The Lower-Order Expectations of High-Stakes Tests: A Four-State Analysis of Social Studies Standards and Test Alignment

Scott W. DeWitt, Nancy Patterson, Whitney Blankenship, Brooke Blevins, Lorrei DiCamillo, David Gerwin, Jill M. Gradwell, John Gunn, Lamont Maddox, Cinthia Salinas, John Saye, Jeremy Stoddard & Caroline C. Sullivan

a University of Cincinnati–Blue Ash College
b Bowling Green State University
c Rhode Island College
d Baylor University
e Canisius College
f Queens College, City University of New York
g Buffalo State College–SUNY
h University of North Alabama
i University of Texas at Austin
j Auburn University
k The College of William and Mary
l Georgia State University

Published online: 04 Sep 2013.

The Lower-Order Expectations of High-Stakes Tests: A Four-State Analysis of Social Studies Standards and Test Alignment

Scott W. DeWitt
University of Cincinnati–Blue Ash College

Nancy Patterson
Bowling Green State University

Whitney Blankenship
Rhode Island College

Brooke Blevins
Baylor University

Lorrei DiCamillo
Canisius College

David Gerwin
Queens College, City University of New York

Jill M. Gradwell
Buffalo State College–SUNY

John Gunn
Queens College, City University of New York

Lamont Maddox
University of North Alabama

Cinthia Salinas
University of Texas at Austin

Correspondence should be sent to Scott W. DeWitt, Department of Behavior Science, University of Cincinnati–Blue Ash College, 1275 Forest Ct., Cincinnati, OH 45215. Email: scottdewitt@cinci.rr.com
Abstract: This study indicates that the state-mandated high-stakes social studies assessments in four states do not require students to demonstrate that they have met the cognitive demands articulated in the state-mandated learning standards. Further, the assessments do not allow students to demonstrate the critical thinking skills required by the standards. In this study, researchers from four states with high-stakes social studies tests questioned how well their states’ tests measured and aligned with expressed expectations for students related to this objective. They analyzed the nature of four states’ high school social studies standards and high-stakes tests for higher or lower cognitive demands and the alignment between them. The analysis showed that the primary expectation for student learning in the standards is for higher cognitive activity than that represented in the tests, which are dominated by low cognitive level items.

Keywords: critical thinking, social studies standards, standardized testing, high-stakes tests, secondary social studies

Critical thinking is a vital 21st century essential skill (Partnership for 21st Century Skills, 2004, 2008) for the education of an engaged, active citizenry that can make informed decisions in an increasingly interconnected society (National Council for the Social Studies [NCSS], 2010). As such, it would seem that the federal educational accountability system that proposes to promote and measure standards-based student learning should encourage states to teach and assess students’ critical thinking skill levels as they relate to social studies content.

However, research has shown that many of the assessments produced as part of the No Child Left Behind (NCLB) Act lack critical thinking and problem-solving skills and negatively impact student learning. In her report on the critical elements of education in high-achieving nations, Linda Darling-Hammond (2007) found that high-achieving countries organized courses and instruction around a curriculum focused on deep understanding and problem solving. Such nations use examinations that require students to conduct research and scientific investigations that require higher-order cognitive skills. The United States does not utilize this kind of higher-order assessment. The result of this lower-order teaching and assessment and the failure rates...
associated with high-stakes, multiple-choice (MC) tests are having serious negative effects on student learning:

It is hard to imagine any more punishing set of consequences for high school failure than those that exist today... The history of student accountability suggests that if our goal is student improvement, we should be looking for ways to restore a culture of aspiration by decreasing, rather than increasing, the threat of punishment [e.g., not graduating from high school if a test is failed]. (Beadie, 2004, p. 48)

While high-stakes accountability is premised on the fact that these tests measure the extent to which students have met objectives set by state content standards, there have been “surprisingly few empirical studies that explore the correspondence between test scores and the competencies laid out in state standards” (Reich, 2009, p. 326). In response, we analyzed the nature of four states’ high school social studies standards and their corresponding high-stakes social studies tests for higher or lower cognitive demands and then examined the alignment between the standards and the tests. The results of this analysis indicate that although state standards include significant numbers of higher-order cognitive activities, lower-order cognitive level items dominate the high-stakes tests in these states.

REVIEW OF LITERATURE

The U.S. Department of Education website describes the NCLB Act accountability measures in the business language of quality management, stating their central purposes are to both promote and measure student learning:

A key principle of quality management is the importance of measuring what is valued (e.g., production rates, costs of materials, etc.). Such measures enable an organization to identify where and how to improve operations. In the same manner, if schools and school systems are to continuously improve, they must measure growth in student achievement. After all, the core of all activity in schools and school systems is teaching and learning, and the key question is: Are the students learning? (Ed.gov, 2004, Question 7)

Given that “a key principle of quality management is the importance of measuring what is valued” (Ed.gov, 2004, Question 7), we need to consider here both what is valued in social studies as well as how to measure effectively that which is valued. Included in this review of the literature are arguments for the centrality of critical thinking in social studies education, the importance
of assessing student learning related to critical thinking, and the necessity of aligning standards and assessments. A technical argument examines the limitations of current assessment tools for providing information on students’ critical thinking skills.

**Critical Thinking as a Central Value in Social Studies**

There are multiple sets of standards that are available for states to use to develop their own assessment systems and, within them, numerous references to critical thinking skills. Processes for developing and adopting standards, however, vary by state. This section of the literature review is therefore not intended to equate any national guidelines with specific state standards. Rather, it is offered as evidence of the high value social studies educators place on teaching and learning critical thinking skills. The emphasis individual states place on these skills is addressed in the Findings section.

Among these standards are the NCSS standards (2010), the Partnership for 21st Century Skills standards (2008), and the various disciplinary standards (American Geographical Society of New York, Association of American Geographers, National Council for Geographic Education, & National Geographic Society, 1994; Center for Civic Education, 1994; Council for Economic Education, 2010; National Center for History in the Schools, 1994a, 1994b). All of these documents include some expectation of critical thinking and/or problem-solving skills. The National Curriculum Standards for Social Studies, revised in 2010 by the NCSS, are built upon a disciplinary foundation of social studies as inquiry, with the purpose of such studies to develop critically thinking citizens who can avail themselves of the rich content of the social sciences to make informed decisions for the public good. The standards state that the specific aim of social studies is to promote civic competence, which requires that “citizens have the ability to use their knowledge about their community, nation, and world; to apply inquiry processes; and to employ skills of data collection and analysis, collaboration, decision-making, and problem-solving” (NCSS, 2010, p. 9). The standards include 10 thematic strands (what learners need to know) as well as guidelines for processes (what learners will be capable of doing) and products (how learners demonstrate understanding).

Another rigorous set of standards for PK–12 schools that promotes critical thinking is the Authentic Instruction and Assessment standards (Newmann, King, & Carmichael, 2007), with authentic referring to that which is “significant and meaningful as opposed to trivial and useless” (Newmann & Wehlage, 1993, p. 8). While Newmann and colleagues’ (1993, 2007) work is not specific to social studies, much of their research was conducted in social studies classrooms. The authentic intellectual work (AIW) rubrics for classroom observation and task analysis include higher-order thinking skills and construction of knowledge that require students to interpret, analyze, synthesize, or evaluate
information rather than merely reproduce it. In a recent study utilizing the AIW framework, a national research team (Saye & the Social Studies Inquiry Research Collaborative [SSIRC], 2013) found that more challenging instruction in no way harms student performance on high-stakes tests and that authentic pedagogy, as advocated by Newmann and others (1993, 2007), is strongly associated with improved higher-order learning outcomes.

This discussion of standards is not intended to minimize the disagreements in the field and in the nation about the goals of social studies and of education. Addressing those disagreements, however, is outside the scope of this article and has been addressed thoroughly by others (e.g., see Cornbleth & Waugh, 1995; Evans, 2004). For purposes of this study, we accepted individual state standards documents as statements of the content, processes, and thinking that are expected of students while recognizing that those state standards were constructed within the larger national standards and assessment context. We thus turn to assessment in social studies and the alignment between assessment and standards.

Assessing what is valued. Assessment has always been an integral part of the social studies standards. The NCSS standards label all assessments as formative and useful in improving teaching and learning, such that “powerful social studies teaching and learning [should] include assessment components designed to inform instructional planning and thus produce continuing improvements through successive cycles” (1994, p. 171). The NCSS-published teacher professional development program based on the AIW framework, entitled Powerful and Authentic Social Studies (PASS), was intended to “build the capacity of teachers to design curriculum, conduct instruction, and assess student learning in accord with high standards recognized in the field of social studies education” (Harris & Yocum, 2000, p. 1).

