Science
Education

SCIENCE EDUCATION POLICY

# Science Education and Test-Based Accountability: Reviewing Their Relationship and Exploring Implications for Future Policy

KEVIN J.B. ANDERSON
*School of Education, University of Wisconsin–Madison, Madison, WI 53706, USA*

*Received 1 December 2010; accepted 18 May 2011*
*DOI 10.1002/sce.20464*
*Published online 6 November 2011 in Wiley Online Library (wileyonlinelibrary.com).*

**ABSTRACT:** Assuming that quality science education plays a role in economic growth within a country, it becomes important to understand how education policy might influence science education teaching and learning. This integrative research review draws on Cooper's methodology (Cooper, 1982; Cooper & Hedges, 2009) to synthesize empirical findings on the relationship between science education and test-based accountability policies. Current accountability policy, particularly at the federal level, is intended to influence educators to more fully consider the needs of *all* students; however, research suggests that, under these policies, many research-based reform efforts in science become sidetracked, teacher practice becomes more fact based, science is taught less, teachers become less satisfied, and many students' needs are not met. Therefore, a clear understanding of educators' perceptions of the impacts of current test-based accountability policies should guide the development and implementation of the next generation of national science standards and subsequent large-scale assessments. By also delineating the limitations of the research into the perceived connections between test-based accountability and science education, this synthesis reveals further research to be done. Finally, this paper details what the reviewed research suggests for improvements to K-12 science education accountability policies.  *Sci Ed* **96:**104–129, 2012

*Correspondence to:* Kevin J.B. Anderson; e-mail: kanderson8@wisc.edu
Accepted under the editorship of Gregory Kelly.

## INTRODUCTION

World-altering events such as World War II, the launch of Sputnik, and the current global economic downturn have brought both science practice and science education to the forefront of public discourse. Furthermore, policy makers have paid special attention to science teaching in schools as it helps to create and prepare the next generation of innovators. When studies critique the quality of public education, as with the *Nation at Risk* (National Commission on Excellence in Education, 1983) and *Rising Above the Gathering Storm* reports (Committee on Science, Engineering, and Public Policy, 2007), government leaders associate problems with science education with the future economic vitality of the United States and its position as a global leader.

Science education continues to iteratively move through reform efforts, from constructivism to direct instruction, and from local accountability to national standards. This study presents a systematic review of literature on the relationship of one reform effort, test-based accountability, to science instruction, from the viewpoint of practitioners. Although many studies have examined this connection empirically (e.g., Diamond & Spillane, 2004; Hamilton et al., 2007; Settlage & Meadows, 2002; Stecher & Barron, 1999), there has been no recent synthesis of these studies. Looking across these studies enables an understanding of the broad perceptions of test-based accountability and its effects regardless of variations in local and state policy. While some studies briefly review research relating science education and accountability policies, they do not exclusively consider empirical research or discuss the limitations of existing research. Few studies have considered or elaborated on further research needed to address these limitations.

An understanding of test-based accountability continues to evolve as policies change and develop. A National Science Board committee recently met to make federal policy recommendations for the National Science Foundation (NSF) to increase student innovation in science, technology, engineering, and mathematics (STEM) fields (Cavanagh, 2009). A committee convened by the National Research Council (NRC) recently put out a new framework for national science standards (Committee on Conceptual Framework for New Science Education Standards, 2010). The Obama administration enacted a new policy to improve STEM education through connecting nonprofit and corporate leadership with schools. In addition, federal officials recently funded a second round of "Race to the Top" applications and are now considering the reauthorization of the Elementary and Secondary Education Act (ESEA). While the particular nature of government methods of improving science and STEM education will likely change over time, there will always be immediate policy implications for the findings of this research review as long as test-based accountability systems maintain their current theoretical underpinnings. To inform this evolving policy context, we need to determine what we know about educators' perceptions of how accountability policies affect science education and what still needs to be learned.

Therefore, this paper examines correlations that research literature details between test-based accountability policies and reported changes in the following:

1. efforts to improve science education, including teacher professional development,
2. science teachers' current instructional practices and curriculum,
3. the amount of time and focus on science instruction,
4. meeting the needs of all students, and
5. science teachers' attitudes about teaching and professional efficacy.

For the purpose of this review, accountability policies are defined as policies that mandate tests with results made publicly available. Generally, these tests measure yearly progress toward, or attainment of, state or federal goals. The phrase "high stakes" indicates the existence of possible sanctions for not meeting these goals. However, under current federal policy (No Child Left Behind Act [NCLB]), most states only consider tests in mathematics and reading for determining Adequate Yearly Progress (AYP) and possible sanctions. Only a few states factor science and/or other subjects into these determinations, albeit at a much smaller percentage than mathematics or reading. In this sense, science tests are generally not high stakes, although teachers in the reviewed studies generally perceived them as such. Accountability in science assessments most often takes the form of publicly releasing disaggregated results from tests given at three grades during K-12 (once each in 3–5, 6–8, and 9–12).

To provide some additional background on this multifaceted issue, the next section provides a brief overview of federal and state accountability policy. Following that is a description of the methodology and framework used in this review. The bulk of the paper then presents findings from the literature pertaining to each of the five parts of the main research question—themes that were developed from findings within the literature. An extensive table summarizes this literature in Table 1. After reviewing this literature, its limitations are presented along with areas needing further research. Finally, based on the findings, suggestions are provided for reframing accountability systems to better meet the widely accepted goals of science education.[1]

## A BRIEF HISTORY OF ACCOUNTABILITY POLICY

Beginning in the early 20th century, educational leaders and theorists began attaching more quantitative measurements to educational quality; broad federal funding, however, did not require such measurements until 1965. The ESEA of 1965 set guidelines to evaluate how well education systems spent the $5 billion dollars it allocated, but it did not mandate specific data to be included in the required reports. An amendment to the bilingual education section of the ESEA in 1967 added that "all projects must provide for an independent accomplishment audit of the project" (DeNovellis & Lewis, 1974, p. 2). President Nixon used these exact ideas in his educational address to the Congress in 1970, emphasizing the need for "measurable" standards with "an independent audit of results" (Sirotnik, 2004, p. 150). The required audit, however, did not link test scores to programs so much as generate outlines of programs, lists of who was served, and records of expenditures. Beginning in 1969, the National Assessment of Educational Progress (NAEP) also tested students' subject matter proficiency. By 1973, accountability legislation had passed in 27 states, with 15 of these states having testing or other assessment programs (DeNovellis & Lewis, 1974).

During the 1970s, 1980s, and 1990s, the practice of educational testing and holding schools "accountable" through these tests continued to expand. More states established content standards and required yearly examinations to measure educational progress. The 1988 reauthorization of ESEA required annual student testing in schools and districts, which received Title I funding (Penfield & Lee, 2009). With the influence of Goals 2000 and the requirements of ESEA 1994, standards and test-based accountability became common practice (Penfield & Lee, 2009). Some states, such as Texas, made these assessments "high stakes," meaning schools faced sanctions if they did not meet set benchmarks indicating success. Testing continued to boom even though studies of high-stakes accountability did

[1] See, for example, goals elaborated through American Association for the Advancement of Science (1989), Project 2061.

**TABLE 1**
**Thirty-Five Empirical Studies Examining Accountability Policy and Science Education Practice**

| Authors (Date) | Methods | Main Findings (in relation to Accountability and Science Education) |
|---|---|---|
| Aronson (2007) and Aronson and Miller (2007) | Case study of one Maryland high school (HS): two interviews and observations of $>15$ classes of six biology teachers; review of accountability documents from federal, state and district levels | Teachers report:<br>• Feeling stress from accountability requirements<br>• Compromising between their views of high-quality teaching and their perception of accountability mandates<br>• standards = "nonnegotiable authority" |
| Coble (2006) | Open-ended interviews of six third-year HS science teachers in North Carolina: focus on how they implemented science education reforms and contextual challenges to doing so | Teachers report:<br>• Using more inquiry-based curriculum in classes not subject to a high-stakes examinations<br>• Using more "decontextualized factual information" and teacher-centered strategies in courses associated with a high-stakes tests |
| Diamond and Spillane (2001) | Theoretical sampling of eight elementary schools (ES) in Chicago: focus on two highest and two lowest performing; 181 sets of field notes from 3+ days of observations of meetings and conversations; semi-structured interviews of all second- and fifth-grade teachers and school leaders | • Time spent on science has decreased, even eliminated in some instances, with more decrease in schools on probation<br>• Science had fewer instructional specialists and leaders and received less attention from administrators |
| Donnelly and Sadler (2009) | Interviews of 22 science teachers from five purposefully selected HS in Indiana: focus on views of standards-based accountability | • Approximately half of teachers modified instruction to meet tested standards<br>• Mixed views of accountability: positive, negative, or neutral (just part of the job) |
| Font-Rivera (2003) | Survey of 86 middle school (MS) science teachers in Virginia | Teachers report:<br>• Intense administrator pressure on themselves and students to increase scores<br>• Considerable time spent on test preparation |

