FROM NATIONAL MOVEMENT TO LOCAL ACTION

The Status of Standards-Based Science Instruction in Middle School Classrooms

Christopher B. Swanson
The Urban Institute

Stephen B. Plank
Johns Hopkins University

Gina M. Hewes
Johns Hopkins University

Report No. 64

August 2003

This report was published by the Center for Research on the Education of Students Placed At Risk (CRESPAR), a national research and development center supported by a grant (No. R117-D40005) from the Institute of Education Sciences (IES, formerly OERI), U.S. Department of Education. The content or opinions expressed herein do not necessarily reflect the views of the Department of Education or any other agency of the U.S. Government. Reports are available from: Publications Department, CRESPAR/Johns Hopkins University; 3003 N. Charles Street, Suite 200; Baltimore MD 21218. An on-line version of this report is available at our web site: *www.csos.jhu.edu*.

THE CENTER

Every child has the capacity to succeed in school and in life. Yet far too many children fail to meet their potential. Many students, especially those from poor and minority families, are placed at risk by school practices that sort some students into high-quality programs and other students into low-quality education. CRESPAR believes that schools must replace the "sorting paradigm" with a "talent development" model that sets high expectations for all students, and ensures that all students receive a rich and demanding curriculum with appropriate assistance and support.

The mission of the Center for Research on the Education of Students Placed At Risk (CRESPAR) is to conduct the research, development, evaluation, and dissemination needed to transform schooling for students placed at risk. The work of the Center is guided by three central themes, ensuring the success of all students at key development points, building on students' personal and cultural assets, and scaling up effective programs, and conducted through research and development programs in the areas of early and elementary studies; middle and high school studies; school, family, and community partnerships; and systemic supports for school reform, as well as a program of institutional activities.

CRESPAR is organized as a partnership of Johns Hopkins University and Howard University, and is one of twelve national research and development centers supported by a grant (R117-D40005) from the Institute of Education Sciences (IES, formerly OERI) at the U.S. Department of Education. The centers examine a wide range of specific topics in education including early childhood development and education, student learning and achievement, cultural and linguistic diversity, English language learners, reading and literacy, gifted and talented students, improving low achieving schools, innovation in school reform, and state and local education policy. The overall objective of these centers is to conduct education research that will inform policy makers and practitioners about educational practices and outcomes that contribute to successful school performance.

ABSTRACT

This study contributes to the growing body of research on classroom instruction by exploring the possibility of measuring a specific instructional strategy using statistical methods based on item response theory (IRT). We seek to measure teachers' instructional practices using the same rigorous statistical techniques that are now applied to most large-scale assessments of student achievement. We focus specifically on teachers' use of pedagogical techniques consistent with those espoused by the national standards-based reform movement in science. We use data for a nationally-representative sample of public school eighth graders and their teachers from the 1996 National Assessment of Educational Progress (NAEP).

Since NAEP is an omnibus study of student achievement, this database does not offer detailed information regarding the specific reform programs and interventions in which the sampled schools might be actively engaging. Consequently, we are not able to directly examine the relationship between engaging in such explicit reform efforts and implementing a standards-based model of instruction. From a methodological perspective, however, a direct link to reform activities is not necessary. Instead this study takes advantage of the naturally existing variation in the classroom practices of a national sample of teachers and the features of IRT measurement models to address several important questions. First, do we find evidence that a coherent style of instruction akin to the standards-based model actually exists in practice in middle school classrooms? If so, which specific practices appear to be part of, and which are inconsistent with, this standards-based approach? Is there evidence that these practices tend to be incorporated into classroom instruction in a systematic manner or even in a predictable order? Finally, how prevalent is a standards-based approach to science instruction for this national sample of teachers?

This study has an admittedly methodological focus. However, we believe that the kind of solid measurement strategy being explored in this study represents an essential foundation and necessary precursor to subsequent studies of a more substantive and policy-relevant nature. Before it is possible to explore inequities in access to standards-based instructional opportunities for students who are educationally or socio-economically at risk, or to examine the impacts of standards-based instruction on student achievement, it is necessary to first devise a methodologically rigorous strategy for measuring the use of standards-based instruction.

ACKNOWLEDGMENTS

The first author completed much of his work on this project while holding a Spencer Foundation postdoctoral fellowship at Johns Hopkins University.

The authors thank Dr. John Smithson for his thoughtful feedback on an earlier version of this report, as well as Mary Maushard and Barbara Colton for their critical reading and skillful editing.

TABLE OF CONTENTS

INTRODUCTION

Virtually all contemporary educational reform efforts espouse improving student learning and raising student achievement levels as a central goal. There are, of course, numerous ways in which this goal could be promoted and, ideally, attained. Accordingly, advocates for change have devised and implemented many such strategies, and have met with varying degrees of success. Since the time of the Great Society programs of the late 1960s, most models of educational reform can probably be described as *regulatory* in nature. That is, these are strategies that attempt to promote fundamental educational change (e.g., improved student achievement) by monitoring and manipulating system inputs and outputs, broadly defined. Examples of such regulatory approaches might include compensatory funding and the more general redistribution of fiscal and other resources, site-based management, improving teacher quality, and performance-based accountability. Although these reforms differ greatly in many respects, they do share at least one thing. Namely, they seek to promote change in ways that do not directly address the core internal processes of schooling—teaching and learning. (See Swanson, forthcoming, for a more extended discussion of process and regulatory reforms.)

There are two main reasons why the schooling process, and in particular classroom instruction, has come to be treated as a "black box" from the point of view of regulatory reforms. First, advocates of input-output oriented reform models often contend that real change can, in fact, occur in school systems without fundamentally altering the process of teaching and learning. This would assume, for instance, that teachers already possess the resources and professional knowledge necessary to deliver high-quality instruction. The primary hurdles to implementing such effective instruction, however, are associated with insufficient or inefficiently distributed resources (which would suggest a compensatory solution) or a lack of motivation to break out of routine, but ineffective, practices (for which stronger accountability might serve as a catalyst).

A second reason that educational process has often been sidestepped by regulatory approaches relates to the issue of local autonomy over schooling. Here it might be acknowledged, at least implicitly, that an insufficient knowledge base regarding effective instructional practices contributes to poor student achievement and learning disparities among groups of students. Proponents of local autonomy, however, would contend that because schooling conditions and educational challenges vary so dramatically from one district, school, or even classroom to another, devising a uniform or one-size-fits-all remedy is not a feasible (or even desirable) solution. Rather, specific changes to classroom instruction and other fundamental schooling activities are best left to be designed and implemented by local decision-makers. Detailed visions of reformed instruction, consequently, tend not to be explicitly elaborated in regulatory models of reform.

During the 1990s, process-oriented approaches to school improvement such as standards-based and systemic reform came to occupy an increasingly prominent place in national educational reform agendas. As discussed in more detail later in this report, these strategies were founded in part on new visions of the learning process that revolved around the acquisition of higher-order thinking skills. Standards-based reform advocates argued that the key step in promoting such learning is a fundamental reformation of the teaching process involving both academic content and the pedagogical techniques through which that content is presented to students. While additional resources and stronger accountability might facilitate student achievement gains, such regulatory measures in themselves are unlikely to be a sufficient driver for change in the absence of meaningful instructional reforms.

Efforts to study this latest, standards-based generation of reforms face some substantial methodological challenges. Evaluating the impact of a reform or other intervention, for instance,

requires that analysts be able to effectively measure both the putative causal factor (e.g., instruction) and the outcome of interest (e.g., student achievement). Highly sophisticated psychometric methods for measuring academic proficiency have been available for decades and remain the object of considerable technical innovation. The state-of-the-art for characterizing the process of classroom instruction, however, is at a much earlier stage of development. Much of the existing work in this area has been strongly influenced by international educational comparisons and focuses on the issue of curricular content (see, Council of Chief State School Officers (CCSSO), 2000; Porter & Smithson, 2001; Schmidt, 2001; Schmidt, McKnight, Valverde, Houang, & Wiley, 1997). More recently, greater attention has turned towards teacher activities and instructional practices, with some analysts beginning to address the methodological issues involved in attempting to empirically characterize coherent approaches to teaching (for example, see Stecher et al., 2002; Swanson & Stevenson, 2002).

This study attempts to contribute to this growing body of research on classroom instruction by exploring the possibility of measuring a specific instructional strategy using statistical methods based on item response theory (IRT). In effect, we measured teachers' instructional practices using the same rigorous statistical techniques that are now applied to most large-scale assessments of student achievement. We focused specifically on teachers' use of pedagogical techniques consistent with those espoused by the national standards-based reform movement in science. This vision of reformed instruction is described in detail below. We obtained data for a nationally representative sample of public school eighth graders and their teachers from the 1996 National Assessment of Educational Progress (NAEP). Since NAEP is an omnibus study of student achievement, this database does not offer detailed information regarding the specific reform programs and interventions in which the sampled schools might be actively engaging. Consequently, we did not directly examine the relationship between engaging in such explicit reform efforts and implementing a standards-based model of instruction.

From a methodological perspective, however, a direct link to reform activities is not necessary. Instead this study took advantage of the naturally existing variation in the classroom practices of a national sample of teachers and the features of IRT measurement models to address several important questions. First, do we find evidence that a coherent style of instruction akin to the standards-based model actually exists in middle school classrooms? If so, which specific practices appear to be part of, and which are inconsistent with, this standards-based approach? Is there evidence that these practices tend to be incorporated into classroom instruction in a systematic manner or even in a predictable order? Finally, how prevalent is a standards-based approach to science instruction in this national sample of teachers? This study has an admittedly methodological focus. However, we believe that the kind of solid measurement strategy being explored in this study represents an essential foundation, a necessary precursor to subsequent studies of a more substantive and policy-relevant nature. Before it is possible to explore inequities in access to standards-based instructional opportunities for students who are educationally or socio-economically at risk or to examine the impacts of standards-based instruction on student achievement, it is necessary to devise a methodologically rigorous strategy for measuring the use of standards-based instruction.

THE STANDARDS MOVEMENT

The standards movement in American education has been a focus of public and scholarly attention for more than a decade (Borman, Cookson, Sadovnik, & Spade, 1996; McLaughlin & Shepard, 1995; Ravitch, 1995; Smith & O'Day, 1991). In its most general form, standards-based reform seeks to improve student achievement by setting rigorous expectations for performance in academic subjects

(McLaughlin & Shepard, 1995). Proponents argue that school systems, teachers, and students are best able to achieve at high levels when clear and challenging expectations for performance exist and are supported by integrated curricular content, instructional strategies, assessment techniques, professional development for teachers, and systemic coordination.

Standards-based reform efforts grew out of concern about the low level of achievement demonstrated by U.S. students, particularly compared to their peers in other industrialized nations (Wattenberg, 1995-96). Swanson and Stevenson (2002), for instance, describe a sequence of changes in the policy arena that followed the release of the influential report, *A Nation At Risk,* by the National Commission on Excellence in Education in 1983. First, in reaction to the perceived failure of prior federal compensatory solutions and with a shift in the political climate favoring a diminished federal role, leadership in educational policy-making during the 1980s moved decidedly toward individual states. Strategies pursued by state-level actors emphasized developing minimum competency standards for student performance, more stringent requirements for high school graduation, and performance-based school accountability systems (Fuhrmann, Clune, & Elmore, 1988; Stevenson & Schiller, 1999).

While these state-led efforts were pursued with considerable energy and good intentions, they might be criticized for sharing a weakness with earlier educational initiatives of the 1970s and 1980s that sought to improve student performance by imposing funding formulas or formal regulatory requirements. That is, it can be argued that all of these efforts lacked a clear articulation of the educational process through which student outcomes were to be generated. They lacked an explication of how regulations, accountability systems, and resource allocations would translate into improved pedagogical practices and student learning within individual classrooms.

Reacting to this perceived weakness, various national professional organizations and advisory groups coalesced in the late 1980s and early 1990s to initiate what has come to be called the standards movement in education. In contrast to the more indirect mechanisms described above, standards-based reform is founded on a concrete model of educational practice with specific recommendations for the curricular content and instructional techniques within individual classrooms. As such, the standards movement seeks to bring about changes at that most local level of educational organization—the classroom.

Whether the tenets of the standards movement are well-founded, whether educators have the incentive and will to embrace them, and whether they can ultimately improve student achievement are all important questions. Regardless of the answers to these questions, it can be said with some certainty that the hallmark of a standards-based approach to reform is its attempt to improve student learning by changing the core productive technologies of schooling—the academic content and instructional practices of classrooms.

Another important conceptual and empirical question is how (or whether) a vision of reform that began mainly at a national level can maintain its basic tenets and integrity as it filters through the various levels and branches of educational organization. National professional organizations and advisory panels such as the National Council of Teachers of Mathematics, the National Research Council (of the National Academy of Sciences), the National Council of Teachers of English, the National Center for History in the Schools, and the National Council for the Social Studies were among the groups that took on the daunting task of drafting standards for various age or grade levels in the various subject areas. These listed groups are just a subset of those that have attempted to create national educational standards or guidelines. Additionally, there are other organizations, such as Project 2061 of the American Association for the Advancement of Science (AAAS), that do not necessarily use the word "standards" to describe their contributions or developments, but that clearly have been influential in the movement.

In response to vigorous advocacy on the part of national actors, states (and sometimes school districts) have adapted the tenets of standards-based reform to serve the needs of their own educational constituencies. Although state and local approaches have varied considerably in the strength of their responses and their specific routes to high standards, five coordinated policy drivers can be described as comprising the core of curriculum-driven agendas for change:

a. *Content Standards* — detailed statements of what a student is expected to know or to have experienced after participating in a particular course of study,

b. *Performance Standards* — statements about levels of mastery students should be able to demonstrate over this content,

c. *Instructional Standards* — statements about the educational philosophies and activities that will be most effective in guiding students through an inquiry-based process of learning,

d. *Assessment Standards* — frameworks for measuring progress toward meeting content and performance goals (to be used both formatively and summatively), and

e. *Professional Standards* — training and certification requirements to ensure that teachers are sufficiently skilled as both pedagogists and subject-matter specialists to guide students through a challenging and engaging curriculum.

THE CASE OF SCIENCE

Each subject area—whether mathematics, science, English, history, social studies, or another—has had its own encounters with, and evolution through, the standards movement. In this report, we are focusing on science education. While a coherent and widely accepted vision for standards-based reform did not emerge as quickly for science as it did for mathematics (National Council of Teachers of Mathematics (NCTM), 1989, 1991), several prominent groups worked throughout much of the 1990s to develop a new paradigm for science education. In the following description, we focus mainly on developments and noteworthy publications that had appeared as of 1996. We impose this timeframe on our description to maintain consistency with the data upon which we are reporting in this study, the 1996 administration of the National Assessment of Educational Progress (NAEP) in Science.

First, the American Association for the Advancement of Science's Project 2061 has been highly influential. The first publication of this working group was *Science for All Americans*, released in 1989 (Project 2061, 1989). That report offers a detailed statement about what constitutes adult science literacy. In this statement, the report outlines what all students should know and be able to do in science, mathematics, and technology by the time they graduate from high school.

A companion document, *Benchmarks for Science Literacy*, appeared four years later (Project 2061, 1993). *Benchmarks* makes statements about how students should progress toward the ultimate goal of science literacy. The document's recommendations are organized by topics and subtopics, and also by grade levels. For example, a chapter on "the human organism" is further divided into sections on human identity, human development, basic functions, learning, physical health, and mental health. Within each of these sections, benchmarks are specified for each of the following grade spans: Kindergarten through Grade 2, Grades 3 through 5, Grades 6 through 8, and Grades 9 through 12. The benchmarks are described as thresholds—that is, levels of understanding that all students should have reached by a given grade in the course of a more extended progression toward science literacy.

To illustrate, the section on basic functions of the human organism offers six benchmarks for students in Grades 6 through 8. Two of these are as follows:

- By the end of the 8th grade, students should know that hormones are chemicals from glands that affect other body parts. They are involved in helping the body respond to danger and in regulating human growth, development, and reproduction.
- By the end of the 8th grade, students should know that interactions among the senses, nerves, and brain make possible the learning that enables human beings to cope with changes in their environment. (Project 2061, 1993, p.137)

By its authors' own admission, *Benchmarks* sheds only partial light on how to achieve the goals it recommends. The Project 2061 staff wanted to encourage common goals without dictating uniform curricula, teaching methods, and instructional materials. Publications released by Project 2061 subsequent to *Benchmarks* have offered more concrete recommendations for professional development, assessment, and instruction (Project 2061, 1997, 1998, 2001). Nevertheless, the Project 2061 staff clearly advocates for state and local educators to make their own decisions about curriculum and instructional strategies as they design "learning experiences for students that take into account state and district requirements, student backgrounds and interests, teacher preferences, and the local environment" (Project 2061, 1993, p. XII).

The National Research Council (NRC) was heavily influenced by the project's work as it developed the *National Science Education Standards*, or *NSES* (National Research Council (NRC), 1996, p. 15). In the introduction to their 1996 document, the NRC authors stated:

> The many individuals who developed the content standards sections of the *National Science Education Standards* made independent use and interpretation of the statements of what all students should know and be able to do that are published in *Science for All Americans* and *Benchmarks for Science Literacy*. The National Research Council of the National Academy of Sciences gratefully acknowledges its indebtedness to the seminal work by the American Association for the Advancement of Science's Project 2061 and believes that use of *Benchmarks for Science Literacy* by state framework committees, school and school-district curriculum committees, and developers of instructional and assessment materials complies fully with the spirit of the content standards.

Thus, the NRC made explicit use of Project 2061's recommendations as it articulated its own content standards for science education. A third-party evaluation of Project 2061 has also noted the influence of *Science for All Americans* and *Benchmarks for Science Literacy* on the *NSES* (Zucker, Young, & Luczak, 1996).

Despite the direct line of influence, however, differences in focus and specificity can be found when one compares the products of Project 2061 with the *NSES*. Most obvious is the fact that the *NSES* offers a vision that goes well beyond science content. The *NSES* document offers explicit standards for each of the following parts of the educational endeavor: teaching, professional development, assessment, content, education program (focusing on school-wide and district issues), and education system (meaning the broader infrastructure responsible for providing schools with necessary financial and intellectual resources).

Tables 1 through 4 summarize some emphases and activities encouraged by the *National Science Education Standards*. These tables respectively summarize the *NSES* vision for teaching, assessment, content, and the promotion of inquiry-based learning. In each case, encouraged practices are explicitly contrasted with discouraged practices.

