Impact of the Design of U.S. History Textbooks on Content Acquisition and Academic Engagement of Special Education Students: An Experimental Investigation

Mark K. Harniss, Jennifer Caros and Russell Gersten

J Learn Disabil 2007 40: 100

DOI: 10.1177/00222194070400020101

Impact of the Design of U.S. History Textbooks on Content Acquisition and Academic Engagement of Special Education Students:

An Experimental Investigation

Mark K. Harniss, Jennifer Caros, and Russell Gersten

Abstract

We used randomized controlled trials to compare the impact of the designs of 2 United States history textbooks on the content acquisition and behavior of 8th-grade students identified for special education services or identified as low achieving. We also investigated whether teachers differed in their use of instructional activities and questioning strategies based on the type of text used. Our findings suggest that students learned more history content, were more actively engaged, and answered more questions correctly when using the experimental textbook. Teachers used different activities depending on which textbook they used, but did not differ in types of questions asked.

In the past 5 years, there has been renewed emphasis on the central role that curricula play in student learning (e.g., Ball & Cohen, 1996). Curriculum selection plays a large role in the *Reading First* program (U.S. Department of Education, 2006a) and will, in all likelihood, play a large role in the new *Mathematics Now* initiative (U.S. Department of Education, 2006b). Problems in most history texts and their detrimental effects on average-performing students were documented almost 20 years ago (Beck, McKeown & Gromoll, 1989). In their analysis of fifth-grade texts, Beck et al. found two major problems. The first was that the texts assumed far more background knowledge than the student possessed. For example, the texts assumed that students understood that each voting citizen has an indirect voice in setting taxes, in that she or he votes for representatives who pass (or fail to pass) tax legislation. So students did not comprehend what the common phrase "no taxation without representation" meant, let alone why the lack of representation would anger colonists to such an extent that they would revolt. The other major failing was that the "presentation of history content in the programs was not oriented toward developing a coherent chain of events. Text presentations lacked the coherence needed to enable students to draw connections between events and ideas" (McKeown & Beck, 1994, p. 7). Subsequent interviews with fifth graders who had read these texts revealed that most students recalled that most colonists were upset about paying taxes and were not satisfied when Britain lowered the price of tea. However, virtually none of the students understood that the colonists were upset because tax laws were passed by the British Parliament, a body in which "they had no voice. Yet this represents the fundamental issue underlying the colonists' struggle for independence" (McKeown & Beck, 1994, p. 7). In other words, the texts failed to help students understand the underlying motives for actions taken by governments or groups of people, the implications of events, and the tensions and conflicts that underlie events.

Method

The research presented in this article investigated the differential effects of an experimental United States history text (Carnine, Crawford, Harniss, & Hollenbeck, 1994) and a traditional American history text (Garraty, 1991). As much as possible, the experimental text was organized around the causal text structure. The specific principles underlying the design of the text have

been described in greater detail elsewhere (Carnine, Miller, Bean, & Zigmond, 1994; Harniss, Hollenbeck, Crawford, & Carnine, 1994). In essence, to help students understand the content, the authors aimed to (a) present coherently organized and linked information, (b) support and guide students in their initial use of this information, and (c) provide students with appropriate review and practice. The major objective of the text was to teach history as a series of *related* events and actions and to make the relationships explicit. Prior research has demonstrated that many students see history as a series of isolated facts and are rarely able to discern reasons for decisions taken by national leaders or groups of people. This is particularly true for special education students (see Gersten, Baker, Smith-Johnson, Dimino, & Peterson, 2006). By using consistent conceptual frameworks to cut across eras and regions, causal relationships could be made explicit. The goal was for students to begin to use these frameworks on their own as they developed interpretations of key historical events.

Research Questions

The primary research question addressed content acquisition. In particular, we asked whether the experimental text, which explicitly taught students about relationships between events, would lead to significantly better outcomes on three measures of content acquisition. The first was a test developed to capture the content of the units covered in both the experimental and comparison texts. The second measure was a selection of relevant items from the *National Assessment of Educational Progress* (NAEP) American history tests. The third was weekly progress monitoring probes similar to those used by Espin and her colleagues (see, e.g., Espin, Busch, Shin, & Kruschwitz, 2001).

Furthermore, we explored whether students' engagement and accuracy during history instruction would be different depending on the type of text they used. Specifically, we asked whether students would be more on task and less off task and whether students would answer questions more accurately during classroom instruction depending on the type of text they used. Finally, we were interested in the impact of the texts on teaching practice. Specifically, we asked whether teachers used different types of activities during instruction and asked different types of questions based on the type of text they used.

Setting

Two middle schools that administered resource and self-contained programs for students identified with the label of either *serious emotional disturbance* (SED) or *learning disabilities* (LD) in two medium-sized school districts in the Pacific Northwest were selected based on their willingness to participate in the project. The two districts were located in adjacent cities with populations numbering approximately 130,000 and 45,000, respectively. Both schools were about the same size (approximately 500 students) and scored around the average on statewide measures of reading and math. These tests are graded on a scale from 150 to 300. Scores higher than 216 in reading and 221 in math are considered proficient at the eighth-grade level. One school scored at 226 in reading and 227 in math. The other school scored at 228 and 229, respectively. The schools drew students from areas of low to moderate socioeconomic status (SES). In a statewide ranking by family income, parent education, student mobility, and student attendance of 336 middle schools, one school was ranked 155th, and the other was ranked 29th. Although the schools' statewide rankings were quite different, we dealt with these differences by ensuring that an experimental and a comparison group were selected from each school, thus ensuring that cross-school differences were shared across experimental and comparison conditions.

