRADE 6	GRADE 7	GRADE 8
Cluster I	Structure & Function of Cells		1		1
Cluster II	Levels of Organization & Development				
Cluster III	Human Body Systems		2		1
Cluster IV	Disease				1
Cluster V	Reproduction		1	1	1
Cluster VI	Heredity			2	1
Cluster VII	Response, Behavior & Adaptation	2		1	1
Cluster VIII	Populations & Ecosystems	2	3	1	1
Cluster IX	Energy use in Ecosystems	1		1	2
Cluster X	Classification of Organisms	4	2	3	
Cluster XI	Extinction & Fossil History				

Note. Numbers in cell represent the number of test items per year and the corresponding initial weight on the curriculum plan.

alignment and coverage, but certainly not mastery and understanding.

Specificity, Transparency, and Professionalism

In this article, we presented our search for alignment and its problematic relationship to the persisting test gap. We argue that, amongst all *instructional* issues facing science education, the one that exerts the most substantial impact on the lasting achievement gap is the "mile-wide, inch-deep" curriculum. This problem is created by superficial alignments among standards, tests, and instructional materials. It squashes opportunities to innovate, experiment, and plan. As our nation's science education crosses the threshold of accountability testing, it is imperative to build, at whatever level feasible—by state, by district, by school, or by science department if need be—a coherent and operational system of alignment among everyday teaching, content standards, and assessments. Such a solution needs to account for and address the issues of specificity and transparency we have raised. Rhetorical arguments and marketing slogans are simply not useful in the search for such an alignment. We need to do the grunt work. We need to plan lessons topic by topic, measure progress assessment by assessment, and track performance grade by grade, in order to narrow the achievement gap using our scarce resources and ever more precious time.

Acknowledgements

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.

Resources

- The complete set of analytical tools and alignment manuals described in this article will be available at http://www.psy.cmu.edu/lessonplans
- The complete set of released TIMSS items and benchmarks from 1995 to 2003 are available at http://timss.bc.edu. The complete set of released NAEP items, scoring sheets, and benchmarks from 1996 and 1999 are available at http://nces.ed.gov/nationsreportcard/itmrls/.

References

American Association for the Advancement of Science (1990). *Science for all Americans*. New York: Oxford University Press.

American Association for the Advancement of Science (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Anderson, C. W. (2004). Science education research, environmental literacy, and our collective future. *NARST News*, 47 (2). National Association for Research in Science Teaching.

Anderson, R. D., and J. V. Helms. (2001). The ideal of standards and the reality of schools: Needed research. *Journal of Research in Science Teaching*, 38 (1), 3-16.

Bauer, H. (1992). *Scientific literacy and the myth of the scientific method*. Urbana & Chicago: University of Illinois Press.

Barton, A. C. (1998). Reframing "science for all" through the politics of poverty. *Educational Policy*, 12, 525-541.

Black, P., and D. William. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappa*, 80 (2), 139-148.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives Handbook 1: Cognitive Domain*. New York: Longman, Green & Co.

Chen, Z. and D. Klahr. (1999) All Other Things being Equal: Children's Acquisition of the Control of Variables Strategy. *Child Development*, 70 (5), 1098 - 1120.

CTB/McGraw-Hill (2001). *Terra Nova CAT complete battery plus level 15, form C*. Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill (2001). *Terra Nova CAT complete battery plus level 16, form C*. Monterey, CA: CTB/McGraw-Hill.

Donmoyer, R. (1995). The rhetoric and reality of systemic reform: a critique of the proposed National Science Education Standards. *Theory into Practice*, 34 (1), 30-34.

Hewson, P., J. B. Kahle, K. Scantlebury, and D. Davies. (2001). Equitable science education in urban middle schools: Do reform efforts make a difference? *Journal of Research in Science Teaching*, 38, 1130-44.

International Association for the Evaluation of Educational Achievement (1998). *TIMSS science items: Released set for population 2 (seventh and eighth grades)*. Retrieved on September 16, 2004 from http://timss.bc.edu/timss1995i/TIMSSPDF/BSItems.pdf

Kesidou, S., and J. E. Roseman. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching*, 39(6), 522–549.

Klahr, D. and J. Li. (2005). Cognitive research and elementary science instruction: From the laboratory, to the classroom, and back. *Journal of Science Education and Technology*, 4, 217-238.

