CRESST REPORT 788

David Silver
Mark Hansen
Joan Herman
Yael Silk
Cynthia L. Greenleaf

IES INTEGRATED LEARNING ASSESSMENT FINAL REPORT

MARCH, 2011

The National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Sciences
UCLA | University of California, Los Angeles

IES Integrated Learning Assessment Final Report

CRESST Report 788

David Silver, Mark Hansen, Joan Herman, and Yael Silk
CRESST/University of California, Los Angeles

Cynthia L. Greenleaf
WestEd

March, 2011

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under a subcontract (S05-059) to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) funded by the US Depart of Education, grant number R305M050031 and administered by the WestEd Agency.

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the US Dept of Education or the WestEd Agency.

To cite from this report, please use the following as your APA reference: Silver, D., Hansen, M., Herman, J., Silk, Y., & Greenleaf, C.L. (2011). *IES integrated learning assessment final report.* (CRESST Report 788). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

TABLE OF CONTENTS

IES INTEGRATED LEARNING ASSESSMENT FINAL REPORT

David Silver, Mark Hansen, Joan Herman, Yael Silk
CRESST/University of California, Los Angeles

Cynthia L. Greenleaf
WestEd

Abstract

The main purpose of this study was to examine the effects of the Reading Apprenticeship professional development program on several teacher and student outcomes, including effects on student learning. A key part of the study was the use of an enhanced performance assessment program, the Integrated Learning Assessment (ILA), to measure student content understanding. The ILA instruments included multiple components that assessed student content knowledge, reading comprehension, metacognition, use of reading strategies, and writing skills in applied knowledge. An analysis of student scores using the ILA found little or no significant effects from the Reading Apprenticeship program on class-level student outcomes. However, the researchers found a significant positive effect on teachers' literacy instruction.

Introduction

Project Background

The major aim of this study was to examine the effects of assignment to the Reading Apprenticeship (RA) professional development program on several teacher and student outcome variables. In other words, the study sought to evaluate the effects of the RA professional development program on teacher practices and student learning. Biology and history high school teachers were recruited for the study and the effects on students' literacy within the subject area were of particular interest.

Large scale assessments, such as the California Standards Test (CST) and Arizona's Instrument to Measure Standards (AIMS), may honor breadth over depth of student knowledge and comprehension; thus, we developed a supplementary, more detailed measure in order to examine the potential effects of the RA on student learning. This new performance-based assessment is called the Integrated Learning Assessment (ILA). The ILA integrates adaptations of CRESST's model-based assessments for measuring content understanding (Baker, Freeman, & Clayton, 1991; Chung, Harmon, & Baker, 2001; Herman, Baker, & Linn, 2004; Baker, Aschbacher, Niemi, & Sato, 1992) with WestEd's Strategic Literacy Initiative's Curriculum Embedded Reading Assessment (CERA). As a starting point, we defined acquisition of conceptual understanding in biology and history to include the mastery of particular concepts and ideas through engagement with texts as well as the ability to effectively integrate concepts into the formulation of explanations. The purpose of this report was to provide information about the development process and the preliminary results of the ILA, including the reliability and distribution of scores for our study.

Why Develop the ILA?

The ability to read and write in a discipline-specific context is increasingly recognized as a critical skill for high performance in an academic setting. Despite widespread acknowledgement that disciplines such as science and history demand multi-faceted literacy skills, few validated instruments have been developed that evaluate the impact of discipline-specific literacy instruction on student outcomes. The ILA was designed to evaluate both discipline-specific content knowledge and literacy skills integral to the successful access and display of content knowledge. As an evaluation tool for skills and knowledge, the ILA was designed to measure how students use RA-guided interactive reading skills, as well as how these skills influence student achievement. Specifically, the ILA was developed as a measure to examine the extent to which students utilize cognitive and meta-cognitive skills considered essential for substantial engagement with scientific and historical texts.

In the RA instructional model, students are frequently exposed to texts in both primary and secondary source types (e.g., textbooks, web resources, articles, data tables); furthermore, students are regularly provided with a variety of comprehension strategies to help access the unique content in these materials. With the use of RA reading comprehension strategies, students are increasingly prompted and expected to *explain* their understandings or questions through elaborated and detailed descriptions. Over time, as students attempt to discuss and explain increasingly complex concepts, they begin to need more academic language in order to communicate these ideas effectively.

Rationale for the ILA Format

National standards for history and science education emphasize the importance of providing students with opportunities to present their understanding as well as use knowledge and academic language to communicate explanations and ideas (see National Research Council, 1996). Due to the close relationship between reading and listening (input) and writing and speaking (response) in discipline-specific literacy as well as in the RA instructional model, we included both reading and writing in the ILA to measure the effectiveness of RA instruction on knowledge and literacy acquisition.

While the written explanation genre was not an explicit component of RA, it is expected that students in RA classrooms would have sufficient experience with this genre through their utilization of a variety of writing formats. Moreover, through their exposure to scientific and historical texts, these students are expected to have developed sufficient familiarity with academic language to have high functionality on this measure.

CRESST's model-based assessment uses standard architectures embedded in disciplinary content to assess core types of learning—basic knowledge, conceptual understanding, problem solving, communication, and teamwork. Two different templates have been developed and extensively validated as a way to evaluate content understanding and communication. One requires students to generate written explanations given primary source materials; the other utilizes computer-based knowledge mapping to display comprehension (see Baker, 1994; Chung, O'Neil, & Herl, 1999). Given the current study's focus on the integration of discipline-specific literacy and content knowledge, CRESST's explanation architecture provided us with the ideal format to evaluate this array of skills simultaneously. The use of an explanation and argumentation task with reading prompts and a writing component provided us with an integrated approach to measuring students' ability to understand and communicate their knowledge. In addition, explanation tasks are a dominant genre of school-based writing (Martin & Miller, 1988) and are well-suited for on-demand assessment conditions.

The explanation and argumentation architectures may have particular relevance for the assessment of scientific literacy— given that science is about the construction of theories that *explain* how the world operates. Scholars have noted that discourse, explanation, and argumentation are at the heart of science learning (Boulter & Gilbert, 1995; Duschl & Osborne, 2002; Erduran, Simon, & Osborne, 2004; Pontecorvo, 1987). The explanation architecture provides a format that examines whether students are able to integrate a complex structure of biological as well as other related concepts, the relationships between these concepts, the reasons for these relationships, and ways to explain and predict other natural phenomena.

Separate ILA instruments were developed for administration to biology and history students. Copies of these instruments are provided in Appendices A and B. In the following sections of this report, we describe the design of the measures for the respective subject areas. Specifically, we present the particular content focus of each ILA, review test specifications, and describe the process of item generation.

Development of the Biology ILA

Content Focus Selection

After reviewing California state content standards for biology and life sciences, we created an ontology—a systematic arrangement and categorization of concepts in a field of discourse. Developing this type of organizational structure allowed us to uncover the relationships between different biology concepts (e.g., what concepts encompass the precursor knowledge-set needed to understand a specific standard). This involved unpacking and elaborating the standards to create a hierarchy of conceptual information. This hierarchy of information was then used to: (a) create a framework for content understanding; (b) shape the design of the ILA; and (c) guide the development of the content rubric. Using the California Science Teachers Association's *Making Connections: A Guide to Implementing Science Standards* (Bertrand, DiRanna, & Janulaw, 1999) as a guide, we examined the standards in two specific science content areas—genetics and physiology. Based on preceding units, content standards and sub-standards, as well as science standards from earlier grade levels (e.g., California Science Standards for Grade 7), we determined that prior content knowledge that would be necessary to understand target standards.

The section of biology content targeted for the ILA was the unit in genetics, which is well represented in the California Standards Test (CST) for biology. We chose the topic of genetics for the ILA because we expected teachers to spend more instructional time covering this content area than other units—given its emphasis on the CST. Overall, the review of content standards and the development of a biology ontology provided us with the context for determining the specific content targeted in the ILA.

Test Specification

The first phase in developing the test specification for each ILA was to provide a detailed description of what was to be tested. Based on a review of the standards in the two subject areas, it was determined that the ILA should incorporate high-level cognitive skills such as analysis, interpretation, evaluation, and synthesis of the information presented in the ILA texts—combined with the material learned in class. The tasks in the ILA were aimed at eliciting students' use of higher-level cognitive skills when engaged in reading, analyzing, evaluating, and synthesizing the documents through writing. Figures 1 through 3 show the target standards for content knowledge, reading, and writing.

BIOLOGY/LIFE SCIENCES STANDARDS: GENETICS

Standard 5: The genetic composition of cells can be altered by incorporation of exogenous DNA into the cells.

Sample basis for understanding this concept:
5a. Students know the general structures and functions of DNA, RNA, and protein.
5b. Students know how genetic engineering (biotechnology) is used to produce novel biomedical and agricultural products.

Figure 1. Biology/life sciences content standards targeted in the ILA.

READING COMPREHENSION STANDARDS

Standard 2.0: Read and understand grade-level-appropriate material. Analyze the organizational patterns, arguments, and positions advanced.

Structural Features of Informational Materials:
Standard 2.5: Extend ideas presented in primary or secondary sources through original analysis, evaluation, and elaboration.

Figure 2. Reading comprehension standards targeted in ILA.

WRITING STANDARDS

WRITING STRATEGIES:
Standard 1.0: Write coherent and focused essays that convey a well-defined perspective and tightly reasoned argument. The writing demonstrates students' awareness of the audience and purpose.
Standard 1.1: Establish a controlling impression or coherent thesis that conveys a clear and distinctive perspective on the subject and maintain a consistent tone and focus throughout the piece of writing.
Standard 1.2: Use precise language, action verbs, sensory details, appropriate modifiers, and the active rather than the passive voice.

WRITING APPLICATIONS:
Standard 2.0: Combine the rhetorical strategies of narration, exposition, persuasion, and description to produce texts of at least 1,500 words each. Student writing demonstrates a command of standard American English and the research, organizational, and drafting strategies.
Standard 2.3: Write expository compositions, including analytical essays and research reports
a. Marshal evidence in support of a thesis and related claims, including information on all relevant perspectives.
b. Convey information and ideas from primary and secondary sources accurately and coherently.
c. Anticipate and address readers' potential misunderstandings, biases, and expectations.
d. Use technical terms and notations accurately.

WRITTEN AND ORAL ENGLISH LANGUAGE CONVENTIONS:
Standard 1.0: Write and speak with a command of standard English conventions.

Grammar and Mechanics of Writing
Standard 1.1: Identify and correctly use clauses (e.g., main and subordinate), phrases (e.g., gerund, infinitive, and participial), and mechanics of punctuation (e.g., semicolons, colons, ellipses, hyphens).
Standard 1.2: Understand sentence construction (e.g., parallel structure, subordination, proper placement of modifiers) and proper English usage (e.g., consistency of verb tenses).
Standard 1.3: Demonstrate an understanding of proper English usage and control of grammar, paragraph and sentence structure, diction, and syntax.

Figure 3. Writing standards targeted in the ILA.