Participants

Initially, 50 middle school students were selected as participants in this study. One student was a seventh grader; all others were eighth graders. There were 26 participants in one school and 24 in the other. Students were selected in two different ways. All students whose primary academic or behavioral placement was in a special education classroom (resource or self-contained) or who were identified with the label of SED or LD based on district and federal criteria were selected to participate in the study. Moreover, low performers who were not identified for special education services were selected by the school principal or vice-principal to participate. In each school, students were randomly assigned to either a target or comparison group, resulting in approximately 12 to 13 students per group per school.

Attrition was high in this study, in part due to the nature of the sample and in part due to the length of the treatment (approximately 20 weeks). A total of 21 students left their school placement during the course of the study. Of the initial sample, 6 students did not remain in their initial school setting long enough to complete pretesting. Another 2 students left soon after the initiation of the study, and the remaining 13 left over the course of the study because they moved to different special education settings or to other schools. The final sample consisted of 29 students. Although we do not have overall attrition rates for these schools for comparison, teachers reported that this level of mobility was not uncommon.

Table 1 shows demographic information for the final sample. The comparison group included all the students with SED, whereas the experimental group included more students with LD.

Teachers

At each school, both special education teachers (i.e., those teaching a resource

TABLE 1
Student Demographics by Group

Measure	Experimental	Comparison
Gender		
Girls	8	4
Boys	7	10
Ethnicity		
African American	1	0
Hispanic	0	3
European American	14	11
Special education		
SED	0	4
LD	5	2
General education	9	9
Total	15	14

Note. SED = serious emotional disturbance; LD = learning disability.

room pullout program and those teaching a self-contained class) participated. One teacher taught the experimental curriculum, and the other the traditional text. Teachers were not selected randomly; rather, two teachers with recent connections to the university were recruited to teach the experimental text. Based on their school placement, two additional teachers were recruited for the comparison condition. Three of the four teachers had master's degrees in special education, and all teachers had special education endorsements from the University of Oregon. The two comparison teachers had more years of experience, with 7 and 20 years, respectively. The two experimental teachers had fewer years of experience, with 1 and 5 years, respectively.

Instructional Materials

This study compared the effectiveness of an experimental textbook to a comparison, "traditional" textbook. The major differences between the two textbooks relate to issues of (a) organization of the content and (b) the types of study strategies suggested in the text. A detailed analysis of one chapter from each text can be found in Harniss (1996). The texts are briefly described in the following sections.

Experimental Text. The experimental text, *Understanding U.S. History: Volume I—Through the Civil War* (Carnine, Crawford, et al., 1994) covers the period of time from early Native American civilizations prior to European settlement up to the Civil War. The four authors wrote, field-tested, and modified chapters based on field-testing with eighth graders. Furthermore, a history expert evaluated the accuracy of the material presented in the text.

The text was a prepublication, post–field-testing version. It was double spaced, copied double sided, and bound. Pictures and graphics were black-and-white line art selected from computer clip art or generated using computer-based graphics programs. The text has been described in more detail in previous articles (Carnine, Caros, Crawford, Hollenbeck, & Harniss, 1996; Crawford, Carnine, Harniss, Hollenbeck, & Miller, 2007; Harniss, Dickson, Kinder, & Hollenbeck, 2001; Harniss, Hollenbeck, Crawford, & Carnine, 1994; Harniss, Hollenbeck, & Dickson, 2004).

The key principle underlying the design of the experimental text was that students need consistent frameworks to help them understand the array of topics and issues in various historical eras. The goal was to present students with a set of coherent conceptual frameworks for understanding history and developing perspectives on issues raised in the text. The authors used several "big ideas" to help students develop a rudimentary understanding of important relationships. Designed primarily around the cause-and-effect text structure, the text presents history as a series of problems that people have encountered in the past and describes how they attempted to solve them and what the intended and unintended consequences of these attempts were. The authors used the terms "problem–solution–effect" (PSE) throughout the entire text. PSE was reiterated in textual passages, graphics, and in many different questioning situations. Other big ideas that helped shape the text include a framework for understanding how groups of people work together (the *stages of cooperation*; i.e., identifying a problem, occasional cooperation, regular voluntary cooperation, and legally binding cooperation) and the *four factors of group success* (i.e., motivation, leadership, capacity, and resources).

Furthermore, the text discussed both cultural and economic rationales for decisions made by governments or their citizens. One major focus was on factors that influence economic development beyond basic subsistence needs (i.e., the three factors of climate, geography, and natural resources affect the economic development of a group of people in terms of their agriculture, manufacturing, and trade). These same factors were reviewed across a wide array of contexts, such as the northern versus southern colonies.

Several activities were included to enhance student learning from the text. First, while reading the text, students answered interspersed questions (Andre, 1979) after every one to two paragraphs. Designed to highlight important information and ensure that students were actively engaged in learning the content, these brief questions were primarily factual. Their goal was simply to have the students stop and think aloud for a minute or so about what they had just read without

breaking the flow of reading. In addition to interspersed questions, students answered frequent embedded discussion questions, which required them to go beyond merely reiterating factual information to combining previously learned information with new information. Finally, as students read, they encountered critical information in the form of "test" questions that required students to remember complex, interrelated information, often in the form of the aforementioned big ideas or conceptual structures.