*(Continued)*

**TABLE 1**
**Continued**

| Authors (Date) | Methods | Main Findings (in relation to Accountability and Science Education) |
|---|---|---|
| | | • Three main instructional influences: reviewing the content and skills on state tests, preparing students for state tests, and adjusting the curriculum sequence based on state tests<br>• More frequent use of multiple-choice assessment strategies<br>• State tests do not accurately portray learning |
| Galton (2002) | Surveys of students and teachers and observations of classrooms in Year 6, England (*N* not provided) | Owing to the high-stakes examination, teachers report:<br>• Having little time to pursue students' interests<br>• Relying on more direct teaching methods<br>• Using less experimentation<br>• Observing less student enthusiasm |
| Goetz Shuler et al. (2009) | Study of NSF-funded, science education centers around the country: interviews of staff and review of reports on their work with school districts and teachers (*N* is unclear) | • Staff found state-level tests detracted efforts to help schools implement research-based assessments and curricula<br>• Staff found it tough to get districts to invest in learning new assessment methods due to traditional assessment methods of state tests<br>• Centers in Delaware and Washington helped improve the state standards and testing to better reflect deeper-level thinking (failed in other locations) |
| Hamilton et al. (2007) | Stratified, representative sample of 89 districts from three states: surveys of ES and MS teachers from 301 schools; case studies of 21 ES and nine MS, which included administrator and teacher interviews | Owing to the high-stakes testing:<br>• Teachers and administrators reported nature of tests led to better alignment of curriculum with state standards and assessments<br>• Teachers reported increased focus on student achievement and increased rigor of curriculum<br>• Science teachers reported more focus on students near the cut score and limited opportunities given to high-achieving students<br>• Teachers reported teaching to the tests |

*(Continued)*

**TABLE 1**
**Continued**

| Authors (Date) | Methods | Main Findings (in relation to Accountability and Science Education) |
|---|---|---|
| Jenkins (2000) | Random national sample of 500 secondary schools in England and Wales in 1998: 296 science teacher respondents to survey, only 239 with >10 years experience used—they had experience teaching before nationalized curriculum | Teachers attribute the following to implementation of the national curriculum and aligned assessments:<br>• Clarified objectives<br>• Increased collaboration with other teachers<br>• Increased personal frustration with their work<br>• Decreased time on science laboratories<br>• Feeling that students enjoy science less |
| Jennings and Rentner (2006) | Summary of empirical studies includes surveys of 100s of state and district leaders; case studies of a representative sample of U.S. schools/districts | • Education leaders say accountability increases attention to achievement gaps and needs of traditionally marginalized students<br>• Schools spending more time on reading and mathematics at the cost of other subjects, such as science |
| Jones et al. (1999) | Survey of 236 ES teachers from a stratified, random sample of North Carolina schools | Owing to testing, teachers report:<br>• Teaching less science, especially in schools not meeting test targets and near test times<br>• Changing instruction (with mixed results): 37% report doing more class-level testing and 2% less; 33% report more hands-on work and 15% less; 14% report more inquiry and 13% less<br>• Feeling testing had negative impact on students' love of learning (reported by 48.5%) |
| Katzmann (2007) | 1.5-year case study of four HS biology teachers in Colorado: two novice, two experienced; three interviews each, repeated classroom observations (*N* unclear); classroom artifacts | • Teachers report and observations confirm testing influences instructional decisions, assessment practices, and use of time—specifically report less inquiry-based teaching |
| Kersaint et al. (2001) | 46 interviews of principals from varying size/context schools in Chicago, El Paso, Memphis and Miami (all part of systemic reform effort in math and science); ES, MS, and HS included | • Most principals reported the hands-on, constructivist approach of the mathematics and science reform competed with standardized tests—balancing the two was very difficult; they felt testing policies would drive instruction, not reform efforts |

*(Continued)*

**TABLE 1**
**Continued**

| Authors (Date) | Methods | Main Findings (in relation to Accountability and Science Education) |
|---|---|---|
| | | • Some principals discouraged teachers missing days for reform workshops, as doing so might limit achievement on standardized tests |
| Lee and Luykx (2005) | Research in classrooms of 75 third to fifth-grade teachers in six representative ES within one school district: research done as part of collaborative reform effort; specifics of methods not specified | • Principals much more supportive of science curriculum reform with explicit connections to state standards in mathematics, reading, and writing<br>• Some teachers told not to teach what is not tested for the 2–3 months before the test<br>• Some ELL students did not receive science instruction, instead receiving more English |
| Louis et al. (2005) | Case studies of three HS, one each in Minnesota, Iowa, and North Carolina: total of 57 one hour, open-ended interviews of teachers (not just science teachers) | Owing to standards-based testing:<br>• Science teachers report doing fewer labs and inquiry-based projects and covering more content with less depth<br>• Mathematics and science teachers reported more impact than any other discipline<br>• Teachers report some enhanced professional collaboration—increased discussion on how to meet standards |
| McMurrer (2008) | Survey of leadership in random, stratified sample of 349 school districts | • Time for English and mathematics teaching increased in 53% and 48% of districts, respectively; time for science decreased in 28% of districts |
| Mintrop (2004) | Case studies of four ES and three MS in Maryland on probation and two ES and two MS in Kentucky on probation: from districts with most schools on probation; more than 30 observations and eight interviews over 2-year span | • Schools on probation had higher staff turnover<br>• Teachers reported feeling that they were being treated like children<br>• Actual instructional changes varied depending on the principal |
| Perlstein (2007) | Ethnographic look at one ES; observations and interviews of classrooms, meetings, etc., over 1-year time period | • Teacher reports and observations showed less time and emphasis on science; more on reading and mathematics<br>• Before testing science instruction was minimal |

*(Continued)*

**TABLE 1**
**Continued**

| Authors (Date) | Methods | Main Findings (in relation to Accountability and Science Education) |
|---|---|---|
| Pinder (2008) | Interview of four secondary teachers: two in science, two in mathematics; review of the content and data from past state examinations | • Teachers report a narrowing of pedagogy to a "teach-to-the-test" functionality, and being required to give state test practice questions |
| Pringle and Martin (2005) | Interview of ES teachers at multiple sites within one district in Florida (uncertain *N* or randomness) | • Teachers report fear of students failing mathematics and reading tests prevents science teaching.<br>• Teachers report science teaching becoming more fact based |
| Rentner et al. (2006) | Survey of education leadership from all 50 states: nationally representative survey of 299 school districts; case studies of 38 geographically diverse districts and 42 schools | Owing to NCLB, teachers and leaders report:<br>• Greater efforts being made to align curriculum and instruction with state standards and tests<br>• Less instructional time given to science<br>• Less creativity in teaching and learning, and use of fewer activities that might aid engagement<br>• Higher expectations for all students<br>• Instances of increased teacher stress and decreased morale |
| Rodgers (2006) | Interviews of 14 teachers and five administrators at two ES and one MS in Texas; classroom observations conducted of each teacher | • Lower grades (1–4) changed little in time spent on science /with or without administrative leadership<br>• In fifth grade, year of state science testing, time on science increased, but methods often focused on increasing test scores, not best-practices<br>• Most teachers accepted personal accountability, but claimed high-stakes tests damaged students emotionally and decreased their love of learning science |
| Saka (2007) | Case studies of two beginning, secondary science teachers: three surveys, nine questionnaires, 20 interviews, 13 observations of each + interviews of coworkers and administrators | • Science teachers asked to drill mathematics and reading skills during their science class<br>• Teachers' implementation of reform-based curriculum depends on personal and contextual factors, with one pursuing this curriculum in spite of perceived restrictions of accountability |