Text Selection

To inform our text passage selection for the Biology ILA, we examined what linguistic resources are used to create scientific meaning and the level of reading comprehension proficiency that is required at the high school level. To gain this understanding, we conducted a linguistic analysis of high school biology text books—Prentice Hall's *Biology* (Miller & Levine, 2006) and BSCS's *BSCS Biology: A Molecular Approach* (Greenberg, 2001). The results of the analysis were used as a basis for selecting and modifying the texts used in the ILA.

Overall, we found that high school science textbooks displayed high technicality and abstractness. This was evidenced by frequent occurrences of technical vocabulary and abstract nouns. In addition, various instances of "grammatical metaphor" (see Halliday, 1994) were identified in biology textbooks.[1] For example, experiential information (i.e., what is happening in the text) was frequently expressed in nominal groups through nominalization (e.g., forming the noun "invasion" from the verb "invade"). These nominal groups were further expanded through the addition of an embedded clause, an adjective, or a prepositional phrase, which resulted in high lexical density. Relationships between experiential elements were marked through various connectors including conjunctions but verbal groups often subsumed the marking of conjunctive relationships (e.g., "to be followed by" instead of "and then").

This comparative analysis helped us select the final passage to include in the ILA. The passage was similar to textbook passages in terms of linguistic difficulty. The final text passage was selected from an internet site[2] listed as a supplemental resource for students in state-adopted biology textbooks (Miller & Levine, 2006).

After the text passage selection, we conducted an external review of text passages. This step involved consulting with genetic scientists at the University of California, Los Angeles (UCLA) for content accuracy, as well as communicating with the authors of the text passages to obtain permission for use and to confirm that the content of the passage could still be considered current and accurate. The text passages were also reviewed and rated by current high school biology teachers for level of difficulty and the content's appropriateness for the study sample.

[1] A grammatical metaphor is a process whereby meaning is constructed in the form of incongruent (i.e., metaphorical) grammar. Incongruence is characteristic of written discourse in relatively formal registers.

[2] http://sciencenow.sciencemag.org/cgi/content/full/2006/113/2

Item Generation

The generation of items for the ILA (i.e., writing prompts, reading comprehension, and metacognitive questions) followed the document selection and was a multi-step process. After four of the five text selections, we included a reading comprehension section as a way to determine whether the quality of students' written responses was influenced by their reading comprehension levels. The multiple-choice questions were developed with three categories of reading comprehension in mind: factual, inferential, thematic and scientific. After the generation of various reading comprehension questions for each text in the text set, two to three questions were selected per document section.

The development of the ILA also involved creating two candidate essay prompts based on the content of two text sets and on the biology curriculum—with the requirement that they elicit higher-order thinking skills. Both prompts required students to synthesize information in the reading with prior knowledge. The first prompt was limited to an explanation task while the second prompt required students to both explain a biological process and develop an argument for one process over another. We used the original essay prompt from an existing Biology ILA developed for a National Science Foundation (NSF) study and then created a new prompt to serve as the comparison.

In the late fall of the 2008-2009 school year, the two ILA prototypes were field tested with biology teachers to verify the appropriateness of the texts; reading comprehension and metacognition questions; and writing prompts. The results from this process indicated that students of varying competency levels would most likely be able to respond to the various sections of the ILA. We reviewed the student responses and determined that students were able to address the more difficult prompt. The second prompt was selected since it met the criteria of requiring students to engage in higher order thinking processes.

Development of the History ILA

Content Focus Selection

Based on a survey of several RA history teachers, we found that World War II was an important content area that would be covered in the spring. Within this broad California state standard (11.7), we reviewed the eight sub-standards and identified one (11.7.5) as the ILA target content standard. We chose this standard for several reasons: First, this standard deals with social history and women's history, which are both commonly addressed on document-based questions (DBQs), since textbooks (especially older ones) often focus more extensively on political and military history.[3] Targeting social history on DBQs exposes students to a wider range of historical issues than those usually included in textbooks. Second, the teacher-identified standard textbook, McDougal Littell's *The Americans*, includes enough coverage of these topics for students to include as prior knowledge. Finally, the textbook includes multiple primary sources for the 11.7.5 standard, exposing students to the types of document genres used in the ILA.

The California standard 11.7.5 states:

> Discuss the constitutional issues and impact of events on the U.S. home front, including the internment of Japanese-Americans (e.g., *Fred Korematsu v. United States of America*) and the restrictions on German and Italian resident aliens; the response of the administration to Hitler's atrocities against Jews and other groups; the roles of women in military production; and the roles and growing political demands of African Americans.

Since the study includes Arizona teachers, we also took the Arizona state standards into consideration. Arizona state standard 1SS-P15, PO 2 states that instruction on World War II should emphasize:

> Events on the home front to support the war effort (including war bond drives, the mobilization of the war industry); women and minorities in the work force (including Rosie the Riveter); [and] the internment of Japanese-Americans (including the camps in Poston and on the Gila River Indian Reservation, Arizona).

Both the California and Arizona standards still cover a broad content area; hence, we narrowed the ILA target to women and African-Americans on the home front. Recognizing that the aim of the ILA is to provide students with opportunities to demonstrate disciplinary thinking and reading comprehension skills in an area of instruction of which they have had some (but not extensive) exposure, we chose this sub-point for several reasons. First, we ruled out addressing Japanese-American internment since this particular topic is traditionally heavily covered in both California and Arizona classrooms. The California and Arizona standards include specific details concerning internment; furthermore, the standard textbook includes a breakout section on the *Korematsu* case. Therefore, students would already have received significant instruction on this content topic. Similarly, it would be difficult to find text prompts containing information that would be entirely new to students.

Next, the sub-point of German and Italian resident aliens was eliminated from consideration in view of the fact that this topic is only briefly covered in the standard

[3] Stovel, J.E. (2000). Document analysis as a tool to strengthen student writing, *The History Teacher 33* (4),: 501-509.

textbook and students would likely not have enough prior knowledge to apply to the essay. Thus, the final decision to focus on African-Americans was made because students would not have previously spent a significant amount of class time developing every relevant theme related to the topic; thus, they would still have enough prior knowledge to potentially use in their essays. Additionally, there was a large pool of documents for this topic from which we could confidently select text prompts that fit ILA specifications.

Test Specification

The first phase in developing the test specification for the ILA was to provide a detailed description of what skills were to be tested. Based on a review of the standards, it was determined that the ILA should incorporate high-level cognitive skills such as analysis, interpretation, evaluation, and synthesis of the information presented in the ILA primary source documents combined with the material learned in history class. The tasks in the ILA were aimed at eliciting students' use of higher-level cognitive skills when engaging in reading, analyzing, evaluating, and synthesizing the documents through writing. Figures 4 through 7 depict the target standards related to content, analysis, reading, and writing.

HISTORY/SOCIAL SCIENCE STANDARDS

Standard 11.7: Students analyze America's participation in World War II.

5. Discuss the constitutional issues and impact of events on the U.S. home front, including the internment of Japanese Americans (e.g., *Fred Korematsu vs. United States of America)* and the restrictions on German and Italian resident aliens; the response of the administration to Hitler's atrocities against Jews and other groups; the roles of women in military production; and the roles and growing political demands of African Americans.

Figure 4. History/social science standards targeted in the ILA.

HISTORICAL AND SOCIAL SCIENCE ANALYSIS SKILLS STANDARDS

Historical Research, Evidence, and Point of View
2. Students identify bias and prejudice in historical interpretations.
4. Students construct and test hypotheses; collect, evaluate, and employ information from multiple primary and secondary sources; and apply it in oral and written presentations.

Historical Interpretation
1. Students show the connections, causal and otherwise, between particular historical events and larger social, economic, and political trends and developments.
3. Students interpret past events and issues within the context in which an event unfolded rather than solely in terms of present-day norms and values.
4. Students understand the meaning, implication, and impact of historical events and recognize that events could have taken other directions.

Figure 5. Historical and social sciences analysis skills standards targeted in the ILA.

READING COMPREHENSION STANDARDS

2.0 Read and understand grade-level-appropriate material. Analyze the organizational patterns, arguments, and positions advanced.

Structural Features of Informational Materials
2.1 Analyze both the features and the rhetorical devices of different types of public documents (e.g., policy statements, speeches, debates, platforms) and the way in which authors use those features and devices.

Comprehension and Analysis of Grade-Level-Appropriate Text
2.4 Make warranted and reasonable assertions about the author's arguments by using elements of the text to defend and clarify interpretations.
2.5 Analyze an author's implicit and explicit philosophical assumptions and beliefs about a subject.

Figure 6. Reading comprehension standards targeted in the ILA.

WRITING STANDARDS

Writing Strategies:
1.0 Write coherent and focused essays that convey a well-defined perspective and tightly reasoned argument. The writing demonstrates students' awareness of the audience and purpose.

Organization and Focus
1.3 Structure ideas and arguments in a sustained, persuasive, and sophisticated way and support them with precise and relevant examples.

Writing Applications:
2.0 Combine the rhetorical strategies of narration, exposition, persuasion, and description to produce texts of at least 1,500 words each. Student writing demonstrates a command of standard American English and the research, organizational, and drafting strategies outlined in Writing Standard 1.0.
2.4 Write historical investigation reports:
a. Use exposition, narration, description, argumentation, or some combination of rhetorical strategies to support the main proposition.
b. Analyze several historical records of a single event, examining critical relationships between elements of the research topic.
c. Explain the perceived reason or reasons for the similarities and differences in historical records with information derived from primary and secondary sources to support or enhance the presentation.
d. Include information from all relevant perspectives and take into consideration the validity and reliability of sources.

Written and Oral English Language Conventions:
1.0 Write and speak with a command of standard English conventions.
1.1 Demonstrate control of grammar, diction, and paragraph and sentence structure and an understanding of English usage.
1.2 Produce legible work that shows accurate spelling and correct punctuation and capitalization.

Figure 7. Writing standards targeted in the ILA.

Text Selection

Documents were chosen for the History ILA based on several factors. First, the language, images, or data had to be presented in a clear and accessible way that met a grade 11 high school reading level. Second, the documents needed to be directly related to the target history content standard and to the essay prompt. Third, the documents had to point to larger themes embedded in both the content standard and essay prompt that students should develop in their essays.

From a review of relevant literature, which largely focused on the Advanced Placement (AP) U.S. History Exam and the New York (NY) State U.S. History and Government Regents Exam, we determined that including a combination of written and visual texts constitutes a standard Document-Based Question (DBQ) writing practice. To demonstrate their disciplinary thinking skills, students should be able to read, understand, and analyze a wide variety of historical genres—both written and visual. Since the ILA target content standard focuses on social history, relevant documents were chosen to connect to students' prior knowledge of the social aspects and effects of WWII on African-Americans on the home front.

In a review of their language aspects, we generally found that high school history textbooks displayed high technicality and abstractness. This was evidenced by the frequent occurrences of historically specific vocabulary and the use of abstract nouns. In addition, various instances of grammatical metaphor (Halliday, 1994) were identified in history textbooks[4]. For example, experiential information (i.e., what is happening in the text) was frequently expressed in nominal groups through nominalization (e.g., forming the noun "migration" from the verb "migrate"). These nominal groups were further expanded through the addition of an embedded clause, an adjective, or a prepositional phrase, which resulted in high lexical density. Relationships between experiential elements were indicated through various connectors including conjunctions but verbal groups often subsumed the use of conjunctive relationships (e.g., "to be followed by" instead of "and then").