Comparison Text. The comparison text, *The Story of America* (Garraty, 1991), was published and nationally distributed by Harcourt Brace Jovanovich Holt. For one district, it was the preferred text; for the other, it was one of three adopted texts. The text covers the period of time from early Native American civilizations to modern times and was designed as a narrative approach to history, telling the stories of different peoples at different times in history. Two components of the text are designed to enhance the narrative structure and increase students' interest and engagement in learning history. First, the use of visuals and primary source documents provides students with the words and pictures of the people who lived in historical times. Second, different points of view and historical interpretations of history are presented to demonstrate to students that there are many ways of understanding historical events.

The comparison text is designed with three additional components to enhance students' learning of history. First, the connection between geography and history is developed through the use and interpretation of maps. Second, essential skills such as reading timelines and graphs are taught within the context of understanding history. Third, chapter and unit reviews are provided to enhance retention of the material.

Measures

Oral Reading Fluency. Because students' ability to read was an important part of this intervention, the oral reading fluency of the experimental and comparison group students were assessed at pre- and posttest. Three passages were randomly selected from the experimental text, the traditional text, and the Macmillan sixth-grade basal reader, for a total of nine passages. Passages were administered using standardized directions. Students were timed for 1 minute, and the number of words read correctly per minute and the number of errors per minute were calculated. Students' median scores from each of the three passages were selected as the best estimate of their reading fluency in each of the three texts (see Table 2). Average reading fluency scores at pretest were not significantly different between groups.

Measures of Content Acquisition. *NAEP.* Relevant items from the *National Assessment of Educational Progress* (NAEP) American history tests were used as one measure of content acquisition. These multiple-choice items were drawn up by teams of history experts and constructed by NAEP's measurement experts. Some of these items were used on previous exams and are open to the public; others were "secure" items, available only to researchers. We used relevant items from the 3rd-, 8th-, and 11th-grade versions of the test that dealt with American history. Initial reliability analysis of the NAEP showed that internal consistency reliability was problematic at pretest (α = .517), although acceptable at posttest (α = .635). Twenty-five items with weak item-to-total correlations were removed, resulting in 24 remaining items. This new measure had a higher alpha at pretest (α = .725) than, and approximately the same alpha at posttest (p = .635) as, the 49-item measure. The NAEP items were administered at both pre- and posttest.

Content acquisition measure. This experimenter-constructed measure covered the content of both the experimental and comparison group texts. Items were short answer, matching, and multiple-choice questions selected equally from the experimental and traditional texts. One teacher from each condition selected questions that most accurately represented the information covered in their respective curricula. By having teachers select the content of the test, we assumed that the measure reflected the content covered in their classes, not just any content found in the text. Moreover, this resulted in a test that had question types familiar to both groups. We assumed that any difference in question types (i.e., multiple choice versus matching) was balanced out across conditions. These questions were then combined in alternating order into a test with 32 questions administered at posttest to all students in both treatments. The standardized item α for this measure was .91.

Progress monitoring measure. We used a progress monitoring system in both experimental and comparison classrooms (Caros, 1996). The progress monitoring measure was based on the concept of *generalized indicators of performance* or *general outcome measures* (Deno, 2003; Espin, Busch, Shin, & Kruschwitz, 2001; Espin & Foegen, 1996). Given the nature of the population engaged in this study, we wanted to en-

TABLE 2
Oral Reading Fluency Scores by Text Type and Group

	Experimental		Comparison	
Text type	***M***	***SD***	***M***	***SD***
Macmillan	105.8	37.1	100.6	51.26
Traditional	89.08	34.22	82.29	45.41
Experimental	94.47	31.43	87.86	41.81

Note. All p values are nonsignificant.

sure that the measure was sensitive to growth for students at the lower end of the ability range and could be used by special education students. Terms from a list of 147 critical content vocabulary words (with definitions) were randomly selected to generate 22 alternate-form weekly vocabulary probes. These terms were randomly sampled from an annual vocabulary test consisting of 90 vocabulary terms and their meanings taught in the first six chapters of the experimental history text. Each probe contained 20 items in a typical matching format. Administration directions were standardized, and students were given 5 min to match vocabulary meanings to terms. Criterion-related validity was .66 for the NAEP items used in this study and .56 with the *Iowa Test of Basic Skills* Social Studies test. Mean alternate form reliability was .70.

Classroom Observation Measure. A classroom observation measure was used in experimental and comparison classrooms to help ascertain differences in engagement and instructional practice between the two conditions. The observation instrument included 18 codes: (a) 5 activity variables, (b) 7 teacher variables, and (c) 6 student response and engagement variables. Interrater reliability data were collected in the classroom for approximately 15% of observations. In all cases, the principal investigator served as the reliability coder. Because of the relatively low sample size and the resulting nonnormality of the data, nonparametric statistics were used to assess reliability. A Spearman rank correlation was used to determine the equivalency of the rank ordering of subjects between the two observers. Correlations were mostly significant at the .05 level and fell generally in the moderate range (M = .64). Only two codes produced nonsignificant correlations (i.e., student-led and non-history questions). Both of these codes occurred at low frequencies. Because correlations do not take into account level differences (e.g., if one coder were to consistently code higher than another, correlations could still be high), a Wilcoxon matched pairs signed-ranks test was used to determine whether mean levels differed between the two observers. None of the variables showed significant differences between observer estimates and reliability estimates noted. This is especially important because differences in mean levels were the main comparison of interest.