Table 1. Tenets of a standards-based approach science education: Teaching standards

As articulated by the *National Science Education Standards*, Teaching Standards encompass...

Less emphasis on:	*More emphasis on:*
Treating all students alike and responding to the group as a whole	Understanding and responding to individual student's interests, strengths, experiences, and needs
Rigidly following curriculum	Selecting and adapting curriculum
Focusing on student acquisition of information	Focusing on student understanding and use of scientific knowledge, ideas, and inquiry processes
Presenting scientific knowledge through lecture, text, and demonstration	Guiding students in active and extended scientific inquiry
Asking for recitation of acquired knowledge	Providing opportunities for scientific discussion and debate among students
Testing students for factual information at the end of the unit or chapter	Continuously assessing student understanding
Maintaining responsibility and authority	Sharing responsibility for learning with students
Supporting competition	Supporting a classroom community with cooperation, shared responsibility, and respect
Working alone	Working with other teachers to enhance the science program

Source: *National Science Education Standards* (NRC, 1996), p.52.

Table 2. Tenets of a standards-based approach science education: Assessment standards

As articulated by the *National Science Education Standards*, Assessment Standards encompass...

Less emphasis on:	*More emphasis on:*
Assessing what is easily measured	Assessing what is most highly valued
Assessing discrete knowledge	Assessing rich, well-structured knowledge
Assessing scientific knowledge	Assessing scientific understanding and reasoning
Assessing to learn what students do not know	Assessing to learn what students do understand
Assessing only achievement	Assessing achievement and opportunity to learn
End of term assessments by teachers	Students engaged in ongoing assessment of their work and that of others
Development of external assessment by measurement experts along	Teachers involved in the development of external assessments

Source: *National Science Education Standards* (NRC, 1996), p.100.

Table 3. Tenets of a standards-based approach science education: Content standards

As articulated by the *National Science Education Standards*, Content Standards encompass...

Less emphasis on:	*More emphasis on:*
Knowing scientific facts and information	Understanding scientific concepts and developing abilities of inquiry
Studying subject matter disciplines (physical, life, earth sciences) for their own sake	Learning subject matter disciplines in the context of inquiry, technology, science in personal and social perspectives, and history and nature of science
Separating science knowledge and science process	Integrating all aspects of science content
Covering many science topics	Studying a few fundamental science concepts
Implementing inquiry as a set of processes	Implementing inquiry as instructional strategies, abilities, and ideas to be learned

Source: *National Science Education Standards* (NRC, 1996), p.113.

Table 4. Tenets of a standards-based approach science education: Toward inquiry-based learning

As articulated by the *National Science Education Standards*, efforts to Promote Inquiry encompass...

Less emphasis on:	*More emphasis on:*
Activities that demonstrate and verify science content	Activities that investigate and analyze science questions
Investigations confined to one class period	Investigations over extended periods of time
Process skills out of context	Process skills in context
Emphasis on individual process skills such as observation or inference	Using multiple process skills -- manipulation, cognitive, procedural
Getting an answer	Using evidence and strategies for developing or revising an explanation
Science as exploration and experiment	Science as argument and explanation
Providing answers to questions about science content	Communicating science explanations
Individuals and groups of students analyzing and synthesizing data without defending a conclusion	Groups of students often analyzing and synthesizing data after defending conclusions
Doing few investigations in order to leave time to cover large amounts of content	Doing more investigations in order to develop understanding, ability, values of inquiry and knowledge of science content
Concluding inquiries with the results of the experiment	Applying the results of experiments to scientific arguments and explanations
Management of materials and equipment	Management of ideas and information
Private communication of student ideas and conclusions to teacher	Public communication of student ideas and work to classmates

Source: *National Science Education Standards* (NRC, 1996), p.113.

DATA ON TEACHER INSTRUCTIONAL PRACTICES

The purpose of this report is to empirically explore the degree to which the kinds of educational values or goals described above as central tenets of the national standards-based reform movement have been applied as a coherent instructional strategy in actual middle school science classrooms. To accomplish this, we draw data from the National Assessment of Educational Progress (NAEP), the only continuing and representative source of data for the academic proficiency of students in the United States in a variety of subject areas. Congressionally authorized and established in 1969, the NAEP program is administered by the National Center for Education Statistics (NCES) of the U.S. Department of Education. NAEP is the source of the department's familiar "Nation's Report Card" publications, which report on U.S. students' performance in a variety of subjects.

The main components within the NAEP program are the Long-Term Trend Assessment and a Main Assessment consisting of separate national and state data collections. In addition to conducting student assessments, NAEP collects background surveys from participating students, teachers, and schools that can be used to characterize the broader schooling environment in which learning occurs. Detailed information on the design of NAEP can be found in technical documentation prepared by the National Center for Education Statistics (1999, 2000). This study employs data from the 1996 National NAEP Assessment in Science for the eighth grade. The National NAEP employs a multi-stage probability sampling design to draw a nationally representative sample of eighth-grade students (see NCES, 2000, for details). The science teachers of the students sampled in NAEP are asked to provide information on a variety of instructional practices and pedagogical techniques used in their classrooms.

To empirically study reform-oriented modes of science education, we must begin by identifying a pool of items that captures classroom instructional practices aligned conceptually with recognized models of standards-based science, such as those articulated by AAAS and NRC, as described above. The Long-Term Trend component of NAEP has remained essentially unchanged since its inception in 1969 to reliably track trends in student performance levels over time against a constant standard of comparison. The two components of the more forward-looking Main Assessment of NAEP, however, have been periodically redesigned to capture evolving trends in national reform and educational priorities, as well as changing understanding about best educational practices. A major wave of redesigns for the NAEP survey program was initiated in the early 1990s by the National Assessment Governing Board (NAGB), an independent body with oversight of NAEP. Since this time, the designs of both the student assessments and the content of teacher background questionnaires in science have been strongly shaped by the vision of science literacy and instruction articulated by the standards movement. A similar, perhaps even stronger, influence of the standards movement on NAEP can also be seen in the case of mathematics (Kenney & Silver 1997; National Association of Educational Progress (NAEP), 1988; National Assessment Governing Board (NAGB), 1999).

With respect to science, NABG states that the redesigned NAEP framework "takes into account current reforms in science education (2000: 10)" and cites the work of AAAS and NRC, among other leading reform advocates. The vision of contemporary science education that NABG drew upon includes many of the tenets of standards-based science described above. It includes an "emphasis on development of such thinking processes as organizing factual knowledge around major concepts, defining and solving problems, accessing information and reasoning with it, and communicating with others about one's science results and understandings (2000:10)." The design of the teacher background surveys is also intended to capture the kinds of instructional activities needed to enact this broader vision of science education, which would encompass "approaches that encourage active student involvement and participation, such as participating in hands-on science

activities; learning in small, cooperative groups; reflecting orally and in writing upon experiences; and completing sustained projects" (2000:10).

Table 5: Items about Science Instructional Practices from the 1996 NAEP Eighth-Grade Teacher Surveys

Label	Item Description (*Response Categories*)
	About how often do your science students do each of the following? (*never or hardly ever / once or twice a month / once or twice a week / almost every day*)
HOTB	Read a science textbook
HORB	Read a book or magazine about science
HODN	Discuss science in the news
HOOS	Work with other students on a science activity or project
HOOR	Give an oral science report
HOWR	Prepare a written science report
HOHO	Do hands-on activities or investigations in science
HOTK	Talk about measurements and results from students' hands-on activities
HOTS	Take a science test or quiz
HOLI	Use library resources for science
HOCO	Use computers for science
	When you teach science, about how often do you do each of the following? (*never or hardly ever / once or twice a month / once or twice a week / almost every day*)
DMTK	Talk to the class about science
DMDM	Do a science demonstration
DMVT	Show a science videotape or science television program
DMCO	Use computers for science (e.g., science software, telecommunications)
DMCD	Use CD's or laser disks on science
SCFT	**About how often do your science students go on a science field trip?** (*never or hardly ever / 1 or 2 times a year / 3 or more times a year*)
SCGU	**About how often do you bring a guest science speaker to talk to your science students?** (*never or hardly ever / 1 or 2 times a year / 3 or more times a year*)
WKPF	Do you save your students' science work in portfolios for assessment (*no / yes*)
WKPJ	**Do you ever assign individual or group science projects or investigations in school that take a week or more?** (*no / yes*)
	Think about your plans for science instruction during the entire year. **About how much emphasis will you give to each of the following objectives for your students?** (*little or no emphasis / moderate emphasis / heavy emphasis /*)
EMFA	Knowing science facts and terminology
EMCN	Understanding key science concepts
EMPS	Developing science problem-solving skills
EMST	Learning about the relevance of science to society and technology
EMCM	Knowing how to communicate ideas in science effectively
EMLS	Developing laboratory skills and techniques
EMIN	Developing students' interest in science
EMDA	Developing data analysis skills
EMTC	Using technology as a scientific tool

(Table 5 cont.)

	How often do you use each of the following to assess student progress in science? (*never or hardly ever / once or twice a year / once per grading period / once or twice a month / once or twice a week*)
ASMC	Multiple-choice tests
ASWR	Short or long written responses (e.g., a phrase or sentence; or several sentences or paragraphs)
ASIP	Individual projects or presentations
ASGP	Group projects or presentations
ASPF	Portfolio collections of each student's work
ASEY	In-class essays
ASSE	Self-evaluations or peer evaluations
ASLB	Laboratory notebooks or journals
ASHW	Homework
ASHO	Hands-on activities
GRHO	**What proportion of a student's evaluation in science (final grade) is based on performance with hands-on activities?** (*none of the grade / very little of the grade / about half of the grade / most or all of the grade*)
	How do you use computers for instruction in science? (*no / yes*)
CODP	Drill and practice
COGA	Playing science/learning games
COSM	Simulations and modeling
CODA	Data analysis and other applications
COWP	Word processing
	As part of their work in this science class, do students produce any of the following records of their work? (*no / yes*)
RWNB	Notebooks or reports of laboratory work
RWPJ	Reports or other written records of extended science projects
RWIS	Written reports on specific topics or issues in science
RWFT	Reports or records of science field trips
RWJL	Journals, diaries, or logs of ideas about science or work done for science class
RWPH	Photographic or pictorial records of projects or other science activities
RWAV	Audiotape or videotape records of science activities
RWIN	Reports of personal interviews about science
RWMO	Three-dimensional scientific models
RWMM	Computer-generated multimedia science projects

The teacher survey items upon which we draw are in Table 5, which reports the survey question wording along with the item response categories. These practices capture signature elements of a standards-based approach to science education, including an emphasis on higher-order skills and reasoning, use of extended written assignments and hands-on science activities, incorporation of performance-based tasks into assessment methods, and more innovative and applied methods for teachers to deliver science knowledge to students. As will be discussed below, this pool of items also includes several questions that ask about more traditional approaches to instruction.

The descriptive analyses conducted for this study provide a statistical portrait of middle school science instruction nationwide. Here we treat students as the unit of analyses, with results weighted to be representative of the science instruction received by eighth graders across the nation.

Analyses are limited to public school students (N=7257). In the formal measurement analyses that follow, our objective is to explore the coherence of standards-based instruction as it exists in practice. In this case the relevant unit of analysis is classroom teachers and we based our analyses on the responses of the science teachers of sampled public school NAEP students. We have data for 547 such teachers. Most survey items reference teachers' general instructional practices across their eighth-grade science classes. In situations where questions ask teachers to provide information about instruction on a classroom-by-classroom basis, we assign teachers the highest level of instructional use reported across their classrooms. The nature of the student-centered NAEP sampling design does not ensure that this sample of teachers is statistically representative of all middle school science teachers in the nation. However, a sample of this size is likely to encompass much of the diversity in teacher backgrounds, working conditions, and approaches to instruction that exists across local school systems nationwide.

A DESCRIPTIVE PORTRAIT OF MIDDLE SCHOOL SCIENCE INSTRUCTION

Table 5 lists 55 items from the 1996 NAEP questionnaire administered to eighth-grade science teachers pertaining to their classroom instructional practices. These items are grouped according to the prompting questions or stems that preceded them in the questionnaire. For instance, the first series asked, "About how often do your science students do each of the following?" Eleven specific practices are listed under this prompt. The four response categories allowed teachers to report their frequency of usage as ranging from "never or hardly ever" to "almost every day." Other questions asked teachers about their use of other instructional activities, including the frequency of science field trips and guest speakers, the degree of emphasis given to various student objectives, particular methods of assessment, uses of computers in the classroom, and various records of academic work that students might produce.

This set of items represents virtually all questions that were asked of NAEP science teachers regarding assessment techniques, instructional practices, and areas of emphasis in terms of objectives for students. This questionnaire did not probe for detailed information about curricular content covered or materials used (e.g., particular science topics addressed or textbooks assigned). Therefore, in our analyses we focus not on curricular topics, but rather on the extent to which teachers' general emphases, instructional practices, and assessment techniques display a coherent structure across the sample of teachers. If a coherent structure is present, we then explore the organization or interrelation of these practices and make statements about whether this structure is consistent with the pedagogical style promoted by advocates of standards-based reform. Findings that confirm the existence of such standards-based instruction (SBI) could be viewed as preliminary evidence that the standards movement may be exerting influence upon the day-to-day activities in eighth grade. (In the context of this study, however, we do not attempt to link the use of SBI with specific policy or reform initiatives.) Finally, if a coherent structure consistent with the tenets of standards-based instruction is revealed, we are able to make statements about the distribution of teachers along a continuum of standards-based orientation.

Table 6: Science Instructional Practices, Arranged by Instructional Element

Instructional Practice (abbreviation)	Prevalence [a]
Teacher's emphasis on objectives for students	Heavy emphasis (%)
Key science concepts (EMCN)	88.5
Science problem-solving skills (EMPS)	68.7
Interest in science (EMIN)	68.4
Relevance of science to society and technology (EMST)	47.4
Laboratory skills and techniques (TMLS)	42.5
Communicating ideas in science effectively (EMCM)	42.3
Science facts and terminology (EMFA)	38.5
Data analysis skills (EMDA)	23.8
Technology as a scientific tool (TMTC)	14.7
Classroom Assessment	
	Monthly (%)
Frequency of tests or quizzes	96.6
Hands-on activities (ASHO)	92.8
Homework (ASHW)	89.7
Short or long written responses (ASWR)	87.5
Multiple-choice tests (ASMC)	76.0
Lab notebooks or journals (ASLB)	43.6
Individual projects or presentations (ASIP)	29.9
In-class essays (ASEY)	27.9
Group projects or presentations (ASGP)	27.7
Self-evaluation or peer-evaluation (ASSE)	18.2
Portfolios of student work (ASPF)	18.0
	Half or more (%)
Proportion of student final grade based on hands-on activities (GRHO)	33.6
	Yes (%)
Student work saved in portfolio for assessment (WKPF)	33.6
Written Assignments	
	Monthly (%)
How often students prepare written science report (HOWR)	68.8
	Yes (%)
Notebooks or reports of lab work (RWNB)	83.5
Reports on specific science topics or issues (RWIS)	65.8
Reports or records of extended projects (RWPJ)	62.4
Journals, diaries or logs of science ideas or class work (RWJL)	38.7
Student-Centered, Hands-On, or Inquiry-Based Activities	
	Weekly (%)
Hands-activities or investigations (HOHO)	82.6
Work with other students on activity or project (HOOS)	69.4
Talk about measurements or results of hands-on activities (HOTK)	67.1
Give oral science report (HOOR)	3.2

(Table 6, cont.)

	Yes (%)
Projects or investigations that take a week or more (WKPJ)	82.1
Three-dimensional scientific models (RWMO)	54.6
Photographic or pictorial records of projects (RWPH)	27.4
Audiotape or videotape records	15.5
Reports of personal interviews (RWIN)	10.3
Computer Use in the Classroom by Students	
	Weekly (%)
Frequency of computer use (HOCO)	7.3
	Yes (%)
Simulations and modeling (COSM)	26.3
Word processing (COWP)	21.1
Data analysis (CODA)	20.2
Playing science or learning games (COGM)	20.1
Computer generated multimedia projects	16.6
Drill and practice (CODP)	8.5
Teacher Initiated Delivery of Science Content	
	Weekly (%)
Talk to class about science (DMTK)	98.5
Read a science textbook (HOTB)	75.1
Do a science demonstration (DMDM)	59.2
Discuss science in the news (HODN)	48.4
Show science videotape or television program (DMTV)	23.0
Read a book or magazine about science (HORB)	16.1
Use CD's or laser disks (DMCD)	13.7
Use library resources (HOLB)	12.7
Use computers (e.g., software, telecommunications) (DMCO)	8.1
	Annually (%)
Bring guest speaker to talk to class (SCGU)	42.1
Go on science field trip (SCFT)	40.8

[a] Prevalence refers to the percentage of students experiencing the specified instructional practice in their eighth-grade science class. Annually= at least once or twice a year, Monthly = at least once or twice a month, Weekly = at least once or twice a week. See Table 5 for abbreviations and complete questionnaire item wording.
Source: 1996 National Assessment of Educational Progress in Science, Eighth grade. Data are weighted to produce nationally-representative estimates for students.

Before we describe the creation of an empirical scale of instructional practices using item response theory (IRT) methods, it is useful to gain a more basic understanding of teachers' responses to the survey items. What techniques and activities are used by these teachers with relatively high frequency? What techniques and activities are used only rarely? Can we identify items that we would expect to be quite central to a measure of standards-based instruction? Can we identify other items that are either ambiguously worded or, in fact, antithetical to expectations about standards-based instruction? Items that fall into these latter categories would be poor candidates for constructing an empirical scale for standards-based approaches to science instruction.

In Table 6 we examine the extent to which NAEP students experience these forms of instruction in their science classes. Here items have been grouped into a set of broader elements of classroom organization or activities. Although there would be other defensible ways of grouping these practices, this way offers one useful scheme. The items are arranged as follows:

- Items on the teachers' emphases on objectives for their students;
- Items on assessment;
- Items focused on written records of work produced by students, which do not explicitly mention assessment;
- Items on student-centered, hands-on, and/or inquiry-based learning activities;
- Items on the use of computers, which might be viewed as a specific form of hands-on activity;
- Items on teacher-initiated learning activities.