Assessments acknowledging the importance of critical thinking in the social studies require that students demonstrate not only knowledge across the many social science disciplines but also the higher-order skills of application, analysis, evaluation, and synthesis (Anderson et al., 2001) so that they may show the ability to use their knowledge in service to community. As Secretary of Education Arne Duncan stated,

The challenge of how we assess student learning in social studies is critical because . . . goals for students are so much larger than any bubble test could measure. [Teachers] are creating contributing and responsible citizens. [Teachers] are unleashing initiative, creativity, and problem-solving. (2011, p. 125)

The intellectual challenge is for students of social studies to undertake rigorous, intellectual work that has value outside the classroom and then to be
assessed accordingly. There is considerable question as to whether the current high-stakes assessments meet this challenge:

In many states requiring exit exams, a growing body of evidence suggests that high school graduates (i.e., those who passed the exams) are unprepared for the academic challenges of college, despite the insistence of officials that, with the test, the diploma means real achievement. For example, since Texas instituted its graduation exam, it has seen a rise in the percentage of its high school graduates requiring remedial classes at public colleges or universities. (Foote, 2007, p. 360)

This concern over the impact of standards-based high-stakes assessment is particularly relevant as a new generation of teachers enters the profession. The schooling experience for this new generation was in a high-stakes testing context, and, as a result, there are many who uncritically accept this context as the way schooling should be (van Hover, Hicks, Stoddard, & Lisanti, 2010). This priority given to tests in curricular decision-making raises the issue of the alignment of these tests with state standards, which are the documents designed to form the basis of classroom instruction.

Aligning assessment with standards. The objectives in the NCLB Act indicate that assessments are intended to both promote and measure student achievement, the implication being that assessments are tied to students’ classroom experiences. Polikoff, Porter, and Smithson (2011) stated: “The primary rationale underlying standards-based reform, the most prominent K–12 education policy of the past 20 years, is coherence” (p. 966). Mainstream thought in educational assessment aligns with this notion, advocating that purposeful and appropriate assessment be developed prior to the instructional planning process. In this view, learning and assessment are aligned with state standards and that alignment promotes greater student achievement (Shepard, 2000).

Standards are most useful when the assessments address the extent to which students have met the performance requirements set by those standards. Closely aligned standards and assessments have been shown to improve student achievement significantly (Black & Wiliam, 1998). Black and Wiliam (1998) argued that one important difficulty with effective assessment is the issue of effective learning, positing that assessment does not accurately reflect achievement when the standards and the tests do not match, i.e., when standards advocate for developing understanding but tests measure rote and superficial learning. They suggest that teachers often appear unaware of the misalignment of standards-based instruction and assessment.

In an effort to provide a teacher-friendly approach to alignment, Wiggins and McTighe (2006) developed a research-based framework called Understanding by Design. The guiding purpose of this work is to deepen student understanding by helping teachers systematically align goals
(standards), assessments, and instruction. The authors argued that appropriate assessment is best realized when students are required to complete complex, authentic tasks that include the six facets of explaining, interpreting, applying, shifting perspective, empathizing, and self-assessing. Beginning by clarifying goals and designing assessments

Helps to avoid the ‘twin sins’ of typical instructional design in schools: activity-focused teaching and coverage-focused teaching. Neither case provides an adequate answer to the key questions at the heart of effective learning: What is important here? What is the point? (p. 3)

Wiggins and McTighe offered a continuum of assessment strategies from which to choose based on intended objectives, from lower-order traditional paper-and-pencil quizzes and tests with selected-/constructed-response items to higher-order performance tasks and projects that are more complex, open-ended, and authentic.

There are relatively few research studies that examine alignment between state standards and high-stakes tests, and most address content areas other than social studies. Generally, studies have addressed both content and level of intellectual challenge, or “cognitive demand level” (Polikoff et al., 2011, p. 973). These studies consistently show that the cognitive level of high-stakes tests is lower than the requirements in the corresponding state standards: “Memorization does generally show less emphasis in the standards and more on the tests, especially for ELAR [English Language Arts and Reading] and science” (Polikoff et al., 2011, p. 988). One possible explanation for this discrepancy is the prevalence of MC items on high-stakes tests and the difficulties in constructing MC questions that require students to demonstrate higher-order cognitive skills.

Technical Limitations of Conventional Assessment for Establishing Students’ Levels of Higher-Order Thinking

There is agreement that MC questions can range along a continuum of cognitive skills from lower to higher order, but there is also recognition that this is not an exact science. Chudowsky and Pellegrino (2003) affirmed the problematic nature of using standardized tests to measure higher-order objectives, asserting that “state tests cannot adequately assess all of the content standards, and thus tend to focus on those that are easiest to test” (p. 78). An analysis of numerous statistics textbook chapters related to the creation of test items found that while the science of MC item-writing is advancing, it is “still largely a creative act that we inexorably link to content standards and instruction” (Haladyna, Downing, & Rodriguez, 2002, p. 329), with no set rules for producing test items targeted to particular cognitive skill levels. In other words,
“cognitive targets are based on curricular expert judgments and are in fact hypotheses about what the items may assess” (Ercikan, Seixas, Lyons-Thomas, & Gibson, 2012, p. 3).

This macro-level view of the problematic nature of MC item-writing is reinforced by a recent study that examined alignment of high-stakes test questions and cognitive demand in testing history knowledge. Reich (2009) showed that “the assumptions made by adults about what items measure is only partially correct” (p. 325), and that while the state history standards were organized around higher-order historical concepts, “the items answered by the small sample of students . . . yielded performances that did not draw upon these concepts” (p. 348). The higher-order thinking ostensibly required by the standards was not readily employed by these students, indicating that elements of the standards and tests were misaligned.

The problems with standardized testing that relies largely on MC items are reinforced by concerns over the psychometric practices involved in developing those tests. The assessment industry’s reliance on practices of selecting test items based on linking individual items to total scores as a way of eliminating items that do not “discriminate” between high and low achievers and norming test scores guarantees that some students will not succeed whether they have met stated criteria or not (Wineburg, 2004). Additionally, recent research conducted by the Historical Thinking Project (Ercikan & Seixas, 2011; Ercikan et al., 2012; Seixas, Ercikan, Gibson, & Lyons-Thomas, 2012) investigated the complexities of assessing students’ declarative and procedural knowledge:

There was evidence of difference between the MC and CR [Constructed Response] tasks in terms of how they related to the overall latent trait of historical thinking, with CR tasks having a closer relationship with the target construct. (Ercikan et al., 2012, p. 24)

The Historical Thinking Project studies also demonstrated the necessity for ongoing, rather than a single-point-in-time, assessment. There were “many instances where the same student performed at different levels in respect to the same cognitive demand, even in response to the same document” (Seixas et al., 2012, p. 23). Taken together, the combination of the complexity of testing cognitive processes in history and psychometric principles involved in developing test items create an environment where scores that result from tests provide no useful information to teachers, parents, or students as to the strengths or weaknesses of students’ declarative or procedural knowledge in history.

Given this lack of information, it should not be a surprise that there may be a mismatch between teachers’ perceptions of the demands of high-stakes tests and the expectations of political leaders and national standards writers. This apparent lack of alignment between teachers’ perceptions of test requirements,
the test structures, and curricular expectations calls for an examination of
the relationship between curriculum guidelines as set out in state standards
documents and the cognitive demands of high-stakes tests.

**Social Studies and High-Stakes Testing**

Since 1998, social studies achievement tests have been present in about
half of the states, with various states adding, dropping, or changing their struc-
ture over time (Grant & Salinas, 2008). An added complexity is that among that
number, not all are linked with a specific consequence, such as a requirement
that the test be passed for a student to graduate. As of 2010, of the 28 states that
had exit exams, 11 of them mandated a social studies exam (Dietz, 2010). With
each state developing their standards and accompanying assessments individ-
ually, there is significant variation from state to state, and this variation has
made it difficult for social studies researchers to evaluate the impacts of testing
on social studies teaching and learning.

Darling-Hammond’s (2007) concern with current U.S. K–12 testing prac-
tices was that “as teaching looks more like testing, U.S. students are doing less
writing, less science, less history, reading fewer books, and even using com-
puters less in states that will not allow their use on standardized tests” (p. 71).
Wiggins and McTighe (2006) reinforced this concern with their observation
that conventional tests are low level and therefore not useful for assessing
higher-order understandings. Given the proposed critical thinking outcomes for
social studies education and the companion recommendations for assessment,
it would appear that conventional testing that includes predominantly MC and
short answer (SA) questions may be a problematic assessment format for social
studies (Grant, 2001b; Horn, 2006).

While there is evidence that high-stakes standardized tests narrow curricu-
lum and mitigate against teaching for higher-order thinking skills (see, e.g.,
Au, 2007; Fischer, Boi, & Pribesh, 2011), Grant (2007) believed that most
Americans, teachers among them, accept the validity of conventional tests as
a means of judging student performance, and think that “tests are useful and
that multiple-choice questions and essay prompts represent reasonable ways
of judging what students know and understand” (p. 251). However, as Ravitch
(2010) pointed out,

Testing experts frequently remind school officials that standardized test
scores should be used not in isolation to make consequential decisions
about students, but only in conjunction with other measures of stu-
dent performance, such as grades, class participation, homework, and
teachers’ recommendations. (p. 149)
This review has illustrated the prominent role of critical thinking in social studies education and the many problems with assessing student learning in the field. We also established the importance of aligning assessment with standards. A simple question is at the heart of the complex issues addressed thus far: Is the public confidence in test scores as a measure of standards-based student learning supported by the evidence available from the tests? This question leads to our study, which explores the extent that high-stakes tests in four states are aligned with the cognitive demands of their corresponding standards documents.