Using criteria developed as part of the RA instructional model, we also looked at potential text passages more holistically for consideration. This selection criterion was conveyed to teachers during their training to help them select appropriate text for classroom use. We utilized the following criteria to select text:

- contains illustrations or graphics;
- has internal coherence;
- identifies a scientist/team or history authority;
- explains the inquiry (use of evidence);

[4] A *grammatical metaphor* is when one grammatical structure is substituted for another, such as with nominalization (i.e., when a verb is used in the form of a noun). This is characteristic of written discourse in relatively formal registers.

- contains technical vocabulary;
- is exposition instead of narrative;
- has data for students to interpret; and
- has description of methodology.

After the initial selection, we conducted an external review of the visual and written texts. Current high school history teachers holistically rated and reviewed the documents for level of difficulty and applicability to the essay prompt (see Appendix C for the teacher feedback survey).

In its entirety, the final document set—composed of three primary sources (i.e., newspaper article, letter, and population data table) and one secondary source (i.e., excerpt from a historical journal article)—presented multiple aspects of the content standard topic that allowed students to make generalizations, analyze cause and effect, discuss contrasting viewpoints, and evaluate the historical impact of the content standard topic. The document set included documents and readings that students would most likely not have seen; moreover, it might have introduced specific historical information that students had not discussed in their classes. Students should have applied their disciplinary thinking skills to analyze and interpret new information in the documents in order to integrate this data with their related prior knowledge and to construct an evidence-based historical narrative.

Item Generation

The generation of items for the ILA (i.e., writing prompts, measures of reading comprehension, and metacognitive questions) followed document selection and was a multi-step process. We included a reading comprehension section after each text in the ILA as a way to determine whether the quality of students' written responses was influenced by their reading comprehension levels. The multiple-choice questions were developed with three categories of reading comprehension in mind: factual, inferential, and thematic and historical. Many reading comprehension questions were generated for each text. From these candidate items, three were selected to be included after each of the texts in the ILA.

The development of the ILA also involved creating two candidate essay prompts based on the content of two text sets as well as the history curriculum—with the requirement that they elicit higher-order thinking skills. The essay prompt requires students to synthesize information in multiple documents with prior knowledge as well as elicit more disciplinary-specific skills through documentary analysis of change over time and historical cause and effect. We first evaluated DBQ questions from the past nine years of AP U.S. History exams and the past six years of NY State Regents exams; this added up to a total of 15 AP questions and 16 Regents questions. AP questions routinely direct students to “analyze” historical change; “assess the effectiveness” of policies, reforms, etc.; assess the “extent” of historical change; or evaluate the “accuracy” of historical interpretations. Conversely, the NY Regents exams overwhelmingly ask students to “discuss” historical issues or changes. While both the AP and the Regents DBQs are challenging tasks that require higher-order thinking, the AP is expectedly more difficult. Therefore, we used the NY Regents exam as a model for developing the essay prompts.

One potential ILA topic focused on African-Americans on the home front during WWII, while the other centered on American women during the same time period. Our desire was to create prompts related to the documents that could be adequately answered by utilizing content learned that year in U.S. History class, together with information directly gathered from the documents. In the fall of the 2007-2008 school year, the two ILA prototypes were field tested with several history teachers in the Los Angeles area to verify the appropriateness of the texts, reading comprehension and metacognition questions, and writing prompts. The results from this process indicated that students of varying competency levels would most likely be able to respond to the various sections of the ILA. We reviewed the student responses and determined that the African-American ILA would elicit the best student responses, since students had more prior knowledge to apply and seemed to demonstrate greater understanding of the texts.

Structure of the ILA

Overview

The ILA instruments for biology and history (provided in Appendices A and B, respectively) each consisted of three parts. The first was an assessment of students’ knowledge of the subject matter. The second part presented students with a series of documents (e.g., narrative texts, graphs, illustrations, data tables). The goals of this section were to examine students’ reading comprehension, metacognition, and use of reading strategies. Reading comprehension was measured by multiple choice questions that could be answered using information presented in the texts. Metacognition was assessed by asking students to describe their reading process. The use of reading strategies was evaluated by reviewing students’ test forms for evidence of note-taking or other annotations. In the third part of the ILA, students were asked to write an essay that drew upon information garnered from the texts as well as their prior content knowledge and skills. These writing samples were rated with respect to both language and content. Additional details concerning the structure and scoring of each section of the ILA are described next.

Part 1: Content Knowledge

Prior to reading the text passages, students completed a short test consisting of 10 multiple choice questions intended to measure students' existing content knowledge. For the Biology ILA, these items were selected from the CST Biology test, the SAT II exam, the AP Biology exam, and preparation resources for these tests. The History ILA consisted of 10 multiple choice questions relating to African-American history of the late-nineteenth to mid-twentieth centuries(particularly to African-American involvement during WWII) and more generally to WWII social history. The items were selected from a pool of publicly released CST History items, AP U.S. History items, N.Y. Regents U.S. History items, and related test preparation resources. The questions in this first section of the ILA drew upon students' knowledge of the particular subject areas and were administered to aid the interpretation of scores on the passage-based multiple choice questions in the subsequent section.

Part 2: Reading Comprehension, Metacognition, and Reading Strategies

In the second part of the ILA, students were asked to read a series of passages, answer multiple choice questions related to those passages, and reflect on their reading process. In addition, students' test booklets were examined for evidence of their utilization of reading strategies.

Reading Comprehension. The multiple choice questions in Part 2 of the ILA were intended to measure students' reading comprehension. Questions were aligned with the passages in such a way that that it would have been possible for students to find relevant information within the passage and provide a correct response—regardless of their prior knowledge of the subject-matter. However, due to the fact that the questions still draw on content knowledge, it should be noted that students could perhaps provide correct answers by relying primarily on their prior knowledge—not only on the particular knowledge gained by reading and comprehending the text at hand. In other words, a student might compensate for low reading comprehension with high prior knowledge or compensate for low prior knowledge with high comprehension. As such, scores on these items are best interpreted alongside students' content knowledge scores from Section I of the ILA. This point will be further addressed in our analysis and discussion of the data.

Metacognition Scoring Rubric. After completing the multiple choice questions, students were encouraged to reflect on their thought process and describe how they approached the reading passages. This metacognition item was designed by WestEd with input from CRESST. The question was designed to measure the degree to which students were aware of the thought processes they had utilized in reading the documents. Students were asked to respond to the following question:

> Parts of this document were complex. What did you do as you were reading to improve your understanding? Please be as detailed as possible.

The metacognition scoring rubric (see Appendix D) was adapted from previous RA work. Students' responses were rated on a 4-point scale. The profile of a score point could be broken down into three main criteria: the degrees to which the student (a) engages with complexities in the text or with the ideas that require attention; (b) describes thinking processes that occur while reading; and (c) explains an approach to how he or she thinks about the reading. Additionally, raters considered how aware students were of their thinking, their degree of self-monitoring, and lastly, their executive control.

Reading Strategies Scoring Rubric. In developing the reading strategies rubric (see Appendix E), we modified the NSF Reading Process rubric that was based on the Strategic Literacy Initiative's CERA assessment. In particular, we extended their work to produce a rubric geared towards use in large-scale scoring sessions. The key points of the rubric address students' reading engagement, based on the reading strategy dimensions identified by the RA approach to content area reading. The rubric was applied to annotations made on the texts presented in Part 2 of the ILA.

The Reading Strategies rubric was based on a 4-point scale. The profile of a score point could be broken down into three main criteria: consideration of the frequency of annotations, the variety in the annotations, and the types of reading strategies used (i.e., general versus discipline-specific). The strategies assessed were drawn from the RA theory of content area reading. Table 1 provides additional information about the evidence that raters looked for while rating as well as the types of reading strategies utilized by students.

Table 1

Descriptions of Annotations and Reading Strategies

Text annotations	General reading strategies	Biology reading strategies	History reading strategies
• Markings • Underlines • Highlights • Circlings/boxings • Connecting lines and arrows • Symbols • Comments • Questions • Statements	• Identifying key vocabulary • Identifying unknown vocabulary • Attempting to define unknown vocabulary (e.g., through identifying root words, looking ahead in the text for a definition) • Identifying the main ideas of the text • Paraphrasing • Summarizing • Predicting the content of text sections • Identifying confusions • Using context clues to build understanding	• Connecting to/applying prior biology knowledge • Questioning scientific methods • Attending to and evaluating evidence • Analyzing graphs, diagrams and other visual aids, including organizing/representing data • Considering the implications of science beyond the text's scope	• Making connections to prior history knowledge • Linking ideas together within a document and/or across documents (intertextual reading) • Evaluating the source of a document • Determining bias or point of view • Considering the document in historical context • Identifying cause and effect

Note. Evidence for text annotations found only on text passage.

A student who received a score of 4, for example, would have displayed a strong use of reading strategies demonstrated through annotations throughout the set of texts; employed a variety of annotations; and shown evidence of using at least one discipline-specific reading strategy. In contrast, a student receiving a score of 1 would have shown little or no evidence of the use of reading strategies. In this case, the annotations may have been minimal, disconnected, or indiscriminate (e.g., large sections of the passage highlighted or underlined lacking an apparent purpose).

Part 3: Writing Assessment

Parts 1 and 2 of the ILA were administered together. During the following ILA administration, half of the students moved onto Part 3; whereas, the other half completed a different assessment called the Degrees of Reading Power (DRP). Part 3 of the ILA is a writing task that directs students to write an essay integrating information from the documents with knowledge they have learned in their biology or history class. For the biology test, in order to help students approach the task as one of scientific explanation and argumentation, students were instructed to imagine that they were biologists advising a farmer about preventing crop destruction. Students were specifically directed to include an explanation of the recombinant DNA process, a description of the safety concerns this process presents, and an argument supporting either traditional cross-breeding or genetic engineering. For the history test, in order to help them approach the task as one of historical explanation, students were instructed to imagine that they were journalists writing about African-Americans' experiences on the home front during WWII. Students were specifically directed to include discussions of labor discrimination, migration, and racial violence; develop larger themes; and provide analysis in their essays.

Writing Rubrics. The scoring rubrics for Part 3 of the ILA (see Appendices F and G) address issues of language and academic writing within the science or history genre. We adapted previously developed and validated rubrics (from NSF Biology ILA scoring), which evaluated student content and language knowledge along two separate dimensions. Our language rubric followed a linguistic analysis of academic language and writing practices and also reflected grade 11 English language arts standards. Both the writing content and writing language rubrics utilize a 4-point rating scale. Through the characterization of their respective score points, both rubrics describe various aspects of writing proficiency. Each score point within a given rubric provides a portrait of students' explanations as they may appear at a given proficiency level.