Within the first instructional element (teachers' emphases on various science objectives and skills), by far the most frequently reported area of heavy emphasis is "understanding key science concepts." The teachers of 88.5% of public school eighth graders students reported putting heavy emphasis on this objective. Advocates of the standards movement would support this focus, to the extent that it contrasts with an emphasis on the more mechanical learning of facts and terminology. The latter is heavily emphasized in the science classes of 38.5% of students. "Developing science problem-solving skills" and "developing students' interest in science" are the next most frequently reported areas of emphasis. These are fairly general and broadly encompassing objectives, and in this sense they are similar to "understanding key science concepts." Some of the more narrow, more specific, and more concrete areas of emphasis presented in the questionnaire were reportedly used considerably less frequently. For example, developing data analysis skills and using technology as a scientific tool were given heavy instructional emphasis for just 23.8% and 14.7% of students, respectively.

The second element of Table 6 involves the assessment practices used in the classroom. Teachers were asked how often their students take science tests or quizzes. In addition, they reported on the frequency with which they use 10 specific methods to assess student progress. Reflecting the questionnaire's strong attention to hands-on learning, teachers were also asked about the weight they placed on hands-on activities when assigning student grades. Finally, teachers were asked whether they saved their students' science work in portfolios for assessment.

Our descriptive results show that nearly all students (96.6%) take science tests or quizzes on a monthly (or more frequent) basis. About half of the students are tested weekly. The specific methods of assessment reportedly used with the greatest frequency were hands-on activities, homework, short or long written responses, and multiple-choice tests. At least three-fourths of the NAEP students experienced each of these techniques at least once a month. For instance, more than 92% of the students engaged in hands-on forms of assessment monthly. Since proponents of the standards movement call for inquiry-based learning and the encouragement of higher-order thinking, they would commend the frequent use of hands-on activities and written responses as assessment opportunities. In contrast, standards advocates probably would be concerned if multiple-choice tests were used too frequently and not in conjunction with other assessment methods. It is more difficult to state whether homework as assessment would be consistent with a standards-based approach. To make that kind of determination, more would need to be known about the format and substance of the homework. Unfortunately, the survey item about using homework for assessment is not specific and, as a result, has limited value for characterizing the pedagogical philosophy and activities of a given classroom.

Table 6 shows that for 60.9% of students at least half of their final grade is based on performance with hands-on activities. This fairly high percentage suggests that reliance on hands-on activities for assessment had become a widespread practice in eighth grade classrooms by 1996. Somewhat less prevalent was the compilation of students' science work in portfolios for assessment, an activity experienced by 33.6% of students. The creation of assessment portfolios is undeniably a time-consuming activity for teachers. We probably would expect a teacher to dedicate time and energy to this method of assessment only if he or she had developed a strong commitment to non-

traditional methods of evaluation or worked within a school culture that provided support for this activity.

The third element of Table 6 involves various records of student work that students might be assigned in their science classes. As they appear on the teacher questionnaire, these practices do not explicitly refer to use for assessment. The table shows that notebooks or reports of laboratory work are the most common form of written assignment, an activity in which 83.5% of eighth-grade science students engage. Writing reports on specific topics or issues in science or about extended science projects are also relatively common (65.8% and 62.4% of students, respectively). Produced with considerably less frequency were journals, diaries, or logs of ideas about science or work done for science class (38.7%) and reports or records of science field trips (13.6%). The prevalence of the three most commonly reported of these activities might suggest that a relatively great amount of writing was occurring in these classrooms. However, we know only whether or not these practices were being employed, not how often. So it would be more accurate to state that the majority of classrooms featured some writing opportunities over the course of a year. For most students, writing reports is probably not something that happens on a daily, or even weekly, basis. Although about two-thirds of students prepare written reports monthly, less than 10% do so once a week or more.

The fourth element of Table 6 deals with the classroom activities that are at the heart of a standards-based approach to science—student-centered, hands-on, or inquiry-based learning opportunities. These activities would seem to require that the teacher cede considerable control of topics for exploration or methods of inquiry to students. Some of these activities were reported to be used rather frequently, a result consistent with the goals of the standards movement and its call for deep engagement and student initiative in science learning. Among the most frequently featured of these activities were (generically) doing hands-on activities or investigations in science, working with other students on a science activity or project, and talking about measurements and results from students' hands-on activities. Fully 82.6% of students engage in hands-on activities at least once a week. A similar number of students also engage in extended science projects during the course of the school year. About two-thirds of students work collaboratively with their classmates to complete science projects or discuss the results of hands-on activities. Although most students engage in some form of hands-on science, certain kinds of activities appear to be encountered only rarely (e.g., compiling audio-visual records for projects, conducting interviews about science, or presenting oral reports).

Despite the standards movement's strong emphasis on incorporating technology into school science, student-centered activities involving computers were quite infrequent. Only 7.3% of NAEP students used computers once a week or more. We find that only about one-fifth to a quarter of students tend to have access to the more specific kinds of computer-based activities included on the teacher questionnaire (e.g., simulations, word processing, data analysis, and science or learning games). Even fewer engage in computer-based multimedia projects or use computers for drill and practice. The infrequency of the latter application of computers, however, would be rather consistent with the standards movement's negative view of learning through repetition.

The final element of Table 6 focuses on classroom activities that are generally directed by the teacher, and are less likely to be characterized as student-centered, hands-on, or inquiry-based. We find teacher-initiated discussion or lecture is ubiquitous in science classrooms (experienced on a weekly basis by 98.5% of students). Students also work from science textbooks with great regularity. These results suggest that most prominent methods of delivering science knowledge tend to be rather traditional. As suggested earlier, more interactive forms of student engagement and more varied learning materials are the ideal under a standards-based model of instruction. It should be noted, however, that because of the general framing of these survey questions, we do not have more detailed information such as the amount of time the teacher spends talking to students in a typical class

period, the extent to which this talk takes the form of lecture or monologue versus a more interactive dialog between teacher and students, or the emphasis placed on inquiry-based science in textbooks. In addition to these modes of communicating science content, the majority of students (59.2%) also experience demonstrations in their science class and about half discuss science topics in the news. Relatively infrequent ways of obtaining information for science class include reading a book or magazine about science, using library resources, watching a science videotape or television program, or using computer applications.

In general, Table 6 paints a picture of eighth-grade science instruction in 1996 as being diverse in the emphasized student skills, assessment techniques, and instructional settings and strategies. There is evidence of a fairly strong commitment to hands-on activities as modes of instruction and assessment. However, many of the most frequently used instructional practices are rather traditional in the sense of relying on standardized materials (e.g., textbooks) and teacher-centered modes of presenting information. The use of computers was also strikingly minimal, suggesting that instructional use of computers has yet (as of 1996) to make significant inroads into the science classroom. With that general overview as a starting point, we continue in our attempt to develop a statistical measurement model that will reveal whether a coherent underlying structure of science instruction exists across this sample of teachers and students.

THE ITEM-RESPONSE THEORY (IRT) APPROACH — A PRIMER

As the descriptive analysis of the preceding section testifies, we have at our disposal a rich source of data on several major aspects of middle-school science instruction. Although a handful of the practices about which the NAEP teachers report might be considered "traditional," most correspond quite closely to the tenets of standards-based reform as described above. This is not a mere coincidence. As noted earlier, the NAEP teacher questionnaire was designed with the expressed intention of capturing the use of teaching methods that reflect a contemporary vision of best practices in science education, a vision that has been strongly shaped by the work of groups such as AAAS and NRC. In this section, we proceed from a consideration of individual instructional practices to an analysis of broader pedagogical strategies. Specifically, we are interested in determining whether the collection of discrete practices described is coherent in a systematic way in teachers' everyday classroom instruction. For the remainder of the study, we focus on teachers as our unit of analysis.

We are essentially interested in determining whether it is possible to take information on the way teachers report engaging in a series of individual instructional practices and use it to construct a single, valid indicator that captures their overall use of standards-based science instruction as a coherent pedagogical style. This kind of construct development is commonly pursued in social scientific research for several reasons, particularly to summarize a large set of items thought to be related to the same underlying latent factor, and to generate an empirical measure that possesses higher reliability and less measurement error than the individual constituent items. Familiar approaches to such measure construction might range from creating a basic additive scale to factor analytic techniques. This study, on the other hand, adopts a measurement method most commonly used in large-scale cognitive assessments—item response theory or IRT (Embretson & Reise, 2000; Lord, 1980).

Because item response theory is a rather specialized and technically sophisticated method, a full explication of the underlying mathematics is beyond the scope of this report. In this section, we provide a brief primer on IRT with an eye toward describing features of the method that will be used later when we analyze and interpret the organization of standards-based instructional practices as

they have been implemented in science classrooms. In explaining our methodological strategy for measuring standards-based instruction (SBI), we draw heavily on the analogy of an individual testing situation, which is the most familiar application of the IRT method and the source of much of its terminology. Naturally, teachers' self-reports about the kinds of instructional practices they use can be likened to a cognitive test only up to a certain point. This comparison, nonetheless, offers a useful way to introduce several key IRT concepts.

The Basic Data Matrix

The object of any test is to obtain an estimate of an individual's level on some unmeasured, underlying trait. In a science assessment administered to individual students, for instance, the construct or latent trait being measured might be described as *science proficiency*. The empirical basis for this proficiency estimate in an IRT framework is a test-taker's pattern of correct and incorrect responses on a set of items or tasks that constitute the assessment. If a test is well-designed, that is, if the individual items all tap into the same underlying trait (science proficiency in this case), we can expect to find a systematic relationship between responses on the individual items and the student's overall score on the test. For instance, on a test with items of varying difficulty we would anticipate that a test-taker who gets a moderately difficult question correct will also be likely to answer easier items correctly. Thus, there is a progressive logic inherent to an IRT approach to measurement—an individual with a higher proficiency score will be increasingly likely to answer increasingly difficult test items correctly.

Table 7: Basic Data Matrix - Subtest for Emphasis on Standards-Based Values

		Raw Score					
		(low)					(high)
Item [a]	% correct [b]	0	1	2	3	4	5
EMCN	.85	.00	.85	.92	.96	.99	1.0
EMPS	.66	.00	.12	.82	.94	.99	1.0
EMCM	.39	.00	.04	.15	.67	.91	1.0
EMDA	.26	.00	.00	.07	.26	.84	1.0
EMTC	.15	.00	.00	.05	.17	.28	1.0

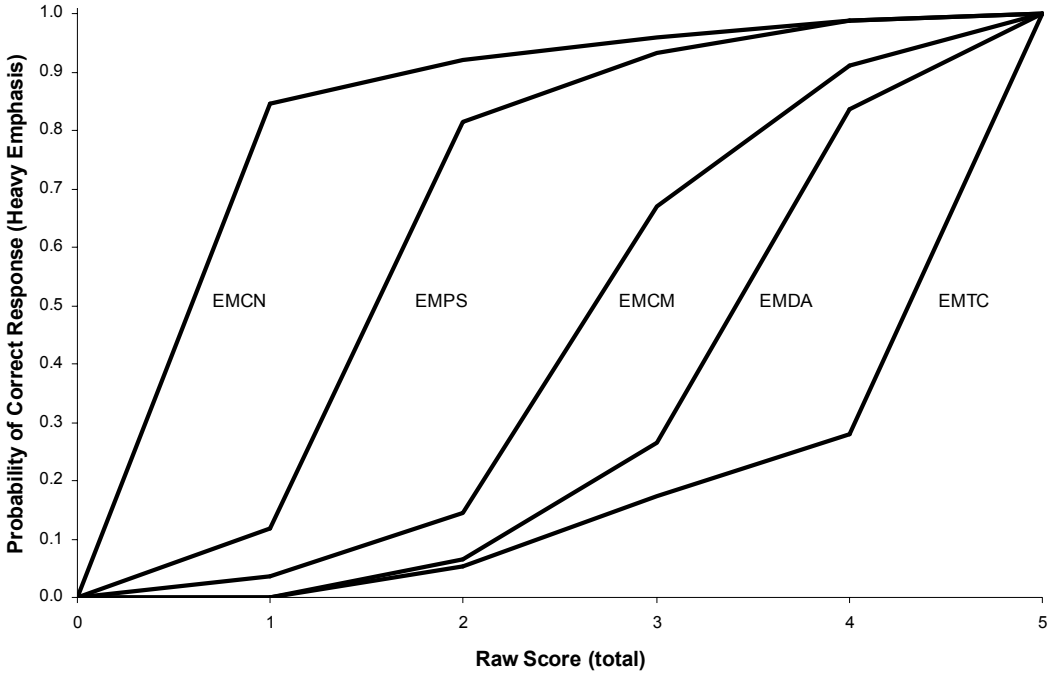
[a] See Table 5 for item label abbreviations.

[b] A correct item response is defined as "heavy" emphasis versus "moderate" or "little or no" emphasis on the respective instructional element. Cell values indicate the proportion of NAEP teachers at a give raw score level who provided a correct response to the respective item.

In this study, we are concerned with measuring a different kind of latent trait—a teacher's level of standards-based instruction in science (SBIS). Nevertheless, the analogy with cognitive testing should hold, as should the progressive nature of IRT noted above. To extend our analogy between testing and survey-based information on instructional practices, the "test-takers" are the NAEP science teachers, the "test problems" are questionnaire items about classroom practices, and item "difficulty" is typically expressed in terms of the frequency with which a teacher reports using a particular practice. A demonstration of the progressiveness property of IRT as it applies to instructional practices is presented in Table 7.

To illustrate this, we draw a set of five items from the NAEP teacher questionnaire that deals with instructional emphases consistent with standards-based reform. (For full descriptions of these items and their response categories see Table 5). We treat these items as if they were a short test. Teachers are assigned a score of either 1 (if they place a "heavy" emphasis on that instructional objective) or 0 (if they indicate a lesser amount of emphasis). These can be thought of as being analogous to "correct" and "incorrect" responses to test questions. For each teacher in the NAEP sample we calculate a raw score on this mini-test for standards-based instructional emphases by totaling up the number of correct responses, a score that may range from 0 to 5. This raw score represents a very basic trait estimate, the trait of interest here being standards-based emphases. A data matrix can then be constructed by arraying the five test items by the six possible raw scores. Test items are ranked from top to bottom in order of increasing difficulty. Item difficulty here is *inversely* proportional to the percentage of teachers who provided a correct response (i.e., who reported a heavy emphasis). So, fewer teachers are expected to answer the more difficult items correctly. Cell entries in the table indicate the percentage of teachers at a given raw score level who provided a "correct" response for the respective item. By definition, teachers that fall into the lowest and highest score groups, respectively, answered either none or all of the items correctly.

Figure 1: Illustrated Data Matrix for Emphasis Subtest - Empirical Item Response Curves (IRC)



Reading the data matrix from left to right across a row, we find that as the raw score for the test increases the probability of providing a correct item response also increases. This pattern holds true for all five items. Following a data column from top to bottom, we see that for teachers with a given test score the probability of a correct response consistently decreases as item difficulty increases. The same trend is observed within each of the non-extreme score categories. For instance, a teacher with a moderate raw score of 3 has a 96% chance of a correct response on the easiest item in this illustrative subtest (emphasis on concepts, EMCN) and a 17% chance of a correct response on the most difficult item (emphasis on technology, EMTC). This data matrix can also be depicted graphically, as seen in Figure 1.

We note that the easier items consistently approach a 100% correct response rate much more quickly (i.e., at lower test score levels) than do the more difficult items (cf. EMCN and EMTC). Each item, however, displays the same characteristic "S-shaped" curve with a trajectory that rises along with the raw score level. The kind of orderly relationship between trait level (raw score) and the probability of providing a correct response on items of increasing difficulty that we observe here signals a high level of correspondence between the test and the latent trait that this test is intended to measure. That is, the test items empirically define a coherent construct.

An Introduction to IRT Item Parameters

The example above illustrates some fundamental properties of a good test, a test that in our case consisted of teacher responses to five questions about their science instruction. In that exercise, the raw score total of correct items represents a very basic kind of constructed measure that summarizes a teacher's overall level on the estimated trait—in this case, standards-based emphases. IRT and other measurement techniques attempt to mathematically formalize this empirical relationship by estimating a statistical model that is intended to accurately capture the observed pattern of item responses. In the analyses presented later in this report, for instance, we attempt to construct a statistical model that estimates teachers' levels on a broader construct that encompasses a wider array of practices (i.e., a general standards-based style of science instruction). The formulation of an IRT model can be expressed in its most basic form as follows:

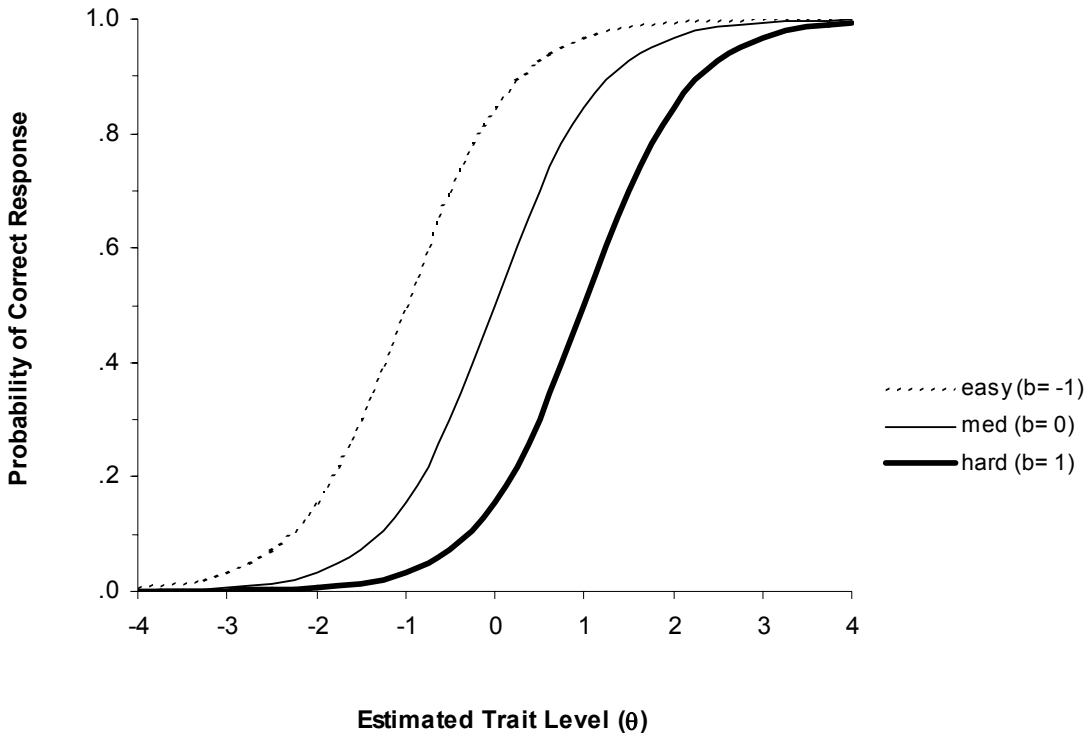
$$\textit{Item Response} = fn(\textit{Trait Level, Item Properties}). \qquad [1]$$

An individual's response to a particular item is defined here as a *function* of both his or her underlying *trait level* and a set of psychometric *item properties*. So, for instance, a person with a higher level of science ability (i.e., trait level) will be more likely to provide a correct answer (i.e., item response) to a difficult problem (i.e., item property) on a science test. The particular functional form the IRT expression takes is a logistic curve, which captures the expected S-shaped relationship between trait level and the probability of a correct item response, a pattern illustrated empirically in the example above (Figure 1). Two specific item properties are represented in our statistical models—difficulty and discrimination—making them two-parameter IRT models. The instructional items that we obtained from the NAEP teacher surveys have several different response formats. They range from dichotomous items indicating whether a practice is used (vs. not) to questions asking teachers to report the frequency with which they use a practice by selecting one of up to five ordered response categories. The presence of such multiple-category or polytomous response items will necessitate the use of a particular form of IRT analyses known as the partial credit model (Masters, 1982; Muraki, 1992, 1993). The properties of the item parameters for a two-parameter partial credit IRT model are described briefly below.