THE STUDY

In this study, we analyzed four sets of high-stakes state high school social studies standards and tests for higher or lower cognitive demands and examined the alignment between them. This project emerged from a larger study which examined the relationship between secondary social studies teaching and student performance (Saye & the SSIRC, 2013). To explore the alignment issue, researchers from four states analyzed the nature of their state social studies standards for degree of difficulty on a dichotomous rubric of higher- and lower-order cognitive skills. The results for the standards were then compared to the states’ high-stakes tests to determine how well the tests align with the standards from which they are purportedly drawn. We do not question whether or not the students are learning, but rather we examine what they are being asked to learn, given research evidence that the graduation tests largely reflect an emphasis on lower-order knowledge and are far from what state curriculum standards establish as the goals for student cognition. Thus, the research questions for this study are narrower than the above-mentioned studies that address the alignment of both content and cognitive demand. Our overall research question is: To what extent do high-stakes tests hold students accountable for demonstrating the cognitive skills required in the corresponding state academic standards in social studies? Underlying this question is a recognition that testing is intended to influence curriculum and instruction (Grant, 2001b) and that if the states’ high-stakes tests do not require students to demonstrate the higher-order thinking skills contained in the standards, then those skills are less likely to be the focus of classroom instruction.

To answer this question, we asked a number of supporting questions, including

1. What is the cognitive level of the standards in each state?
2. What is the cognitive level of the test items in each state?
3. Across the four states, how well are the tests aligned with the standards?
Participating States

The research team represents four states with high-stakes tests in high school social studies: Ohio, New York, Texas, and Virginia. These states represent over 20% of all high school students in the United States (see Table 1).

Data Collection and Analysis

Framework. We addressed the standards and testing contexts in our respective states and analyzed our states’ social studies standards and testing documents based on a framework adapted from Bloom’s Taxonomy (Anderson et al., 2001). We used Bloom’s Taxonomy as the basis for analysis for two reasons. First, the AIW criteria (Newmann et al., 2007) do not map well with assessment tasks that are primarily MC. For example, two of the four analytic categories in the standards require students to generate written communications. MC questions are, by definition, given the lowest rating on the four-point scale in these categories. Second, Bloom’s Taxonomy is widely used by standards writers and test developers to differentiate item expectations and intended cognitive difficulty. Note that our use of the Taxonomy in this analysis represents an acknowledgement of the realities of standards and assessment writing practices. It is not an endorsement of Bloom’s Taxonomy, nor is it reflective of a belief that the Taxonomy accurately describes the relative difficulty of individuals’ learning processes.

The action words in the standards (e.g., identify, analyze, or compare) and in each individual test item were compared with the descriptors in the cognitive process dimension of the revision of Bloom’s Taxonomy of Educational Objectives (hereafter the Taxonomy or Bloom’s Taxonomy; Anderson et al., 2001). This revision lists six cognitive processes with sub-processes in a hierarchical listing of increasing cognitive complexity: Remember, Understand,

### Table 1. High School Student Population of Study States

<table>
<thead>
<tr>
<th></th>
<th>Grade 9–12 enrollment in fall 2008</th>
<th>Enrollment as % U.S. total</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>14,980,000</td>
<td></td>
</tr>
<tr>
<td>New York</td>
<td>897,512</td>
<td>6.0</td>
</tr>
<tr>
<td>Ohio</td>
<td>577,669</td>
<td>3.9</td>
</tr>
<tr>
<td>Texas</td>
<td>1,305,637</td>
<td>8.7</td>
</tr>
<tr>
<td>Virginia</td>
<td>380,787</td>
<td>2.5</td>
</tr>
<tr>
<td>Four-state total</td>
<td>3,161,605</td>
<td>21.1</td>
</tr>
</tbody>
</table>

*Note.* Data from National Center for Education Statistics (2010).
Apply, Analyze, Evaluate, and Create. As with the original Taxonomy, cognitive processes are described as being on a continuum “from simple to complex and from concrete to abstract” (Krathwohl, 2002, p. 212), from lower- to higher-order cognitive processes with multiple functions frequently involved in a single assessment task.

The revised Taxonomy acknowledges that there is overlap in cognitive complexity between categories. For example, Explaining, which is associated with Understand, is “cognitively more complex than at least one of the cognitive processes associated with Apply” (Krathwohl, 2002, p. 215). Thus, questions calling for a skill in a higher cognitive category may actually be easier than questions that fit a lower category. This overlap complicates the process of rating an item’s difficulty. At the same time, it is also possible that a question designed to require students to remember may be more difficult for a student to answer correctly than one that requires analysis. For example, a trivial question such as “How much did the *Enola Gay* weigh?” defies analysis but is extraordinarily difficult to answer accurately. Thus, analyzing the state assessments for cognitive level is not intended to determine the difficulty of any particular test. Rather, the goal is to ascertain the extent to which the tests match the expectations in the states’ standards and the extent that higher-order thinking skills in social studies are promoted as part of the high-stakes testing environment.

For purposes of analysis in this study, these six categories are grouped into lower order (Remember and Understand) and higher order (Apply, Analyze, Evaluate, and Create). To facilitate analysis and eliminate as much ambiguity as possible in the scoring, the break point between categories was set between the Understanding and Applying levels of the Taxonomy. While the dividing line between these categories is necessarily arbitrary, the clearest distinction between cognitive functions occurs between these categories. In describing the rationale for using this analytic dichotomy, Vockell (n.d.) stated:

> Bloom’s use of the term application differs from our normal conversational use of the term. When working at any of the four highest levels of the taxonomy [emphasis added], we “apply” what we have learned. At the application level, we “just apply.” At the higher levels, we “apply and do something else.” (Para. 5)

This definition of lower and higher order has also been used in other assessment contexts (e.g., National Assessment Governing Board, 2001, p. 47).

**Procedures.** We collected the standards documents and high-stakes tests administered to high school students in New York, Ohio, Texas, and Virginia. The Ohio and Texas tests were given in 2009. The Virginia test was given in 2007 but not released to the public until 2009. Members of the research team from New York used the June 2010 test administration for their state.
We provide contextual information on each state’s testing requirements, consequences, and policies regarding releasing test materials. These states have revised, or are in the process of revising, their standards and assessment procedures. Testing formats, however, have not changed.

Then, the first two authors independently analyzed the four states’ standards documents. Standards with language indicative of remembering and understanding, e.g., describe or identify, were categorized as lower order. Standards with terms indicative of applying, analysis, evaluation, or creation, e.g., analyze or justify, were labeled higher order. For example, the Ohio standard which requires students to “Explain how the United States has been affected politically, economically and socially by its multicultural diversity (e.g., work force, new ideas and perspectives, and modifications to culture)” (Center for Curriculum and Assessment, 2002, p. 175) was rated as lower order. This rating was based on the verb “explain” and its classification in the Understanding category of the revised Bloom’s Taxonomy (Anderson et al., 2001). There was 100% interrater agreement on classifying standards. We attribute this high rate of agreement to the thoroughness of the Taxonomy in identifying categories for verbs used in the standards and the frequent appearance in the four states’ standards of particular verbs, e.g., analyze, evaluate, explain, and identify.

For each of the states, the test materials we analyzed were the most recently released versions of the high-stakes test. Prior to individually rating the questions from each state, we discussed the Revised Bloom’s Taxonomy (Anderson et al., 2001) and its expectations for cognitive tasks and analyzed several sample MC questions to establish interrater agreement. At that point in the process, we developed several guidelines for item analysis.

1. Questions with some form of document (map, table, reading passage) that do not require application of skills beyond comprehension to determine the correct answer are lower order. In other words, if students use map-reading or graph-reading skills to understand a graphic but are not required to implement a formula or otherwise execute a particular procedure, then the question is lower order.

Example: The federal lands shown [in Figure 1] were set aside as a response to concerns about:

A. border conflicts
B. community development
C. environmental conservation
D. property values (Texas Education Agency, 2009b, p. 31)

The spaces indicated on the map provide no information that can assist students in answering the question (unless environmental conservation is seen as a concern inherent to the western United States).
question could be written and answered just as easily using only the map title without the visual, which may actually be a distractor for some students.

2. Chronology questions, e.g., “which happened first” or “which event led to . . .,” are lower order, since the primary expectation is for students to remember events in sequence.

Example: What was a result of the Industrial Revolution in Europe?

A. The growth of the middle class
B. An increase in nomadic herding
C. A decline in urban population
D. A decrease in international trade (The University of the State of New York, 2010a, p. 6)

This question calls for, at most, comprehension of the impacts of the Industrial Revolution. A student who remembers the sequence of urban development/labor specialization but who does not have an understanding of the Industrial Revolution may also answer it correctly.

3. Questions for which answers provide only one response related in any way to the prompt are generally lower order. The factor that could move such a question to higher order is whether information in the prompt can be analyzed to guide students’ decisions.
Example: Which heading below best completes the partial outline below?

a. Magna Carta
b. House of Burgesses
c. Town meetings
d. John Locke

A. Ideas of Social Darwinism
B. Basis of British Mercantilism
C. Contributions to American Literature
D. Influences on United States Constitutional Government (The University of the State of New York, 2010c, p. 2)

Only one answer relates to any of the topics in the prompt. A student who knows one of the topics and the terms used in the potential answers has only one possible choice, rather than having to reason out the “best” choice.