Rationale for Two Writing Rubric Dimensions. Our evaluation of commonly used performance assessments revealed that language expectations are often implicitly embedded within the assessment criteria. Based on a review of performance assessments used in high school biology and history settings, we found a reoccurring discrepancy between assessment scoring criteria and performance expectations. For example, in the AP exam scoring guidelines, points are awarded to student writing based on the inclusion of certain content information. However, the AP scoring guidelines also specify that high scoring essays will be "well organized" and "well written," without further discussion of the specific features that constitute these writing characteristics. The final score is the accumulation of these points. The scoring rubric for the NY Regents U.S. History exam combines aspects of essay organization (e.g., inclusion of an introduction and conclusion) with content-focused criteria of document analysis and the incorporation of relevant outside knowledge. Similar problems were found in the scoring of routine, in-class writing tasks. For example, in the 2006

Prentice Hall biology textbook, students are asked to complete writing assessments called “Writing in Science” as part of the end-of-chapter assessments. This task entails writing a paragraph or group of paragraphs on target biology content. Like the AP Biology writing exam, Prentice Hall’s writing assessment criteria explicitly refer only to the scoring of content. For example, in one prompt, students are asked to write a paragraph that includes (a) an explanation of a polymer; (b) a description of organic compounds; and (c) how these organic compounds are used in the human body. Notably, the evaluation criteria relate only to biology content (e.g., one of the criteria requires that students “explain that a polymer is a macromolecule made up of monomers joined together”). None of the evaluation criteria pertain specifically to the language features needed to successfully provide an explanation of the content. As with the AP Biology exam, students are expected to communicate science concepts using academic language, though these literacy skills are only implicitly evaluated as part of the assessment score.

Since the scoring guidelines for tests and writing tasks often conflate content and language, it is unclear whether raters’ scores measure content understanding or a combination of content understanding and students’ literacy skills for describing, analyzing, and explaining. Without explicit (and separate) scoring criteria to evaluate language and literacy skills, it is difficult to determine the extent to which writing quality should reflect literacy/writing skills versus content knowledge. In order to measure student performance on the written explanation task, we developed two separate rubrics to evaluate biology content knowledge and academic language proficiency in the student written explanations, with both constructs expected to be impacted by RA instruction and students’ use of RA strategies.

Writing Content Scoring Rubric. For the content rubric (see Appendix F), criteria were formed, in part, by using the previously developed and validated CRESST rubrics; AP scoring guidelines; and NY Regents test rubrics as guides. Our goal was to measure students’ conceptual knowledge; ability to connect principles and concepts; and capability to extend prior knowledge of concepts (beyond the limited contexts in which they were acquired), in order to create well-developed explanations. Based on this goal, we developed a list of four initial key points upon which to base our rubric: (a) understanding of the target discipline-specific content; (b) clarity of explanation; (c) use of supportive evidence from the provided texts; and (d) inclusion of prior knowledge.

Both writing rubrics (language and content) were rated on a 4-point scale, with each score reflecting different aspects of writing proficiency. The rubrics provide a portrait of a student’s biology explanation as it may appear at a given proficiency level.

A student response receiving a high writing content score had to satisfy most or all of the scoring criteria, which were elaborated in the rubric's 4-point description. Specifically, the response demonstrated well-developed understanding of the target content. In addition, this content was clear, focused, thoroughly explained, and elaborated with strong supportive evidence. The content dimension also encompassed whether or not a student demonstrated relevant knowledge that extended beyond information explicitly given in the text passage (i.e., whether or not a student incorporated prior knowledge). Lastly, this dimension focused on the extent to which students incorporated relevant information from the texts into their responses. The specific content raters were to look for in student responses was elaborated in the supplemental documents for the writing content rubric.

Together, these aspects of the rubric were collectively expected to measure content understanding and students' ability to successfully meet a fairly demanding cognitive challenge. Specifically, in addition to possessing the necessary content knowledge, in order to score well on this task, students needed to apply complex cognitive skills, such as textual analysis and synthesis of historical information, from multiple sources.

Writing Language Scoring Rubric. In developing the ILA Writing Language Rubric (see Appendix G), we modified the language dimensions that were previously developed and validated in earlier CRESST work (see Aguirre-Muñoz et al., 2005) in order to align them with the RA instructional model; a discipline-specific setting; and the explanation genre. Key points were used to evaluate students' academic language proficiency on the ILA, based on the dimensions identified as significant in academic writing. The language rubric specifically focuses on assessing students' linguistic command of grammatical structures that are directly related to the explanation genre and that are also aligned with the California Content Standards in writing. Additionally, the measured language features include those that students frequently become aware of during their analyses of text schemas and text structures in the RA instructional model. For students in RA classrooms, the language rubric also implicitly measures how well students are able to transfer the academic language they have become familiar with in the Reading Strategies into their writing process. Specifically, the language rubric measured three concepts that define the overall qualities of a historical or scientific explanation. These include: (1) appropriate text cohesion, (2) varied and precise word choice, and (3) a formal, impersonal tone.

As we looked for text cohesion, we checked for sentence structure variety and the use of expressions of causality through the use of nominalization (i.e., noun phrases used in place of verb form), causative verbs (e.g., led to, resulted from), and/or transitional expressions. In looking for precise and varied word choice, we checked for discipline-specific vocabulary, as well as everyday terms used with subject-specific meanings. In both cases, we looked for these words to be organized as part of expanded noun phrases (e.g., *because of racial discrimination, many blacks decided to pack up and get out of the rural south*). For evidence of an impersonal and authoritative tone, we looked for use of third person, passive voice, and for the presence of few or no speech markers (e.g., “well”, “you know”, “like”). While some debate exists in the field as to whether an authoritative tone is necessary for good written communication, it remains the standard for academic writing; thus, it is a key aspect of how we have defined and measured appropriate academic language use in our language rubric.

Based on previous CRESST work (see Aguirre-Muñoz et al., 2005), we knew that most students in the early years of high school do not have the academic language proficiency to produce high-quality academic explanations. For this reason, the language rubric was structured to sensitively measure a range of academic language proficiency levels in science and history writing. We related the ideas of abstraction, informational density, and technicality to three systemic functional linguistic concepts. *Mode* (the manner in which ideas are communicated) refers to students’ ability to create appropriate text cohesion in their writing. *Field* (the linguistic elements used to communicate those ideas) signifies students’ ability to use varied and precise word choice. *Tenor* (the tone of that communication) refers to students’ ability to establish a formal, impersonal tone in their writing.

In order to receive a high score on the language dimension, a student’s explanation had to meet most or all of the following criteria: demonstration of very good text cohesion through regular use of sentence structure variety (specifically, through use of marked themes); consistent use of precise and varied word choice (specifically, through use of expanded noun phrases); and use of an impersonal and authoritative tone with few or no speech markers. The length of a student’s paper was taken into consideration to the extent that the writing needed to be long enough to provide evidence of academic language proficiency. Further discussion of the writing language scoring rubric is provided in Appendix H.

Methods

Sample

Sixty-one biology teachers (i.e., 20 men and 41 women), representing 54 public high schools in California, agreed to participate in the study. Their length of teaching experience at the onset of the RA training ranged from 1 to 36 years, with an average of 11 years.

Teachers in the treatment group participated in the initial RA professional development during the summer of 2007 and then attended follow-up sessions during the school year. The Biology ILA was administered at the end of the 2008-2009 school year. A total of 825 ILA Part 1, 798 ILA Part 2, and 383 ILA Part 3 student assessments were collected from 47 biology teachers.

Sixty-three history teachers from 56 California public high schools participated in the study. The sample included roughly an equal number of male (31) and female (32) teachers. Their length of teaching experience at the onset of the training year ranged from 2 to 37 years— with an average of 12 years. Two cohorts of history teachers were trained in the RA program. The first cohort participated in the initial professional development during the summer of 2006 and administered the History ILA at the end of the 2007-2008 school year. Teachers in the second cohort began their training during the summer of 2007 and administered the History ILA at the end of the 2008-2009 school year. A total of 869 ILA Part 1, 850 ILA Part 2, and 391 ILA Part 3 student assessments were collected from 49 history teachers.

The Scoring Session

CRESST researchers trained teams of raters to score Parts 2 and 3 of the ILA during the summers following their administration. The training and scoring sessions were held over several days. To minimize rater bias, all identifying information (student names, teacher names, school names) was removed from the student papers. In addition, the test booklets did not include any markings related to treatment group assignment. Responses were randomly distributed into packets containing approximately 20 responses each.

All raters underwent intensive training to learn and practice implementing the scoring procedures. These sessions also provided opportunities to address raters' questions and ensure that the scoring rubrics were clear. Raters received two days of training on the content and language rubrics and a half day of training for the reading strategies and metacognition rubrics. The training was followed by a scoring session. Within each scoring session, students' responses were read and scored by two different raters. The final scores were obtained by taking the arithmetic mean of the scores assigned by two raters, thereby reducing the influence of rater variability.

Reliability of ILA Scores

A series of generalizability studies were conducted in order to examine the reliability of the ILA components. Generalizability theory (see e.g., Cronbach et al., 1972; Shavelson & Webb, 1991) explicitly acknowledges that some universe of acceptable observations exists that is larger than the set of test conditions within a given study. Moreover, we would view any sample of observations drawn from that universe as being equally acceptable. In the case of the ILA, this means that we would not want scores to depend greatly on the particular test items that students were given or the particular raters who assigned scores. Generalizability theory, then, provides a framework for understanding the extent to which variability in observed scores can be attributed to various aspects of the measurement design. Importantly, it allows simultaneous treatment of these design features (though, in the case of the student ILA scores, only single facet designs were used). This is in contrast to more classical approaches, in which only a single source of measurement error is considered at a time, leading to the calculation of multiple reliability coefficients (inter-rater reliability, internal consistency, test-retest reliability, etc.), which can make it difficult to assess the overall dependability of a measure. Here, we describe findings from generalizability studies for the metacognition and reading strategies items in Part 2 as well as the writing language and content scores from Part 3. In addition, we present an examination of students' scores on the multiple choice tests in Parts 1 and 2 of the ILA. For each measure, we present estimates of the reliability coefficients based on the measurement designs used in this study. However, it should be noted that generalizability studies provide valuable information that could inform the design of future assessments, including use of the ILA in future studies.