The Location Parameter (β)

The IRT location parameter reflects the relative difficulty of an item. Under the expectations of item response theory, an individual of a given trait level should have a higher probability of a correct response for an easier rather than for a more difficult item. Figure 2 provides illustrations of three item response curves (IRCs), which demonstrate the effects of changes in the value of this difficulty parameter (β). We should note that the format of this figure generally resembles that of the empirical item response curves shown above (Figure 1). The horizontal axis measures the estimated trait level (θ) on a standard normal scale, while the vertical axis indicates the probability of a correct response. In this figure, however, we have estimated response curves for three hypothetical dichotomous items that differ only with respect to their levels of difficulty. For correct-incorrect dichotomous items such as these, the value of the location parameter (β) corresponds to the trait level (θ) at which there is a 50% chance of a correct response. So in this illustration, an individual with a trait level (θ) of -1 has a 50-50 chance of getting a correct response on the easy item (β=-1). For that person, the probability of a correct response would decrease to about 15 and 3%, respectively, for the items of moderate and high difficulty (where β = 0 and 1). Holding other item properties constant, as is the case here, varying the difficulty of an item alters only its location along the horizontal axis, not the fundamental shape or the steepness of the curve.

Figure 2: Illustration of IRT Location Parameter (β)



The Slope Parameter (α)

Items generally do differ, however, with respect to properties other than difficulty. The second IRT parameter to be considered captures a characteristic of item responses often referred to generally as discrimination. We may think of this slope parameter (α) as the steepness of the item response curve, which reflects the rate at which the probability of a correct item response changes as the estimated trait level increases. Items are said to have higher discrimination when the likelihood of a correct response changes more rapidly around its inflection point, the point on the IRC where the probability is 50%. Figure 3 illustrates the effects of changes in the slope parameter for a set of hypothetical dichotomous items. The easy and hard items from the previous figure have both been replicated here. We note that these two items have the same slope parameter value (α=1), which accounts for their identical shape and steepness. Their different slope parameter values (β= -1 and 1) account for their respective locations along the horizontal axis. The third IRC represents an item with a high difficulty (β=1) but with a lower discrimination value (α=.5). Compared to the other "hard" item, the curve for this lower discrimination item is flatter or more gradual. Since these two items are equally difficult, however, they do inflect around the same coordinates—the point at which an individual with a trait level of 1.0 has a 50% chance of a correct response. Varying the slope parameter, therefore, changes the steepness of an IRC but not its horizontal location or its point of inflection.

Figure 3: Illustration of IRT Slope Parameter (α)

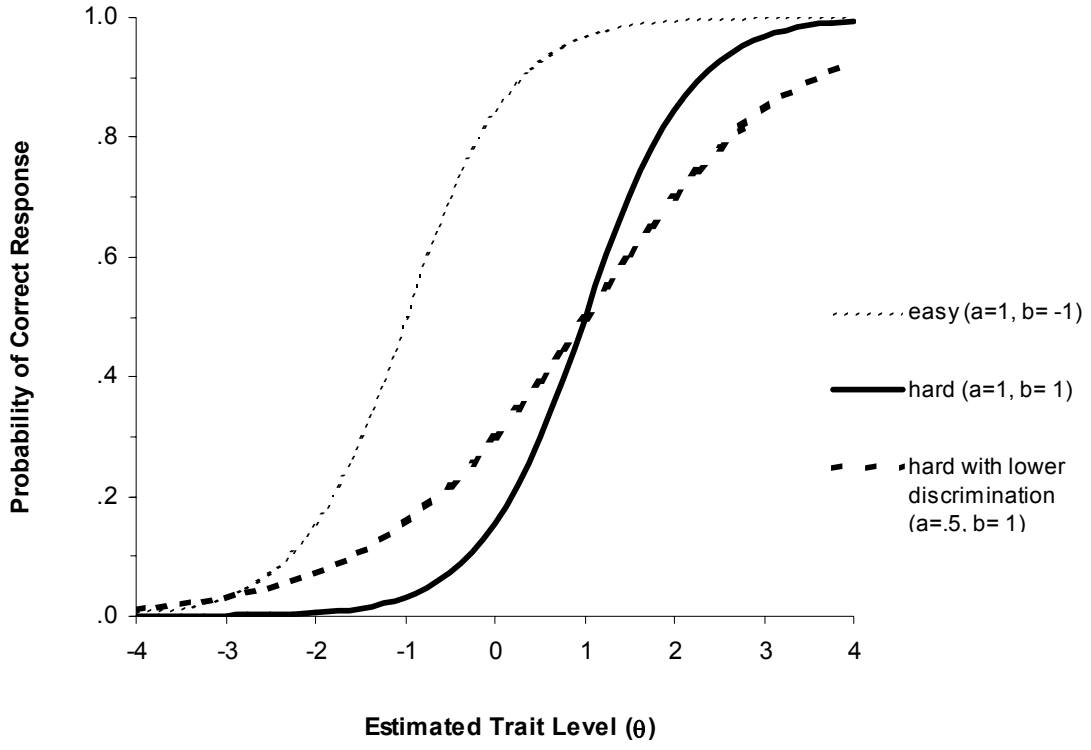
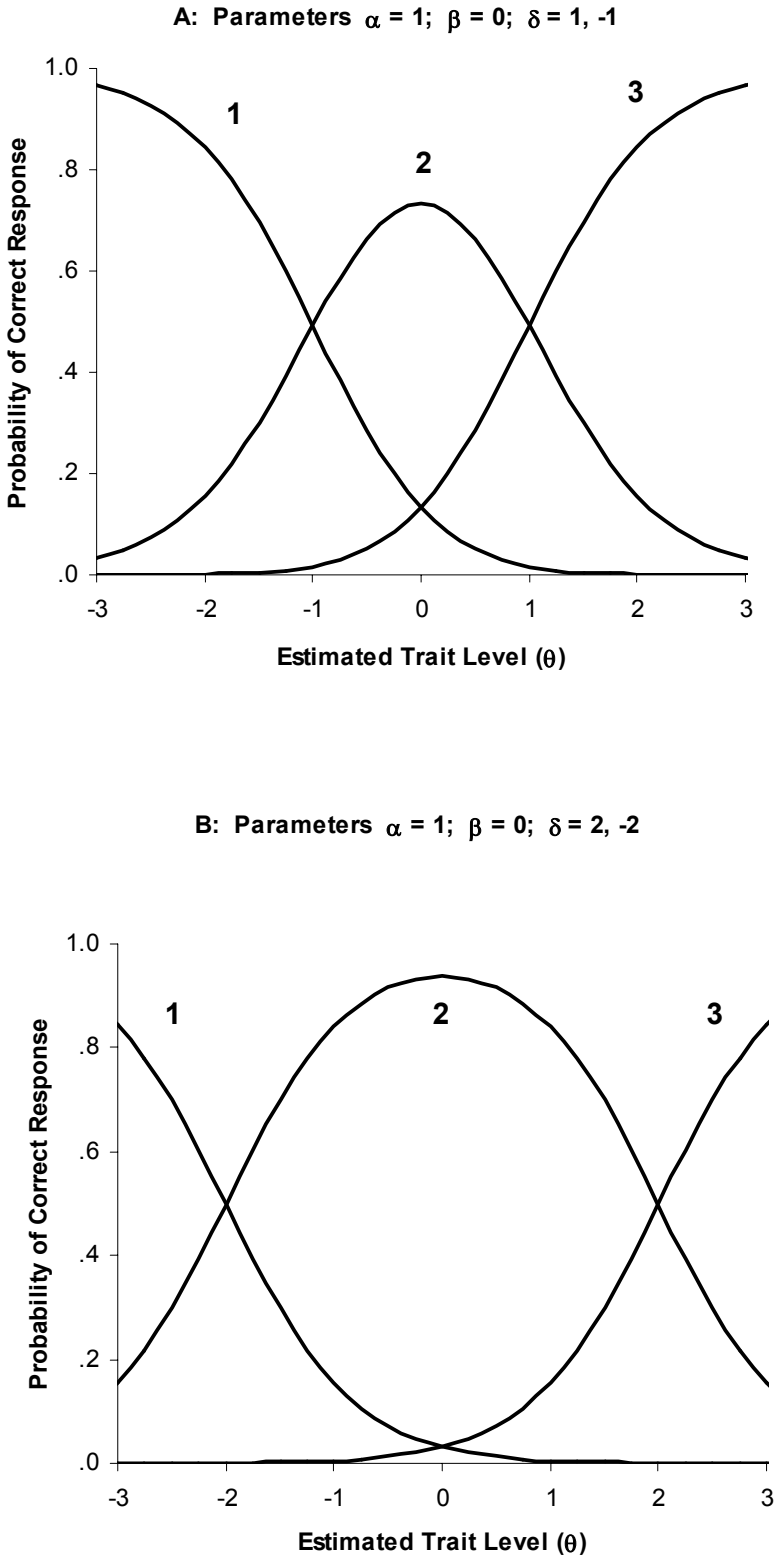


Figure 4: Illustration of IRT Step Parameter (δ)



The Step Parameter (δ)

Our discussions of item properties above have used dichotomous items for illustrative purposes. IRT analysis, however, is not limited to items that can be represented only as correct-incorrect or yes-no responses. In fact, many of the NAEP questionnaire items considered in this study ask teachers to choose a response from a set of three or more predefined categories. The partial credit IRT analyses used to estimate the models of standards-based science instruction below were designed to accommodate items with more than two *ordered* response categories. This variant of IRT can be thought of as an extension of the model for dichotomous responses. In the polytomous context, however, the model estimates the probability of an individual's response falling into a particular category (versus the category below) rather than estimating the probability of a correct (versus incorrect) response. In addition to the difficulty (β) and slope (α) parameters, the partial credit model introduces a new item-step parameter (δ) that captures the trait level at which an individual has an equal probability of appearing in adjacent response categories. This can alternatively be thought of as the transitional point where the higher category becomes the relatively more likely response. The step parameter for the first category of a polytomous item (δ_1) is fixed at a value of 0 as a mathematical property of the IRT estimation procedure (Muraki & Bock, 1997). Therefore, this first non-informative step parameter value is generally not reported with the results of IRT analyses.

In the top panel of Figure 4 we find an illustration of the item category response curves (ICRCs) for a single item with three ordered categories. (The dichotomous IRCs shown earlier could also be described as the upper category step for a two-category item.) For a given item *i*, we can calculate the trait level that corresponds to the transition to a particular category *j* by subtracting the step parameter value from item difficulty value ($\beta_i - \delta_{ij}$). In this example, a response in category 2 becomes equally as likely as a category 1 response for an individual with an estimated trait level of -1 (i.e., $\beta - \delta_2 = 0 - 1 = -1$). This corresponds to the point at which the ICRCs for these two categories intersect in the figure. Using the same procedure, we find that the curves for categories 2 and 3 intersect at a trait value of 1 (i.e., $\beta - \delta_3 = 0 - [-1] = 1$). The ICRCs in this example are symmetrical due to the equal size (but opposite signs) of the step parameters. According to the mathematics of the estimation procedure, the step parameters must sum to zero. This symmetry, however, is not a necessary feature of a polytomous item, as shown in the empirical examples presented later in this report.

Discrimination in the polytomous item response context is a property of both the slope parameter and the step parameter. Lower slope (α) values and greater differences or distances between the step parameters (e.g., $\delta_1 - \delta_2$) both result in less discriminating items, which are characterized by less peaked item category response curves. The second panel of Figure 4 plots the ICRCs for a three-category item with the same difficulty and slope parameter values as the illustrative item in Panel A, but with larger (and more widely spaced) step parameters. Compared to the previous example, the item appearing in the lower panel has a wider profile with more gradual, less steep ICRC slopes. In addition, we note that category 2 is the expected response for trait values ranging from $\theta = -2$ to 2, a much wider span than for the more highly discriminating polytomous item in the upper panel (with span from -1 to 1). The item with more widely spaced category steps is, therefore, less able to discriminate trait levels based on changes in item responses.

The Formal Model

The procedures for estimating an IRT measurement model are iterative. The IRT model first takes data on individuals' responses to a set of items and calculates estimated values for the item parameters specified in the formal model. These parameters can be used to formally characterize the

item category response curves, illustrated empirically in Figure 1 and derived mathematically in the sections immediately above. The next stage of the iterative estimation process uses these estimated item parameters and data on the observed pattern of item responses for the sample of respondents to derive a latent trait estimate (referred to as theta, θ) for each member of the sample. In the analyses below, theta will be a teacher's level of usage of standards-based science instruction. The specific type of IRT analyses used in this study, a two-parameter partial credit model, can be expressed formally as follows:

$$\Pr(X_{pix} = 1) = \frac{\exp \sum_{j=0}^{x} \alpha_i (\theta_p - \beta_i + \delta_{ij})}{\sum_{r=0}^{m} [\exp \sum_{j=0}^{x} \alpha_i (\theta_p - \beta_i + \delta_{ij})]} \qquad [2]$$

where,

$Pr(X_{pix} = 1)$	is the probability that the response for respondent p to item i will fall into category x
j	indexes response categories
m	is the maximum response category in an ordered series
θ_p	is the trait estimate for respondent p item
α_i	is the item slope parameter for item i
β_i	is the item location parameter for item i
δ_{ij}	is the step parameter for category j of item i

The item parameters produced by an IRT analysis are central elements of this methodological approach that distinguish it from other scaling techniques. Properties of item performance, as represented in these parameter values, provide the analyst with crucial information for evaluating the fit of an estimated model—that is, whether or not individual items are coherent with the broader latent trait of interest. In later sections of this report, we further explore the implications of differential item parameter values in the context of an actual IRT measurement analysis. There we discuss the model-building process and in particular the ways in which IRT item parameters allow us to diagnose problems with the design of the empirical measurement model.

Advantages of an IRT Approach

An item response theory approach to estimating teacher levels of standards-based science practice has a variety of advantages relative to other possible choices for a scaling technique. Some of these benefits derive from its more rigorous statistical properties compared to classical testing theory, the more desirable properties of the derived IRT scales, and its greater flexibility in accommodating items displaying a variety of different categorical response formats within a single empirical model. Space does not permit an extended consideration of beneficial features of IRT analysis compared to alternative measurement methods. However, we will briefly discuss two aspects of the IRT approach that prove especially useful in developing and interpreting an empirical model as a representation of a theoretical concept: (1) rich information on item performance, and (2) a dual measurement scale that puts item and traits on the same metric. The interested reader is referred to Embretson and Reise

(2000) for more detailed technical discussions of item response theory as it compares to alternative statistical strategies for measurement.

The objective of many scaling exercises is to produce an estimate of an underlying trait of substantive or theoretical interest. An empirical measure of this trait can then be used as a putative causal factor or outcome in subsequent analyses. Developing such a measure is one goal of the study described in this report. Another goal of this study, however, is to gain some understanding of the coherence and internal organization of instructional practices that constitute a particular (standards-based) approach to science education. The estimated parameters and summary information produced by many statistical scaling techniques are not especially helpful in guiding interpretations of the substantive importance and empirical fit of individual items vis-à-vis a broader construct or latent trait. By contrast, IRT analyses produce estimates of several item properties that can be used to characterize, in concrete and interpretable terms, the anticipated relationship between an individual's level on a latent trait and some more specific behaviors, namely his or her responses to a set of items. In this study, these behaviors would be a teacher's likelihood of using particular instructional practices. In developing and refining an empirical measurement model, questions will arise about whether particular items align or fit with the larger construct. With IRT methods, we are able to base our answers to these analytic questions on model statistics (e.g., item parameters) that can be interpreted in relation to substantive behaviors. This diagnostic approach to scale construction and issues of item and model fit helps to enlighten, rather than obscure, our understanding of the construct of interest.

The iterative estimation procedure of an IRT analysis produces dual, linked scales for the latent trait and for the constituent items. As we found in the examples above, the common metric of these two scales allows one to draw a direct extension between a particular trait level and the likelihood of a particular behavior (item response). For instance, a teacher with a trait value of 1 on the IRT scale would have a 50% chance of using (versus not using) a standards-based instructional practice with a difficulty score of 1. The IRT scale is therefore criterion-referenced in the sense that one can interpret the estimated latent trait scale in terms of concrete behaviors. By contrast, norm-referenced techniques provide relative measures of a trait, often expressed as the number of standard deviation units above or below the mean level for the analytic sample. The latter measurement can be disadvantageous to the extent that one is interested in more substantive interpretations of latent trait levels. Knowing that a teacher has a level of standards-based instruction a half standard deviation below average says little about what specific kinds of practices that teacher is likely to use and with what level of intensity. This sample-dependent property of norm-referenced measurement techniques also poses difficulties with regard to linking measurement scales across separate groups or time points, where the empirical distributions of the latent trait may differ or change dramatically (Embretson & Reise, 2000). The present report examines standards-based instruction for a national sample of teachers at a single point in time. The IRT methods used here, however, will afford us the opportunity to extend the work of this project into new areas, such as rigorously measuring differences in instruction across samples (e.g., the U.S. states) or changes for a particular group over time (e.g., using the 1996 and 2000 NAEP assessments) according to a common scale that captures absolute rather than relative trait levels.