4. Assume when reading the question that students do not have the content knowledge necessary to answer by remembering. Any question that provides students information that creates a possibility of figuring out the answer, e.g., through analysis, is higher order.

Example: In recent years, information about foods and food preparation around the world has had a wide audience on U.S. television and on internet websites. People who are eager to try new recipes have created a demand for ingredients not usually produced in the United States. To meet this demand, U.S. supermarkets have been stocking goods formerly unknown to U.S. buyers. What has been the cultural effect of the trend described above?

A. a decrease in religious restrictions on particular foods
B. a reduction in the numbers of people “eating out” in restaurants
C. the inclusion of additional international food choices in the U.S. diet
D. people in foreign countries eating more foods produced in the United States (Ohio Department of Education, 2009a, p. 9)

Students who do not know whether the possible responses are real “trends” can analyze the information in the question to determine the correct answer.

With these premises, we independently rated the questions from each state. Four hundred four questions were rated. Examples of questions from each state that were rated lower order and higher order are shown in Appendices A and B, respectively. There was 93% initial agreement between raters, and agreement was reached on all questions after discussion. The results of the analysis of the questions for each test were then put into tables in the form of number of
The Lower-Order Expectations of High-Stakes Tests

higher-order questions, lower-order questions, and total questions. To allow for comparison across the different states, the raw scores were then converted to percentages. A final calculation was done to combine all of the ratings into a single, composite percentage of higher-order and lower-order questions across all of the test materials.

**FINDINGS FOR INDIVIDUAL STATES**

Each state’s standards and test structure will be explained first, followed by the analyses of the test questions. The test from each state was based on different, state-specific standards and testing regimens, though the tests represent the high school exit/graduation requirement in each state. As shown in Table 2, the test structures differ significantly, although all rely either exclusively or predominately on MC questions. After addressing the individual state standards and tests, the analysis will turn to the entire group.

It is important to note that each of these states has conducted extensive statistical evaluations of the validity and reliability of their tests. For example, Texas reports an internal consistency of .76 for its overall test, with individual category values ranging from .73 to .80. We do not contest these results. Rather, we suggest that they provide an incomplete representation of the degree to which high-stakes test expectations match the standards that students and teachers are expected to meet.

In all four states, committees of teachers, state education employees, and test company advisors initially develop test questions. Several factors influence which items are chosen for final inclusion on a test. Among those factors is the overall relationship of the questions to the standards. See Table 3 for an example of the ways question content is distributed as a result of matching

<table>
<thead>
<tr>
<th>Test subject</th>
<th>New York</th>
<th>Ohio</th>
<th>Texas</th>
<th>Virginia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global History and Geography</td>
<td>50</td>
<td>50</td>
<td>32</td>
<td>55</td>
</tr>
<tr>
<td>U.S. History and Government</td>
<td></td>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>Social Studies</td>
<td></td>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>World History I</td>
<td></td>
<td></td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>World History II</td>
<td></td>
<td></td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>U.S. and Virginia History</td>
<td></td>
<td></td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

Note. All of the SA questions for the New York exams are based on documents. MC = multiple choice; SA = short answer; ER = extended response; DBQ = document-based question.
Table 3. 2010 Ohio Graduation Test: Social Studies

<table>
<thead>
<tr>
<th>Number of questions by standard</th>
<th>History</th>
<th>Geography</th>
<th>People in Society</th>
<th>Economics</th>
<th>Skills and Methods</th>
<th>Citizenship Rights and Responsibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

questions to standards. The tests are based on a plan that is derived directly from all or part of the state standards. In Virginia and New York, the standard that focuses on higher-order thinking skills is explicitly excluded from the test plan. In Texas and Ohio, the “Skills” category has a limited number of questions that are distinct from those that target the standards’ content categories. Thus, in two states, the tests focus exclusively on content. In two others, a limited number of questions are designed to address higher-order abilities. No state, however, aligns its test as a whole with the cognitive level demands of the respective state standards. While each state’s testing context is different, our study provides evidence that the result—an overemphasis in testing on lower-order cognitive processes as compared to the standards that students are expected to meet—is consistent across these four states.

New York

New York standards. The New York Learning Standards for Social Studies include five core content areas: history of the United States and New York; world history; geography; economics; and civics, citizenship, and government. The standards are organized further into key ideas and performance indicators. An analysis shows the 32 key ideas as divided fairly evenly between higher order (44%) and lower order (56%). One higher-order example from the world history standards is:

Students will analyze different interpretations of important events, issues, or developments in world history by studying the social, political, and economic context in which they were developed; by testing the data source for reliability and validity, credibility, authority, authenticity, and completeness; and by detecting bias, distortion of the facts, and propaganda by omission, suppression, or invention of facts. (New York State Education Department, 1996, p. 13)

New York core curriculum. The New York State Learning Standards articulate broad goals that guide instruction, but they are not the basis for the New York Regents Examination. Rather, the Social Studies Core Curriculum (New York State Education Department, 1999) includes curriculum guidelines for global and U.S. history and geography in high school. The document is
divided into four columns. The first lists content. The document contains a note that the Regents Examination will be “based on the content column in this core curriculum” (p. 89). This is significant, for the higher-order thinking elements are found in the untested “Connections” column. It is only on the Connections side that questions are raised, such as the inevitability of the war, the possibility of compromise, and the presence of factors besides slavery. The questions in the Connections column raise issues of multiple interpretation and analysis found in the learning standards. Those questions, however, are not used in formulating the Regents Examination, despite the requirement that New York teachers address all of the standards in their classroom instruction.

**New York graduation tests.** New York has separate graduation tests for world and American history. The tests analyzed for this study were the June 2010 Regents Examinations in global history and geography (GHG; The University of the State of New York, 2010a) and in U.S. history and government (USHG; The University of the State of New York, 2010c). Each test had 50 MC questions, one thematic essay question, and a two-part document-based question (DBQ). The first part of the DBQ included SA open-ended questions about specific documents. The GHG exam had 13 questions in this section. The USHG exam had 14 such questions. The second part of the DBQ was an extended response (ER) essay based on the full collection of documents provided in the first part of the question (The University of the State of New York, 2010a, 2010c).

A scaled score of 65 was necessary to pass each of the tests, and the score was determined by combining the raw scores from the MC and DBQ SA sections into one score and the ER scores from the thematic essay and DBQ part two into a “total essay score.” The combined MC and DBQ SA score was determined by awarding 1 point for each correct MC response and 1 point for each correct SA response. Students could thus achieve a raw score of 0 to 63 on this part of the GHG test and 0 to 64 on the USHG test. The thematic essay and DBQ essay were each rated on a 0- to 5-point rubric. The total essay score was derived by adding those two scores. Thus, students could achieve a score

<table>
<thead>
<tr>
<th>Question type</th>
<th>Global History and Geography</th>
<th>U.S. History and Government</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>44</td>
<td>45</td>
</tr>
<tr>
<td>SA</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>DBQ</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Thematic ER</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note. MC = multiple choice; SA = short answer; ER = extended response; DBQ = document-based question.*
of 0 to 10 for their total essay score. These scores were combined to generate scaled scores.

**New York State Regents Examination: GHG.** The New York GHG Regents Examination included 65 questions in MC, SA, and ER formats. As illustrated by Table 4, it was possible for students to achieve a passing score on the test with 55 correct answers on the combined MC and SA sections. Thus, a student could pass the GHG Regents Examination without demonstrating any higher-level cognitive ability.

**New York State Regents Examination: USHG.** With 52 correct answers on MC questions necessary to achieve a passing scaled score of 65, it was possible for a student to pass the New York Regents USHG Examination without demonstrating the ability to use higher-order cognitive skills. The DBQs and thematic essay ER questions were rated higher order. For example, the DBQ included a requirement that students “discuss the extent to which the [student-chosen reform] movement was successful in achieving its goals” (The University of the State of New York, 2010c, p. 21). This question requires students to make an evaluative judgment about a reform movement, which is a higher-order cognitive process. In addition, the scoring guide requires responses scoring a 4 or 5 on a 5-point scale to be “analytical (applies, analyzes, evaluates, and/or creates information)” (The University of the State of New York, 2010d, p. 13).

**Conclusion: New York.** Of the total items (both MC and SA), 59/64 (70%) on the U.S./geography test were lower order, and 57/65 (70%) were lower order on the GHG. With 52 correct answers necessary to achieve a passing scaled score of 65 on either test, it is possible for a student to pass either New York examination without demonstrating the ability to use higher-order cognitive skills.

The majority (66%) of the New York standards were rated as lower order. Even so, there is a considerably greater emphasis on lower-order knowledge on the tests than there is in the state standards.