*(Continued)*

**TABLE 1**
**Continued**

| Authors (Date) | Methods | Main Findings (in relation to Accountability and Science Education) |
|---|---|---|
| Settlage and Meadows (2002) | Personal conversations with and observations of urban secondary, science teachers (*N* unclear, methodology informal) | • Teachers report that standards and assessments undercut professionalism, diminish relationships with students and water down science curriculum<br>• Teachers see disconnects between district curriculum and state tests |
| Shaver et al. (2007) | Questionnaire followed up with focus groups of 43 third- and fourth-grade teachers in a large, urban school district with a large percentage of ELLs, located in the southwest | • Few studies have looked at how accountability policies impact the learning of ELLs<br>• Teachers report less hands-on and other research-based teaching methods due to accountability measures, even though teachers know they work better for ELLs |
| Shepard and Dougherty (1991) | Teacher surveys ($N = 360$) of a random sample of third-, fifth-, and sixth-grade teachers in two large districts: one district in the southeast and one in the southwest; both had standardized testing programs described by district personnel as "very high stakes" | • 53% felt "great pressure" and 26% felt "substantial pressure" from the district or board to raise test scores; few (5.5% and 4.7%) felt great pressure from parents and other teachers<br>• 42% said they read more books about science; 42% say that emphasis has not changed<br>• 64% said they spend less time on science<br>• Many reported giving less attention to higher-order thinking and extended projects |
| Smith and Southerland (2007) | Case studies of two ES teachers' science teaching and related beliefs; including repeated interviews and observations (*N* unclear) | • Teachers reported feeling constrained in using project or inquiry-based learning because of the need to cover all tested material<br>• Teachers reported the tests assess knowledge of isolated facts rather than deep content knowledge or critical thinking |
| Smith (1991) | Case studies of two ES in Arizona: 15 months of data collection; interviews of teachers and administration, observations of classrooms and meetings, and document analysis (*N* unclear) | • Teachers reported decreased to no science taught due to the standardized tests, and decreased use of active-learning methods within science, especially approaching testing time<br>• Teachers reported increased stress and frustration |

*(Continued)*

**TABLE 1**
**Continued**

| Authors (Date) | Methods | Main Findings (in relation to Accountability and Science Education) |
|---|---|---|
| Stecher and Barron (1999) | Stratified, random sample of ES and MS in Kentucky: fourth- to seventh-grade teachers within those schools were contacted by mail with survey, 479 teachers responded | • Professional development (PD) was more likely when subjects were tested: 50% of fourth-grade teachers reported significant science PD (science is tested); 28% of fifth-grade teachers report significant science PD (science not tested)<br>• Time per subject and curriculum covered was highly influenced by what was tested there<br>• Teachers report teaching to the test more than underlying standards for understanding |
| Stuart Hammer (2004) | Survey of 265 randomly selected MS science teachers in Florida | Owing to the state science test (Florida Comprehensive Assessment Test [FCAT]):<br>• 92% reported increased stress<br>• 86% reported loss of autonomy and control over their classes<br>• 54% reported loss of freedom and creativity in curriculum and lessons<br>• 83% believe that time on test prep has come at expense of more important items<br>• 19% believe that the science FCAT has brought about improvement in curriculum, instruction, and student learning in science<br>• 25% believe the reform efforts (includes standards) will improve their school |
| Taylor et al. (2008) | Semistructured interviews of 37 scientists from various disciplines and 21 secondary school science teachers | • Teachers and scientists claimed that test-based accountability is reducing science to facts, making inquiry/creative activities more difficult, decreasing the quality of science instruction, and constraining efforts to improve science education |
| Tye and O'Brien (2002) | Survey of 114 random individuals who received teaching credential from Chapman in California during 1991–1992, 1994–1995 (not only science) | • Top reason why individuals no longer teaching reported leaving the profession in 2001 was accountability; second reason was increased paperwork |
| Vogler (2002) | Survey of a stratified, random sample of 10th-grade Massachusetts mathematics, science, and English teachers ($N = 257$, 84 in science) | • Science teachers report aligning practice with state assessment, and using more research-based pedagogy (such as problem-solving activities and rubrics) |

*(Continued)*

**TABLE 1**
**Continued**

| Authors (Date) | Methods | Main Findings (in relation to Accountability and Science Education) |
|---|---|---|
| Wideen et al. (2007) | Study of random, stratified sample of 10 districts in British Columbia, Canada; two random schools per district; interviewed and observed one teacher per school from Grades 8, 10, and 12 ($N = 80$); interviewed principals, students, and district administrators in these 10 districts ($N$ unclear); pilot case studies of two districts informed research | • Majority of teachers felt the Grade 12 tests discouraged inquiry and active learning, diminished the quality of classroom discourse, and reduced creativity and depth of learning<br>• Teachers reported that high-stakes examinations in Grade 12 decreased use of research-based methods in that grade compared to other grades taught by same teacher<br>• Most teachers in grades where testing is not seen as high stakes reported it had minimal influence |
| Wood (1988) | Observed 4 seventh-grade science teachers and 165 students in six classes; interviewed the four teachers and their principal | Teachers stated that accountability policy:<br>• Constrained and routinized their behavior, leading them to violate their own standards of good teaching<br>• Pressured them to get through the material and cover facts instead of ensuring understanding<br>• Inhibited students' opportunities to ask questions and express curiosity<br>• Limited ability to invite guest scientists to class |

ES: Elementary school, HS: high school, MS: middle school, and PD: professional development.

not find it necessarily led to improved student learning outcomes (Costigan & Crocco, 2004).

In 2001, the NCLB renewed the ESEA with an increased focus on test-based accountability and public reporting of results. Arguably, NCLB "is much stronger and more far-reaching than previous federal efforts to raise education standards" (Gamoran, 2007, p. 4). It is also more diligently enforced (Loveless, 2007). While all states had yearly assessments prior to NCLB, only a handful were high stakes.

## METHODOLOGY AND FRAMEWORK

Following the methodology outlined by Cooper (1982), this integrative research review synthesizes and analyzes *empirical* studies dealing with test-based accountability and science education to make generalizations about the relationship between the two.

Methods also include aspects of research synthesis as detailed more recently by Cooper and Hedges (2009). Specifically, the aim of this study is to "integrate empirical research" to infer generalizations about what changes in science education are perceived as being due to accountability policies (Cooper & Hedges, 2009, p. 5). This literature review also reveals further research needs and educational policy implications. As Cooper (1982) suggests, threats to the validity of the findings will be addressed through descriptions of how literature was found, criteria for accepting this literature, and limitations of this literature.

Methods included searching a wide range of databases and academic journals for connections between science education and accountability policies including assessment. Databases included Google Scholar; JSTOR; Academic Search; EBSCO Host, including ERIC; Web of Knowledge; LexisNexis; and ProQuest, including the national dissertations database. Past issues of highly cited science education and policy journals, including the *Journal of Research in Science Teaching*, *Science Education*, *Educational Policy*, and *Educational Evaluation and Policy Analysis*, provided lists of accountability terminology used in further literature searches. To focus on the modern era of accountability ushered in through the *Nation at Risk* report, this review used only articles published after 1983. While NCLB changed the nature of federal accountability, including reactions to accountability policies throughout recent history broadens the context for understanding these policies and their relationship to science education. Searches employed the following terms and combinations of terms: science, science education, No Child Left Behind, NCLB, test-based, accountability, assessment, testing, high-stakes, standardized, state, federal, policy, and standards. Searches included a wide range of topics to capture broad definitions and descriptions of accountability. Bibliographies of relevant studies provided other relevant studies to investigate. Notably, this review only looked at studies relating to systemwide accountability testing used to comply with state or federal policies, and did not consider studies on the ACT, SAT, or high school exit examinations.

Each scholarly work included in this review met three criteria. First, each appeared in a peer-reviewed journal or conference proceeding, was an accepted dissertation, or was a report of a large-scale empirical study by government-funded educational organizations.[2] Second, each article included empirical data and the authors supported their assertions with these data. Editorials, position papers, or research reviews were not included. Very few rigorous studies of stratified, random samples of teachers or schools could be found; even fewer of those made comparisons across school systems. Therefore, this review included studies within one school or a small number of teachers. This wide range of studies, while not generalizable to every context, reflects an extensive range of policy practices and geographical areas. Third, each study specifically addressed the connection between K-12 test-based accountability and science instruction and/or learning. For example, some studies discussed accountability measures, but did not differentiate findings about standards-based reform from findings about testing, potentially conflating two policies that are perceived very differently (Donnelly & Sadler, 2009).

International articles on accountability within science education that met the above criteria were included in this review as the relationship between accountability and science in other countries provides a broader picture of this interaction. While the national curriculum and standards of these countries contrast with current U.S. educational policy, continued

[2] However, informative anecdotes about science education in an accountability context can be found in non-peer-reviewed sources. See, for example, the discussion of decreasing science fair involvement in Albuquerque, New Mexico (Peter, 2008) or the *Time* article on a school in Iowa, where teachers no longer take their students on field trips to the natural history museum or to see bald eagles on the Mississippi River (Ripley, 2004).

movement toward common core standards and national examinations in the United States make an international perspective relevant.