IRT ANALYSES FOR STANDARDS-BASED SCIENCE INSTRUCTION

A Model-Building Strategy

The purposes of engaging in a formal statistical measurement analysis in this study are three-fold: (1) to determine whether standards-based instruction exists as a coherent and empirically identifiable teaching strategy in the actual practices of a national sample of teachers, (2) to identify specific practices that align (or do not align) with this larger construct, and (3) to accurately measure individual teachers' levels of standards-based instruction. Applications of *multi-dimensional* scaling techniques (such as factor analysis) are often exploratory, in the sense that analysts aim to identify a limited *set of constructs* that explain the systematic variation that exists among a larger group of constituent items. Item response theory, on the other hand, is a *unidimensional* technique to develop a measurement scale for a *single latent trait*. As such, we follow a largely confirmatory strategy for building a measurement scale for standards-based science instruction. In an IRT framework, we approach an observed pattern of item responses with the assumptions or hypotheses that these responses define a coherent construct (e.g., standards-based instruction) and that the individual items align with this same latent trait. Based on the empirical results of the IRT analysis, we can identify situations where our initial assumptions were mistaken. In other words, we can identify items that do not align or fit empirically with the larger construct. This information will allow us to incrementally refine the statistical model to more accurately represent the construct of interest by removing or modifying "misfitting" items.

The pool of items about teacher classroom practices in our initial IRT analysis can be grouped into two categories. (See Table 5 for a complete listing of items). First, the majority of instructional practices about which NAEP teachers report could be classified a priori as *standards-consistent*. These are practices that we have good reason to believe reflect the principles of a standards-based approach to science education as articulated by leading voices in the national reform movement. A second smaller set of items captures the use of practices that we would expect to be *inconsistent* with or *unrelated* to a standards-based approach. For example, instructional methods that rely on teaching from textbooks (HOTB), emphasizing facts (EMFA), and assessing student performance using multiple-choice questions (ASMC) clearly conflict with central tenets of a standards-based model of education, such as inquiry-based learning, hands-on activities, and the authentic demonstration of higher-order thinking skills. Two additional survey items provide rather limited insight regarding the precise nature of an instructional practice—the frequency of testing (HOTS) and use of homework to assess student learning (ASHW). In the absence of additional information, such as the kinds of tests given or the tasks included in homework assignments, it is difficult to determine with much confidence whether or not these practices are standards-consistent.

If the measurement strategy we have adopted is appropriate and methodologically sensitive, we would expect the results of the IRT analysis to be able to distinguish between these two sets of practices. Namely, the standards-consistent practices that make up the bulk of the instructional practices of interest should collectively define a coherent measurement scale (construct). The remaining standards-inconsistent (or standards-inconclusive) practices should be clearly and empirically identifiable as items that do not fit with the measurement scale estimated by the IRT model. Such items are likely to be prime candidates for removal from the analytic model. By intentionally including both kinds of practices in the initial pool of items in the analyses below, we provide an important methodological safeguard and a test for the sensitivity and utility of the IRT approach.

Developing an adequate measurement model is a process for which there are no definitive guidelines (see Embretson & Reise, 2000: pp. 226-246). This study takes a conservative approach by relying on multiple criteria to assess the adequacy of both the overall fit of the measurement model

with the observed data and the alignment of individual items with the general scale. The PARSCALE software used to estimate the IRT models reported below calculates the level of agreement between the pattern of observed responses for a particular item and the pattern that would be expected based on the specified measurement model (Muraki & Bock, 1997). This fit statistic is chi-square distributed and can be used to empirically test whether an item aligns with the broader construct captured by the measurement scale. These statistics can be cumulated across the full set of items to provide an empirical measure of the overall fit or coherence of the entire model. This cumulative statistic captures the likelihood that the estimated model fits the observed data.

A careful examination of IRT item parameters also provides a nuanced insight into the behavior of the specific instructional practices that will help us determine whether an item is, in fact, standards-consistent. In particular, instructional practices either antithetical or unrelated to a standards-based approach should display low levels of item discrimination. That is, a teacher's level of standards-based instruction should not bear a strong or systematic relationship to his or her likelihood of using a standards-inconsistent practice. We can diagnose this symptom of misfit empirically by identifying practices with very low slope parameter estimates (a criterion level of $\alpha < .100$ is used in this study). Poorly fitting polytomous items may also display very large and widely spaced step parameter values, which also results in poor item discrimination. In the analyses below, however, a diagnosis of slope parameters alone proves to be a sufficient method for detecting misfitting polytomous items. By engaging in an iterative process of model estimation, diagnosis and removal of misfitting practices from the item pool, and then reestimation of the revised model, we eventually arrive at an empirical measurement model that is able to adequately explain the pattern of observed data. In the following sections, we describe the main stages in this analytic process and summarize the empirical results.

Stages in Model-Building

Developing an IRT measurement model that both defines a coherent construct conceptually and provides an adequate explanation of observed item responses statistically can be a rather complex and technical undertaking. For this study, for instance, we began the model-building process with a pool of 55 separate instructional items. Relying heavily on the assumptions and diagnostic strategies outlined above, we eventually winnowed this initial pool down to the 42 items that constitute our final measurement model. This final model captures standards-based instruction as a coherent teaching strategy and also displays a very strong fit to the observed data. Before examining the results from this final measurement model in detail, we provide an overview of the model-building process. Particular attention is given to several major stages, or decision points, in the analysis that provided substantive insights into how certain practices prove difficult to incorporate into a standards-based instructional strategy. From a methodological perspective, this discussion also illustrates the utility of adapting IRT techniques for the measurement of systematic instructional practices. A synopsis of results and model revisions appears in Table 8.

Stage 1: The Base Model

Our initial IRT analysis employs the full complement of 55 science practices spanning several major domains of classroom instruction—emphasis on certain topics or skills, teacher activities related to the presentation or delivery of science content, student activities (written and hands-on), assessment practices, and use of computer technology in the classroom. As noted earlier, we expect that empirical analyses are likely show that several of these practices are actually inconsistent with, or unrelated to, standards-based models of instruction. The first column of Table 8 summarizes the

results of the IRT measurement analysis for our base model. Nine out of the 55 items display a significant degree of misfit with the estimated scale. That is, for these nine items a formal chi-square test using a 5% criterion value leads us to reject the null hypothesis that teachers' *actual* item responses from the observed data are not statistically different from those *estimated* on the basis of the IRT model.

Table 8: Summary of IRT Model Development Stages

	Model-Building Stage			
	1	**2**	**3**	**4**
Total items	55	50	50	42
Misfitting Items [a]	9	6	1	1
Probability of Model Fit [b]	.000	.000	.271	.871
Item Parameter Diagnosis	Very poor overall model fit, with a number of items showing very low discrimination	6 of 8 assessment items show "reversals" in steps, possibly suggesting problem in format of response categories (ASWR, ASPF, ASEY, ASSE, ASLB, ASHO)	3 of 6 reversals corrected after recoding (ASEY, ASSE, ASHO), remaining 3 retain reversal (ASWR, ASPF, ASLB)	Very good fit between model and observed pattern of responses, misfit rate of individual items in chance range
Revision for Next Stage	Items with very low discrimination ($\alpha <$ .100) removed: HOTB, HOTS, EMFA, ASMC, ASHW	All assessment items recoded to collapse categories 2 and 3 (once/twice a year, once per grading period)	Removal of items that refer to specific uses of computer technology (RWMM, DMCO, DMCM, CODP, COGA, COSM, CODA, COWP)	(Final Model)

[a] Results are derived from chi-square tests for individual item fit to observed data. An item is classified as misfitting if there is a statistically significant
difference between the estimated (from IRT model) and observed (from data) item response patterns (criterion value = .05).

[b] Probability of total model fit is derived from a cumulative chi-square test across individual items

An examination of item parameters estimated by this model reveals that 5 of the 55 items display very low levels of discrimination. These prove to be the same set of practices we identified earlier as likely to be either inconsistent with, or unrelated to, a standards-based approach to teaching: use of textbooks, frequency of testing, emphasis on facts, and assessments using multiple-choice tests or homework assignments. These practices display very low levels of discrimination with respect to their slope parameters. Their slope values range from .005 to .090, all below the criterion value established earlier as a diagnostic threshold ($\alpha = .100$). These are all also polytomous items, having between three and five response categories. Model results also show very large gaps between step parameters (δ). Each of these practices contain at least one step between categories that spans a range of 10 logits or more on the estimated theta scale for standards-based instructional level. *Logits*, the units in which IRT scales are conventionally expressed, have a roughly standard-normal metric. Therefore, step gaps of the size observed here profoundly reduce the discrimination power of these items.

To illustrate the distinction between items with high and low discrimination in more accessible terms, we can take two questions that ask about the frequency with which teachers employ certain instructional practices—use of hands-on activities and use of textbooks. Since these items come from the same series on the teacher survey, they have the same four response categories (see Table 5). While the former practice represents an integral element of the active and inquiry-focused orientation of a standards-based instruction, we would expect the latter technique to be inconsistent with a standards-based approach. The upper panel of Figure 5 plots the item category response curves for use of hands-on activities (HOHO) derived from the base IRT model. These results indicate that this item has a high discrimination value ($\alpha = .910$) and modest gaps between the category steps ($\delta = 1.533, .125, -1.659$). Graphically we find an orderly series of category steps with sharply peaked ICRCs—an empirical result that closely resembles the hypothetical illustration shown earlier in the upper portion of Figure 4.

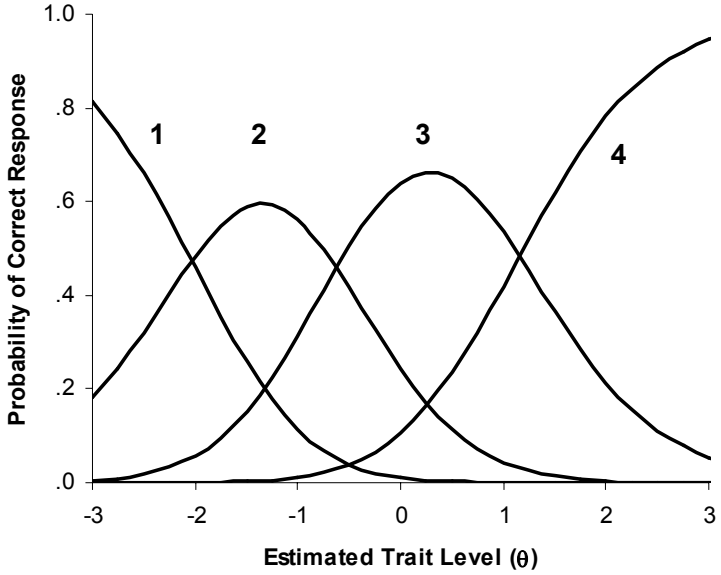
In the lower panel of Figure 5 we use the results from the base model to plot the item category response curves for textbook use (HOTB). A perfectly non-discriminating item would have produced a series of horizontal, non-intersecting ICRCs. The combination of a very low α and very large δ values that we observe for this item on textbook use produces a set of ICRCs that resembles that extreme rather closely. The four curves, for example, are nearly parallel to one another (only categories 1 and 4 intersect within the 6-logit theta range depicted). In addition, we find that category 3 will be the expected or most probable item response associated with any plausible value on the estimated trait scale. That is, regardless of a teacher's overall level of standards-based instruction, he or she would be expected to use textbooks once or twice a week (the third response category for this series of items). Essentially, trait level shows no predictive power or discrimination with respect to an individual's expected response on this item. Similar, although somewhat less extreme, patterns are also displayed by the other four instructional practices that have very low item discrimination.

Stage 2: Eliminate Poorly-Discriminating Items

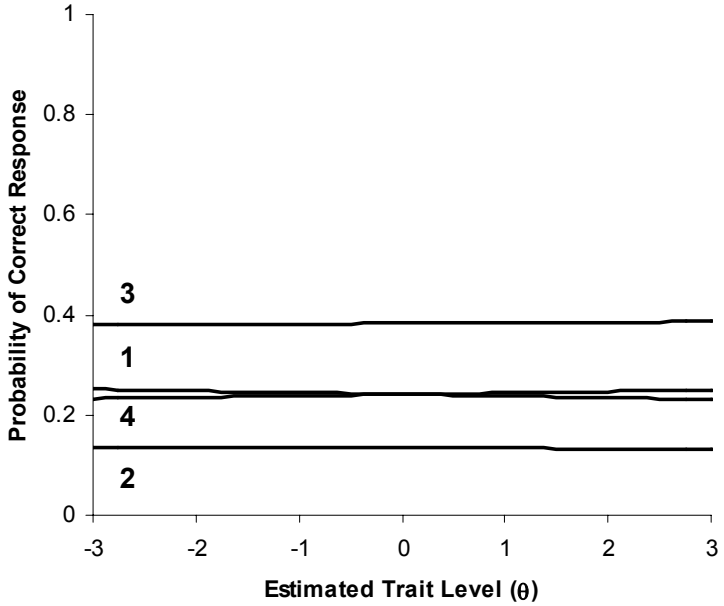
Based on the results from the initial measurement model, we removed the five instructional practices found to have very low item discrimination. These are items that can be concluded to have little or no empirical relationship with the overall estimated measurement scale. They are construct-inconsistent. After eliminating these items, we estimate a new IRT model with the remaining 50 instructional practices. An examination of the item and model fit statistics for this revised analysis suggests some improvement. Six items show significant misalignment with the larger construct (compared to nine in the base model). As before, however, there remains a poor statistical agreement between the overall model and the observed pattern of item responses ($p = .000$).

Figure 5: Illustrations of Item Discriminations from Base IRT Model

A: High Discrimination - Use of Hands-on Activities
(α = .910, β = -.498, δ = 1.533, -.125, -1.659)



B: Poor Discrimination - Use of Textbooks
(a = .050, b = .000, d = -66.782, 123.097, -54.315)



An examination of the results from this second-stage IRT model uncovers multiple instances of an irregularity in step parameter ordering known in IRT parlance as a *reversal*. A reversal occurs when the progressive step parameter values (δ) for a polytomous item are not strictly ordered. When the step parameters are systematically ordered (decreasing in value), each category will prove to be the most likely response for individuals at some trait level. Individuals with very low trait levels will fall into the lowest response category (1). As trait level increases, a point will be reached where the next higher category (2) will become the more likely response. This point is represented mathematically by the second step parameter (δ_2) and graphically by the point where the ICRCs for categories 1 and 2 intersect. A similar relationship would be found for the step between categories 2 and 3, and so forth.

In the case of a reversal, however, this progressive stepwise pattern is not found. The upper panel of Figure 6 illustrates a reversal found in the ICRCs for assessment using hands-on activities (ASHO) drawn from the results of this second IRT analysis. Here category 2 (once or twice a year) displays a reversal because the step parameter for category 3 is greater than the one for category 2 ($\delta_2 = .116$, $\delta_3 = .698$). Of the five possible responses for this item, category 2 is the only one that is not the most likely response for some trait level. It should be pointed out that category reversals of this kind do not necessarily indicate that an item is conceptually inconsistent with the larger construct being measured. The present example might suggest, for example, a pattern of incorporating a standards-consistent assessment practice into the larger instructional strategy whereby teachers tend to "skip" a step. They might jump from not using the practice at all (category 1) to using it once a grading period (category 3). This accelerated implementation could be a valid means of developing a standards-based instructional style.

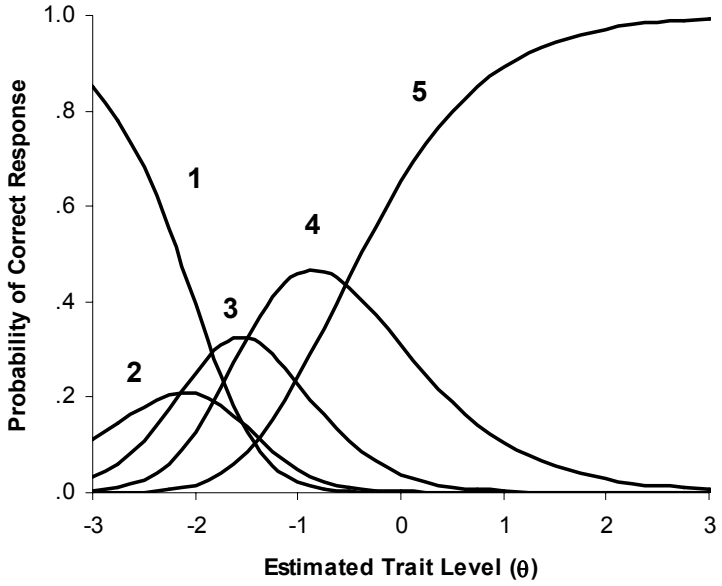
In this model, however, we find a similar reversal involving category 2 (once or twice a year) for six of the eight items from this assessment series that we have retained for this stage of the analysis. Regardless of whether reversals are construct-consistent, the presence of a recurrent reversal pattern isolated among a certain set of items might be symptomatic of other problems that could affect the fit of a model. A closer examination of the five response categories for this series suggests a possible cause for this systematic irregularity. Namely, there appears to be a considerable amount of overlap in the frequencies defined by categories 2 and 3—"once or twice a year" and "once per grading period." For example, in a situation where there are two major grading periods in an academic year (e.g., semesters), using a form of assessment once per grading period would also correspond to two administrations annually. So here, categories 2 and 3 would not be mutually exclusive. Some teachers might choose category 1 while others choose category 2, but both would be appropriate responses in this situation. The pattern of responses underlying the reversals among these assessment items, therefore, may be a product of a sub-optimal questionnaire design rather than actual differences among teachers in their trait levels (i.e., their use of standards-based science).