**Ohio**

**Ohio standards.** The Ohio standards are organized into grade-level benchmarks that span 2–3 years and accompanying “grade level indicators” that provide measurable objectives for each benchmark. Of the 42 benchmarks in the high school segments of the Ohio Academic Content Standards: Social Studies, 24 (57%) include higher-order tasks such as analyzing or evaluating. One example is
Analyze the reasons for the rise and growth of labor organizations in the United States (i.e., Knights of Labor, American Federation of Labor and Congress of Industrial Organizations) including unregulated working conditions, laissez-faire policies toward big business, violence toward supporters of organized labor. (Center for Curriculum and Assessment, 2002, p. 165)

The Ohio Standards also include “application,” a higher-order skill, in a stand-alone process standard called “Social Studies Skills and Methods” that is listed separately to emphasize its application throughout the social studies program. These skills are meant to be taught as students are learning the content associated with the other standards. Students need to use these skills on a regular basis as they learn content knowledge (Center for Curriculum and Assessment, 2002, p. 2). Among these standards are higher-order benchmarks such as “Critique evidence used to support a thesis” (p. 170).

The Ohio Department of Education explains in its background material for the Ohio Graduation Tests that alignment between the tests and the content standards is both a goal and inherent in the development of the tests.

Following the adoption and integration of the Ohio Academic Content Standards into the school curricula, item and test specifications were developed to make sure that the tests and their items are aligned to the standards, benchmarks, and grade-level indicators they are intended to measure. (American Institutes for Research, n.d., p. 6)

*Ohio graduation test.* The 2009 Ohio Graduation Test in Social Studies (OGTSS) was based on seven Ohio Academic Content Standards for Social Studies: history, geography, people in societies, economics, citizenship rights and responsibilities, government, and social studies skills and methods (Center for Curriculum and Assessment, 2002). Each of these standards was addressed with discrete questions on the test. The History standard had the most questions (10), while the Citizenship Rights and Responsibilities standard had the fewest (3). The 2009 OGTSS was comprised of 38 items (Ohio Department of Education, 2009a). Thirty-two of the items were MC. Four items were identified as SA. Two items were identified as ER. See Table 5 for the breakdown of items by cognitive level.

Students were given 1 point for a correct answer on a MC question, up to 2 points for each SA response, and up to 4 points for each ER response. Thus, two-thirds of the potential points (32 of 48) on the test are from forced-response questions and one-third (16 of 48) from open-ended questions. Partial credit is given on SA and ER questions. A raw score of 21.5 out of the 48 possible points was required to achieve a passing score (Ohio Department of Education Office of Assessment, 2009).
Table 5. Ohio State Test Analysis

<table>
<thead>
<tr>
<th>Question type</th>
<th>Lower order</th>
<th>Higher order</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>29</td>
<td>3</td>
</tr>
<tr>
<td>% overall test score</td>
<td>60%</td>
<td>6%</td>
</tr>
<tr>
<td>SA</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>% overall test score</td>
<td>8%</td>
<td>8%</td>
</tr>
<tr>
<td>Thematic ER</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>% overall test score</td>
<td>17%</td>
<td>0%</td>
</tr>
</tbody>
</table>

*Note. MC = multiple choice; SA = short answer; ER = extended response.*

Categorizing the ER questions as lower order is based on the language of the question and the expectations of students as demonstrated in the scoring guide (Ohio Department of Education, 2009b). For example, one question requires students to “Identify four factors (social, political, economic and/or environmental) that contributed to the migration of large numbers of people from Europe to the United States in the late 19th century” (Ohio Department of Education, 2009a, p. 5). The Taxonomy labels identify as a lower-order cognitive process. The scoring guide reinforces this expectation by providing the following as the “exemplar” response:

U.S. industrialization offered opportunities for jobs. Better wages were available in the United States. Land was abundant in the United States. Ethnic minorities came to the United States to escape persecution. (Ohio Department of Education, 2009b, p. 15)

While it may be possible for students to use higher-order processes to derive a response meriting full credit, the language of the question and of the exemplar response shows clearly that the test writers’ expectation is for students to generate responses using lower-order cognitive skills.

**Conclusion: Ohio.** The 2009 OGTSS dealt with a vast amount of information in 38 questions. Seven state standards representing broad disciplinary fields were represented on the test. The great majority of the test (85% of the available points), including open-ended questions, requires students to demonstrate recall or comprehension of information. Only five questions, representing 15% of the test score, pushed students to engage in higher-order thinking. With a minimum 21.5 passing score out of 48 possible points, there was no expectation on the OGTSS that students demonstrate higher-order cognitive skills to pass the test. Also, the cognitive level of the test was not consistent with the Ohio Academic Content Standards, 57% of which require demonstrations of higher-order thinking skills.
Texas

Texas standards. The Texas Assessment of Knowledge and Skills (TAKS) exam is based upon the Texas Essential Knowledge and Skills (TEKS) standards (Texas Education Agency, 1998). The 1998 standards were in effect from 1998 through 2010 and form the basis for the standards and test included in our analysis. While a new test is now administered to Texas students, that test has not been released to the public.

The 11th-grade TAKS exit exam for social studies is based on a total of 106 individual standards covering U.S. and World History (33); Geography (6); Economics (15), Government (10); Citizenship (9); Culture (11); Technology, Science & Society (6); and Social Studies Skills (16). Of the 118 indicators in the 106 TEKS, 62 (53%) were coded as lower-order skills and 56 (47%) classified as higher-order skills. Although many of the individual content objectives were classified as lower order, the Social Studies Skills TEKS were evenly balanced between lower-order (8) and higher-order skills (8), and the standards explicitly direct teachers to teach these skills in all social studies units (Texas Education Agency, 1998). Thus, the overall ratio of higher- to lower-order standards indicators was 64 (48%) higher order to 70 (52%) lower order.

According to the Texas Education Agency, the high-stakes tests given to students are closely aligned with these standards:

The tests are designed to reflect the range of content and level of difficulty of the skills represented in the TEKS; include only those items judged to be free of possible gender, ethnic, and/or cultural bias and deemed acceptable by the educator review committees; and reflect problem-solving and complex thinking skills. (Texas Education Agency, 2010, p. 20)

TAKS High School Social Studies Exam. The TAKS assessed four TEKS objectives, labeled issues and events, geographic influences, economic/social influences, and political influences, with specific sub-objectives. The sub-objectives were divided by discipline, with a sub-objective for the economic/social influences objective being,

History. The student understands the effects of reform and third party movements on American society. The student is expected to . . . (B) evaluate the impact of reform leaders such as Susan B. Anthony, W.E.B. DuBois, [and Robert LaFollette] on American society. (Texas Education Agency, 2009a, p. 3)

Sub-objectives were divided into several categories: History; Economics; Culture; Science, Technology, and Society; Geography; Government; and Citizenship. A fifth objective was “The student will use critical thinking skills
Table 6. Texas State Test Analysis

<table>
<thead>
<tr>
<th>Question type</th>
<th>Lower order</th>
<th>Higher order</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>49</td>
<td>6</td>
</tr>
<tr>
<td>% overall test score</td>
<td>89%</td>
<td>11%</td>
</tr>
<tr>
<td>SA</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Thematic ER</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Note. MC = multiple choice; SA = short answer; ER = extended response; TAKS = Texas Assessment of Knowledge and Skills.

to analyze social studies information” (Texas Education Agency, 2009a, p. 6). Sub-objectives for this category were skills such as

The student applies critical thinking skills to organize and use information acquired from a variety of sources. . . . The student is expected to . . . explain and apply different methods that historians use to interpret the past, including the use of primary and secondary sources, points of view, frames of reference, and historical context. (Texas Education Agency, 2009a, p. 6)

A scaled score of 2,100 was required to pass the TAKS. For the test analyzed here, a 2,100 scaled score was achieved with 27 correct answers. See Table 6 for the analysis of the TAKS examination.

Conclusion: Texas. The TEKS standards were divided nearly evenly between higher-order (48%) and lower-order expectations for students. The TAKS exam, however, was predominately lower-order questions (89%), with just 11% of the questions rated as higher order. With a raw score of 27 required to pass the TAKS and 49 lower-order questions on the test, there was no expectation or requirement that students correctly answer, or even attempt, any of the higher-order questions in order to pass the test and meet the graduation requirement.

Virginia

Virginia standards. The Virginia History and Social Science Standards of Learning (SOL; Virginia Department of Education, 2001) include the content areas of history, civics, and geography. The SOLs are divided by course content, i.e., Virginia and U.S. History, World History I, and World History II. These course standards are then further divided into core standards and substandards. Analysis of the 173 total standards from the three courses showed
that 143/173 (83%) were lower order and 30/173 (17%) were higher order. An example of a higher-order standard in Virginia and U.S. History was:

The student will demonstrate knowledge of economic, social, cultural, and political developments in the contemporary United States by analyzing the effects of increased participation of women in the labor force. (Virginia Department of Education, 2001, n.p.)

The SOL includes a skills category that applies across content areas. According to the test technical report, any standard not specifically excluded from the test by the test blueprint is “eligible for inclusion on each form of an SOL test” (Virginia Department of Education, n.d.-c, p. 5). The World History and Geography I and II test blueprints do not exclude any of the standards (Virginia Department of Education, n.d.-a, n.d.-b). The Virginia and U.S. History blueprint lists three standards that are not included. These require that students demonstrate the following abilities.

1. Formulate historical questions and defend findings based on inquiry and interpretation.
2. Communicate findings orally and in analytical essays and/or comprehensive papers.
3. Develop skills in discussion, debate, and persuasive writing with respect to enduring issues and determine how divergent viewpoints have been addressed and reconciled (Virginia Department of Education, 2001, p. 38).