Two coefficients are calculated for each score. The first, $\hat{\rho}^2$, describes the reliability of the score for relative decisions and is roughly equivalent to the squared correlation between the observed scores and those that might be obtained by averaging over many repeated observations (the universe score). It is calculated as the proportion of expected variance in observed scores ($\hat{\sigma}_s^2 + \hat{\sigma}_{\mathrm{Re}l}^2$) that is due to variance in universe scores ($\hat{\sigma}_s^2$). This coefficient can be considered the extent to which the measure provides a consistent rank ordering of students. The second coefficient, $\hat{\phi}$ (also known as the index of dependability; Brennan & Kane, 1977), describes the proportion of *total* variance in observed scores ($\hat{\sigma}_s^2 + \hat{\sigma}_{Abs}^2$) that is attributable to variability in the universe score. It reflects the reliability of the scores for absolute decisions (when the magnitude of the score of interest and not only the rank ordering of students). Formulas for both $\hat{\rho}^2$ and $\hat{\phi}$ are shown below.

$$\hat{\rho}^2 = \frac{\hat{\sigma}_s^2}{\left(\hat{\sigma}_s^2 + \hat{\sigma}_{\mathrm{Re}l}^2\right)}, \; \hat{\phi}^2 = \frac{\hat{\sigma}_s^2}{\left(\hat{\sigma}_s^2 + \hat{\sigma}_{Abs}^2\right)}$$

As evident in the formulas, the two indices differ only in their denominator. In both cases, the denominator is expressed as a sum of "true" variance ($\hat{\sigma}_s^2$) and error variance (either $\hat{\sigma}^2_{\mathrm{Re}l}$ or $\hat{\sigma}^2_{Abs}$). The difference between the two is simply in how the error variance is calculated. In the case of $\hat{\sigma}^2_{\mathrm{Re}l}$, only variance components that represent interactions with students (and thus affect rank ordering) are considered. For $\hat{\sigma}^2_{Abs}$, both interactions and main effects are considered. Thus, $\hat{\sigma}^2_{Abs}$ is always equal to or larger than $\hat{\sigma}^2_{\mathrm{Re}l}$. As a consequence, $\hat{\rho}^2$ is always equal to or greater than $\hat{\phi}$.

For the ILA, two measurement designs were utilized. In regards to the multiple choice tests for content knowledge and reading comprehension, scores reflect an averaging across the items of each test. This corresponds to a students-by-items (S x I) design. Here, variance components for students ($\hat{\sigma}^2_s$) and items ($\hat{\sigma}^2_i$) are estimated, along with a residual term ($\hat{\sigma}^2_{si,e}$). The subscript of the residual reflects the fact that this term is actually a sum of the variance due to the interaction of students and items (*si*) and additional unexplained random variance (*e*). Since this is a design with only one facet, the variance $\hat{\sigma}^2_{\mathrm{Re}l}$ is equal to $\hat{\sigma}^2_{si,e}$, divided by the number of items; $\hat{\sigma}^2_{Abs}$ is the sum of $\hat{\sigma}^2_i$ and $\hat{\sigma}^2_{si,e}$, divided by the number of items. The scores for reading strategies, metacognition, writing content, and writing language were based on averages of the scores assigned by multiple raters, a students-by-raters (S x R) design. The variance components estimated for these scores include those for students ($\hat{\sigma}^2_s$), raters ($\hat{\sigma}^2_{raters}$), and the residual term ($\hat{\sigma}^2_{sr,e}$). Here, the variance $\hat{\sigma}^2_{\mathrm{Re}l}$ is equal to $\hat{\sigma}^2_{sr,e}$, divided by the number of raters; $\hat{\sigma}^2_{Abs}$ is the sum of $\hat{\sigma}^2_r$ and $\hat{\sigma}^2_{sr,e}$, divided by the number of raters.

Reliability coefficients were estimated from samples of student ILA responses randomly drawn from the full scoring samples in order to estimate the generalizability coefficients for the content knowledge and reading comprehension scores. For scores obtained from Parts 2 and 3 of the ILA, coefficients were estimated from either random samples from the scoring sample or from independent (calibration) samples scored by multiple raters. Estimates of the variance components for scores on the biology and history assessments are summarized in Tables 2 and 3, respectively. The final column of these tables present the proportions of variance attributed to each component. Larger values for the component attributed to students are desired, as they result in larger reliability coefficients. On the other hand, these proportions should not be directly compared across scores, since the measurement designs differ. Specifically, scores on the content knowledge and reading comprehension tests are obtained by averaging over the test items, while other ILA scores result from averaging over raters. Nevertheless, it is somewhat concerning that the percentages related to $\hat{\sigma}^2_s$ are rather small (relative to those for $\hat{\sigma}^2_i$ and $\hat{\sigma}^2_{si,e}$) for the multiple choice tests of content knowledge and reading comprehension, compared to the other ILA scores. The estimates for the main effect of items ($\hat{\sigma}_i^2$) reflect variation in the difficulty of items, while the large estimates for the residual term suggest substantial person-by-item interaction (i.e., different items give different rank ordering of students), a large amount of unexplained variance in scores, or both. We will return to these tests in the subsequent section. It appears that the estimates are more reasonable for the other measures. The small percentages related to the rater facet ($\hat{\sigma}_r^2$) indicate that the raters were quite consistent in the severity of their ratings. The estimates for the $\hat{\sigma}_{sr,e}^2$ term (and the corresponding percentages), suggest that student-by-rater interactions and unexplained random error contributed more to the observed variability in scores than the main effect of raters.

Table 2

Variance Component Estimates for Biology ILA Scores

Measure	Source of variation	Component	Estimate***	% total
Content knowledge*	Students (*s*)	$\hat{\sigma}_s^2$	.013	5.3
(100 students, 10 items)	Items (*i*)	$\hat{\sigma}_i^2$	.070	27.7
	Residual (*si*,*e*)	$\hat{\sigma}_{si,e}^2$	.169	67.0
Reading comprehension*	Students (*s*)	$\hat{\sigma}_s^2$	.035	14.1
(100 students, 10 items)	Items (*i*)	$\hat{\sigma}_i^2$	.022	8.6
	Residual (*si*,*e*)	$\hat{\sigma}_{si,e}^2$	.194	77.3
Metacognition**	Students (*s*)	$\hat{\sigma}_s^2$	.202	43.6
(20 students, 8 raters)	Raters (*r*)	$\hat{\sigma}_r^2$	.032	6.8
	Residual (*sr*,*e*)	$\hat{\sigma}_{sr,e}^2$	.230	49.6
Reading strategies**	Students (*s*)	$\hat{\sigma}_s^2$	1.114	87.5
(20 students, 8 raters)	Raters (*r*)	$\hat{\sigma}_r^2$	.032	2.5
	Residual (*sr*,*e*)	$\hat{\sigma}_{sr,e}^2$	.128	10.0

Measure	Source of variation	Component	Estimate***	% total
Writing – content**	Students (*s*)	$\hat{\sigma}_s^2$	.963	76.8
(20 students, 9 raters)	Raters (*r*)	$\hat{\sigma}_r^2$	.024	1.9
	Residual (*sr*,*e*)	$\hat{\sigma}_{sr,e}^2$	.266	21.2
Writing – language**	Students (*s*)	$\hat{\sigma}_s^2$	.835	72.4
(20 students, 9 raters)	Raters (*r*)	$\hat{\sigma}_r^2$	.073	6.4
	Residual (*sr*,*e*)	$\hat{\sigma}_{sr,e}^2$	.245	21.2

Note. *Content knowledge and reading comprehension estimates based on groups of 100 students randomly selected from the full scoring sample. **Analyses of scores for Parts 2 and 3 based on reliability samples of 20 students and 8 or 9 raters (depending on measure). ***Variance component estimates obtained using random effects ANOVA.

Table 3

Variance Component Estimates for History ILA Scores

Measure	Source of variation	Component	Estimate***	% Total
Content knowledge*	Students (*s*)	$\hat{\sigma}_s^2$	.016	6.6
(100 students, 10 items)	Items (*i*)	$\hat{\sigma}_i^2$	.038	15.3
	Residual (*si*,*e*)	$\hat{\sigma}_{si,e}^2$	.194	78.1
Reading comprehension*	Students (*s*)	$\hat{\sigma}_s^2$	.025	10.2
(100 students, 12 items)	Items (*i*)	$\hat{\sigma}_i^2$	.040	16.0
	Residual (*si*,*e*)	$\hat{\sigma}_{si,e}^2$	.184	73.8
Metacognition*	Students (*s*)	$\hat{\sigma}_s^2$	.412	79.7
(100 students, 2 raters)	Raters (*r*)	$\hat{\sigma}_r^2$	.003	0.6
	Residual (*sr*,*e*)	$\hat{\sigma}_{sr,e}^2$	.102	19.7
Reading strategies**	Students (*s*)	$\hat{\sigma}_s^2$	1.564	92.7
(5 students, 7 raters)	Raters (*r*)	$\hat{\sigma}_r^2$	.031	1.8
	Residual (*sr*,*e*)	$\hat{\sigma}_{sr,e}^2$	.093	5.5

Measure	Source of variation	Component	Estimate[***]	% Total
Writing – content[**]	Students (*s*)	$\hat{\sigma}_s^2$	.355	49.5
(20 students, 5 raters)	Raters (*r*)	$\hat{\sigma}_r^2$	.000	.0
	Residual (*sr*,*e*)	$\hat{\sigma}_{sr,e}^2$	.362	50.5
Writing – language[**]	Students (*s*)	$\hat{\sigma}_s^2$	.465	57.8
(20 students, 5 raters)	Raters (*r*)	$\hat{\sigma}_r^2$	.057	7.1
	Residual (*sr*,*e*)	$\hat{\sigma}_{sr,e}^2$	.283	35.1

Note: [*]Content knowledge, reading comprehension, and metacognition estimates based on groups of 100 students randomly selected from the full scoring sample. [**]Analyses of scores for reading strategies and writing scores based on reliability study of samples with varying numbers of students and raters (depending on measure). [***]Variance component estimates obtained using random effects analysis of variance. Negative estimates set to zero.

As previously described, coefficients $\hat{\rho}^2$ and $\hat{\phi}$ were calculated from the variance component estimates and the number of observations for each of the design facets (i.e., the numbers of items and raters used in actual scoring); these results are shown in Table 4. As expected from the results in Tables 2 and 3, reliability estimates are somewhat low for the multiple choice tests (content knowledge and reading comprehensions) but in a more acceptable range for the other measures.

Table 4

Coefficients for Relative and Absolute Decisions for Biology ILA Scores

Measure	n_{levels} (items or raters)	Relative decisions	Absolute decisions
Biology ILA			
Content knowledge	10	.44	.36
Reading comprehension	10	.65	.62
Metacognition	2	.64	.61
Reading strategies	3	.96	.95
Writing–content	2	.88	.87
Writing–language	2	.87	.84
History ILA			
Content knowledge	10	.46	.41
Reading comprehension	12	.62	.58
Metacognition	2	.90	.89
Reading strategies	3	.97	.96
Writing–content	2	.66	.66
Writing–language	2	.77	.73

Note. Based on estimated variance components (Tables 2 and 3) and number of facet levels (items or raters) in the measurement design.

It should be noted that the generalizability coefficient $\hat{\rho}^2$ for the multiple choice tests is equivalent to Cronbach's alpha (internal consistency), which may be viewed as a measure of the average correlation among items on a test. However, this index is most interpretable for uni-dimensional tests. The presence of multiple dimensions (i.e., multiple constructs influencing test responses) could result in biased estimates of reliability, though the direction of such bias would depend on the nature of the relationships between dimensions. Thus, we consider possible violations of uni-dimensionality in the tests of knowledge and reading comprehension.