Stage 3: Correcting Category Reversals

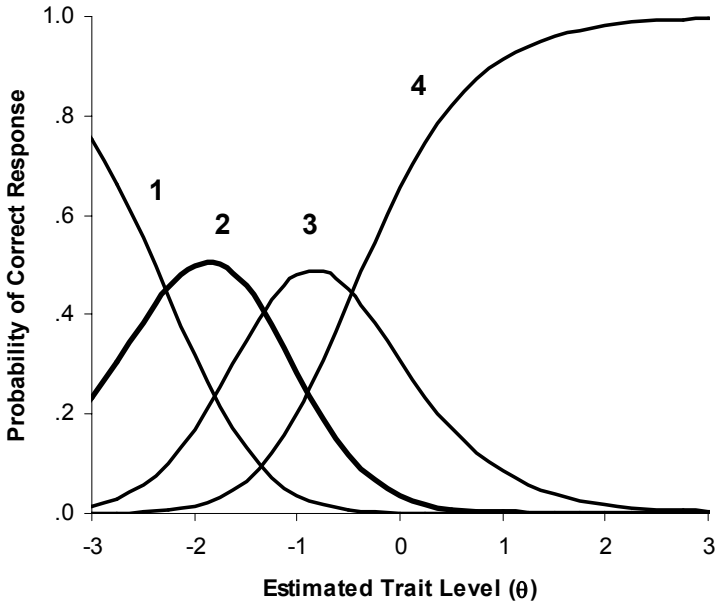
Flaws in item design may introduce error or noise into response patterns, which can in turn negatively affect several aspects of the measurement model including item-trait alignment, the goodness-of-fit of the larger analytic model, and properties of the estimated measurement scale itself. To determine whether this is indeed the case in the present situation, we can recode the entire series of assessment items by collapsing the two overly similar responses (categories 2 and 3) and leaving the others unaltered. The revised assessment items used in this third-stage IRT analysis will, therefore, have only four categories. Because all other items remain unchanged from the prior analysis, any differences in model and item performance can be attributed to the revised formatting of these assessment item response categories.

Figure 6: Illustration of Category Reversal - Assessment using Hands-On Activities

A: Five Original Response Categories (Stage 2 Model)
(α = .817, β = -1.427, δ = .116, .698, .074, -.888)



B: Four Recoded Response Categories (Stage 3 Model)
(α = .954, β = -1.355, δ = .917, -.027, -.890)



The results of this third model indicate that the recoding procedure described above eliminated the reversals in three of the six affected assessment items. To illustrate this correction, the lower panel of Figure 6 plots the revised ICRCs for assessment using hands-on activities (HOHO) from the third-stage analysis. After collapsing response categories, the step parameters are consistently ordered ($\delta_{2\text{-}4}$ = .917, -.027, -.890), and this item no longer displays a reversal. Similar corrections were achieved for the items reporting assessment using essays and self-evaluations. However, the reversals remain for three of the assessment practices—written responses, portfolios, and lab notebooks. This result suggests that these techniques may be incorporated into classroom teaching in a somewhat uneven fashion.

We find that recoding the assessment items substantially improves the performance of the measurement model. In this third stage of the analysis only 1 of the 50 practices included in the model displays a non-significant fit with the overall measurement scale—a degree of empirical misfit consistent with chance levels in an item pool of this size. Accordingly, the model provides a statistically adequate fit with the observed data. The probability that the discrepancies between the observed pattern of responses and those expected on the basis of the full analytic model are due to chance alone is .271, a level that exceeds the conventional 5% threshold used in most tests of statistical significance.

Stage 4: The (Special) Case of Computing Technology

In the series of IRT analyses described above, our objective has been to develop a coherent and methodologically rigorous measurement scale for standards-based instruction. The criteria guiding decisions to eliminate or alter particular items have been largely technical in nature—item and model fit and the problematic performance of certain items as diagnosed by their parameter values. Based on these diagnoses, we have made a progressive series of refinements to the analytical model. At this stage, we have developed a model in which the constituent items define a coherent latent construct and which provides a statistically adequate explanation for the observed pattern of teacher responses about instructional practices. This measurement model appears to successfully define a broad construct of standards-based science that encompasses a range of distinctive instructional practices. On technical grounds, therefore, we would be justified in concluding our model development at this stage. Before making that determination, however, it is useful to also view standards-based instruction from a more practical or substantive perspective, one that might recommend further refinements to the measurement model.

Using particular classroom practices may require access to certain resources. Students cannot read from textbooks if they have no textbooks, nor can they perform library research for a report if such facilities are lacking. Access to adequate laboratory facilities, manipulatives, or computing technology can likewise affect a teacher's ability to use instructional techniques that involve these resources. Resource considerations of this kind are common to many instructional approaches, not unique to a standards-based strategy. In developing conceptual and analytic models for a standards-based instructional strategy, therefore, we may wish to consider issues of substantive importance (like resource availability) along with the technical performance of a model.

The decision to include in our empirical model instructional practices to which all teachers may not have (equal) access may be justifiable on the conceptual grounds of wanting to represent as comprehensively as possible the practices consistent with a standards-based vision of science education. As the argument goes, because resource-dependent practices are consistent with (even integral to) this instructional approach, teachers who do not use them are by definition using a less complete form of standards-based instruction. The reasons why teachers fail to use these practices are irrelevant from this point of view. Teachers are treated the same whether they have access to necessary resources (but decide not to use a practice) or lack these resources.

There might also be a valid counterargument that supports the decision to remove strongly resource-dependent practices from consideration. From this point of view, it may be desirable to capture a more generalized form of standards-based science that is relatively broad in scope but that does not include practices for which there are substantial barriers to access. The goal of this latter approach would be to produce a construct capturing a form of standards-based instruction that is as widely accessible to teachers as possible. Taking these two perspectives together suggests that it is possible to conceive of legitimate variants to a theoretical construct that are alternately more *comprehensive* in the scope of items encompassed and more *generalizable* in terms of their practical applicability to a wider array of respondents. Ideally, we would want to develop a measurement model that incorporates and balances these properties.

The sorts of conditions or resources that could potentially affect a teacher's ability to pursue a given instructional practice are too numerous either to catalogue or to explore empirically. In this study, however, one issue that would appear to be particularly relevant is the availability of computer technology. As technology becomes an increasingly central element of the practice of science in both academic settings and in the workplace, advocates of standards-based reform have argued that technology like computers must also be incorporated into science classes. Teachers with limited access to computers, however, will be less able to incorporate this form of technology into their classroom practices than would teachers with ready access. To the extent that a comprehensive scale includes such resource-specific practices, its generalizability might be compromised.

In the NAEP teacher surveys, there is a substantial number of items that deal with various uses of computers in the classroom. This presents a dilemma. On one hand, systematic differences (or biases) in teachers' answers to such a large set of items have the potential to affect the results of our measurement model and the resulting scales for standards-based instructional use. On the other hand, the use of technology, and specifically computers, is a legitimate and even important component of a standards-based science education. In the final stage of revision of the measurement model, we arrive at a compromise. This last model excludes eight items that relate to specific uses of computers for instructional purposes (RWMM, DMCO, DMCD, CODP, COGA, COSM, CODA, COWP). However, we retain a more generally worded item in which teachers report how often their students use computers for science (HOCO). This modeling approach strikes the middle ground of incorporating some information on the use of computer technology without unduly influencing the construct by including a large set of practices that are all similarly dependent on access to computers. Results of this IRT model are reported in the final column of Table 8.

In this reduced model of 42 items, only a single practice displays a statistically significant level of misfit with the measurement scale. This high degree of item alignment with the construct is comparable to the previous analysis. The current model excluding practices that involve specific uses of computers, however, appears to provide a substantially better overall fit to the observed data than does the more comprehensive model including the computer items ($p = .871$ vs. $.271$). Additional analyses confirmed that this improvement in model fit is also statistically significant. These results suggest that although the models estimated in these two stages of the analysis both adequately fit the data, the more generalized version of standards-based science (excluding specific computer activities) does a better job of explaining the observed pattern of teacher responses than does the more comprehensive scale (including these practices).

Choosing a Final Model

Before concluding that the generalized version of the model is preferable, we should also consider the impact that removing the specific computer practices has on the other side of the dual measurement scale. That is, how much will our estimates of a teacher's level of standards-based science instruction differ depending on whether it is based on a measurement model that includes

specific instructional uses of computers or on a model that excludes them? Figure 7 plots the trait estimates (θ) for a teacher's level of standards-based science instruction under these two conditions. Theta values derived from the comprehensive model including computer practices appear on the horizontal axis, with values for the generalized model plotted along the vertical dimension. We find a very strong correlation, approaching unity (r = .992), between the θ estimates generated from the two model variants. This implies that on average these two measures are virtually identical.

To view this finding another way, removing this block of eight items produces a very negligible impact on our measurement of instructional practices. This might be the case for two major reasons. First, the use of computer technology may represent a less central or less tightly integrated element of a broader standards-based approach to science. Second, these practices may be used infrequently and therefore contribute relatively little weight in the empirical calculation of teacher trait values (because few teachers use them). A closer examination of the IRT results from Stage 3 provides some support for both explanations. Six of the eight specific computer practices have slope parameters below the mean level for the model (α = .499). This is consistent with the interpretation that practices that are less central to standards-based reform should display lower discrimination. In addition, these computer practices also tend to be among the most difficult (i.e., least frequently used) of the practices examined, with an average difficulty level (β_{avg} = 2.857) far exceeding that of the average practice (.575). Six of the eight computer items are among the most difficult of the practices in the model ($\beta > 2.0$). By comparison, only 7 of the remaining 42 items fall into this extreme difficulty range. Taken as a whole, these results recommend the analytical model excluding specific uses of computers for instruction (Stage 4) as the preferred empirical measurement of standards-based science instruction.

Figure 7: Comparison of Estimated Thetas from Alternative Model Specifications

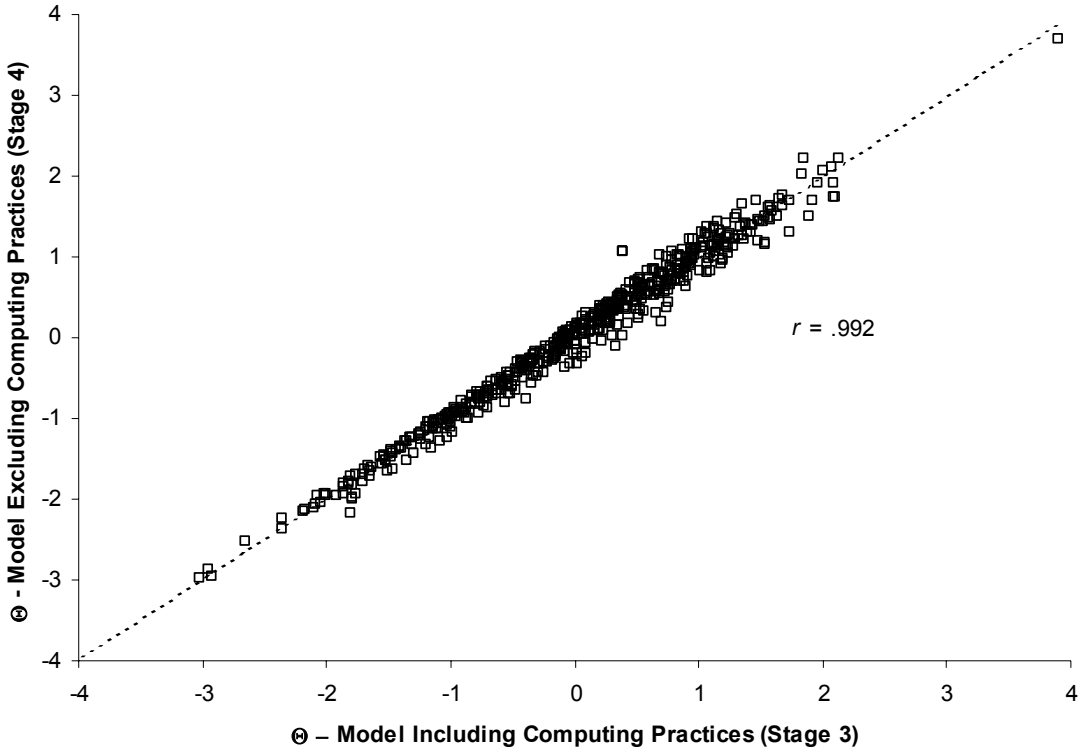


Table 9: IRT Item Parameters from Final (Generalized) Model of Standards-Based Science Instruction

Item	Slope (Discrimination) α	Location (Difficulty) β	Category Steps $\delta 2$	$\delta 3$	$\delta 4$
HORB	.317	1.884	3.365	-.564	-2.801
HODN	.230	-.211	4.422	-.464	-3.959
HOOS	.804	-.430	1.483	-.133	-1.350
HOOR	.495	2.531	2.275	-.797	-1.479
HOWR	.590	1.284	2.073	-.755	-1.318
HOHO	1.094	-.488	1.484	.093	-1.576
HOTK	.957	-.164	1.438	.208	-1.646
HOLI	.376	2.704	3.725	-.874	-2.851
HOCO	.304	2.826	1.182	.072	-1.254
EMCN	.385	-4.425	1.549	-1.549	
EMPS	.837	-1.708	1.061	-1.061	
EMST	.265	-2.899	2.851	-2.851	
EMCM	.522	-.947	1.432	-1.432	
EMLS	.989	-.332	.759	-.759	
EMIN	.570	-2.271	1.316	-1.316	
EMDA	.699	-.283	1.330	-1.330	
EMTC	.412	.653	1.514	-1.514	
ASWR	.247	-1.481	1.060	2.027	-3.087
ASIP	.593	.480	2.951	-.650	-2.301
ASGP	.587	.483	2.123	-.507	-1.615
ASPF	.213	1.768	.085	-.903	.818
ASEY	.211	1.709	2.132	-.118	-2.015
ASSE	.310	1.632	1.408	-.693	-.715
ASLB	.330	.090	.311	-.625	.314
ASHO	1.075	-1.312	.915	-.012	-.903
GRHO	.905	-.339	2.336	-.154	-2.183
WKPF	.393	.515			
WKPJ	.806	-1.226			
RWNB	.694	-1.432			
RWPJ	.640	-.532			
RWIS	.251	-1.163			
RWFT	.507	2.030			
RWJL	.480	.576			
RWPH	.451	1.460			
RWAV	.439	2.731			
RWIN	.579	2.753			
RWMO	.616	.163			
DMTK	.231	-5.143	.659	-.341	-.319
DMDM	.500	-.293	2.663	.245	-2.908
DMVT	.244	2.323	5.833	-1.009	-4.824
SCFT	.253	2.506	2.044	-2.044	
SCGU	.452	1.859	1.141	-1.141	

Note: For information on model fit, see the final column of Table 8.

A Generalized Model of Standards-Based Science Instruction

In this section we present the results of our final IRT model of a generalized standards-based instructional strategy for middle school science (Table 9). By examining the item properties of individual practices and the behavior of broader elements of instruction, such as hands-on activities or assessment techniques, we may gain important insights into the internal organization of this teaching strategy in the classroom. We focus particularly on the implications of the IRT slope and difficulty parameters estimated by this model. The former provides an indication of which elements tend to be more (or less) centrally integrated into a standards-based approach. The latter suggests the ways in which standards-based science instruction could be incrementally built from a set of classroom practices.

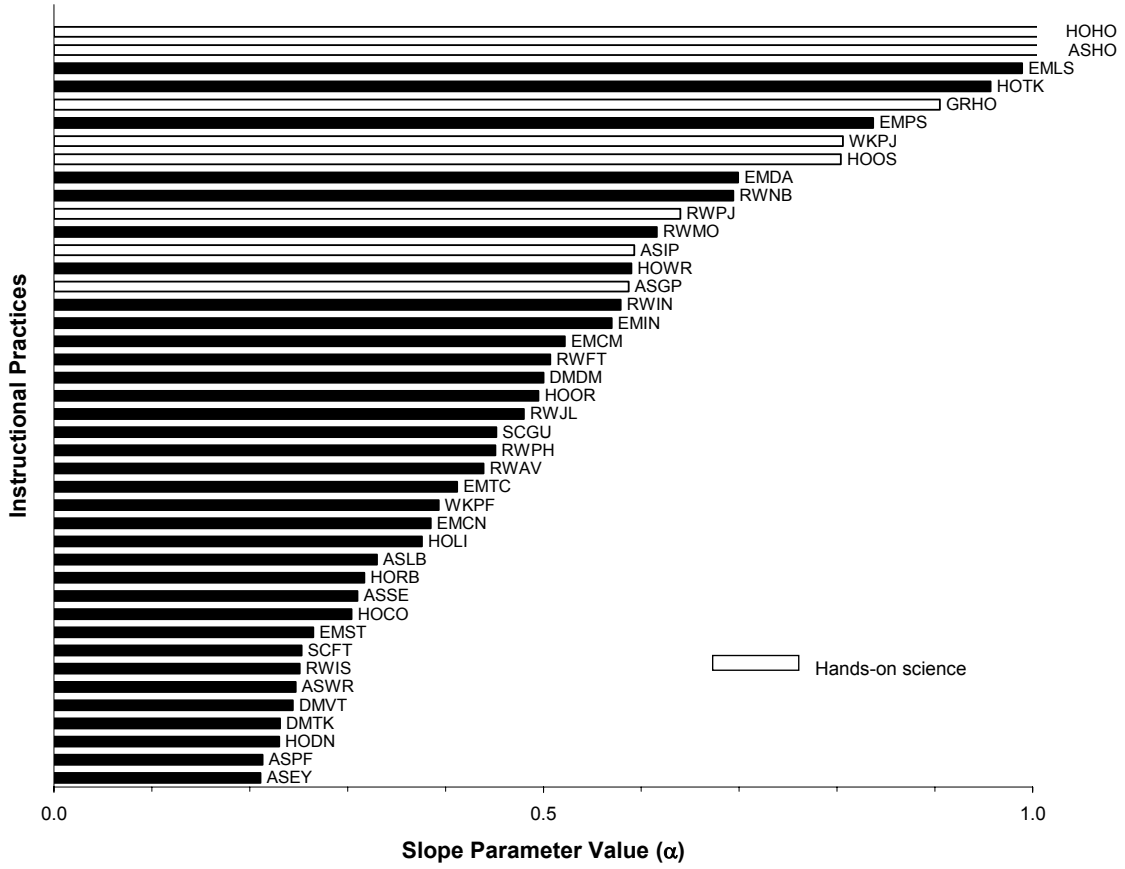
Item Discrimination and the Centrality of Particular Practices to Standards-Based Science

The first column of Table 9 reports the IRT slope parameters from our final model of standards-based science instruction. Consistent with an analysis showing excellent fit for both individual items and the model as a whole (see Table 8 and the discussion above), we find that all items have slope parameter values over .200, with an average score of .520 across the 42 practices. The slope parameters, however, do vary considerably from a low of .211 for assessment using essays (ASEY) to a high of 1.094 for frequency of hands-on activities (HOHO). This suggests that some practices are, in effect, more strongly aligned with the latent construct captured by the measurement scale than are others.