Standards are excluded because they “cannot be appropriately assessed in the multiple-choice format” (Virginia Department of Education, n.d.-c, p. 6). All other standards are thus assumed to be appropriate for assessment through MC questions.

Virginia History and Social Science end-of-course assessments. Virginia requires multiple end-of-course examinations for graduation, three of which are social studies exams analyzed for this study: World History and Geography I, World History and Geography II, and Virginia and U.S. History. The tests analyzed here were given in 2007. These tests are generally given sequentially in 9th, 10th, and 11th grades. Each test consisted of 60 MC questions. New tests are now in place in Virginia, but they have not been released to the public.

A scaled score of 400 or above was required to pass each of the tests. The scale was consistent between the tests. Twenty-nine correct answers on the Virginia and U.S. History and World History II tests and 30 correct answers on the World History I test were the cut-offs for a score of
Table 7. Virginia State Test Analysis

<table>
<thead>
<tr>
<th>Question type</th>
<th>World History and Geography I</th>
<th>World History and Geography II</th>
<th>Virginia and U.S. History</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower order</td>
<td>Higher order</td>
<td>Lower order</td>
</tr>
<tr>
<td>MC</td>
<td>57</td>
<td>3</td>
<td>59</td>
</tr>
<tr>
<td>Percent of overall test score</td>
<td>95%</td>
<td>5%</td>
<td>98%</td>
</tr>
<tr>
<td>SA</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Thematic ER</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Note. MC = multiple choice; SA = short answer; ER = extended response.

400 (Virginia Department of Education, 2007a, 2007b, 2007c). Question categories included chronological descriptions of eras, e.g., “Postclassical Civilizations” (Virginia Department of Education, 2007b) and “Age of Revolutions” (Virginia Department of Education, 2007c) in World History, and “Expansion, Reform, Civil War and Reconstruction” in Virginia and U.S. History (Virginia Department of Education, 2007a). In addition, question categories include disciplinary divisions, e.g., geography, civics, and economics. See Table 7 for the Virginia test analysis.

Conclusion: Virginia. The Virginia World History I, World History II, and Virginia and U.S. History tests require students to correctly answer either 29 or 30 questions correctly. With respectively 57, 59, and 55 questions at the lower level of Bloom’s Taxonomy, students clearly are not expected to demonstrate higher-order cognitive abilities in order to pass these high-stakes tests.

The Virginia standards of learning emphasize lower-order goals for student learning, with 81% of the standards statements across the three courses utilizing terms that are lower order. The Virginia state tests place an even higher emphasis on lower-order tasks, with 95% of test questions at that level.

Individual State Findings Conclusion

A number of the standards that guided test creation in the states under study included higher-order content and concept objectives, with expectations that students demonstrate critical thinking and problem solving. It is apparent in the analysis of the test items from those states that there is a significant disconnect between the expectations expressed in the state standards and the ways those expectations are operationalized through the high-stakes tests given to students. Two approaches emerge from our analysis. First, some states exclude...
particular standards from the testing process, which eliminates or minimizes the expectation that students will demonstrate higher-order cognitive abilities. A second approach is to define critical thinking as a distinct testing criterion, which allows those writing and evaluating the tests to ignore the cognitive demands of a state’s “content” standards. In addition, it is clear from this analysis of individual test questions on the high-stakes tests in New York, Ohio, Texas, and Virginia that students are neither expected nor required to demonstrate higher-order cognitive abilities in order to pass the required exams.

CROSS-STATE FINDINGS

The standards and accompanying high-stakes tests given in these four states shared some characteristics but not others. The structure of learning outcomes in each state was similar, in that there were broad standards with more specific guidelines. The MC questions on each state test were formatted similarly, in that each had four possible answers from which students could choose.

Content varied greatly between states. For example, New York tested students only on knowledge of history and geography, while Ohio tested content across all of the social studies disciplines. The level of specificity of the tests also varied, as indicated by the number of questions in a particular area. Virginia’s tests included 60 questions on U.S. and Virginia history and 120 questions on world history, while Ohio’s test included 10 questions dealing with all areas of history and 38 questions total. New York’s and Ohio’s tests included opportunities for students to provide constructed response answers. The discussion of the study findings is organized below by research question.

Cognitive Level of State Standards

As shown in Table 8, higher-order expectations ranged from 17% to 57% of the state standards, with three of the four states showing 44% or more. There is no apparent pattern in the relationship between the number of standards and the percentage of standards that point toward higher-order outcomes.

Cognitive Level of the Test Items in These States

This analysis showed that higher-order questions make up from 2% to 12% of the MC questions on these tests (see Table 9). In each state, students can meet the passing standard for their graduation tests without demonstrating the
Table 8. Cross-State Comparison of Standards Ratings

<table>
<thead>
<tr>
<th>Standard</th>
<th>New York</th>
<th>Ohio</th>
<th>Texas</th>
<th>Virginia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number higher order/total</td>
<td>14/32</td>
<td>24/42</td>
<td>64/134</td>
<td>30/173</td>
</tr>
<tr>
<td>% higher order</td>
<td>44%</td>
<td>57%</td>
<td>48%</td>
<td>17%</td>
</tr>
</tbody>
</table>

Table 9. Cross-State Comparison of MC Question Ratings

<table>
<thead>
<tr>
<th>Test grade/topic</th>
<th>New York</th>
<th>Ohio</th>
<th>Texas</th>
<th>Virginia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global History and Geography</td>
<td>6/50</td>
<td>5/50</td>
<td>3/32</td>
<td>6/55</td>
</tr>
<tr>
<td>U.S. History and Government</td>
<td>12%</td>
<td>10%</td>
<td>9%</td>
<td>5%</td>
</tr>
<tr>
<td>Social Studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History I</td>
<td>3/60</td>
<td>1/60</td>
<td>5%</td>
<td></td>
</tr>
<tr>
<td>History II</td>
<td></td>
<td></td>
<td></td>
<td>2%</td>
</tr>
<tr>
<td>U.S. and Virginia History</td>
<td></td>
<td></td>
<td></td>
<td>8%</td>
</tr>
<tr>
<td>Social Studies</td>
<td>1/60</td>
<td></td>
<td>2%</td>
<td></td>
</tr>
</tbody>
</table>

capacity to perform any higher-order cognitive task. Across the entire sample of 367 MC questions, only 29 (8%) required some demonstration of higher-order thinking on the part of students.

In addition to the MC questions, New York and Ohio include open-ended SA questions. All of the New York SA open-ended questions were rated as lower order. Two of the four Ohio SA questions were rated higher order and two were lower order. The New York ER questions, both DBQ and thematic, were rated as higher order, while the Ohio ER questions were lower order.

High-Stakes Test and Standards Alignment with Cognitive Demands

It is clear from comparing the cognitive requirements of the high-stakes tests given by these four states that students are held accountable for very different levels of cognitive difficulty than that described in the state standards (see Table 10). The most dramatic difference was in Ohio, with 57% of its standards being higher order and just 9% of MC questions at that level. Note that the constructed response questions on the Ohio test were also primarily lower order. The state with the most congruence between its standards and its high-stakes test was Virginia. A contributing factor to this, however, is the low percentage (17% across the three content areas) of standards that require higher-order cognitive processes, the lowest percentage in the sample.
Levels of Accountability for Meeting State Academic Standards in Social Studies

There is clearly a disconnect between the cognitive expectations of state standards writers and the demands of high-stakes tests. Not only are the tests written at a much lower cognitive level than the standards, but it is possible in each state to meet the passing requirement of its test without demonstrating any higher-order cognitive ability. Each of the four states includes processes in its standards documents, suggesting that students are expected to demonstrate analysis, evaluation, etc. The tests that provide the accountability to those standards, however, provide little opportunity for students to demonstrate that capacity and no requirement that they do so.

CONCLUSION

This study illustrates that high-stakes social studies graduation tests in at least four states are not challenging students intellectually. Although NCSS standards and many state standards specifically state that students should be provided with active, meaningful curricula that have value beyond high school, this analysis shows the tests are not measuring these criteria, at least in regard to higher-order cognitive skills. Instead, they are illustrating whether or not students can answer lower level questions about social studies content. Studies have shown that assessment will fail to accurately reflect student learning when the standards and the tests do not match, i.e., when standards advocate for developing understanding, but tests measure rote and superficial learning (Black & Wiliam, 1998).

National standards ostensibly establish assessment expectations for a system intended to improve student achievement. Each of these states has failed, having created high-stakes tests that have uniformly low expectations for cognitive challenge. There is considerable evidence that social studies teachers adjust curricular decisions away from standards-based practice in high-stakes testing contexts (Au, 2007; Misco, Patterson, & Doppen, 2011; Vogler, 2005). The current study suggests that students’ higher-order thinking abilities are not valued by the states. Thus, a narrowing of content and emphasis on teacher-centered instructional strategies noted in the research literature is a
rational reaction on the part of teachers to emphasize lower-order knowledge, despite what a standards document may state. These states are thus far from reaching the civic competence goals set forth by the field.