Procedure

At the beginning of the study, each student was given her or his own copy of the experimental or comparison U.S. history text. All teachers were given a 2-hour professional development session that described the text they would be using, demonstrated the instructional strategies designed into the text, and provided suggestions for effectively teaching from the text. Teachers taught history in block periods of 90 min per day, 5 days per week, and were not directed to cover a specific number of pages per day. Rather, they were allowed to vary their pace based on student needs. Although this creates a threat to validity due to unequal content coverage, this flexibility made the intervention more acceptable to teachers and allowed us to determine whether text difficulty influenced content coverage. In the end, teachers covered roughly the same number of chapters (between four and five). As we noted in the Measures section, we compensated for differences in coverage by having teachers select the questions for the criterion-related test that most closely related to the information that students had learned.

Teaching and Classroom Management Practices

In both conditions, teachers' instructional and behavioral approaches were monitored but not controlled. Three of the teachers were interviewed at the end of the study and asked about their teaching practices and behavior modification programs, and whether they made any modifications to the history curriculum. One experimental condition teacher was unavailable for interview. Summaries of these interviews are provided in Harniss (1996) and suggest that teachers were similar in their approaches to teaching and behavior management and in the degree to which they modified the curricula.

Data Collection

NAEP measures were administered to the classroom as a group and were collected at the beginning and end of the study. The content acquisition measure was also group administered, but was used only at posttest. Observational data were collected from late April to late May. Most students were observed at least three times during this period, although a few were observed fewer times due to chronic absenteeism. The three observations for each student were combined and averaged to provide an estimate of teacher and student behavior. Finally, the brief vocabulary general outcome measures were administered to students in both experimental and comparison classrooms once a week.

Data Analysis

The primary question relates to students' acquisition of history knowledge and is answered through three measures of history knowledge—the NAEP, the content acquisition measure developed by the researchers, and the vocabulary progress monitoring measures. In developing a data analysis plan, we tried to take into account that although participants were randomly assigned to intervention conditions, they were taught in four different classes. In such a situation as this, there remains some controversy as to which is the most appropriate unit of analysis—the student or the teacher. The student was the unit of assignment to condition, and would, according to the What Works Clearinghouse guidelines, for example, be the most appropriate unit of analysis. Yet some have argued (e.g., Peckham, Glass, & Hop-

kins, 1969) that it is most important to analyze data at the level at which instruction was received—in this case, the class. Contemporary analysis guidelines would call for the use of hierarchical linear modeling with an examination of effects at both the classroom and the individual student level (Murray, 1998). However, with only two classrooms per condition, such analyses were not feasible. The design possessed insufficient power for any analyses to be conducted at the classroom level. Thus, we followed the What Works Clearinghouse guidelines by accounting for the nesting of classrooms in conditions and explored effects at the classroom and student level, taking into account the nesting.

For the first two measures, ANOVAs were conducted with classroom nested within condition; a follow-up analysis using a Scheffé test was conducted on the content acquisition measure to look at the difference between conditions by classroom.

Three additional analyses were conducted on the NAEP and the content acquisition measure. First, a univariate, repeated measures ANOVA was used to evaluate differences on growth from pretest to posttest of the NAEP. Second, analyses of the NAEP items were conducted for both the long form and a streamlined form (adjusted for higher coefficient alpha reliability). Third, the content acquisition measure was divided into two separate tests based on whether the questions were drawn from the experimental text or from the comparison text. Because there were more items for the experimental portion of the test, ratios were calculated for each test and compared using two separate ANOVAs. These ANOVAs were conducted to determine whether students in the two conditions performed differently on one item cluster than on the other, as one might anticipate.

For the vocabulary progress monitoring measures, the data were graphed and visually analyzed. The two questions assessing teacher and student behavior were addressed through direct observational data. These data are reported as means and standard deviations. Univariate tests of significance were calculated to determine differences between conditions. An alpha level of .05 was used for all statistical tests.

Results

Measures of History Knowledge

ANOVAs with classroom hierarchically nested in condition were conducted to evaluate the difference between groups on the NAEP and content acquisition measures. The results of these F tests showed significant differences between groups on the content acquisition measure, $F(3, 18) = 6.14$, $p = .005$. A post hoc Scheffé test was conducted on the content acquisition test to determine the degree to which groups differed by classroom. These tests showed that Experimental Group 2 was significantly different from both comparison groups. Experimental Group 1 was significantly different from one of the comparison groups, but not the other.

In contrast, the results for the NAEP were nonsignificant, $F(3, 18) = 0.46$. Secondary analyses using time of test as the repeated factor and treatment condition as the between-students factor revealed no main effect for time of test (pre/post), $F(1, 23) = 0.157$, $p = .70$; no main effect for condition (experimental/comparison), $F(1, 23) = 0.437$, $p = .52$; and no interaction, $F(1, 23) = 0.218$, $p = .64$. These results indicate that neither group of students improved on NAEP performance. When the NAEP items were adjusted to increase its reliability, the modified, more reliable NAEP did not increase its ability to discriminate between conditions.