Figure 8 arrays the science instructional practices by their slope parameter values. Here we notice that practices dealing with hands-on activities in general and science projects in particular are consistently among the most highly discriminating items on the standards-based science scale. Therefore, we find here empirical confirmation that the hands-on science and inquiry-based investigations emphasized in national models of standards-based reform are also central components of this approach as it is implemented in the classroom. Represented by white bars in the figure, these hands-on practices all have above-average slope values and they also constitute a large share of the most discriminating items in the model as a whole. These practices span several broader domains of instruction, including not just student engagement in individual or group activities and projects (HOHO, HOOS, WKPJ, RWPJ), but also the use of assessments that incorporate these types of performance (ASHO, ASGP, ASIP) and the weight placed on hands-on activities by teachers when assigning grades (GRHO).

Other practices displaying high levels of item discrimination also capture commonly cited elements of a standards-based model of instruction. These include an emphasis on higher-order skills such as lab techniques (EMLS), problem solving (EMPS), and data analysis (EMDA). Among items with lower slope values, we find practices that reflect more passive, less student-centered techniques for communicating science knowledge, such as teacher talk or lecture about science (DMTK) and the use of video tapes or television programs (DMVT). Assessment practices that do not specifically involve hands-on activities also tend to show relatively low discrimination levels. Within the context of implementing a standards-based science strategy, teachers may use these latter (more traditional) methods of gauging student progress to supplement the hands-on assessment techniques central to the standards-based approach. This supporting role and the likelihood that individual teachers use different combinations of assessment techniques would together account for the generally low discriminations found among this set of items.

Figure 8: Slope Parameter Values from Final IRT Model



Item Difficulty and the Progressive Development of Standards-Based Science

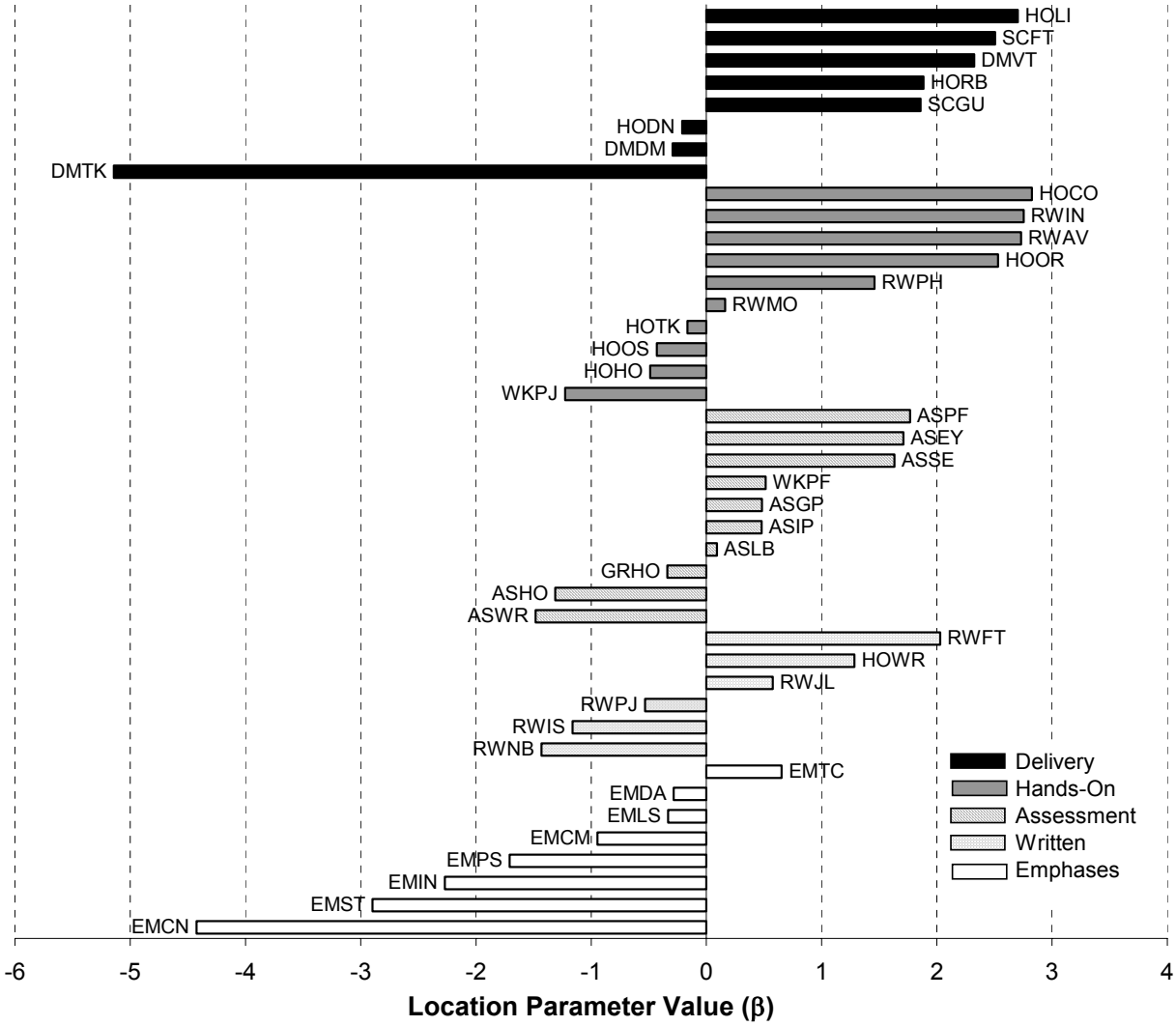
The data on teachers' use of instructional practices analyzed in this study were collected at a specific point in time. Because this information is not longitudinal, we cannot know with certainty when teachers adopted certain practices or whether they used them more in 1996 than they had in the past. We can speak only to the question of whether (or how much) teachers were using these practices in their science classes in 1996. As discussed earlier, a well-designed IRT measurement model should embody a systematic, progressive relationship between trait level and the likelihood of a correct response for items of varying difficulty. This should also be true of our final model for standards-based science. In the present case, for instance, we would expect that teachers who use moderately difficult standards-based instructional practices would be very likely to use the less difficult practices as well. The strength of our empirical model, combined with this progressive property of IRT measurement methods, will afford us some latitude in drawing insights about how standards-based instruction might be implemented by teachers over time on the basis of a cross-sectional analysis. We would propose that, in the aggregate, the rankings of practices according to their item difficulty scores will reflect a sequence of steps through which a typical teacher might build a standards-based instructional strategy in science. This developmental trajectory characterizing the modal pattern of implementation would also tend to be mirrored rather closely in the activities of the large majority of individual teachers.

The second column of Table 9 reports the item difficulty parameters from the final IRT model of standards-based science. We find that item difficulties (which here generally reflect frequency of use) range widely, but rather evenly, across the entire set of practices, a desirable property for the collection of items constituting a measurement scale. Much as would be the case for

a cognitive assessment, the inclusion of very easy and very difficult items helps to minimize floor and ceiling effects, as the performance levels of very few individuals will tend to fall outside these boundaries. An even coverage of items across the full range of the scale also helps to ensure that any given individual's (latent) trait level will tend to fall close to the difficulty level of one or more items. An item's discrimination power will be greatest at the point on the measurement scale corresponding to its difficulty parameter. Taken together, these scale properties give us confidence that we will be able to accurately estimate a teacher's use of standards-based science across a wide range of "ability" levels, reflecting here the degree of implementation for this particular instructional strategy.

To simplify the presentation of these results, Figure 9 illustrates the difficulty or location parameters (β) for the 42 instructional items in the final generalized model. These are grouped by five major domains of instruction described earlier: topic and skill emphases, teachers' modes of presenting science content knowledge, assessment methods, and two types of student-centered learning opportunities—hands-on and written activities. These domains are ordered by their overall level of difficulty, with practices ranked within domains by their individual item location parameters. The average difficulty parameter value across all items is .188.

Figure 9: Difficulty of Instructional Practices, Arrayed by Element



We first note the very low difficulties of two practices—teacher delivery of content by talking to students (DMTK) and an emphasis on science concepts (EMCN). As the earlier descriptive analyses indicated, these practices are nearly ubiquitous. Most teachers lecture to their students on a regular (even daily) basis and nearly all teachers believe that it is important to communicate key scientific concepts to their students. Both of these practices, while consistent with a standards-based teaching strategy, are reminiscent of fairly traditional approaches to instruction when compared to other items examined here. For instance, emphasizing concepts would be more consistent with standards-based reform than concentrating on facts and terminology (EMFA), but less unique to this reformed approach than emphasizing problem solving (EMPS) or communicating math ideas (EMCM). One implication we could draw from this observation is that lecturing and attention to concepts represent a more generalized pedagogical foundation upon which a more specific and clearly articulated standards-based model can be built. Given the widespread use of these two practices as well as some of those eliminated in earlier model-building stages (e.g., use of textbooks), another related implication would be that middle school science teachers rely heavily on relatively traditional instructional techniques.

Of the five major instructional domains, the easiest for teachers to implement is clearly emphasizing skills and topics consistent with standards-based reform. Among these possible emphases, the most frequently endorsed tend to be the ones reflecting the more generalized objectives or values of standards-based science, such as emphasizing the social and technological importance of science (EMST), encouraging students' interest in science (EMIN), and problem-solving (EMPS). By comparison, teachers tend to place less emphasis on the acquisition of skills that are more concrete or advanced. These include fostering students' science communication abilities (EMCM), data analysis skills (EMDA), and their facility with applied laboratory techniques (EMLB) or the use of technology (EMTC). The low difficulty ranking of instructional emphases as a general element of teaching is consistent with the notion that expressing support for the general principles and goals of standards-based reform should be the first step in incorporating them into actual classroom practice. The more fine-grained ranking of practices within this domain lends further support for this view of implementing standards-based reform as a progressive enterprise. Specifically, teachers tend to begin by emphasizing the most basic or general elements of the reformed practice and then move on to other, increasingly more elaborate, topics and skills over time.

The instructional domains reflecting written student work and assessment practices both display moderate levels of difficulty on average (β_{avg} = .127 and .355 respectively). However, the frequency of use for individual practices in these two areas does vary considerably, with difficulty scores ranging from about -1.5 to 2 logits. Among the most commonly used types of written work are laboratory notebooks (RWNB) and assignments that ask students to write about specific topics or issues (RWIS) or report on an extended project (RWPJ). Ongoing written assignments like routinely logging class work in journals (RWJL) tend to be moderately difficult practices. Large assignments, such as formal written reports (HOWR) or write-ups of science field trips (RWFT), that mark major milestones in a science course are generally the least frequently used and, therefore, the most difficult forms of written work to incorporate into science instruction.

In a similar fashion, we find that the most difficult classroom assessment methods include practices that are cumulative in nature, such as compiling portfolios (ASPF, WKPF), or activities that are especially time-consuming, such as in-class essays (ASEY) and self- or peer-evaluations (ASSE). Teachers use specific hands-on activities, such as projects, presentations, and laboratory notebooks (ASIP, ASGP, ASLB), to assess student performance with moderate frequency. As would be expected, the use of general hands-on activities (i.e., of an unspecified kind) for assessment purposes proves to be comparatively easy (ASHO). It is also interesting to note that assigning substantial weight to hands-on activities in science grading is a moderately easy aspect of assessment practice.

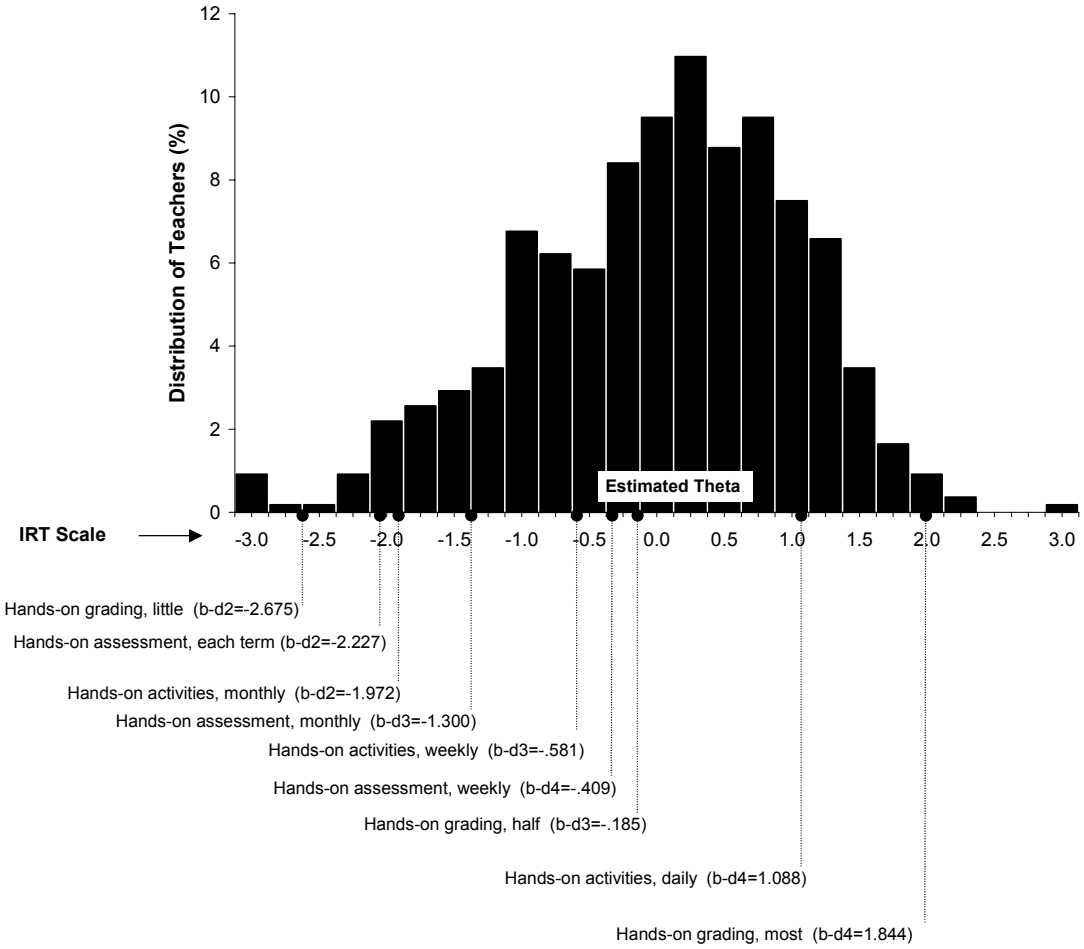
These results suggest that teachers not only use hands-on activities frequently in class but also view them as an important basis for evaluating student performance and assigning grades. Many teachers are, in effect, "putting their money where their mouths are" when it comes to hands-on science. Among the practices examined here, the easiest form of assessment to incorporate into a standards-based instructional style proves to be written responses (ASWR). Ranging anywhere from a phrase to several paragraphs, these writing tasks could be incorporated into science quizzes or tests that also contain more traditional types of items like multiple-choice questions. They may provide a point of entry into standards-based science for teachers with more traditional pedagogical orientations.

Although the results above have suggested a relatively favorable orientation toward certain aspects of hands-on science, the practices that make up our instructional domain of hands-on and interactive student activities tend to be substantially more difficult to enact on average ($\beta_{avg} = 1.016$). Looking beyond this general tendency, however, we find within this area of instruction individual practices that range from the rather easy to the very difficult. It is fairly common, for instance, for teachers to make use of extended projects (WKPJ) and hands-on activities (HOHO). Engaging students in interactive work, such as collaborating on projects (HOOS) or discussing the results of hands-on activities (HOTK), are also relatively common instructional strategies. Other kinds of hands-on work, however, are found with much less regularity in science classrooms. As was the case for writing assignments and assessments, these difficult activities include culminating assignments such as oral reports (HOOR) and those that incorporate more innovative methods of collecting and presenting scientific information using interviews (RWIN), audio-visual records (RWAV, RWPH), and computer technology (HOCO).

On the whole, the area of science instruction in which it appears to be most challenging to incorporate standards-based methods is the delivery of content or scientific information to students ($\beta_{avg} = 1.539$, excluding DMTK). This aspect of science education continues to maintain a rather traditional cast. As noted above, teacher-initiated discussion or lecture still occupies a central place in the classroom and may at times be reminiscent of the drill-and-practice pedagogical style that inquiry-oriented standards-based approaches have tried to supplant. Somewhat more interactive methods of communicating content knowledge, such as teacher demonstrations (DMDM) or discussions of science issues in the news (HODN), are used with some regularity. Only teachers deeply immersed in standards-based instruction, however, are likely to use more innovative strategies that present scientific inquiry in its more applied expressions. Activities of this kind might include studying books other than textbooks or magazines about science (HORB), using library resources (HOLI), or showcasing applied science through guest speakers (SCGU) or field trips (SCFT).

Throughout this discussion, we have found that a number of the more difficult-to-enact elements in this model of standards-based instruction involve practices that require access to particular forms of technology, such as computers and audio-visual equipment, or the expenditure of resources for field trips. On one hand, this suggests that access to certain resources may facilitate the implementation of a standards-based instructional strategy. It should be no surprise that the provision of a high-quality education (standards-based or otherwise) requires the commitment of substantial resources to schools and classrooms. We should also point out, however, that among the ranks of the most difficult practices we do find other activities that make few resource demands (e.g., oral reports). Conversely, some activities that might require special instructional supplies, facilities, or substantial material resources of other kinds (e.g., lab assignments and hands-on projects) are among the easier aspects of standards-based instruction to implement. So while resource availability may influence a teacher's choice of classroom practices, the enactment of a credible form of standards-based science in the classroom does not appear to be overly dependent on access to exceptional instructional resources.