This paper does not make causal claims. It reveals correlations between test-based accountability policy and educators' perceptions of changes in instructional practice. While studies in this review consistently implied or claimed accountability and testing *affect* science education practices, their methods do not support causal arguments as described in standards from the American Educational Research Association (AERA, 2008). These studies do, however, fit under the AERA's definition of scientifically based research, which requires the use of rigorous, systematic, and objective methodologies. Furthermore, as suggested by the AERA, these studies use appropriate methods and provide data to validly support their findings.

Teachers and administrators in these studies do use causal language, reporting that testing or NCLB has affected them or their teaching in some way. In this review, these educator sentiments will be described, while also acknowledging the limitations of self-reporting. Even with these limitations, however, policy studies repeatedly cite the importance of a stakeholder perspective in understanding policy implementation and outcomes—a perspective this reviewed literature abundantly provides (Elmore & McLaughlin, 1988; Phillips, Freeman, & Wicks, 2003).

Underlying this paper is a theoretical framework of how policy influences practice. Teachers are constantly inundated with implicit and explicit messages about what they should teach (Coburn, 2001). Science teachers may get messages that they perceive to be incongruent, such as the importance of inquiry from professional societies versus the need to cover all the standards from their administration. Based on their own background, beliefs, and knowledge, at least some of which is constructed through interactions with others, educators go through a personal sensemaking process to construct meanings for these policies (Cohen & Ball, 1990; Goldstein, 2008). Institutional theory tends to hold that teachers' actual classroom work is decoupled from policy (Coburn, 2001); however, detailed curriculum guides tied to testing and accountability measures may be challenging that traditional institutional framework, giving increased relevance to cognitive, sensemaking frames (Sykes, O'Day, & Ford, 2009). An institutional theory suggesting that teachers can close their classroom doors and ignore many policy requirements does not hold up as well under intense accountability frameworks. With individualized reporting of test score data and expectations to strictly follow curriculum related to these tests, teachers must make sense of how these tests and curriculum fit within their understanding of effective teaching and meeting student needs.

## REVIEW OF LITERATURE FINDINGS

Generalizations from the literature will be presented within the five research themes found within that literature. As mentioned above, these findings include connections between test-based accountability and (1) efforts to improve science education, including teacher professional development, (2) science teachers' current instructional practices and curriculum, (3) the amount of time and focus on science instruction, (4) meeting the needs of all students, and (5) science teachers' attitudes about teaching and professional efficacy. A detailed listing of the empirical studies discussed in these sections, including methodology and main findings, can be found in Table 1.

### Relationship With Other Reform Efforts

Research suggested that accountability policies alter the practice of science education because they limit the effectiveness of other reform efforts. The American Association for

the Advancement of Science (AAAS), NRC, NSF, and other scientific organizations support the use of inquiry-based instruction, constructivist learning, project-based learning, student-centered teaching, and other pedagogical approaches shown to engage students in science and advance their learning. Teachers, administrators, and educational consultants largely believed that reform efforts aimed at implementing these types of teaching strategies get sidelined by efforts to improve standardized test scores (Goetz Shuler, Backman, & Olson, 2009; Kersaint, Borman, Lee, & Boydston, 2001; Lee & Luykx, 2005). Researchers also noted a clear disincentive for states to adopt the rigorous standards proposed by these national organizations: it would make meeting standards more difficult (Aronson & Miller, 2007).

In some studies, researchers found that NSF-funded work to implement research-based science reform competed with accountability-driven efforts, because school leaders did not think the research-based science reform was the best means to improve critical test scores. Principals, in particular, reported difficulty balancing accountability policies with building a constructivist approach to science learning (Kersaint et al., 2001). Kersaint et al. (2001) interviewed 46 principals supported by NSF-funded science education centers in four cities across the nation. Most of these principals expressed concern that extended teacher participation in NSF-supported professional development would hurt student achievement on standardized tests. The principals felt that testing policies, not reform ideals, would, by necessity, drive instruction. Another study found that the staff at these and other NSF-funded centers struggled to convince district leaders to invest in new, research-based assessment methods because state standardized tests evaluated schools in a much more traditional, fact-based method (Goetz Shuler et al., 2009). Notably, staff at centers in Delaware and Washington did make inroads after influencing the states to make testing and standards better reflect deeper-level thinking. An in-depth case study of two elementary teachers' work and beliefs further noted disconnects among different systems of standards (Smith & Southerland, 2007). These teachers felt the deeper science literacy and inquiry skills emphasized in national standards contradicted state curriculum and accountability tests. Similarly, in their work as science consultants, when Lee and Luykx (2005) made recommendations to schools, leaders generally only implemented their recommendations when these science reforms explicitly connected with the standards assessed in high-stakes subject areas (i.e., mathematics and reading).

### Instructional and Curricular Changes

In addition to disrupting new research-based reform efforts, many teachers perceived accountability as altering current science instruction. Frequently, teachers felt that they must teach to the test (Pinder, 2008). Teachers also asserted that state tests, not teacher professional opinion, frequently became the primary influence in instructional planning (Font-Rivera, 2003). While teachers and administrators reported that this test focus leads to positive outcomes such as improved instructional alignment and greater attention to the learning outcomes of all students (Hamilton et al., 2007), policy makers likely did not intend other reported consequences. Teachers reported that they no longer teach the way they think is best (Aronson, 2007; Galton, 2002). Teachers saw science curriculum becoming more fact based (Pringle & Martin, 2005; Settlage & Meadows, 2002; Smith & Southerland, 2007; Taylor, Jones, Broadwell, & Oppewal, 2008). Inquiry-based lessons reportedly happened less frequently (Coble, 2006; Katzmann, 2007; Wideen, O'Shea, Pye, & Ivany, 1997), as did other active-learning strategies such as laboratory work (Jenkins, 2000; Perlstein, 2007; Smith, 1991).

Science teachers reported compromising their teaching practice to accomplish what they perceived as accountability mandates. Aronson (2007) observed and interviewed science

teachers in a case study of one high school. The teachers cited accountability requirements as the reason for change in their practice; they reported that their previous high-quality practices did not mesh with the new directives of standards-based accountability. A British study found similar results (Galton, 2002). Teacher and student surveys and classroom observations indicated that teachers relied on more direct-teaching methods and incorporated less experimentation during Year 6 as students prepared for the high-stakes, national examinations that occur at the end of primary school. Teachers further reported having insufficient time to pursue students' interests due to the need to cover material on the tests.

The reviewed studies generally indicated that accountability measures emphasize isolated facts rather than higher order thinking (Pellegrino, Chudowsky, & Glaer, 2001). Even when tests try to assess higher order thinking skills, they do not necessarily influence teachers to teach these skills (Mintrop, 2004). In any case, most state science assessments rely on multiple-choice formats and test vocabulary and factual knowledge rather than application of concepts or problem-solving skills (Pellegrino et al., 2001). In depth conversations with four urban science teachers and their own experiences as science education consultants convinced Settlage and Meadows (2002) that the science curriculum had become diluted by the necessity of covering long lists of unrelated topics. Teachers declared that leaders actively discouraged them from teaching anything that did not help students decode standardized test questions. Teachers maintained that "science is being reduced to a myriad of facts" due to the implementation of NCLB policies (Taylor et al., 2008, p. 1072).

As previously mentioned, national scientific research organizations encourage the use of inquiry-based lessons in science; however, teachers reported that accountability testing discourages the use of these teaching strategies. Through observations and interviews of science teachers in Colorado, Katzmann (2007) found that state science assessments altered the use of class time and decreased the amount of inquiry activities. In a Canadian study, researchers found that the amount of inquiry, active student learning, and effective classroom discourse all diminished from Grades 8 to 12. Teachers' attributed this decrease to the examinations given in the 12th grade (Wideen et al., 1997). Even though 12th-grade tests are not a required part of NCLB test-based accountability, these findings do suggest a broad influence of test-based accountability. In a North Carolina study, Coble (2006) interviewed six teachers who each taught classes associated and not associated with state testing. Teachers stated that they included much more inquiry-based curriculum in classes not connected with tests, whereas classes connected to the tests were more fact based.

Based on teachers' views, accountability measures can impact other engaging instructional practices. Wood (1988) found that after accountability policies began, teachers felt they no longer had time to invite outside scientists into the school to share their work, so they stopped inviting them. Admittedly, teachers may have cited testing as the primary reason for changes in their practice, even when other factors may also be influential.