Tables 5 and 6 presents descriptive statistics for each item in the tests of content knowledge, including the percent of respondents with correct answers and the correlation between item and total score on the remaining items in the test. Here, it is evident that these tests include items that reduce the internal consistency of the scale (resulting in a smaller generalizability coefficient). Specifically, items 1, 2, 5, and 6 from the Biology ILA and item 10 from the History ILA have rather weak correlations with other items on the test. Analyses of item responses suggest that the poor performance of these items may be due to students having difficulty choosing between available response choices. It is notable that the percentages of students answering these items correctly were low for each of these five questions. This may create a floor effect of sorts, where even high achieving students (as demonstrated in their responses to other questions) seemed to do no better on these items than what might be expected if they were simply guessing. An alternative explanation could be that these items measure abilities that are qualitatively different from the remainder of the test (i.e., the test is multidimensional). Whatever the cause, the internal consistency of the test can actually be increased if the four problematic items are removed. The last columns of Tables 5 and 6 show that the item-test correlations are generally larger once the problematic items are removed.

Similar analyses were conducted for reading comprehension tests. Tables 7 and 8 present descriptive statistics for these tests. Item 6 from the Biology ILA and item 3 from the History ILA both appear to be problematic. The correlation between the scores on these items and the total scores on the remaining items are rather close to zero. Reanalyzing the tests without these items produces very little change in the item-test correlations.

Response data for the content knowledge and reading comprehension tests were also analyzed within an item response theory (IRT) framework. Appendix G presents a summary of the results for the full- and reduced-length tests for both the Biology and History ILA instruments. A three-parameter logistic (3PL) model with two correlated factors (corresponding to the two tests, content knowledge and reading comprehension) was used. The 3PL model estimates discrimination, intercept, and guessing parameters for each item.

Table 5

Descriptive Test Statistics for the Biology Test of Content Knowledge (Biology ILA Part 1)

Item	% of students answering correctly	Corrected item-total correlation	
		Full test	Reduced test [a]
1	31.4	.02	NA
2	6.4	-.07	NA
3	42.8	.24	.23
4	73.7	.21	.28
5	13.6	.02	NA
6	18.9	-.04	NA
7	82.4	.18	.26
8	54.5	.23	.29
9	36.1	.21	.26
10	61.1	.18	.22

Note.[a] Reduced test excludes items 1, 2, 5, and 6.

Table 6

Descriptive Test Statistics for the History Test of Content Knowledge (History ILA Part 1)

Item	% of students answering correctly	Corrected item-total correlation	
		Full test	Reduced test [a]
1	79.3	.23	.24
2	48.6	.32	.33
3	67.1	.34	.33
4	67.9	.14	.15
5	43.6	.27	.27
6	38.3	.13	.12
7	64.7	.29	.27
8	56.4	.18	.19
9	81.6	.14	.14
10	26.7	.13	NA

Note.[a] Reduced test excludes items 10.

Table 7

Descriptive Test Statistics for the Biology Test of Reading Comprehension (ILA Part 2)

Item	% of students answering correctly	Corrected item-total correlation	
		Full test	Reduced test [a]
1	69.5	.27	.28
2	56.8	.23	.24
3	51.3	.28	.28
4	53.4	.21	.21
5	50.4	.29	.31
6	20.1	.01	NA
7	50.4	.29	.30
8	71.6	.31	.32
9	40.0	.19	.18
10	66.4	.23	.24

Note.[a] Reduced test excludes item 6.

Table 8

Descriptive Test Statistics for the History Test of Reading Comprehension (History ILA Part 2)

Item	% of students answering correctly	Corrected item-total correlation	
		Full test	Reduced test[a]
1	32.7	.20	.19
2	56.2	.24	.25
3	20.8	.09	NA
4	59.6	.24	.24
5	69.6	.29	.29
6	72.7	.32	.32
7	64.1	.18	.19
8	81.6	.33	.33
9	57.5	.19	.20
10	74.6	.29	.30
11	73.3	.29	.29
12	41.4	.28	.27

Note.[a] Reduced test excludes item 3.

The discrimination parameter (or slope) is analogous to the item-test correlations presented in Tables 5 through 8; it represents how well the item discriminates between individuals who differ on the latent trait. The intercept parameter is related to both the slope and the difficulty of an item (i.e., the percentage of students correctly answering a question). The guessing parameter accounts for the fact that even individuals with low ability levels have some nonzero probability of choosing the correct response.

When a confirmatory factor model was fit to the test data with single factors for each of the two 10-item tests, there was evidence of bias in the item parameter estimates due to the same items that appeared problematic in the descriptive analyses. When these items were removed, the resulting parameter estimates were in a more reasonable range. Table 9 shows estimates of the reliability coefficients $\hat{\rho}^2$ and $\hat{\phi}$ for the full- and reduced-length tests. The reliability coefficients increase for each test when the problematic items are excluded, though in some cases the change is quite small.

Table 9

Coefficients for Relative and Absolute Decisions for the Biology ILA Tests of Content Knowledge and Reading Comprehension, Based on Estimated Variance Components.

Measure	n_{items}	Relative decisions	Absolute decisions
Biology ILA - Content knowledge			
All items	10	.44	.36
Reduced test*	6	.50	.47
Biology ILA - Reading comprehension			
All items	10	.65	.62
Reduced test**	9	.66	.65
History ILA - Content knowledge			
All items	10	.46	.41
Reduced test***	9	.46	.42
History ILA - Reading comprehension			
All items	12	.65	.62
Reduced test****	11	.68	.67

Note. *Omits items 1,2,5,6. **Omits item 6. ***Omits item 10. ****Omits item 6.

Taken together, the descriptive and IRT analyses suggest that scores from the reduced tests may be preferable to the full-length tests. In the subsequent section, analyses are conducted using these shorter versions of the tests, in which the problematic items are omitted.

Theoretical and Statistical Models of RA Effects

A possible model for the effects of the RA program is presented in Figure 4. Participation in the program is expected to result in certain changes in teachers' instructional practices. These, in turn, may affect how students approach reading. The ILA metacognition and reading strategies scores are intended to measure such changes. Both the use of particular reading strategies and improved metacognition may contribute to students' reading comprehension and other desired outcomes. Additional measures were used to examine variations in instruction. Although the development and properties of these measures are beyond the scope of this report, in order to more fully examine the plausibility of the model and to aid the interpretation of the student-level variables, some results based on their use are presented here.

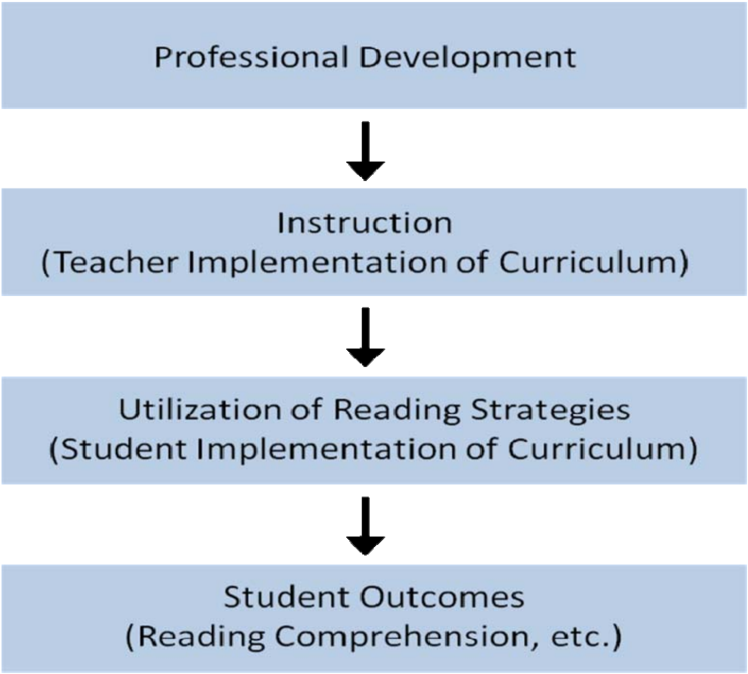

Professional Development
Instruction
(Teacher Implementation of Curriculum)
Utilization of Reading Strategies
(Student Implementation of Curriculum)
Student Outcomes
(Reading Comprehension, etc.)

Figure 4. A possible model for the hypothesized effects of the program.

To estimate the effect of treatment group assignment on the student-level variables, we fit a series of hierarchical linear models to the ILA scores. A multi-level approach is needed in order to acknowledge the fact that students in the study are not independent; rather, they are nested within classrooms. All the models follow the same basic structure. The level-1 (student-level) equation relates the observed ILA score (Y_{ij}) to a class-level mean (β_{0j}), plus a residual term (r_{ij}):

$$Y_{ij} = \beta_{0j} + r_{ij},\ r_{ij} \sim N\left(0, \sigma^2\right)$$

The level-2 (class-level) equation presents the class-level mean (β_{0j}) as a sum of a grand mean (γ_0), the product of the treatment effect (γ_1) and an indicator of class-level treatment status ($TREATMENT_j$, a variable with value 0 or 1 for the control and treatment groups, respectively), and a class-level residual term (u_j):

$$\beta_{0j} = \gamma_0 + \gamma_1\left(TREATMENT_j\right) + u_j,\ u_j \sim N\left(0, \tau^2\right)$$

Linear regression models were used to examine the relationship between treatment assignment and teacher-level variables. The form of these models is essentially the same as the level-2 equation in the multi-level models. Specifically, teachers' implementation scores (Y_j) are modeled as the sum of a grand mean (β_0), the product of the treatment effect (β_j) and treatment status ($TREATMENT_j$), and a residual term (e_j):

$$Y_j = \beta_0 + \beta_1\left(TREATMENT_j\right) + e_j,\ e_j \sim N\left(0, \sigma^2\right)$$

In addition to fitting the various multi-level and regression models, we calculated the Pearson correlations between the study variables. To account for the nested data structure, student ILA scores were first averaged within classrooms.

Effects of treatment assignment on teacher instruction variables (literacy instruction, content coverage) were estimated via ordinary least squares regression. Effects on student implementation variables (metacognition, reading strategies) and other student outcomes (content knowledge, reading comprehension, writing content, and writing language) were estimated using hierarchical linear models.

Results

Estimated Treatment Effects and Correlations Among Study Variables

Tables 10 and 11 present the results of the analyses described above for the biology and history samples, respectively. For the Biology ILA, the estimated effect of assignment was positive for literacy instruction ($\beta_j = .20$, $p<.01$). No effect was observed on the measure of content coverage. For the history sample, assignment to the RA program had positive effect on both literacy instruction ($\beta_j = .21$, $p<.01$) and content coverage ($\beta_j = .10$, $p<.05$). The positive effects on literacy instruction are consistent with the intended goals of the RA curriculum. However, it would be possible for that emphasis to come at the expense of other aspects of the curriculum. Based on this measure of content coverage, however, coverage was similar for the treatment and control groups. Although the estimated effects of assignment on metacognition and reading strategies were positive for both the biology and history samples, none of these effects were statistically significant. Estimated effects on other student outcomes varied in direction; that is, there were some positive and negative effects. However, these also were not significant. In sum, there is evidence that assignment to the RA group had a positive effect on the intended teacher practice, which is literacy instruction. However, no significant effects were observed on more distal variables.