As noted earlier, there were significant differences between conditions on the content acquisition measure. Additional analyses were conducted to determine whether there were differences between the groups' performance on questions selected from the experimental text versus those selected from the comparison text. Two content acquisition scales were created, one with items from the comparison text ($n = 7$) and another with items from the experimental text ($n = 25$). The difference in the number of items was due to the type of questions selected by teachers from their respective curricula. Teachers in the experimental condition selected short answer and matching questions, which often had more than one part. In contrast, teachers in the comparison condition selected one-part multiple-choice questions.

Cronbach's alpha was recalculated for each content acquisition scale. Coefficient alpha for the comparison text scale was relatively low ($\alpha = .39$), but alpha for the experimental text scale was quite high ($\alpha = .94$). The Pearson product-moment correlation between the two scales was small and nonsignificant, $r = .16$, $p = .44$, so separate t tests were used to analyze group differences on each of the two measures.

To address the problem of comparability between scales with disparate numbers of items, ratios were calculated by dividing the number of questions answered correctly by the number of questions in each scale. These ratios were used in the t tests. The results showed that students were not significantly different on the items selected from the comparison text (experimental group, $M = .38$, $SD = .21$; comparison group, $M = .41$, $SD = .25$), $p = .754$, but were significantly different on the items selected from the experimental text (experimental group, $M = .87$, $SD = .22$; comparison group, $M = .38$, $SD = .26$), $p = .000$, in favor of the experimental group.

Progress Monitoring Measures

Results from the vocabulary progress monitoring measures are shown in Figure 1. Students in the experimental group increased from 3 to 16 correct meanings per 5 min, an average of 0.2 meanings per minute each week. Comparison students decreased slightly,

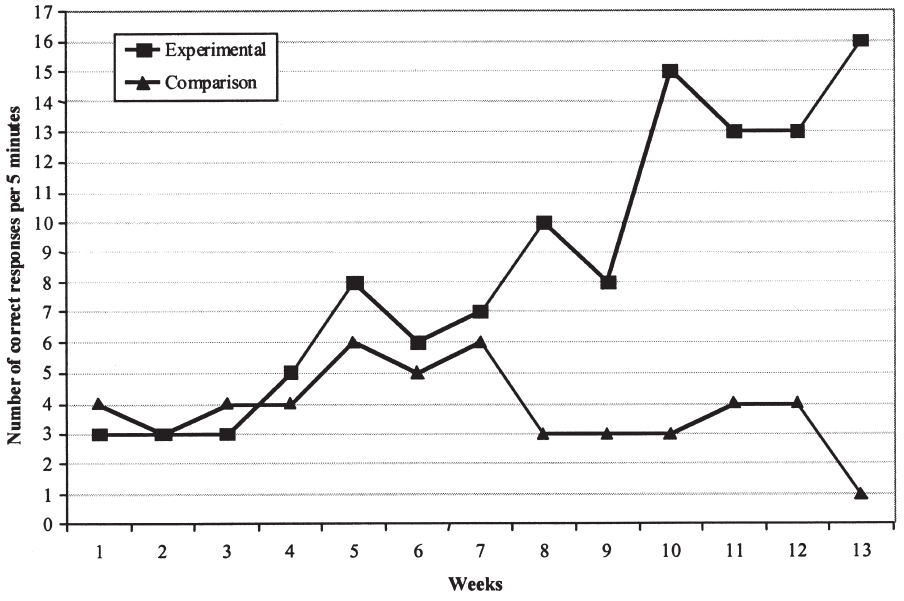

FIGURE 1. Experimental and comparison group performance on U.S. history vocabulary progress monitoring measures across 13 weeks

from 4 to 1 correct meanings per 5 min. Thus, the slope for the experimental group was positive; for the comparison group, the slope was essentially zero.

Oral Reading Fluency

A secondary analysis was conducted on the oral reading fluency data collected at the beginning and end of the study. These data were analyzed to assess differences between groups and over time. Three repeated measures ANOVAs were conducted on each of the three passages. All three passages showed only a significant effect for time, $F(1, 22) = 0.075$, $p = .787$, for the Macmillan passage; $F(1, 22) = 24.39$, $p = .000$, for the traditional history textbook passage; and $F(1, 22) = 13.30$, $p = .001$, for the experimental text passage. These results suggest that there were no differences between conditions, but that there was significant growth in oral reading fluency for both groups from the beginning to the end of the study on both types of history texts.

The relationship between students' scores on measures of history knowledge and oral reading fluency were investigated by correlating students' average oral reading fluency scores at the end of the study to measures of history knowledge collected at the end of the study using a Spearman rank correlation. The results of this analysis showed high, significant correlations between students' oral reading fluency on all reading passages and their own scores on the NAEP (Macmillan, $r = .749$, $p = .000$; traditional, $r = .817$, $p = .000$; experimental, $r = .625$, $p = .001$). In contrast, there were low to moderate correlations between students' oral reading fluency and the content acquisition measure (Macmillan, $r = .346$, $p = .106$; traditional, $r = .446$, $p = .033$; experimental, $r = .400$, $p = .052$). Only the correlation between the comparison text and the content acquisition measure was significant at the .05 level, although the correlation between the experimental text and the content acquisition measure approached significance.