The emphasis on lower-order knowledge transmits a very shallow and inaccurate view of social studies and of history. Teaching historical thinking encompasses a variety of higher-order processes, e.g., developing empathy and analyzing the implications of differing perspectives (Barton & Levstik, 2004). Learning history as a collection of events in time significantly hinders students’ development of the citizenship skills identified by NCSS and distances the content from students’ lives. Students who are only assessed with MC measures are not provided the opportunity to demonstrate whether they are developing the critical thinking and historical thinking skills necessary to become productive citizens.

A further implication of the study is the clear message to teachers that cognitive expectations for students are quite low. As such, and as demonstrated by case studies in the literature (see, e.g., McNeil, 2000; Segall, 2003; van Hover & Pierce, 2006), an argument can be made that teachers should reduce their focus on the content addressed in high-stakes tests and increase the cognitive demands of their classroom instruction. As Saye and SSIRC (2013) demonstrated, more challenging instruction does not harm performance on high-stakes tests. Authentic pedagogy is strongly associated with improved higher-order learning outcomes (Parker et al., 2011).

The disconnect between the cognitive requirements of the standards documents and what the high-stakes tests require of students is dramatic. An important policy issue that arises out of this disparity is whether the differences between the standards and the tests are intentional, i.e., testing authors and authorities have a markedly different vision of what social studies education should be than do those who approved the standards. This view is supported by the explicit exclusion of some or all of the process-oriented standards by various states. This exclusion suggests that those overseeing the production of assessments do not place sufficient priority on higher-order standards to include them in the tests given to students.

Alternatively, perhaps the issue of convenience and expense in testing so controls test development as to negate the mandate for higher-order expectations in the standards, as in Virginia, where “the state’s interest in expediency of grading won out over the appropriateness of the measure” (van Hover et al., 2010, p. 92). In either case, it is clear that the policy choices made in creating state standards are not being followed in the construction of the tests that are intended to promote student achievement toward the goal of civic competence in the social studies.

This study shows clearly the need for better tools for authentically assessing secondary social studies students’ learning. State and national standards documents include requirements for students to analyze sources, synthesize information to create meaningful products, evaluate evidence, and justify decisions, among many other critical thinking skills. If assessment of
standards-based learning is to continue, then it must reflect more accurately the standards that students are expected to meet. Tools for assessing the learning of critical thinking skills in the social studies must be improved so that alignment between state social standards and high-stakes assessments can become a reality.

It is important to note that there is no causal argument regarding classroom instruction or teacher practice that can be made from this study. There is considerable evidence of a historic emphasis on lower-order thinking in social studies classrooms (Cuban, 1991; Goodlad, 1984), as well as a recognition that tests are an “uncertain lever” (Grant, 2001a) in changing pedagogical approaches. Despite this context, the study demonstrates clearly that structures now in place are at best inconsistent, and at worst contradictory, in the extent to which they promote highly valued goals in social studies. There are many possible explanations for this inconsistency, e.g., the technical difficulties of assessing historical thinking and the high cost of authentic assessment.

Nevertheless, there is no question that the state-mandated social studies assessments in these four states do not require students to demonstrate that they have met the state-mandated learning standards. In fact, it is not just that the assessments do not require such a demonstration—they do not allow students to demonstrate that level of achievement. These high-stakes assessments institutionalize lower-order expectations for students in our schools.

REFERENCES


The Lower-Order Expectations of High-Stakes Tests


Texas Education Agency. (1998). *Texas essential knowledge and skills for social studies, Title 19, Part II § 113*. C. Austin, TX: Author.


---

**ABOUT THE AUTHORS**

SCOTT W. DEWITT is an Adjunct Instructor in the Department of Behavior Science at the *University of Cincinnati–Blue Ash College*, Cincinnati, OH 45215. He can be contacted at scottdewitt@cinci.rr.com.

NANCY PATTERSON is an Associate Professor of Secondary Education in the School of Teaching and Learning at *Bowling Green State University*, Bowling Green, OH 43403. She can be contacted at ncpatte@bgsu.edu.

WHITNEY BLANKENSHIP is an Assistant Professor of Educational Studies in the Feinstein School of Education and Human Development at *Rhode Island College*, Providence, RI 02908. She can be contacted at wblankenship@ric.edu.

BROOKE BLEVINS is an Assistant Professor in the School of Education at *Baylor University*, Waco, TX 76798-97304. She can be contacted at brooke_levins@baylor.edu.
LORREI DICAMILLO is an Associate Professor in the Teacher Education Department at Canisius College, Buffalo, NY 14208-1517. She can be contacted at dicamill@canisius.edu.

DAVID GERWIN is an Associate Professor in the Department of Secondary Education and Youth Services at Queens College, City University of New York, Flushing, NY 11367. He can be contacted at david.gerwin@qc.cuny.edu.

JILL M. GRADWELL is an Associate Professor in the History and Social Studies Education Department at Buffalo State College–SUNY, Buffalo, NY 14222. She can be contacted at gradwejm@buffalostate.edu.

JOHN GUNN is an Assistant Professor in the Department of Secondary Education and Youth Services at Queens College, City University of New York, Flushing, NY 11367. He can be contacted at John.Gunn@qc.cuny.edu.

LAMONT MADDOX is an Assistant Professor in the Department of Secondary Education at the University of North Alabama, Florence, AL 35632. He can be contacted at lmaddox@una.edu.

CINTHIA SALINAS is an Associate Professor in the Department of Curriculum and Instruction at the University of Texas at Austin, Austin, TX 78712. She can be contacted at cssalinas@austin.utexas.edu.

JOHN SAYE is a Professor in the Department of Curriculum and Teaching at Auburn University, Auburn, AL 36849-5212. He can be contacted at sayejoh@auburn.edu.

JEREMY STODDARD is Associate Professor in the School of Education at The College of William and Mary, Williamsburg, VA 23187-8795. He can be reached at jdstod@wm.edu.

CAROLINE C. SULLIVAN is a Professor in the Department of Middle and Secondary Education at Georgia State University, Atlanta, GA 30303. She can be reached at ccsullivan@gsu.edu.

APPENDIX A

Examples of Lower-Order Questions from Each State Test

New York

Global History and Geography exam:

What was one social change Mao Zedong instituted in China after 1949?

A. granting legal equality for men and women
B. requiring arranged marriages
C. adopting the practice of foot binding  
D. mandating Confucianism as the state philosophy (The University of the State of New York, 2010a, p. 7)

This question requires students to remember policies of Maoist China and provides no opportunity for analysis.

SA question rated lower on Bloom’s Taxonomy:

. . . For several centuries, these contacts [between North Africa and the interior] were limited by the nature of the Sahara itself. More than 3 million square miles in area, the Sahara is the world’s largest desert. Because temperatures during the day can reach as high as 120 degrees Fahrenheit and supplies of water are scant, the 40-day journey across the desert required courage, determination, and careful planning. Travelers who became separated from their companions were seldom seen again. The trans-Sahara trek became somewhat easier after the 4th century A.D., when camels were introduced in place of horses; camels are able to travel long distances without water, and their wider hooves make it easier for them to move through sand. However, intensive contact between North Africa and the interior did not begin until the 7th century, when a revolutionary change took place in the political and religious life of the region. By this time, the old empires of the Mediterranean and the Middle East were in decline or in ruins. In their place was a powerful new force—Islam . . . . (Source: Philip Koslow, Ancient Ghana: The Land of Gold, Chelsea House Publishers)

What is one reason travel across the Sahara Desert was difficult, according to Philip Koslow (The University of the State of New York, 2010a, p. 16)

This question required students to read and understand the passage. To adequately respond to the question, students need to identify a reason that travel across the Sahara was difficult that was mentioned in the passage. No analysis or interpretation of reasons is required, however. For example, the grading document makes it clear that students mentioning either the size of the desert or the extreme heat were to receive full credit for their response (The University of the State of New York, 2010b).

U.S. History and Geography exam:

Which two groups debated the ratification of the new Constitution?

A. loyalists and revolutionaries
B. Federalists and Antifederalists
C. Democratic Party and Whig Party
D. executive branch and judicial branch (The University of the State of New York, 2010c, p. 3)

This question requires remembering and does not provide information that can allow students to engage in higher-order cognition.

SA DBQ question rated lower order on Bloom’s Taxonomy:

. . . The winning of female suffrage did not mark the end of prejudice and discrimination against women in public life. Women still lacked equal access with men to those professions, especially the law, which provide the chief routes to political power. Further, when women ran for office—and many did in the immediate post-suffrage era—they often lacked major party backing, hard to come by for any newcomer but for women almost impossible unless she belonged to a prominent political family. Even if successful in winning backing, when women ran for office they usually had to oppose incumbents [those in office]. When, as was often the case, they lost their first attempts, their reputation as “losers” made re-endorsement impossible. . . . (Source: Elizabeth Perry, “Why Suffrage for American Women Was Not Enough,” History Today, September 1993)

According to Elizabeth Perry, what was one way in which women’s participation in public life continued to be limited after winning suffrage? (The University of the State of New York, 2010c, p. 14)

This question requires students to read and understand the passage. A satisfactory answer can be taken directly from the passage, e.g., “women still lacked equal access with men to those professions, especially the law, which provide the chief routes to political power” (The University of the State of New York, 2010b, p. 5). No analysis or interpretation beyond understanding is needed to correctly answer the question.