In some studies, teachers and administrators report positive instructional and curricular outcomes of standards-based testing implementation. In a stratified, nationally representative sample of districts, Rand Corporation researchers surveyed elementary and middle school teachers and conducted case studies of elementary and middle schools (Hamilton et al., 2007). They found that high-stakes testing in reading and mathematics, along with tests in science and social studies, correlated with greater alignment of all curricula to state standards and assessments. Teachers also reported an increased focus on student achievement and increased rigor of curriculum overall. In another study, Vogler (2002), who surveyed a stratified, random sample of mathematics, science, and English teachers, found that teachers were aligning practice with standards and assessments to ensure that students scored well. Specifically, science teachers reported increasing their use of critical thinking questions,

rubrics and scoring guides, and inquiry-based investigations since test scores began being publicly released.

### Less Emphasis on Science

In addition to noted changes in curriculum and instruction, educators reported that accountability policies emphasizing mathematics and reading achievement result in less overall emphasis on science (Diamond & Spillane, 2004; Smith, 1991). Many studies directly connected less time given to science with accountability-related testing (Hamilton et al., 2007; Jones et al., 1999; Shepard & Dougherty, 1991; Smith, 1991). In a nationwide, representative survey of districts, researchers from the Center on Education Policy studied changes in the time spent on different subjects in elementary schools from 2001 to 2007 (McMurrer, 2008). In 28% of districts, instructional time in elementary school science decreased by an average of 75 minutes per week (or 45 hours per year). Teachers reported that fear of failure and sanctions dissuaded those who felt inclined to teach more science from doing so (Pringle & Martin, 2005). Diamond and Spillane (2001) found that some elementary schools have severely curtailed or completely eliminated science instruction, with teachers citing testing as the reason. As one teacher stated, "I just can't fit it in. [There is] so much math and so much reading that it's hard to fit the science and social studies [in]. So most of the time . . . I begin teaching science and social studies after the test" (as quoted in Diamond & Spillane, 2001, p. 17). At the high school level, Saka (2007) found that some science teachers were asked to drill mathematics and reading skills during their science lessons to prepare for state tests.

With science testing now included as a part of NCLB, it is possible that the emphasis on science instruction will increase. The effect may be small, however, as federal policy does not require science scores to be factored into states' AYP accounting (Paige, 2002). At least one study indicated that accountability measures do appear to increase the amount of science teaching in grade levels where science is tested (Rodgers, 2006).

In an ethnography of one elementary school, Perlstein (2007) described changes in science teaching as the school responded to new accountability policies. First, time spent on science decreased. Students, who were scheduled for 45 minutes of science a day, typically received 20 or fewer. Efforts to teach more science by reading in the content area resulted in the reading of random, disconnected science-related articles. In the third grade, the teacher opened the "plastic kits" in the back of the classroom only three or four times during the year, and even then the projects were "severely abridged—no hypotheses, no data . . . mostly students read from the textbook and did worksheets" (Perlstein, 2007, p. 120). With required state testing of science beginning, the principal expected the teaching of social studies to be "obliterated" (Perlstein, 2007, p. 126). Despite limited science teaching, district officials made an example of this school as reform done right.

With less instructional time for science and increased emphasis on mathematics and reading, educators stated that professional development in science content or pedagogy also sometimes decreased. Stecher and Barron (1999) found that Kentucky teachers in grades where science was tested received significantly more science professional development than those teaching untested grades. In another study, principals claimed that an NSF-sponsored professional development series took science teachers away from their classrooms too often; they worried that test scores might suffer (Kersaint et al., 2001). Moreover, government funding for professional development in science has been altered under NCLB (Peterson, 2002). Specifically, Eisenhower funds, part of NCLB Title II grants, which had been a major funding stream for professional development in science instruction, can now be used

for recruitment, hiring, or professional development that relates to increasing the number of "highly qualified" teachers of *any* discipline.

## Accountability and Meeting Students' Needs

While not typically addressed within literature specifically focusing on science education, test-based accountability appears to have increased attention to achievement gaps, resulting in more focus on the needs of English language learners (ELLs), low socioeconomic status students, minorities, and students labeled with disabilities (Jennings & Rentner, 2006). Schools increasingly use assessment tools to gauge how well all students actually learn the material being taught, and they are held accountable for the test scores of all sizeable subgroups, instead of just showing that on average students do well. Consequently, Linn (2003) and Batt, Kim, and Sunderman (2005) noted that test-based accountability with disaggregated scores has increased expectations for all students, particularly low income and minority students.

On the other hand, some studies suggested that some testing programs may adversely impact particular groups of students more than others. First, Kahle (2004) noted that while girls tend to score lower than boys on NAEP science, the questions tended to favor curricular areas historically preferred by boys (such as chemistry and physics). When tests instead emphasized reading skills and real-world connections (such as new Performance Indicators of Student Achievement [PISA] testing), scores did not differ significantly by gender (Kahle, 2004). While NAEP and PISA are not high stakes at the school or district level, this finding suggests some of the standardized tests given for NCLB requirements should be reviewed for gender bias. Teachers also stated that testing negatively impacts the teaching of ELL students (Shaver, Cuevas, Lee, & Avalos, 2007). However, these authors noted that few studies have analyzed *how* accountability policies impact the learning of ELL students, only suggesting that they *did*. Finally, research suggested that the structure of most accountability test systems led teachers to focus more on students near the scoring cutoff point of meeting standard, who are more likely to be able to move from below to above the critical standard (Hamilton et al., 2007). Researchers also noted that this focus may reduce attention and opportunities for the highest performing students, who will not be able to move up in performance categories.

Other studies conveyed teachers' concerns about the effects of tests on all students. Elementary and middle school teachers reported their belief that tests damage students emotionally and reduce their love of learning science (Rodgers, 2006). Teachers also said that, since the introduction of testing, students enjoy science less (Jenkins, 2000). Taylor et al. (2008) shared teacher views that content-focused tests limit student opportunities for creative expression. However, none of these studies collected data directly from students. Rodgers (2006) conducted classroom observations, but did not indicate directly witnessing these adverse effects on students. One study (Galton, 2002) surveyed student attitudes in relation to testing, but questions were not science specific.

## Accountability and Teachers

Teachers also reported a personal impact of test-based accountability. Primarily, many teachers saw these measures as belittling and constraining their professional judgment (Jenkins, 2000; Mintrop, 2004; Settlage & Meadows, 2002). They felt they had to standardize their behavior and saw fewer avenues for creativity (Stuart Hammer, 2004; Taylor et al., 2008; Wood, 1988). While their understanding of education accountability systems may be limited, a group of 37 scientists interviewed by Taylor et al. (2008) also expressed concerns

that accountability limits science education reform and constrains teachers' ability to make science engaging for their students.

Owing to the nature of accountability tests, some teachers felt increased pressure and stress. In one study, 79% of a random sample of elementary teachers felt "great" or "substantial" pressure from the district specifically to raise test scores (Shepard & Dougherty, 1991). Because these teachers taught all subjects (not only science), this pressure related primarily to the areas of mathematics and reading scores, but the study emphasizes a generalized increase in teacher stress. These findings were substantiated by Smith (1991). In another study, 92% of 265 randomly selected middle school science teachers in Florida attributed increased stress to the new state science assessment (Stuart Hammer, 2004). Yet, only 19% of these teachers felt that the test had improved instruction or curriculum.

In some schools put on probation under high-stakes accountability, teacher turnover increased (Mintrop, 2004). In a general study of teacher turnover (including but not limited to science teachers), Tye and O'Brien (2002) found that teachers cited accountability policies as the number one reason they stopped teaching. While there are potential biases in this self-reported data, these findings lend further support to the claim that accountability testing decreased teachers' job satisfaction.

## CONCLUSIONS, LIMITATIONS, AND FUTURE DIRECTIONS

In conclusion, test-based accountability policies frequently correlated with changes in instructional practice, the amount of science taught, and teacher satisfaction. Importantly, current test reporting requirements reveal gaps in achievement between subgroups of students and may result in the mobilization of greater resources, i.e., time, effort, and funding, to assist students not meeting performance standards. In the studies reviewed here, some teachers and administrators also positively discussed the alignment of curriculum and instruction with standards and assessments within their classes and throughout school systems. Overall, out of 35 studies reviewed, nine (26%) portrayed such positive perspectives on accountability. By contrast, 34 (97%) included discussion of negative impacts of test-based accountability on science education. Teachers and administrators repeatedly expressed the feeling that accountability-based reform disrupts research-based reform efforts in science. They asserted accountability limits time and effort spent on science, drives the remaining science instruction toward memorization of facts, and constrains student learning. Overall, these studies suggested that unlike many policies, many teachers perceived test-based accountability policy as changing their classroom practice, signifying the importance of understanding how educators make sense of this type of policy.