Figure 10: Linked Difficulty and Instructional Scales, Hands-On Science



Placing Standards-Based Instruction in Context

The preceding section has concentrated on explicating the internal organization of standards-based instructional practices by an in-depth examination of the properties of the individual teacher practices used to estimate an empirical measurement scale for this teaching strategy. By discovering which practices were in turn easier and more difficult to apply in the classroom, we have been able to surmise the outlines of a generalized developmental process through which teachers might progressively implement a standards-based form of science instruction. For instance, a teacher would probably begin by emphasizing standards-based educational values, then incorporating more innovative practices into student writing tasks and assessment methods, engaging students in substantial hands-on and interactive activities, and then finally delivering content using more applied interactive methods and real-world sources.

The instructional difficulty scale upon which these observations are based, however, has a flip side—the measurement scale that captures the extent to which teachers use standards-based instruction (i.e., the latent trait or θ scale). A norm-referenced method for measuring standards-based instruction would typically produce a score that indicates a teacher's use of this strategy relative to the average teacher in the sample (often expressed as the number of standard deviation units above or below the sample mean). One of the benefits of the IRT-based measurement strategy used in this study, however, is that the measurement model empirically links the scales for items and for individuals and represents both item difficulties and individual trait estimates on the same metric (conventionally expressed in roughly standard-normal units called *logits*, although transformable to

other more intuitive metrics). This feature of IRT analysis allows us to draw a more concrete, criterion-referenced interpretation for a particular teacher's score by identifying the specific kinds of practices in which a teacher with that level of standards-based instruction is likely to engage.

In Figure 10 we engage in an interpretive exercise to illustrate the criterion-referenced nature of the IRT scale, focusing specifically on practices related to hands-on science activities. The horizontal axis of the figure depicts the estimated IRT scale for standards-based science. We can think of this scale as a ruler to measure both the difficulties of classroom practices and the use of standards-based instructional practices. The vertical bars above this axis form a histogram that represents the frequency distribution of teachers with respect to the standards-based instruction scores derived from the final IRT measurement model (estimated θ). The levels of standards-based instructional usage closely approximate a normal distribution in this national sample of teachers (mean=-.004, s.d.=1.108).

Along the horizontal axis we have marked a number of reference points, each providing a criterion-referenced interpretive linkage between a particular point on the theta scale and the expected level of usage for a particular instructional practice. Our examples are limited to a consideration of different aspects of hands-on science, although other elements of instruction could have been considered. The three questionnaire items are student engagement in hands-on activities in class (HOHO), assessment based on hands-on activities (ASHO), and the proportion of a student's grade based on these hands-on activities (GRHO). Each of these items has a polytomous format, containing four response categories. (See the discussion of model development above for details on the recoding of assessment items, which initially had five response categories.)

The theta score that corresponds to a particular level of use for a given instructional practice can be calculated on the basis of the item difficulty and (in the case of polytomous items) the category step parameters. For dichotomous instructional items that indicate whether a practice is used versus not used, there would be a straightforward correspondence between the item difficulty and estimated theta scales. For instance, a teacher with a instructional score (θ) of 1.0 would have a 50% chance of using a practice with an item difficulty score (β) of 1.0. The interpretation of polytomous items, however, requires an additional step. Here we must also take into account the category step parameter values (δ) which, in conjunction with the difficulty parameter, allow us to identify the most likely *level* of usage (i.e., response category) associated with a given theta score. Specifically, if we subtract a category step parameter from the difficulty parameter, the resulting value corresponds to the point on the trait scale at which that response becomes the most probable.

As a more concrete example, we take the frequency of hands-on activities in the classroom, which has the following parameter values:

α = 1.094

β = -.488

δ_1 = not estimated (never or hardly ever)

δ_2 = 1.484 (once or twice a month)

δ_3 = -.095 (once or twice a week)

δ_4 = -1.576 (every day or almost every day)

Teachers with the lowest theta levels for standards-based instruction are expected to be in the lowest response category for use of hands-on activities—never or hardly ever. To calculate the theta level at which teachers would be expected to transition into a higher level of use, we subtract the category 2 step parameter from the difficulty parameter. The same calculations can also be repeated to identify

the theta levels associated with the thresholds that mark subsequent transitions to higher levels of use for this practice.

No use to Monthly use: $\beta - \delta_2 = -.488 - 1.484 = -1.972$

Monthly use to Weekly use: $\beta - \delta_3 = -.488 - (-.095) = -.581$

Weekly use to Daily use: $\beta - \delta_4 = -.488 - (-1.576) = 1.088$

These values also correspond to the points on the theta scale where the item category response curves intersect, demarcating the expected response for teachers in a particular theta range. The first panel of Figure 5 provides graphic illustration of ICRCs for this survey item from a preliminary IRT analysis, where item parameters differ slightly compared to the final model. In the final analysis, however, we find that teachers with estimated theta values less than -1.972 would be expected not to use hands-on activities, those with scores between -1.972 and -.581 would use this practice on a monthly basis, and so forth. The category transition points or steps for teachers' frequency of using hands-on science activities in the classroom are marked on the IRT scale in Figure 10. The theta scores corresponding to item steps for assessment using hands-on activities (ASHO) and weight placed on hands-on science in grading (GRHO) are also indicated.

By overlaying the standards-based instruction frequency distribution with the IRT scale, we gain a sense of the relative proportion of teachers who fall above and below particular levels of use for these three practices. For instance, nearly all teachers are expected to incorporate some level of hands-on science into their student activities, assessment methods, and grading. That is, few teachers have theta values below (to the left of) the lowest category steps for these practices ($\beta - \delta_2 = -1.972$, -2.227, and –2.675, respectively). By contrast, we would expect a relatively small number of teachers to use hands-on activities daily and even fewer to base student grades mostly or entirely on these activities ($\beta - \delta_4 = 1.088, 1.844$).

This analytic exercise can also be employed to determine the kinds of hands-on instructional practices that a typical middle school science teacher would use. As mentioned earlier, based on the formal IRT model, the average trait level (θ) for this sample of teachers is -.004 logits. Matching this trait level against the instructional difficulty scale (which shares the same metric), we would expect an average teacher to use hands-on student activities weekly and to assess student learning using these kinds of tasks with a similar frequency. Such a teacher would base about half of student grades on these hands-on assignments.

Finally, an analytic exercise of this kind can also shed light on the finer points involved in implementing a standards-based science education as a set of coordinated instructional practices. As an illustration, we will again take the related practices of engaging students in hands-on work and assessing student learning using these hands-on tasks. It is reasonable to assume that teachers cannot assess students on hands-on activities unless they have first introduced such tasks into the classroom. As a result, these paired teaching practices should be linked in a predictable way. Namely, hands-on student activities should be easier to incorporate than assessment at a given level of frequency (i.e., student activities will be implemented before assessment practices). The similar question formats for these items share two response categories in common—monthly and weekly use. We can, therefore, directly test for the presence of the developmental connection between student activities and assessments suggested above. The illustration in Figure 10 confirms our expectation. The IRT scale value associated with monthly use of student hands-on activities is, as hypothesized, lower than the corresponding value for hands-on assessment (-1.972 vs. –1.300). A similar pattern is also found for use of these practices on a weekly basis (-.581 vs. -.409).

Although we could explore similar questions regarding the connections between other clusters of practices (e.g., those involving the use of science projects), a full explication of the model for standards-based instruction in this fashion is beyond the scope of this report. Even the limited illustration using hands-on science practices, however, helps to highlight the richness of an IRT-based approach to measuring instructional practices. An interpretive exercise of this kind is useful for several reasons. First, it places the somewhat abstracted results of the statistical IRT model in a more concrete and practical context by expressing item parameters in terms of teacher behaviors. Overlaying the scales for instructional and teacher scales also helps to illustrate the prevalence of a standards-based approach to middle school science and to express levels of teacher implementation in terms of the likelihood that a teacher will use certain practices with a certain degree of frequency. Finally, this type of analysis also provides a means for further exploring the internal organization and relationships among the constituent items that together represent our theoretical construct. For example, a careful examination of item parameters can tell us which standards-based practices or broader pedagogical domains tend to be easier or more difficult to enact in the classroom. We can also gain important insights into how certain applications (e.g., student work vs. assessment) of standards-based principles (e.g., an emphasis on hands-on science) are related to one another, not just as conceptual elements of reform models or theories but also as elements of instructional strategies in actual science classrooms.

CONCLUSION

An IRT-based approach is, of course, only one way to empirically measure a particular instructional strategy. Future research would benefit from a more systematic comparison of various measurement approaches. Although such comparisons were beyond the scope of this study, the results do suggest that there are a number of benefits to an application of IRT models to instruction. The diagnostic information produced by the IRT analyses provided an empirical means to systematically identify practices that did not appear to align conceptually with the construct of standard-based instruction. In fact, the IRT models proved very sensitive in this regard, able to distinguish each of the practices that we had identified a priori as standards-inconsistent (or -inconclusive) from those expected to be consistent with a standards-based approach.

A careful examination of IRT item parameters also helped us to isolate an irregularity in responses to a series of questions that asked teachers how often they used various techniques to assess their students' progress in science. Specifically, our results suggested that two of the response categories on the questionnaire may have been too similar, resulting in a somewhat undifferentiated response pattern from teachers. The IRT analyses were able to suggest specific ways in which the format of this particular item might be revised to improve future NAEP administrations. For example, these two response categories might be collapsed or labeled more carefully so that they are mutually exclusive categories.

An IRT-based measurement strategy, although methodologically sophisticated, enables an analyst to present the results of complex statistical analyses in ways that can be accessible to a wider audience without an extensive psychometric background. For example, this study drew heavily upon the analogy of cognitive or achievement testing. Our intention was to use this type of psychometric measurement (which would be relatively familiar to a lay audience) to introduce some of the more subtle aspects of our more novel application of IRT to the measurement of instruction. IRT analyses also produce a dual measurement scale for item difficulties and trait levels that enables a more concrete, criterion-referenced interpretation of such values. In this application, we could relate particular teacher scores on the standards-based instructional scale to concrete expectations for the

kinds of classroom practices they were likely to use. Although we did not avail ourselves of this feature of IRT scales in this study, it would also have been possible to convert the estimated IRT measurement scale from its conventional *logit* units to more intuitive or familiar metrics. These might include a 4-point scale reminiscent of a grade point average, a scale approximating percentiles, or even the scales used in familiar achievement tests like the SAT or ACT.

The NAEP data used in this study are cross-sectional, providing a snapshot of a teacher's practices at a particular point in time. All analyses based on a single observation will be subject to certain limitations. The strong psychometric foundations upon which IRT analyses are based, however, allow us to infer more subtle insights from such cross-sectional information. For instance, in an IRT model we make (and test for) the assumption that there is a systematic relationship between a person's trait level, item difficulty, and the likelihood of a "correct" response. We described this earlier as IRT's *progressiveness principle*. In this application, the implication is that teachers who use the more difficult standards-based practices are very likely to also be using the less difficult-to-enact practices. This allows us to extrapolate a series of developmental steps (i.e., practices) through which a typical teacher would be likely to progress when putting standards-based instruction into practice in the classroom. A more conclusive determination of this implementation process must, of course, await confirmation using longitudinal data on teacher practices.

This notion of a developmental implementation process also has more direct implications for policy and practice. For instance, our results suggested that teachers probably start on the path toward standards-based instruction by first espousing and emphasizing the most generalized values or objectives of the standards movement. This is followed by affording increased instructional attention to more specific, advanced, or specialized science topics and skills. Teachers then tend to incorporate standards-based techniques into writing assignments and classroom assessments and make use of basic hands-on activities. Only at more advanced stages of implementing a standards-based instructional style would teachers tend to employ more "enriched" forms of hands-on activities or modes of presenting science content knowledge. It is interesting to note that this progression from general to increasingly more specific and concrete actions mirrors in many ways classic models of the adoption and diffusion of technical innovations (Rogers, 1995).

Reformers seeking to promote standards-based instruction might benefit from an understanding of these progressive stages of implementation. For instance, this kind of information might suggest where scarce resources would be most effectively directed. One possible approach to promoting standards-based science education might involve providing teachers with greater access to computers or specialized forms of audio and video technology that could be used for hands-on activities. Our analyses, however, suggest that for all practical purposes activities involving these kinds of technology tend to play a very small role in standards-based science. Very few teachers have progressed to a stage where they could effectively integrate these elements into their instructional practices. At the other extreme, most teachers report focusing on the basic tenets of standards-based reform, such as understanding key science concepts, encouraging an interest in science, or promoting problem-solving skills. Expending additional resources to promote teacher awareness of these general principles would appear to be unnecessary. The best all-around reform strategy might prove to be focusing on moderately difficult practices that are within the reach of most teachers. A reform strategy of this kind might involve providing professional development, showing teachers how to more effectively incorporate specific kinds of hands-on activities or collaborative (i.e., group) tasks into their science teaching.

Finally, as we noted at the beginning, developing a reliable measure of standards-based instruction represents a first, but essential, step in a larger research agenda. Based on the results of this study, we have reason to believe that standards-based science does exist as an identifiable style of instruction in actual science classrooms, and that we have the means to reliably measure variations

in the strength and coherence of this instructional approach. This will enable us to address a number of critical issues that are central to understanding the progress of standards-based reform in future studies. For instance, we may ask whether the coherence of SBI and its prevalence are affected by policy initiatives at the state level. Using data from the State NAEP Assessment we can explore this connection between policy and practice. Similarly, we could examine future NAEP administrations to track the progress of the standards movement over time, in terms of the degree to which teachers are enacting a standards-based form of science in their classrooms. Finally, as implied by the slogan "high standards for all," one of the goals of the standards movement is to provide all students with the opportunity to receive high-quality instruction. This applies to students who are educationally, socially, and economically at-risk, as well as to their more advantaged peers. Extensions to this study will be able to address both the equity and efficacy of this reform strategy by first determining the extent to which specific groups of students have differential access to standards-based instruction, and then by exploring the extent to which these instructional gaps may be implicated in the achievement gaps found among these groups.

REFERENCES

Borman, K.M., Cookson, P.W., Sadovnik, A.R., & Spade, J.Z. (Eds.) (1996). *Implementing educational reform: Sociological perspectives on educational policy*. Norwood, NJ: Ablex Publishing Corporation.

Council of Chief State School Officers (CCSSO) (2000). *Using data on enacted curriculum in mathematics and science: Sample results from a study of classroom practices and subject content*. Washington, DC: Author.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fuhrman, S.H., Clune, W.H., & Elmore, R.F. (1988). Research on education reform: Lessons on the implementation of policy. *Teachers College Record*, *90,* 237-257.

Kenney, P.A., & Silver, E.A. (Eds.) (1997). *Results from the Sixth Mathematics Assessment of the National Assessment of Educational Progress*. Reston, VA: National Council of Teachers of Mathematics.

Lord, F.M. (1980). *Applications of Item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

McLaughlin, M.W. & Shepard, L.A. (1995). *Improving education through standards-based reform*. Stanford, CA: National Academy of Education.

Muraki, E. (1992). A Generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*(4), 351-363.

Muraki, E. & Bock, R.D. (1997). *PARSCALE: IRT based item analysis and test score for rating-scale data*. Chicago: Scientific Software International.

National Assessment of Educational Progress (NAEP). (1988). *Mathematics objectives: 1990 assessment*. Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (NAGB) (1999). *Mathematics framework for the 1996 and 2000 National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board (NAGB) (2000). *Science framework for the 1996 and 2000 National Assessment of Educational Progress*. Washington, DC: Author.

National Center for Education Statistics (NCES) (2000). *National Assessment of Educational Progress: 1996 National Assessment secondary-use data files user guide*. Washington, DC: U.S. Department of Education.

National Center for Education Statistics (NCES) (1999). *The NAEP 1996 technical report*. Washington, DC: U.S. Department of Education.

National Commission of Educational Excellence (NCEE) (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: Government Printing Office.

National Council of Teachers of Mathematics (NCTM) (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics (NCTM) (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

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

Porter, A.C., & Smithson, J.L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S.H. Furhman (Ed.), *From the capitol to the classroom: Standards-based reform in the states.* Chicago: University of Chicago Press.

Project 2061 (American Association for the Advancement of Science) (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology.* Washington, DC: Author.

Project 2061 (American Association for the Advancement of Science) (1993). *Benchmarks for science literacy.* New York: Oxford University Press.

Project 2061 (American Association for the Advancement of Science) (1997). *Resources for science literacy: Professional development.* New York: Oxford University Press.

Project 2061 (American Association for the Advancement of Science) (1998). *Blueprints for reform: Science, mathematics, and technology education.* New York: Oxford University Press.

Project 2061 (American Association for the Advancement of Science) (2001). *Designs for science literacy.* New York: Oxford University Press.

Ravitch, D. (1995). *National standards in American education: A citizen's guide.* Washington, DC: Brookings Institution.

Rogers, E.M. (1995). *Diffusion of innovations, fourth edition*. New York: Free Press.

Schmidt, W.H. (Ed.) (2001). *Why schools matter:A cross-national comparison of curriculum and learning*. San Francisco, CA: Jossey-Bass.

Schmidt, W.H, McKnight, C.C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics.* Boston: Kluwer Academic Publishers.

Smith, M.S., & O'Day, J. (1991). Systemic school reform. In S.H. Fuhrmann & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-268). Philadelphia: Falmer Press.

Stecher, B., Hamilton, L., Ryan, G., Le, V., Williams, V., Robyn, A., & Alonzo, A. (2002). Measuring reform-oriented instructional practices in mathematics and science. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, April 2002.

Stevenson, D.L., & Schiller, K.S. (1999). State education policies and changing school practices: Evidence from the National Longitudinal Study of Schools 1980-1993. *American Journal of Education*, *197,* 261-288.

Swanson, C.B. (forthcoming). Organizational coupling, control and change: The role of higher-order models of control in educational reform. In L. Hedges & B. Schneider (Eds.), *Social organization of schooling*. New York: Russell Sage Foundation.

Swanson, C.B., & Stevenson, D.L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP State Assessments. *Educational Evaluation and Policy Analysis, 24,* 1-27.

Wattenberg, R. (1995-96). Helping students in the middle: What average students can achieve when standards are high and the stakes are clear. *American Educator*, *19,* 4-18.

Zucker, A.A., Young, V.M., & Luczak, J.M. (1996). *Evaluation of the American Association for the Advancement of Science's Project 2061 (Volume I: Technical report).* Menlo Park, CA: SRI International.