These correlations were broken down by condition to determine whether there were different relationships depending on condition. The results of this analysis showed that the oral reading fluency scores of students in the comparison group were highly correlated with their scores on the NAEP (Macmillan, $r = .853$, $p = .005$; traditional, $r = .923$, $p = .002$; experimental, $r = .850$, $p = .004$) and moderately correlated with the content acquisition measure (Macmillan, $r = .614$, $p = .042$; traditional, $r = .495$, $p = .101$; experimental, $r = .586$, $p = .052$). In contrast, oral reading fluency scores for the experimental group were only moderately correlated with scores on the NAEP (Macmillan, $r = .573$, $p = .047$; traditional, $r = .510$, $p = .077$; experimental, $r = .458$, $p = .112$) and not correlated with the content acquisition measure (Macmillan, $r = .144$, $p = .633$; traditional, $r = .257$, $p = .395$; experimental, $r = .074$, $p = .806$).

Student Academic Engagement

Student behavior was analyzed by condition to determine whether students were more engaged and accurate depending on the type of text. Students in the experimental group ($M = .45$, $SD = .19$) were more actively engaged than students in the comparison group ($M = .26$, $SD = .14$), $p = .011$, and students in the comparison group were more off-task ($M = .21$, $SD = .16$) than students in the experimental group ($M = .08$, $SD = .08$), $p = .028$. The groups did not differ on passive engagement (experimental group, $M = .45$, $SD = .17$; comparison group, $M = .49$, $SD = .13$), $p = .45$, or disruptive behavior (experimental group, $M = .012$, $SD = .02$; comparison group, $M = .026$, $SD = .03$), $p = .173$.

We compared accuracy variables for student responses to questions (i.e., correct, incorrect, partially correct). Experimental group students answered questions significantly more correctly ($M = .91$, $SD = .13$) than comparison group students ($M = .59$, $SD = .40$), $p = .038$. Comparison group students answered questions significantly more incorrectly ($M = .37$, $SD = .37$) than experimental group students ($M = .04$, $SD = .08$), $p = .021$. There were no sig-

nificant differences between groups on partially correct responses (experimental group, $M = .05$, $SD = .12$; comparison group, $M = .06$, $SD = .11$), $p = .902$.

Teaching Practice

A direct observation measure was used to determine whether teachers used different activities and asked different types of questions depending on text type. Five possible activity structures were coded (i.e., teacher led, student reading, student led, independent work, and other). Means, standard deviations, and the results of *t* tests are reported for the proportion of time that teachers in each condition engaged in these activity structures. Levene's test for equality of variance was conducted and, if necessary, the *t* tests were adjusted for inequality of variance. The results of these analyses indicated that teachers using the comparison text engaged in significantly more teacher-led instruction (experimental group, $M = .45$, $SD = .26$; comparison group, $M = .66$, $SD = .24$), $p = .038$. Teachers using the experimental text had students spend more time working independently (experimental group, $M = .43$, $SD = .28$; comparison group, $M = .16$, $SD = .18$), $p = .008$. The groups did not differ on the student reading activity structure (experimental group, $M = .06$, $SD = .071$; comparison group, $M = .13$, $SD = .17$), $p = .207$; the student-led activity structure (experimental group, $M = .00$, $SD = .00$; comparison group, $M = .00$, $SD = .00$), or the "other" activity structure (experimental group, $M = .05$, $SD = .05$; comparison group, $M = .03$, $SD = .04$), $p = .30$.

Four categories of teacher questions were coded (i.e., history task, content question, non-history question, and disciplinary statement). The proportion of questions that teachers asked in each category was compared across conditions using *t* tests. The results of these analyses suggested that there were no significant differences between groups in terms of the types of questions or directions asked. Content questions were also coded in terms of level of complexity (as opinion, simple, or complex) and compared across conditions using *t* tests. The results of these analyses suggested that there were no significant differences in terms of the types of content questions asked between the experimental and comparison groups.

Discussion

It appears that the curricular design principles implemented in the experimental text significantly affected student content acquisition and engagement. These findings are especially important in light of the population used in this study. Students with learning disabilities, those with behavioral disorders, and students who are at risk for failure were able to access and use history knowledge.

The primary question was whether the type of history text used had an impact on students' knowledge of history. Overall, the results were positive. The most relevant measure was the test developed by the teachers from the curriculum with which the students were working. Teachers were instructed to select the items that best represented what they had covered—that is, to make a test that would demonstrate what their students had learned. The fact that experimental group students and comparison group students performed roughly equivalent on the section of the test drawn from the comparison text suggests that comparison group students did not learn much from their own text, even though they covered the same content. In fact, the scores achieved by both groups of students on the multiple-choice items from the comparison text (approximately 30%–40%) were not much higher than one might expect by chance. In contrast, the experimental group students' high performance on questions selected from the content of the experimental text suggests that they were indeed fluent on the material they had learned from their textbook.

Findings on the NAEP items were quite different. No significant differences were found between the two groups, and no significant growth was demonstrated from pretest to posttest. Even after significant modification of the test to improve its internal consistency reliability, no significant differences were found.