Klahr, D. and M. Nigar. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, 15(10), 661-667

Klahr, D., Z. Chen, and E. Toth. (2001). Cognitive development and science education: Ships passing in the night or beacons of mutual illumination? In Carver, S. M. and Klahr D. (Eds.) *Cognition and Instruction: 25 years of progress*. Mahwah, NJ: Lawrence Erlbaum Associates.

Li, J. and D. Klahr. (2005). The psychology of scientific thinking: Implications for science teaching and learning. In J. Rhoton & P. Shane (Eds.) *Teaching Science in the 21st Century.* National Science Teachers Association and National Science Education Leadership Association: NSTA Press.

National Center for Education Statistics (2002). *Qualification of the public school teacher workforce: Prevalence of out-of-field teaching 1987-88 to 1999-2000*.

National Center for Education Statistics (n.d.). *The nation's report card (NAEP): 1996 assessment science public release grade 4*. Retrieved on September 16, 2004 from http://nces.ed.gov/nationsreportcard/itmrls/sampleq/96sci4.pdf.

National Center for Education Statistics (n.d.). *The nation's report card (NAEP): 1996 assessment science public release grade 8*. Retrieved on September 16, 2004 from http://nces.ed.gov/nationsreportcard/itmrls/sampleq/96sci8.pdf.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

Olson, L. (2003). Standards and tests, keeping them aligned. *Research Points*, 1 (1) 2003. Washington, D.C.: American Educational Research Association.

Pennsylvania Department of Education (2002). *Academic standards for science and technology* and *Academic standards for environment and ecology*.

Rodriguez, A. (1997). The dangerous discourse of invisibility: A critique of the National Research Council's national science education standards. *Journal of Research in Science Teaching*, 34, 19-37.

Schmidt, W. H., C. C. McKnight, and S. A. Raizen. (1997). *A splintered vision: an investigation of U.S. science and mathematics education.* Boston/Dordrecht/London, Kluwer Academic Press.

Settlage, J. and L. Meadows. (2002). Standards-Based Reform and Its Unintended Consequences: Implications for Science Education within America's Urban Schools. *Journal of Research in Science Teaching*, 39, 114-127.

Shamos, M. H. (1995). *The myth of scientific literacy*. New Brunswick, NJ: Rutgers University Press.

Shiland, TW (1998). The atheoretical nature of the national science education standards. *Science Education*, 82, 615-617

Stern, L., and A. Ahlgren. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching*, 39(9), 889–910.

Stern, L., and J. E. Roseman. (2004). Can middle-school science textbooks help students learn important ideas? Findings from Project 2061's curriculum evaluation study: Life science. *Journal of Research in Science Teaching*, 41(6), 538–568.

Stigler, J.W., and J. Hiebert. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.

Strand Cary, M. and D. Klahr. (2005). Two roads diverged in the classroom, but did it make a difference? Path independence in learning & transfer. (Cognitive Development Society, Biennial Meeting, 2005, San Diego, CA)

Thomas B. Fordham Institute. (2000). *The state of state standards.*

Thomas B. Fordham Institute. (2005). *The state of state science standards.*

Toth, E., D. Klahr, and Z. Chen. (2000). Bridging research and practice: A cognitively-based classroom intervention for teaching experimentation skills to elementary school children. *Cognition & Instruction*, 18(4), 423-459.

Triona, L. M. and D. Klahr. (2003). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition & Instruction*, 21, 149-173.

Vesilind, E.M. and M.G. Jones. (1998). Gardens or graveyards: Science education reform and school culture. *Journal of Research in Science Teaching*, 35 (7), 757-775.

Wolk, R. A. (1999). Making mid-course correction in standards-based reform. *1999 National Education Summit Briefing Book*. Achieve Inc.

Wolk, R. A. (2004). Perspective: Way off course. *Teacher Magazine*, 6 (2), 5

Junlei Li is a postdoctoral fellow at the Department of Psychology at Carnegie Mellon University. He investigates the practical implications of educational policy and cognitive research by co-teaching with teachers in urban school science classrooms. Correspondence concerning this article can be sent to junlei@andrew.cmu.edu.

David Klahr is professor of Psychology at Carnegie Mellon University. He has written many articles and books on the analysis of complex cognitive processes in such diverse areas as voting behavior, college admissions, consumer choice, peer review, problem solving, and scientific reasoning, and has more recently worked on how to better teach children to design and interpret simple experiments.

Stephanie Siler is a postdoctoral research fellow at the Department of Psychology, Carnegie Mellon University. She investigates motivation, instruction, and conceptual development in science learning.