These findings generally remain consistent across contexts and across time. To be sure, these general findings should not be conceived as indicating that changes are clearly happening due to accountability policies, but that educators *perceive* accountability policies causing these changes. In addition, because these studies occurred throughout the development of test-based accountability policies, they reveal that educators' perceptions of the effects of accountability on science education have not changed significantly under NCLB. Science teachers have felt pressure from accountability and have reported making subsequent changes in their curriculum and instruction for over 20 years. The consistency in findings across contexts also reveals the general perception among teachers of the limitations of a wide range of test-based accountability policies. While these studies do not generalize to all contexts, arguably they do to most. Studies produced by the Center on Education Policy (Jennings & Rentner, 2006; McMurrer, 2008; Rentner et al., 2006) provided data representative of the entire United States and schools from K to 12. Five studies offered comparisons across multiple states (see Table 1). Four considered entire K-12 school

systems. Ten studies focused exclusively on elementary schools, three on middle schools, and five on high schools. Five included elementary and middle schools, and seven middle and high schools.

While the cited studies cover a broad spectrum of states, districts and levels, and even international insight, further research in other contexts might be fruitful, particularly in states or districts where accountability policies or testing practices significantly differ from the norm. For example, research could examine changes in science education in states such as Illinois and Colorado since the inclusion of the ACT in the statewide assessment program.

In addition, although these studies, taken together, provide a window on the history of high-stakes testing and science education, other methodologies could provide further insights. For example, no study reviewed employed a longitudinal analysis of accountability policy. Observations of classrooms before and after the implementation of new testing policies would have provided a more objective view of their impact, but such a study would be difficult now. Nevertheless, local databases or resources such as the Survey of Enacted Curriculum (http://seconline.wceruw.org/secWebHome.htm) could make such as study possible. No studies reviewed here mined National Educational Longitudinal Study (NELS) or other longitudinal data sets for changes in variables such as science pedagogies used, student course-taking patterns, courses offered, science achievement (grade point average [GPA] and standardized tests), career pursuit, or attitudes toward science *before* and *after* implementation of accountability policies—a potentially rich vein for further investigation. Studies of students' patterns of course taking and career objectives do exist, but they do not link changes to accountability policies (e.g., Ingels & Dalton, 2008). One could imagine a study comparing trends in these variables for the 10 years before implementation of high-stakes testing and 10 years after implementation at the district, state, or national level (while controlling for other variables). Using methods such as regression discontinuity models could successfully show such changes over time. As one example of longitudinal analysis, Lee (2006) shows that improvements in state-level mathematics and reading tests do not validly predict corresponding improvements in NAEP; consequently, she debates the effectiveness of NCLB at improving test scores. A similar study could explore NAEP science scores.

Considering NAEP scores brings up another important limitation of these studies: None of them analyze the correlation of student achievement in science with the implementation of accountability policies. Most study changes in curriculum and instruction, but without a direct (or even indirect) connection to any measure of student learning of concepts or skills. Some states and districts had their own science assessments before testing became high stakes. Trends on such tests could reveal whether and how student achievement changed in response to policy changes, such as those required by NCLB. For example, in Maryland and Kentucky, where core content knowledge examinations replaced performance-based assessments in 2002, a longitudinal study of high school graduates' success in college, particularly in STEM fields, could offer insights into the impact of these two testing modalities.

Beyond student achievement data, this body of research is limited by a lack of data collected directly from students. Galton (2002) describes student attitudes about their education, a little of which is science specific, but does not meaningfully connect these data to assessment or accountability policy. Kahle (2004) reviews studies of student attitudes toward science and science assessment but also does not connect her work with accountability policy of any type. Future research could more carefully examine how (or whether) students perceive any impact from test-based accountability in science. Such a study could be done with large-scale longitudinal data, or even by mining attitudinal surveys currently

given each year in many districts, and noting changes before and after implementation of test-based accountability.

Many of the studies reviewed here also rely largely on self-reports of teachers and administrators. Findings are based on their perceptions of practice. In interviews and on surveys, respondents may give a biased description of themselves and their practice to appear competent or correct (Brinthaupt & Erwin, 1992). For example, teachers may say that testing reduces opportunities to use inquiry-based activities even though they rarely, if ever, used them previously. Testing or standards may be used as excuses by teachers to not use inquiry or other approaches, when the actual reasons are different. Indeed, individuals are often unaware of the reasons for the decisions they make, which poses problems for studies attempting to link a particular cause to changes in curriculum and instruction. Some of the studies included observations of teachers and leaders within schools, which provides more reliable evidence (Brinthaupt & Erwin, 1992). However, even in these studies, the authors do not always clearly delineate which findings were unique to or specifically verified by observation. Future studies including rich descriptions of behavior and activities over time in various classrooms, at meetings, and in other staff interactions could offer independent evidence supporting or refuting the testing constraints claimed by teachers.

Finally, both teachers and researchers at times conflate standards-based reform with test-based accountability. These policies have often been enacted together, as tests ideally measure the learning represented in the standards. Thus, teachers may consider standards and testing collectively. This connection is problematic as teachers' views of standards-based reform tend to be more favorable than their views of test-based accountability (Donnelly & Sadler, 2009; Mintrop & Sunderman, 2009). That researchers also conflate the two can be seen in Louis, Febey, and Schroeder (2005), which purports to review accountability, but reports on a great deal of data around standards. Further research must clearly maintain this distinction.

## RECOMMENDATIONS FOR IMPROVING ACCOUNTABILITY POLICY IN SCIENCE

Much has been written on improving test-based accountability generally (e.g., Baker, Linn, Herman, & Koretz, 2002; Pellegrino et al., 2001). Utilizing that research background and the five categories of findings in this paper, the following section discusses several means to improve *science* accountability policies.

First, accountability policies must be compatible with and encourage research-based efforts to improve science instruction and curriculum. Currently, educators perceive that efforts from groups such as NSF to improve science learning often compete with testing requirements. Principals and teachers need help in understanding the benefits of professional development that encourages teaching methods focused on critical thinking skills in addition to understanding standards-based content—that these efforts are not incompatible. In fact, current research shows that effectively integrated, inquiry-based activities can actually improve student scores on standardized accountability tests (Blanchard et al., 2010). Changes to current testing and standards could also help in this endeavor. If tests and standards focused more on scientific reasoning and process skills, educators would see fact-based teaching as less effective. Some current state assessments in science reflect a shortcoming of content standards, emphasizing too broad of an array of content knowledge. Staff at some NSF centers have found success in improving science teaching after influencing state policy makers to make testing and standards better reflect deeper-level thinking (Goetz Shuler et al., 2009). Arguably, some teachers use testing as an excuse not to adopt more rigorous pedagogical practices. To move past excuses and encourage engagement in

research-based professional development, future standards and large-scale tests must be more thoughtfully focused on key skills and ways of knowing instead of broad factual understanding.

Second, because teachers perceive current accountability policy as discouraging their current active-learning and inquiry-based instruction, states need assessments and assessment models that encourage such practice. Most state tests meeting NCLB requirements "do a reasonable job with certain functions of testing, such as measuring knowledge of basic facts" (Pellegrino et al., 2001, p. 26). But these assessments generally do not assess the sorts of complex thinking that shows students' deeper understanding and application of science and science-related cognitive skills. Many assessment models derive from the idea that learning is a "step-by-step accumulation of facts" (Pellegrino et al., 2001, p. 26). Open-ended and performance-based questioning, such as that found in the Massachusetts Comprehensive Assessment System, more effectively measures students' understanding and may even sway teachers to keep or adopt research-based pedagogical approaches to science (Vogler, 2002). The 2009 version of NAEP also provides an example of research-based, large-scale science assessment. It uses concept mapping, clusters of questions, predict—observe–explain items, hands-on performance tasks, and interactive computer tasks to gain a more accurate picture of student knowledge and ability (Fu, Raizen, & Shavelson, 2009). While such assessments cost more in terms of time and money to develop, administer, and score, they provide more accurate pictures of student learning. And, as warned from the beginning of the accountability era, we cannot "let ourselves become preoccupied and distracted by the easily measurable" (Merrill, 1972, p. 23).