Ohio

In the late 19th century, industrialization led to harsh working conditions in the United States. Which policies of the U.S. government allowed such conditions to develop and later led to the growth of labor unions to correct abuses of workers?

A. laissez-faire policies toward big business
B. antitrust policies toward monopolies
C. imperialist policies regarding territorial expansion
D. isolationist policies regarding international alliances (Ohio Department of Education, 2009a, p. 4)

This question requires students to know a variety of important social studies terms, e.g., imperialist, laissez-faire, but at most requires students to comprehend the question as one of economic conditions allowing pro-business practices and categorize the responses as related to that theme. This certainly rises above the Remembering level on Bloom but does not require students to apply historical thinking skills nor analyze or synthesize information.

Example SA question rated lower order:

The United States fulfilled one of its imperialist ambitions in the early 20th century by acquiring land to build the Panama Canal. State two reasons (political and/or economic) why U.S. imperialists wanted to build the canal. (Ohio Department of Education, 2009a, p. 15)

This question requires students to comprehend the prompt and remember imperialist motivations for building the canal. A student with a good grasp of the concept imperialism may generate a correct answer. Absent remembering specific reasons for building the canal, however, a student is just as likely to provide an incorrect response, e.g., “the U.S. needed a more convenient refueling situation for cross-Pacific shipping.” Thus, remembering and understanding are the cognitive functions needed to provide an appropriate answer to this question.

Example ER question rated lower order:

Identify four factors (social, political, economic and/or environmental) that contributed to the migration of large numbers of people from Europe to the United States in the late 19th century. (Ohio Department of Education, 2009a, p. 5)

This is a simple identification problem that calls for students to remember factors influencing immigration at a particular point in time.

Texas

Use the information [below] and your knowledge of social studies to answer the following question.

- Famous Rough Rider
- First American to win the Nobel Peace Prize
Motto was “Speak softly and carry a big stick”
Strengthened the U.S. Navy

Which twentieth-century U.S. leader is described by the list above?

A. Henry Cabot Lodge
B. George Marshall
C. Theodore Roosevelt
D. Harry S. Truman

This question requires students to recall specific information related to Theodore Roosevelt.

Virginia

World History and Geography I test:

What geographic feature did Hannibal cross to invade the Italian peninsula?

A. The Atlantic Ocean
B. The Himalayas
C. The Alps
D. The Sahara Desert

This question requires simple recall, although students can use their prior knowledge of either geography (name of the physical feature blocking access to Italy) or history (their recollection of Hannibal’s invasion).

World History II test:

The Mughal Empire originated in the area now known as:

A. Northern India
B. Southern China
C. the Middle East
D. the Balkan Peninsula

The presence of the map leads to a presumption that the application of map-reading skills will be necessary to answer the question correctly. This could move the question to the higher order on Bloom’s Taxonomy. The wording of the question, however, makes it clear that students are expected to simply recognize the Indian subcontinent as the area shown on the map. The intent of
the question is student reproduction of knowledge, with no application of skills necessary.

Virginia and U.S. History test:

Missouri’s admission to the Union started the debate over—

A. the right of deposit at New Orleans
B. funding for internal improvements
C. the balance between slave and free states
D. the relocation of American Indians (First Americans) (Virginia Department of Education, 2007a, p. 11)
The question requires students to recall the issues involved in Missouri’s admission. No analysis or synthesis of information was required to identify the correct response. Also, understanding general concepts related to admitting states to the Union does not provide sufficient information for a student to derive the answer without remembering Missouri’s specific case.

APPENDIX B

Examples of Higher-Order Questions from Each State Test

New York

Global History and Geography:

‘Reliance on Imported Oil Raises Concerns in China, Europe, and the United States’
‘Coca-Cola Accused of Wasting Precious Water in India’
‘Competition for Control of Cobalt Mines Causes Violence in Congo’

These headlines best illustrate the economic concept of
A. inflation
B. embargo
C. boycott of goods
D. scarcity of resources (The University of the State of New York, 2010a, p. 8)

This question required students to understand the meanings of the three headlines but also to relate those meanings to their understanding of the economic concepts provided as possible responses. To correctly answer the question, students needed to analyze the headlines for a common theme and connect that theme to the appropriate economic concept.

U.S. History and Geography:

- In the 1940s, President Franklin D. Roosevelt made winning World War II a priority over extending the New Deal.
- In the 1950s, President Harry Truman’s focus shifted from the Fair Deal to the Korean War.
- In the 1960s, President Lyndon B. Johnson’s attention to the Great Society gave way to preoccupation with the Vietnam War.
These presidential actions best support the conclusion that

A. Presidents prefer their role as commander in chief to that of chief legislator
B. Domestic programs are often undermined by the outbreak of war
C. Presidents Roosevelt, Truman, and Johnson were not committed to their domestic initiatives
D. Large domestic reform programs tend to lead nations toward involvement in foreign wars (The University of the State of New York, 2010c, p. 7)

To appropriately answer this question, students needed to examine the three statements for common elements and then draw a conclusion based on that analysis. Students could not determine the correct response from the statements, as no language in the possible answers directly mirrored terminology used in the statements.

Ohio

In 1942, President Roosevelt issued a series of executive orders that authorized the military to remove persons of Japanese ancestry from the West Coast and relocate them to internment camps. In 1982, a commission established by Congress to review the reasons for the relocation and internment found that the decisions were not justified by military necessity. Instead, the commission found that the decision to relocate and intern Japanese-Americans was the result of “race prejudice, war hysteria, and a failure of political leadership.”

The commission findings reflect a change in views on what subject?

A. the balance of power between Congress and the judiciary
B. the balance between individual rights and national security
C. the balance of power between the states and the federal government
D. the balance between freedom of the press and compelling government interest (Ohio Department of Education, 2009a, p. 12)

In addition to understanding the information in the prompt, students needed to demonstrate knowledge of the several important terms in the potential answers. That understanding then had to be analyzed to establish the best answer to the question, which could not be determined either by remembering or by only comprehending the passage.

Interestingly, the prompt gives information showing that the preferred answer (B) is incorrect. Since the Congressional commission “found that
the decisions were not justified by military necessity,” there was no apparent change in views on “the balance between individual rights and national security.” Based on the commission’s findings, if the relocation was “justified by military necessity,” then it would have been allowable. This is the same rationale used by Roosevelt in issuing the executive orders. The change that happened over time in this situation was the perception of military necessity.

Example of SA question rated higher order:

What would be the expected effect on consumer spending and on the unemployment rate if the United States government were to increase personal income tax rates? (Ohio Department of Education, 2009a, p. 3)

Students who do not remember this cause and effect scenario can analyze the situation to develop an appropriate response, i.e., reducing the amount of disposable income available to consumers and a resulting reduction in consumer spending leading to increased unemployment.

Texas

Use the information [below] and your knowledge of social studies to answer the following question.

- 500,000 Hispanics served in the armed forces during World War II.
- The Naval Air Station at Corpus Christi refused to treat sick Hispanic veterans.
- Dr. Hector P. García founded the American GI Forum in 1948.

Dr. Hector P. García founded the American GI Forum to ensure that Hispanic veterans received—

A. the same honorary rank as other veterans who served
B. lifetime employment in the military
C. job placement in medical professions
D. the same benefits provided to other veterans (Texas Education Agency, 2009c, p. 20)

It may be possible for some students to remember the answer to this question without the information provided. Those who do not answer from memory must read and comprehend the information provided in the question stem and then analyze that information to make the connection between denial of health care and benefits provided to veterans.
World History I test:

A prince, therefore, ought always to take counsel, but only when he wishes and not when others wish . . . also, on learning that any one, on any consideration, has not told him the truth, he should let his anger be felt. – Machiavelli, *The Prince*

Based on this quote, Machiavelli’s political writings supported the idea of

A. absolute power  
B. limited monarchy  
C. popular sovereignty  
D. parliamentary government (Virginia Department of Education, 2007b, p. 34)

Students may remember Machiavelli’s preference for absolute power. Those who do not will still need to remember several significant social studies concepts, i.e., limited monarchy and parliamentary government, to successfully respond to the question. In addition to knowing these terms and understanding the quote, however, students must interpret the quote’s implication that the ruler should have the authority to “let his anger be felt” and not be constrained in the exercise of power.

World History II test:

Germany does not look to Prussia’s liberalism, but to her power . . . . The great questions of the day are not to be decided by speeches and majority resolutions . . . but by blood and iron! — Otto von Bismarck, 1861

What term is used to describe Bismarck’s beliefs expressed in this quote?

A. Realpolitik  
B. Perestroika  
C. Imperialism  
D. Appeasement

Similarly to the explication of the Machiavelli quote in the World History I test, students who do not remember the link between Bismarck and realpolitik can
still reach the correct answer by combining knowledge of social studies terms and interpretation of the quote’s implications.

**Virginia and U.S. History test:**

All of the following contributed to the trend shown in the graph EXCEPT—

A. increasing industrialization  
B. poor worker-management relations  
C. discontent with working conditions  
D. a decline in the population (Virginia Department of Education, 2007a, p. 19)

This question required students to apply graph-reading skills to know that labor union membership increased dramatically over the time shown in the graph. The content of the question and possible answers require students to understand that union membership increased and that population decline would be irrelevant to, and possibly contraindicative of, such an increase.