Table 10

Analyses of Treatment Effects and Correlations Between Study Variables – Biology Sample

	Trt effect		Pearson correlations of class level means									
Variable	Est	*SE*	Trt	LI	OTL	MC	RS	CK	RC	WC	WL	DRP
Instructional practices												
Literacy instruction (LI)	.20**	.06	.45**									
Content coverage (OTL)	.00	.04	.00	-.14								
Student reading processes												
Metacognition (MC)	.15	.12	.28	.28	.34*							
Reading strategies (RS)	.03	.20	.13	.25	.21	.47**						
Other student outcomes												
Content knowledge (CK)	-.38	.23	-.24	.06	.39*	.40**	.29					
Reading comprehension (RC)	-.38	.32	-.12	.00	.35	.52**	.42**	.69**				
Writing content (WC)	.01	.15	-.12	-.09	.12	.46**	.28	.55**	.65**			
Writing language (WL)	.08	.15	-.04	-.06	.17	.55**	.25	.52**	.66**	.94**		
Degrees of reading power (DRP)	-.77	3.45	.06	.05	.02	.45**	.07	.44**	.48**	.47**	.48**	
Biology CST	-12.34	12.15	-.56*	-.17	.16	.40*	.09	.79**	.70**	.73**	.72**	.62**

*p<.05; **p<.01

Table 11

Analyses of Treatment Effects and Correlations Between Study Variables – History Sample

	Trt effect		Pearson correlations of class level means								
Variable	Est	*SE*	Trt	LI	OTL	MC	RS	CK	RC	WC	WL
Instructional practices											
Literacy instruction (LI)	.21**	.04	.57**								
Content coverage (OTL)	.10*	.05	.36*								
Student reading processes											
Metacognition (MC)	.08	.08	.10	.13	-.09						
Reading strategies (RS)	.42	.22	.23	.26	-.06						
Other student outcomes											
Content knowledge (CK)	.39	.26	.18	.23	.20	.26	.24				
Reading comprehension (RC)	-.01	.31	-.07	-.07	-.40	.44**	.30*	.45**			
Writing content (WC)	.12	.14	.09	.18	-.12	.46**	.43**	.57**	.68**		
Writing language (WL)	.10	.13	.06	.16	-.09	.39**	.35*	.54**	.67**	.94**	
Degrees of reading power (DRP)	-2.02	3.11	-.14	-.08	-.22	.31*	.20	.32*	.71**	.58**	.63**

*p<.05; **p<.01

There are, of course, many reasons that such effects might not be observed. Perhaps the most straightforward interpretation is that the proposed model (Figure 4) is incorrect. Specifically, while participation in RA professional development may lead to enhanced literacy instruction, this change in instruction may not affect student reading processes or other student outcomes. An alternative to this conclusion might be that it is too soon to observe any effects on student outcomes. From this perspective, changes in student variables may indeed be related to instruction (and so the model may be generally correct). However, those changes could take longer to develop and perhaps had not occurred when the ILA was administered. The correlations between the class-level mean scores for these variables are at least suggestive of positive relationships between the steps in the hypothesized model. The correlations between variables that are adjacent in Figure 4 are shaded in gray in Tables 10 and 11. Given these apparent positive relationships, another possibility is that effects of treatment assignment on the student variables are attenuated by multiplication of modest stepwise effects.

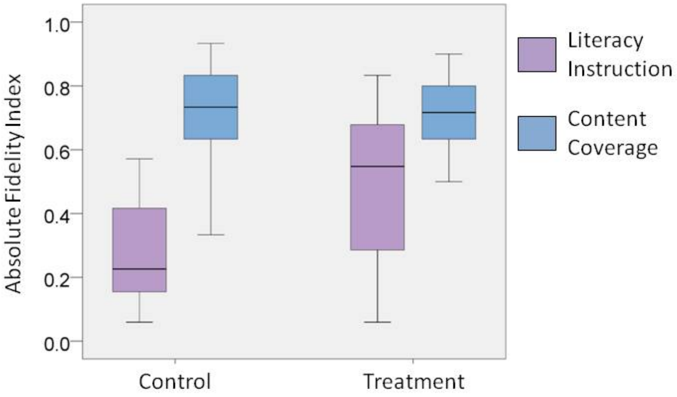


Figure 5. Teacher implementation variables–Biology Sample. Literacy Instruction is shown in purple; content coverage is shown in blue.

In addition, heterogeneity in implementation of the RA curriculum by teachers and utilization of RA strategies by students may further reduce the power of the study to detect overall effects of treatment assignment. As an example, Figure 5 presents boxplots of the biology teacher implementation variables for the two study groups. A similar pattern was observed among history teachers. As described previously, the average scores for content coverage are similar across groups, while the level of literacy instruction is generally higher in the treatment group. That said, there is substantial variability within the groups; in fact, the treatment and control groups display a substantial amount of overlap. As a consequence, it appears that some students in control classrooms may have been exposed to levels of literacy instruction that were comparable to or even exceeded those present in some treatment classrooms.

Annotations are Associated with Higher Performance. While annotating the document sections required of all students, only 26% of the Biology sample and 31% of the History sample annotated at least once.[5,6] Given the importance of Reading Strategies to the RA intervention and predominance of treatment students in this subgroup, we decided to further explore this process measure. Again, raters used students' text annotations as the sole source of evidence for scoring Reading Strategies.

The developers of the RA curriculum strongly believe that annotation is a critical tool for improving reading comprehension; data from this study lend support to this belief. As shown in Tables 12 and 13, on almost all ILA measures (the one exception is the content knowledge score for the biology sample), students in both samples who annotated their ILA text set outperformed those who did not annotate. In addition, the odds of annotating were greater for the treatment group than for the control students—2.6 times greater for the biology sample and 2.4 times greater for the history sample. It is not surprising that RA students are more likely to annotate but this result provides evidence that a strategy emphasized in the RA curriculum was utilized.

[5] An ILA was considered annotated if a single word or symbol (e.g., arrow, underline, circle) appeared on one of the four texts or in the margins of one of the four texts.

[6] Students who did not annotate their texts and clearly completed other items in the ILA Part 2 received a score of 1 for Reading Strategies.

Table 12

Comparison of ILA Scores For Students Who Did And Did Not Annotate Texts, Ignoring Treatment Assignment – Biology Sample

Outcome variable	Annotations	*N*	Mean	*SD*	*SE* of Mean
Writing content*	None	285	1.7	.9	.1
	Some	98	2.1	.9	.1
Writing language*	None	285	1.9	.9	.1
	Some	98	2.3	.9	.1
Reading comprehension*	None	572	5.0	2.2	.1
	Some	226	6.0	1.8	.1
Reading strategies*	None	44	1.0	.0	.0
	Some	226	2.8	.8	.1
Content knowledge	None	618	4.2	1.7	.1
	Some	207	4.4	1.5	.1
Metacognition*	None	572	2.5	.8	.0
	Some	226	3.0	.8	.1

*p < 0.05 for independent samples t-test

Table 13

Comparison of ILA Scores For Students Who Did And Did Not Annotate Texts, Ignoring Treatment Assignment – History Sample

Outcome variable	Annotations	*N*	Mean	*SD*	*SE* of Mean
Writing content*	None	240	1.7	0.7	0.0
	Some	151	2.2	0.8	0.1
Writing language*	None	240	1.9	0.8	0.0
	Some	151	2.3	0.8	0.1
Reading comprehension*	None	581	6.8	2.3	0.1
	Some	275	7.6	2.2	0.1
Reading strategies*	None	399	1.0	0.0	0.0
	Some	275	2.4	1.0	0.1
Content knowledge*	None	589	5.6	2.1	0.1
	Some	269	6.0	2.1	0.1
Metacognition*	None	583	2.2	0.6	0.0
	Some	275	2.5	0.9	0.1

*p< 0.05 for independent samples t-test

Student Annotation Frequency Varies Across Texts. The frequency of annotations by document is presented for the Biology ILA in Table 14. Here, it is apparent that annotations were not equally distributed across the sections of the test. Among the students included in this analysis, 78% (81% in the treatment group and 72% in the control group) showed annotations in document section one, while about 70% of the assessments showed annotations in document sections three and five. Roughly 50% of students showed annotations in document sections two and four. Results for the history sample are provided in Table 15. Approximately 80% of the student assessments included in this analysis showed annotations for documents one and two, while 73% of the assessments showed annotations in document three. Only 34% showed annotations in document four.

It is important to note that the documents varied in format (e.g., paragraph versus data table), length, and language difficulty. In addition, there was a mix of primary and secondary sources. All of these factors may have influenced whether or not students annotated specific texts.

Table 14

Frequency of Annotations Across Documents by Status for ILA Part 2 – Biology Sample

	Treatment (*N*=161)		Control (*N*=107)	
Document	*N*	%	*N*	%
1	131	81.4	77	72.0
2	93	57.8	48	44.9
3	122	75.8	60	56.1
4	88	54.7	45	42.1
5	118	73.3	57	53.3

Table 15

Frequency of Annotations Across Documents by Status for ILA Part 2 – History Sample

	Treatment (*N*=176)		Control (*N*=97)	
Document	*N*	%	*N*	%
1	144	81.8	79	81.4
2	140	79.5	77	79.4
3	131	74.4	60	69.1
4	58	33.0	36	37.1

In addition to recording frequencies for annotation use across the documents, we were also interested in identifying and recording frequencies for types of reading strategies used. Specifically, we focused on identifying annotations that were indicative of discipline-specific reading strategies since these types of strategies may be most useful when reading the biology texts in the ILA and completing the tasks that follow. The discipline-specific strategies were counted as present when it was possible to identify them from the text annotations alone.

The biology teachers who scored the ILAs were successful in consistently identifying when a student was using a discipline-specific strategy, but had little agreement when labeling these strategies. The scoring process the teachers undertook was patterned after the History ILA scoring process. The history teachers were able to label history specific strategies they identified with greater agreement across raters. It is possible that the biology teachers struggled because they are less likely to address reading strategies in their classroom, as compared to their history colleagues. Another possibility is that the training needs to be revised to include more opportunities for biology raters to practice scoring student examples of discipline-specific strategies. Given the lack of agreement among the biology content expert raters, we only included strategies in Tables 15, 16, and 17 if two raters were in exact agreement.

While the number of clear biology-specific strategies was relatively small, some patterns did emerge. As shown in Table 16, students in treatment classrooms more frequently made connections between the text and their prior biology knowledge; whereas, control students more frequently considered science implications beyond the scope of the document sections.

Table 16

Frequency of Discipline-Specific Reading Strategies by Status – Biology ILA

	Treatment (*N*=161)		Control (*N*=107)	
Reading Strategy	*N*	%	*N*	%
Connect to prior knowledge	13	8.1	2	1.9
Questioning scientific methods	3	1.9	2	1.9
Attending to and evaluating evidence	1	.6	0	.0
Analyzing graphs, diagrams, etc.	1	.6	0	.0
Considering science implications beyond text scope	10	6.2	11	10.3

Results for the utilization of reading strategies among the history sample are shown in Table 17. Students in treatment classrooms, in comparison to students in control classrooms, more frequently connected to prior knowledge; conducted intertextual reading; identified bias or point of view; placed the document into a historical context; and identified cause and effect. In the following section, each discipline-specific reading strategy is described.