Third, science education needs to be made more of a priority. Particularly at the elementary level where time devoted to each subject is more flexible, current accountability policies appear to lead to less time devoted to science learning. While elementary teachers report putting more emphasis on the *high-stakes* subjects of mathematics and reading, secondary teachers also see testing pressures pushing them toward more fact-based content. Given improved tests, policy makers could encourage more emphasis on science through making its testing as high stakes as reading and mathematics. Limited evidence suggests that making science more high stakes improves achievement on standardized tests. Specifically, Judson (2010) found that states incorporating science scores into their Annual Yearly Progress calculations showed higher NAEP science scores among fourth-grade students, but no NAEP score correlation among eighth-grade students. Having science tested along with mathematics and reading at all tested grade levels may also result in more time and professional development devoted to science. On the other hand, policy makers could put aside the high-stakes aspect of accountability testing entirely. Mintrop and Sunderman (2009) make a case against the effectiveness of the sanction-driven accountability model. They argue that psychologically, inspiration, not punishment, is the most effective means for changing behavior long term. As seen in the reviewed literature, fear and threats more often increase stress and lead to lower common denominator solutions. There is no clear evidence that sanction systems work to improve student learning (Mintrop & Sunderman, 2009). While poor science performance rarely results in sanctions, current accountability policies likely discourage states from using rigorous assessments, as states do not want to have the majority of their students not scoring well enough to meet standard. Therefore, assessment systems should also be changed to emphasize growth in student learning rather than attaining arbitrary test score cutoff points. Current proficiency steps in science test scoring generally lack any connection to what students actually know and are able to do. As is beginning to happen in the growth models used by some states, assessment models should clearly align goals with growth in the knowledge and abilities desired in students.

Fourth, assessment systems should support efforts to meet the needs of all students. Foremost, they should continue to require closing achievement gaps among all subgroups. While current accountability policies help make teachers aware of gaps in student success, teachers note that these policies make it difficult for them to take the time to meet the needs of many of their nontraditional students, particularly ELLs (Lynch, 2000; Penfield & Lee, 2010). This finding further emphasizes the need to focus on a narrowed set of learning outcomes. Use of the growth model discussed previously could also help with this challenge, as it emphasizes progress of all students, from students with language barriers to students labeled as gifted. Test designers should also consider issues with gender bias; two means of improvement mentioned in the literature being emphasizing real-world connections and situated reading skills. Finally, as some teachers perceive testing as diminishing students' enjoyment of science, tests should connect more to what engages students in science. Assessments might require creativity and ask about science enjoyment. If assessments measure what teachers see as most important, teachers and students could be more successfully coached to see tests as useful tools.

Fifth, it follows that teachers and administrators need tools and assistance to help them view assessments as a legitimate resource to improve practice, not as another stressor (assuming effective assessments). While science assessments typically do not dictate sanctions, teachers appear to associate the science tests with increased achievement pressure under NCLB. Therefore, teaching narrows to these types of tests and their underlying theory of cognition, not necessarily to the actual, varied learning needs of students (Pellegrino et al., 2001). One means to accomplish a reduction in test stress would be to have a true multiple measure accountability system, as opposed to primarily relying on one examination that is perceived as high stakes. Currently, the vast majority of testing systems cannot identify causal reasons for student test scores; tracking organization practices and monitoring the growth of individual students could begin to untangle the multiple factors affecting student test performance (Linn, 2006). Alternative assessments, such as performance or contextual assessment, have the potential to provide additional insights into student learning and ability while also giving a broader picture by which to evaluate school performance; however, more work needs to be done to establish the validity and reliability of such methods (Ellis, Jablonski, Levy, & Mansfield, 2009; Fu et al., 2009; Klassen, 2006). As another assessment strategy, students' attitudes and perceived opportunities may also demonstrate the quality of science teaching. Science teaching could be judged not only on student learning but also on whether attitudes about science and desires to continue to learn about it have improved through a school year. The goal of schooling, of course, goes far beyond instilling content knowledge. Assessment practice should more concretely adhere to Sec. 1111(3)(C)(vi) of NCLB, which requires states' assessments to "involve multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding."

Moving forward, education leaders and policy makers need to continually evaluate whether or not accountability policies inspire teachers and students in science, foster innovation, and increase teachers' ability to use research-based practice. The research strongly suggests that, in general, current accountability policy does not meet these aims and thus needs reform. To ensure that new test-based accountability policy truly supports science teaching and learning for the 21st century, the following key recommendations should be considered:

1. Accountability testing in science should place more emphasis on skills and scientific reasoning found in instructional methods such as inquiry and active learning.
2. Accountability systems should use multiple measures of students' ability, connecting to creativity, and student's enjoyment of learning.

3. Policy makers should eliminate the high-stakes nature of mathematics and reading tests, and science where applicable, to enable more balanced curricular emphases.
4. Federal policy should encourage growth models tied to specific learning benchmarks, instead of arbitrary goals and scoring cut points.
5. Educators need assistance moving beyond seeing test-based accountability as a stressor and instead perceiving and using effective assessments as tools to improve practice and meet the needs of all students; an understanding that inquiry-based teaching can work in a high-stakes testing environment would also be helpful.

I would like to thank L. Allen Phelps, Jim Stewart, John Rudolph, Peter Hewson, UW–Madison science seminar colleagues, Jesse Boyett Anderson, and anonymous reviewers for their constructive feedback on this paper.

## REFERENCES

American Association for the Advancement of Science. (1989). Science for all Americans. New York: Oxford University Press.

American Educational Research Association. (2008). Alternate definition of scientifically based research. Retrieved August 8, 2010, from http://www.aera.net/Default.aspx?id = 6790.

Aronson, I. (2007). Negotiating the terrain of high-stakes accountability in science teaching (Doctoral dissertation). Available from ProQuest Dissertations & Theses Database. (UMI No. 3297419)

Aronson, I., & Miller, J. (2007). Competing horizons. The Science Teacher, 74(7), 64 – 67.

Baker, E., Linn, R., Herman, J., & Koretz, D. (2002). Standards for educational accountability systems (Policy Brief 5). National Center for Research on Evaluation, Standards, and Student Testing. Retrieved January 20, 2010, from http://www.cse.ucla.edu/products/policy/ cresst_policy5.pdf.

Batt, L., Kim, J., & Sunderman, G. (2005). Limited English proficient students: Increased accountability under NCLB. Cambridge, MA: The Civil Rights Project at Harvard University.

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Leonard, L. A., & Granger, E. M. (2010). Is inquiry possible in light of accountability?: A quantitative comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. Science Education, 94(4), 577 – 616.

Brinthaupt, T. M., & Erwin, L. J. (1992). Reporting about the self: Issues and implications. In T. M. Brinthaupt & R. P. Lipka (Eds.), The self: definitional and methodological issues (pp. 137 – 171). Albany: State University of New York Press.

Cavanagh, S. (2009). Science panel seeks ways to fan student innovation. Education Week. Retrieved November 30, 2010, from www.edweek.org/ew/articles/2009/09/02/02stem-2.h29.html/&destination= or http://www.edweek.org/ew/articles/2009/09/02/02stem-2.h29.html/&levelId=2100

Coble, J. (2006). Curricular constraints, high-stakes testing and the reality of reform in high school science classrooms (Doctoral dissertation). Available from ProQuest Dissertations & Theses Database. (UMI No. 3207430)

Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. Educational Evaluation and Policy Analysis, 23(2), 145 – 170.

Cohen, D. K., & Ball, D. L. (1990). Policy and practice: An overview. Educational Evaluation and Policy Analysis, 12(3), 233 – 239.

Committee on Conceptual Framework for New Science Education Standards. (2010). A framework for science education: Preliminary public draft. Board of Science Education, Division of Behavioral and Social Sciences and Education of the National Research Council. Retrieved January 20, 2011, from http://www.aapt.org/Resources/upload/Draft-Framework-Science-Education.pdf.

Committee on Science, Engineering, and Public Policy. (2007). Rising above the gathering storm: Energizing and employing America for a brighter economic future. National Academies. Retrieved August 15, 2010, from http://www.nap.edu/catalog.php?record_ id=11463.

Cooper, H. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52(2), 291 – 302.

Cooper, H. M., & Hedges, L. V. (2009). Research synthesis as scientific process. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 3 – 16). New York: Russell Sage.

Costigan, A. T., & Crocco, M. S. (2004). Learning to teach in an age of accountability. Mahwah, NJ: Erlbaum.

DeNovellis, R. L., & Lewis, A. J. (1974). Schools become accountable: A PACT approach. Washington, DC: Association for Supervision and Curriculum Development.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging of reproducing inequality? Teachers College Record, 106(6), 1145–1176.