Progress-monitoring data clearly indicated growth over time for students in the experimental group. Because the vocabulary terms were drawn from the experimental history text, it is not surprising that the experimental group outperformed the comparison group. It is somewhat surprising, however, that the comparison group showed *no* improvement over time, given the reasonable amount of overlap in the content. This finding could indicate a complete lack of overlap in vocabulary terms between the texts, a lack of vocabulary acquisition, or some combination of both. This is an area that deserves more investigation. We remain convinced that the progress-monitoring measure is sensitive to growth, easy to administer, and reliable. Thus, it can serve a useful function in history instruction, especially for students with LD and behavior disorders.

The results of the classroom observation suggest that the teachers did use different types of activity structures across the two conditions. Teachers in the experimental condition had students work independently more frequently than teachers in the comparison condition, where students spent more time in teacher-led instruction. Given that the process–product literature (e.g., Brophy & Good, 1986) suggests that effective teachers engage in more teacher-led, interactive instruction and less frequent independent work, this finding is surprising. There are two possible explanations for this finding. First, it should be noted that the observations did not discriminate between lecture and interactive

teaching. It may be that the nature of the teacher-led instruction was different between conditions. This explanation is supported by the finding that students in the comparison group were significantly more off task during teacher-led instruction (experimental group, $M = .04$, $SD = .05$; comparison group, $M = .14$, $SD = .15$), $p = .041$, and the experimental group was significantly more actively engaged during independent work (experimental group, $M = .29$, $SD = .20$; comparison group, $M = .08$, $SD = .12$), $p = .003$. Thus, students may have been involved in qualitatively different activities during these time periods.

A possible explanation—supported by the senior author's informal observations—is that students in the experimental group were able to work actively and independently because they had been presented with information in a clear, structured fashion and then provided with review and practice activities that were based on sound principles of review (i.e., sufficient, distributed, cumulative, and varied). In contrast, the activities facing comparison group students were inappropriately distributed and often required them to supply information that had never been explicitly presented to them. Difficult tasks requiring students to use poorly organized information may result in students' frustration, failure to persist in the learning activity, and resultant off-task behavior. Engaging in off-task behavior would be a reasonable reaction for students, necessitating more teacher-guided practice.

We also asked whether students would answer questions more accurately during classroom instruction depending on the type of text they used. Students in the experimental group did answer significantly more questions correctly, whereas the comparison group answered more questions incorrectly. These findings suggest that the curricular design of texts affects the accuracy of students' responses to teacher questions. Students in the experimental group may have been able to answer questions accurately more frequently because of the organization of the content and the design of the instructional strategies in the text. Alternatively, students may have been able to answer questions more accurately because the types of questions asked by teachers were better designed and more clearly linked to what students had learned. A combination of these two factors may also explain the results.

Our findings also suggest that there were no differences between the conditions on oral reading fluency, although both groups did improve over time. This latter finding is encouraging, given that all of these students were low performing in general, and many were poor readers. On average, students in both groups improved approximately 20 words per minute over the course of the study. Although we did not investigate other types of instruction that students may have been receiving, it is likely that many were receiving supplemental reading instruction as part of their special education services.

We suspected that students' reading skills might correlate with their performance on the history measures. Interesting enough, this hypothesis proved true for the comparison group, which had highly significant correlations between all oral reading passages and both measures of history knowledge, but *not* for the experimental group, which had significant correlations between all oral reading passages and the NAEP, but not between any of the oral reading passages and the content acquisition test. It appears that the fluency with which an experimental student read was not related to his or her ability to perform on the content acquisition test. On the other hand, this was not true for experimental students on a measure like the NAEP, which was not closely related to the curriculum and the content covered. One explanation for this finding is that experimental group students were able to acquire history knowledge *despite* their poor reading fluency, because the design of the text facilitated their retention and recall. Gersten et al. (2006) found a somewhat similar relationship, in that the performance of comparison group students was predicted moderately well by scores on a prior knowledge test, whereas prior knowledge did not predict posttest performance for students taught with methods that explicitly pointed out relationships between actions and events.

Limitations

The findings from this study should be interpreted in the context of two threats to validity. First, concerns can be raised about the sample. The extent to which findings are generalizable to other groups depends on the degree to which the sample of students is representative of the larger population. There was unequal representation of special education students between the conditions. The comparison group included all the students with SED, whereas the experimental group included more students with LD. These differences may have affected findings on both the classroom observation measure and the history knowledge measures. Stratification by disability category would have ensured equivalent numbers of students by disability category across conditions. Small subsample sizes precluded secondary analyses of the various subsamples (LD, SED, no disabilities). Finally, we did not collect baseline measures of student behavior and, thus, do not know whether groups initially differed on the variables measured during the study.

Second, we investigated the effect of two different types of texts that were designed in different ways. Each text was made up of many different types of activities and represented an instructional package. The results of this study do not allow us to evaluate the relative effect of any *part* of an instructional package; we can only talk about effect of the entire package. Moreover, curriculum design can only be considered a factor affecting teacher behavior if it is used appropriately—that is, with

fidelity. In this study, implementation was allowed to vary naturally, was observed informally, and was described through teacher interviews. Formal observations and checklists would be other ways to increase confidence that the levels of the independent variable were implemented with fidelity. In fact, the low levels of asking complex questions by the two teachers in the experimental condition suggest that teachers may not have fully implemented the experimental curriculum.