Donnelly, L. A., & Sadler, T. D. (2009). High school science teachers' views of standards and accountability. Science Education, 93(6), 1050–1075.

Ellis, P., Jablonski, E., Levy, A., & Mansfield, A. (2009). High school science performance assessments: an examination of instruments for Massachusetts. Newton, MA: Education Development Center. Retrieved April 30, 2011, from http://archives.lib.state.ma.us/handle/2452/58025.

Elmore, R. F., & McLaughlin, M. W. (1988). Steady work. Policy, practice, and the reform of American education. Santa Monica, CA: Rand Corporation. Retrieved August 15, 2010, from http://www.rand.org/pubs/reports/2007/R3574.pdf.

Font-Rivera, M. J. (2003). A descriptive study of the reported effects of state-mandated testing on the instructional practices and beliefs of middle school science teachers (Doctoral dissertation). Available from ProQuest Dissertations & Theses Database. (UMI No. 3091147).

Fu, A. C., Raizen, S. A., & Shavelson, R. J. (2009). The nation's report card: A vision of large-scale science assessment. Science, 326(5960), 1637–1638.

Galton, M. (2002). Continuity and progression in science teaching at key stages 2 and 3. Cambridge Journal of Education, 32(2), 249–265.

Gamoran, A. (2007). Introduction: Can standards-based reform reduce the poverty gap in education? In A. Gamoran (Ed.), Standards-based reform and the poverty gap: Lessons for No Child Left Behind (pp. 3–16). Washington, DC: Brookings Institution.

Goetz Shuler, S., Backman, J., & Olson, S. (2009). The role of assessments and accountability. In B. B. Berns & J. O. Sandler (Eds.), Making science curriculum matter: Wisdom for the reform road ahead (pp. 49–59). Thousand Oaks, CA: Corwin Press.

Goldstein, L. S. (2008). Kindergarten teachers making "street-level" educational policy in the wake of No Child Left Behind. Early Education and Development, 19(3), 448–478.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., et al. (2007). Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states. Santa Monica, CA: Rand Corporation.

Ingels, S. J., & Dalton, B. W. (2008). Trends among1p1(J.)-1798m&lsTc[(Fu,5senistrngels,)-419729(448)-125.5(–)-1.2(4ge

Linn, R. L. (2003). Accountability, responsibility, and reasonable expectations. Educational Researcher, 32(7), 3–13.

Loveless, T. (2007). The peculiar politics of No Child Left Behind. In A. Gamoran (Ed.), Standards-based reform and the poverty gap: Lessons for No Child Left Behind (pp. 253–285). Washington, DC: Brookings Institution.

Louis, K. S., Febey, K., & Schroeder, R. (2005). State-mandated accountability in high schools: Teachers' interpretations of a new era. Educational Evaluation & Policy Analysis, 27(2), 177–204.

Lynch, S. J. (2000). Equity and science education reform. Mahway, NJ: Erlbaum.

McMurrer, J. (2008). Instructional time in elementary schools: A closer look at changes for specific subjects. Washington, DC: Center on Education Policy (CEP).

Merrill, R. J. (1972). Accountability and the science teacher. The Science Teacher, 39(8), 23.

Mintrop, H. (2004). Schools on probation: How accountability works (and doesn't work). New York: Teachers College Press.

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement—and why we may retain it anyway. Educational Researcher, 38(5), 353–364.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for education reform. Washington, DC: Government Printing Office.

National Research Council. (1996). National Science Education Standards. Washington, DC: National Academy Press.

Paige, R. (2002). Key policy letters signed by the education secretary or deputy secretary. U.S. Department of Education. Retrieved October 20, 2010, from http://www2.ed.gov/policy/elsec/guid/secletter/020724.html.

Pellegrino, J. W., Chudowsky, N., & Glaer, R. (Eds.). (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Penfield, R. D., & Lee, O. (2010). Test-based accountability: Potential benefits and pitfalls of science assessment with student diversity. Journal of Research in Science Teaching, 47(1), 6–24.

Perlstein, L. (2007). Tested: One American school struggles to make the grade. New York: Henry Holt.

Peter, Z. (2008). Science fairs getting left behind. Albuquerque Journal. Retrieved January 18, 2009, from http://www.abqjournal.com/news/metro/303282metro04-28-08.htm.

Peterson, J. (2002). No Child Left Behind—How will it affect science educators? Journal of College Science Teaching, 32(2), 93.

Phillips, R., Freeman, R. E., & Wicks, A. C. (2003). What stakeholder theory is not. Business Ethics Quarterly, 13(4), 479–502.

Pinder, P. J. (2008). A critique analysis of NCLB, increased testing, and past Maryland mathematics and science HAS exams: What are Maryland's practitioner's perspectives? Paper presented at the 16th annual Association for Science Teacher Education Conference, St. Louis, MO.

Pringle, R. M., & Martin, S. C. (2005). The potential impacts of upcoming high-stakes testing on the teaching of science in elementary classrooms. Research in Science Education, 35(3), 347–361.

Rentner, D. S., Scott, C., Kobler, N., Chudowsky, N., Chudowsky, V., Joftus, S., et al. (2006). From the capital to the classroom: Year 4 of the No Child Left Behind Act. Center on Education Policy. Retrieved January 5, 2010, from http://www.cep-dc.org/index.cfm?fuseaction = page.viewPage&pageID=540&nodeID=1.

Ripley, A. (2004). Beating the bubble test. Time, 163 (March 1), 52–53.

Rodgers, P. E. (2006). What goal is of most worth? The effects of the implementation of the Texas Assessment of Knowledge and Skills on elementary science teaching (Doctoral dissertation). Available from ProQuest Dissertations & Theses Database. (UMI No. 3219180)

Saka, Y. (2007). Exploring the interaction of personal and contextual factors during the induction period of science teachers and how this interaction shapes their enactment of science reform (Doctoral dissertation). Available from ProQuest Dissertations & Theses Database. (UMI No. 3312780)

Settlage, J., & Meadows, L. (2002). Standards-based reform and its unintended consequences: Implications for science education within America's urban schools. Journal of Research in Science Teaching, 39(2), 114–127.

Shaver, A., Cuevas, P., Lee, O., & Avalos, M. (2007). Teachers' perceptions of policy influences on science instruction with culturally and linguistically diverse elementary students. Journal of Research in Science Teaching, 44(5), 725–746.

Shepard, L., & Dougherty, K. C. (1991). Effects of high-stakes testing on instruction. Paper presented at the 58th annual meeting of the American Educational Research Association, Chicago, IL.

Sirotnik, K. A. (2004). Holding accountability accountable: What ought to matter in public education. New York: Teachers College Press.

Smith, L. K., & Southerland, S. A. (2007). Reforming practice or modifying reforms? Elementary teachers' response to the tools of reform. Journal of Research in Science Teaching, 44(3), 396–423.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. Educational Researcher, 20(5), 8–11.

Southerland, S. A., Smith, L. K., Sowell, S. P., & Kittleson, J. M. (2007). Resisting unlearning—Understanding science education's response to the United States' national accountability movement. Review of Research in Education, 31(1), 45–77.

Stecher, B. M., & Barron, S. I. (1999). Quadrennial milepost: Accountability testing in Kentucky (Report No. 505). National Center for Research on Evaluation, Standards, and Student Testing. Retrieved January 25, 2010, from http://www.cse.ucla.edu/products/Reports/TECH505.pdf.

Stuart Hammer, K. E. (2004). An assessment of standards-based reform in Florida's middle school science programs (Doctoral dissertation). Available from ProQuest Dissertations & Theses Database. (UMI No. 3146258)

Sykes, G., O'Day, J., & Ford, T. G. (2009). The district role in instructional improvement. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), Handbook of educational policy research (pp. 767–784). New York: Routledge.

Taylor, A. R., Jones, M. G., Broadwell, B., & Oppewal, T. (2008). Creativity, inquiry or accountability? Scientists' and teachers' perceptions of science education. Science Education, 92(6), 1058–1075.

Tye, B. B., & O'Brien, L. (2002).Why are experienced teachers leaving the profession? Phi Delta Kappan, 84(1), 24–32.

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. Education, 123(1), 39–55.

Wideen, M. F., O'Shea, T., Pye, I., & Ivany, G. (1997). High-stakes testing and the teaching of science. Canadian Journal of Education, 22(4), 428–444.

Wood, T. (1988). State-mandated accountability as a constraint on teaching and learning science. Journal of Research in Science Teaching, 25(8), 631–641.