Future Research

Future research should continue to investigate important curricular design principles and their effect on teachers' behavior and students' behavior and achievement. Specifically, researchers should

- continue to evaluate the relative effect of different curricular design principles for special education students, using larger samples and larger numbers of teachers;
- more carefully investigate classroom implementation;
- evaluate the degree to which the types of curricular tools provided to teachers affect teaching;
- consider the use of observational data as potential mediator or moderator variables; and
- investigate alternative, sensitive content area assessments.

ABOUT THE AUTHORS

Mark K. Harniss, *PhD, is an assistant clinical professor in rehabilitation medicine at the University of Washington. His research interests include instructional and assistive technology, instructional design, and effective instruction for students with disabilities.* ***Jennifer Caros****, PhD, is a school psychologist at Vancouver School District and conducts longitudinal program evaluation research in urban public schools in the United States and Canada.* ***Russell Gersten,*** *PhD, is executive director of the Instructional Research Group, a nonprofit educational research institute, as well as professor emeritus in the College of Education at the University of Oregon. His main areas of expertise include instructional research on English Language Learners, reading comprehension research and evaluation methodology. Address: Mark K. Harniss, Box 357920, University of Washington, Seattle, WA 98195.*

AUTHORS' NOTES

1. *Research funding was provided by the United States Department of Education, Office of Special Education Programs, HO23B30019. The opinions expressed in this article do not necessarily reflect the position of the U.S. Department of Education.*
2. *Sincere thanks are due Doug Carnine, who guided the design and development of the experimental text, and helped conceptualize and implement the study, and to Betsy Davis, who guided and conducted the statistical analysis.*

REFERENCES

Andre, T. (1979). Does answering higher-level questions while reading facilitate productive learning? *Review of Educational Research , 49*(2), 280–318.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform. *Educational Researcher, 25,* 6–8.

Beck, I. L., McKeown, M. G., & Gromoll, E. W. (1989). Learning from social studies texts. *Interchange, 17,* 10–19.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Carnine, D. W., Caros, J., Crawford, D., Hollenbeck, K., & Harniss, M. K. (1996). Designing effective United States history curricula for all students. In J. Brophy (Ed.), *Advances in research on teaching, Vol. 6. History teaching and learning* (pp. 207–256). Greenwich, CT: JAI Press.

Carnine, D., Crawford, D. B., Harniss, M. K., & Hollenbeck, K. L. (1994). *Understanding U.S. history, Vol. 1. Through the Civil War.* Eugene, OR: Considerate.

Carnine, D., Miller, S., Bean, R., & Zigmond, N. (1994). Social studies: Educational tools for diverse learners. *School Psychology Review, 23,* 428–441.

Caros, J. (1996). *Content area assessment: Technical adequacy of standardized, short-duration, alternate-form, curriculum-based measures of content area vocabulary for monitoring student progress with eighth-graders.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Crawford, C., Carnine, D. W., Harniss, M. K., Hollenbeck, K., & Miller, S. (2007). Effective strategies for teaching social studies. In E. Kame'enui & D. Carnine (Eds.), *Effective strategies that accommodate diverse learners* (3rd ed., pp. 203–230). Columbus, OH: Merrill.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37,* 184–192.

Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measurement in the content areas: Validity of vocabulary-matching as an indicator of performance in social studies. *Learning Disabilities Research & Practice, 16,* 142–151.

Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62,* 497–514.

Garraty, J. A. (1991). *The story of America.* Orlando, FL: Harcourt Brace Jovanovich.

Gersten, R., Baker, S., Smith-Johnson, J., Dimino, J., & Peterson, A. (2006). Eyes on the prize: Teaching complex historical content to middle school students with learning disabilities. *Exceptional Children, 72,* 264–280.

Harniss, M. K. (1996). *Task requirements of content area textbooks: Effects on the academic achievement and engagement of middle-level students.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Harniss, M. K., Dickson, S. V., Kinder, D., & Hollenbeck, K. L. (2001). Textual problems and instructional solutions: Strategies for enhancing learning from published history textbooks. *Reading and Writing Quarterly, 17,* 127–150.

Harniss, M. K., Hollenbeck, K., & Dickson, S. V. (2004). Direct instruction in the content areas. In N. E. Marchand-Martella, T. A. Slocum, & R. C. Martella (Eds.), *Introduction to direct instruction* (pp. 246–279). Boston: Pearson/Allyn & Bacon.

Harniss, M. K., Hollenbeck, K. H., Crawford, D. B., & Carnine, D. (1994). Content organization and instructional design issues in the development of history texts. *Learning Disability Quarterly, 17,* 235–248.

McKeown, M. G., & Beck, I. L. (1994). Making sense of accounts of history: Why young students don't and how they might. In I. Beck & C. Stainton (Eds.), *Teaching and learning in history* (pp. 1–26). Hillsdale, NJ: Erlbaum.

Murray, D. M. (1998) *Design and analysis of group randomized trials.* New York: Oxford University Press.

Peckham, P. D., Glass, G. V., & Hopkins, H. D. (1969). The experimental unit in statistical analysis. *The Journal of Special Education, 3,* 337–349.

U.S. Department of Education. (2006a). *Math now: Advancing math education in elementary and middle school.* Washington, DC: Author. Retrieved July 29, 2006, from http://www.ed.gov/about/inits/ed/competitiveness/math-now.html

U.S. Department of Education. (2006b). *Reading first.* Washington, DC: Author. Retrieved July 29, 2006, from http://www.ed.gov/programs/readingfirst/index.html