Table 17

Frequency of Discipline-Specific Reading Strategies by Status – History ILA

	Treatment (*N*=176)		Control (*N*=97)	
Reading Strategy	*N*	%	*N*	%
Connect to prior knowledge	42	23.9	12	12.4
Questioning scientific methods	4	2.3	1	1.0
Attending to and evaluating evidence	6	3.4	5	5.2
Analyzing graphs, diagrams, etc.	15	8.5	0	.0
Considering science implications beyond text scope	9	5.1	1	1.0

Connecting to prior knowledge (Biology and History). Students made connections to prior content knowledge and understanding gained from previous learning experiences by writing single words and comments in the margins or embedded in the text. For example, one biology student wrote in the margin next to the diagram of the process of genetic modification in section four, “this experiment was like are experiment with bacteria cultures that we geneticly [*sic*] altered to grow.” This student successfully connected the biological process depicted in the text with a related experiment previously conducted in class.

Another biology student wrote in the margin of section three next to the paragraph describing the process of transferring DNA from one organism to another, “which is called insertion.” One history student underlined the word “lynchings” in the text and then wrote beside the word in the margin “hate crime”. Here the student responded to a section of text with a word that is not included in any of the texts.

Questioning scientific methods (Biology). This strategy included questioning the scientific method and processes presented in the text; student questions ranged from general to critical. One student drew an arrow to the phrase “the most recent technique in biotechnology” in the text and asked “Is the new technique better than the old one?” Another student wrote a series of questions next to the steps of the genetic engineering process diagrammed in section four. First the student asked, “How much time is taken for this process?” Then next to the following step, the student wrote, “Why are reproductive cells harvested?” Finally next to the modified tomato the student wrote, “When was this process first attempted?”

Attending to and evaluating scientific evidence (Biology). Another aspect of disciplinary reading in biology is attending to and evaluating the scientific evidence presented in the texts. Only one instance of evaluating evidence was clearly identified in our sample. This student created a graphic organizer in order to list the points of support and criticism for genetic modification presented in the text. The student’s chart had two columns for criticism and one for support. The student also drew an arrow pointing to one of the criticisms and wrote “bottom line” to evaluate what he/she thought was the most important criticism.

Considering implications beyond text’s scope (Biology). Reading like a scientist requires the ability to consider implications of the content that go beyond its scope. Only one instance of this disciplinary reading strategy was clearly identified in our sample. This student asked multiple questions about topics covered in the text. These questions indicate student thinking related to as well as expanding upon the text’s content. For example, in the first text section, the student wrote in the margin next to the description of how bacteria are used to make different foods like yogurt, “What are some other products that includes [*sic*] using bacteria?” Later, in section five, which discusses criticisms of genetic engineering, the student responded to arguments about genetically modifying corn to resist damaging insects with the following questions, “What other insects can cause damage to corn?” and “What are some beneficial insects?”

Analyzing graphs and diagrams (Biology). This strategy involves analysis of scientific graphs and diagrams presented in the texts. Students annotated the two visual representations of scientific data and processes, including the Mendelian dihybrid cross and the depiction of the steps to genetically modify a tomato. For example, one student analyzed the dihybrid cross and wrote down which genotypes were expressed. Another student circled the crosses with the dominant gene that is expressed. Fewer students annotated the genetic modification process diagram than the dihybrid cross.

Conducting intertextual readings (History). This strategy included identifying information in one of the texts and making a connection to information presented in a different text. Only one instance of intertextual reading was clearly identified from our sample. The student wrote "March on Washington" above the second part of the first sentence in document four, which cited that the racial incidents ranged from "full-scale riots in Detroit, Harlem, and Los Angeles." In this case, the student linked information presented in document four related to the racially motivated violence to the call for a March on Washington information presented in document one.

Evaluating the source of the document (History). An important aspect of disciplinary thinking in history is attending to and evaluating the source of information of a particular document. Thus, one of the reading strategies we looked for was students' ability to evaluate the sources of the documents included in the assessment. Students were able to effectively show their consideration of the source by underlining, circling, or commenting on the source information. For example, one student attended to the fact that some of the information came from a secondary source by underlining "secondary" in the phrase "the following document is a secondary source published in the 1990's". Some students attended to the dates the documents were published, information about the author, or the type of document being read. One student double underlined *The American Newspaper* in the footnote for document two, indicating attention was being paid to the text source.

Identifying bias or point of view (History). Another part of reading like a historian requires the ability to determine the point of view of the author and/or whether a text may be written with a biased perspective. Students most frequently demonstrated this skill by directly questioning sections of the text or the source. For example, one student identified potential bias by drawing a connecting line from the sentence, "Like all true Americans, my greatest desire at this time…"He/She wrote in the margin: "Stereotypical much?" More specifically, the student's connecting line and comment were directed at the portion of the sentence that reads: "true Americans". Students also identified the point of view in the text using connecting lines and comments. One student used this combination of annotations above the sentence “Being an American of dark complexion…these questions flash through my mind: ‘Should I sacrifice my life to live half American?”. The student noted the ethnicity of the author by the use of the sentence segment “dark complexion” and wrote in above those words “‘black’-history of discrimination”. This latter example demonstrates that the student was attending to a possible perspective from which the author was writing.

Placing the document into a historical context (History). This strategy involves considering the place in time of a document’s printing or publication. Using information found in the documents or the source information, students most frequently demonstrated this strategy by commenting on a document’s printing date in relation to the war or questioning a document’s place in time in relation to other historical events. For example, one student attended to document four, which consists of a data table describing the greater Los Angeles population between 1940 and 1950. The student underlined the year in each of the heading columns. Above the underlined years the student wrote, “During WWII”. In a different use of this strategy, a student drew an arrow to the date in the sourcing information for document one indicating the date of press (April 12, 1941) and wrote “before the war ended”. This same student attended to the sourcing information for documents two and three and noted that one was penned “right after we joined the war” and the other “published after riots.”

Identifying cause and effect (History). Determining cause and effect within or between documents is an essential aspect of reading like a historian. In our descriptive analysis, this strategy was more frequently observed being employed with information contained within a document rather than across documents. A sophisticated example of this strategy involved one student’s analysis of document three. The student circled “more than 240 racial incidents” in the first sentence and then drew a connecting arrow to the margin where the student wrote “main idea” and immediately noted beneath “too many racial problems”. Next, the student underlined a section of text that read “such as Harlem, African Americans focused their anger and frustration on property.”, and drew a connecting line to the margin and wrote “Effect! People started protesting”. Finally, this same student double underlined the last sentence of the document “These tensions were exacerbated by wartime migrations, overcrowding in [defense] areas, competition for jobs, and conflict over housing”, and drew a connecting line to the margin space and wrote “Other effects: job competition, housing.”

Conclusion

The aim of the study was to examine the effects of the Reading Apprenticeship (RA) professional development program on teacher practices and student learning. Recognizing the limitations of the existing large scale assessments, a supplementary, more detailed measure was developed to examine the potential effects of the RA program on students' literacy skills specifically embedded in biology and history.

Central to the ILA development process was the notion that subject-specific literacy is demonstrated in both the acquisition of knowledge through the extraction of information from texts and the integration of such information in the formulation of written explanations. These distinct aspects of literacy were examined in separate tasks within the ILA. For example, the RA program emphasizes that reading comprehension is a product of engaging in a thoughtful and strategic reading process. Thus, following the writing task, students were asked to reflect on how they approached each text and the ways they sought to maximize their understanding. Raters examined the extent to which the students' responses showed evidence of some consideration of the thought processes involved in reading. In addition, students' test booklets were evaluated for evidence of specific reading strategies—including note-taking, underlining, and other forms of annotation.

Generalizability theory was used to examine the dependability of the various ILA scores. We found that the multiple choice tests for content knowledge and reading comprehension had rather low reliability that could be improved by dropping problematic items. The scores for metacognition, use of reading strategies, and writing (both language and ability) demonstrated acceptable levels of reliability; thereby, suggesting a fairly consistent application of the scoring rubrics by raters.

The analyses of the ILA scores for the treatment and control groups suggest that assignment to the RA program had little or no effect on class-level student outcomes measured by the ILA. Estimated effect sizes ranged from -.38 (content knowledge) to .29 (reading strategies) but none were statistically significant. That said, there was a significant positive effect of treatment assignment on teachers' literacy instruction; moreover, the level of literacy instruction appeared to have a positive (though not significant) relationship with the student metacognition and reading strategies scores. These proximal student outcomes were, in turn, positively related to scores on more distal outcomes, including content knowledge, reading comprehension, and writing. One possible explanation for the pattern of observed findings could be the possibility that the proposed mechanism that links teacher participation in the RA program to student literacy was flawed. However, we also noted that this model implies multiplication of effects across several steps that may attenuate the relationship between treatment assignment and student outcomes. Heterogeneity in implementation of the curriculum could be another possibility.

Nevertheless, examination of the correlations between variables that are adjacent in the proposed mechanism and other exploratory analyses (e.g., how use of particular reading strategies relate to comprehension and writing scores) suggest that this mechanism is plausible. Specifically, it appears that the particular reading strategies emphasized by the RA program are positively related to subject-specific literacy. Moreover, the use of those strategies seems to be positively related to the sorts of instructional practices that RA teachers are trained to implement; what is more, those instructional practices appear to be related to participation in the RA professional development program.

References

Aguirre-Muñoz, Z., Boscardin, C., Jones, B., Park, J., Chinen, M., Shin, H., et al. (2005). *Consequences and validity of performance assessment for English language learners: Integrating academic language and ELL instructional needs into opportunity- to-learn measures.* (CRESST Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E.L., Aschbacher, P.R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E.L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M.C. Wittrock & E.L. Baker (Eds.), Testing and Cognition (pp.131-153). Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L. (1994). Making performance assessment work: The road ahead. *Educational Leadership, 51* (6), 58-62.

Bertrand, C., DiRanna, K., & Janulaw, A. (1999). *Making connections: A guide to implementing history standards.* Sacramento, CA: California History Teachers Association.

Boulter, C. J., & Gilbert, J. K. (1995). Argument and history education. In P. J. M. Costello & S. Mitchell (Eds.), *Competing and consensual voices: The theory and practice of argumentation.* Clevedon, UK: Multilingual Matters.

Brennan, R.L., & Kane, M.T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277-289.

Chung, G. K., O'Neil, H. F. J., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15* (3-4), 463-494.

Chung, G.K.W.K., Harmon, T.C., & Baker, E.L. (2001, November), The impact of a simulation-based learning design project on student learning. *Education, IEEE Transactions on Computers*, *44*(4), 390-398.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York, NY: John Wiley.

Duschl, R., & Osborne, J. (2002). Supporting and promoting argumentation discourse. *Studies in History Education*, *38*, 39-72.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the use of Toulmin's Argument Pattern in studying history discourse. *History Education*, *88* (6), pp.915-933.

Greenberg, J. (Ed.). (2001). *BSCS history: A molecular approach* (8th ed., Blue Version). Chicago, IL: Everyday Learning.

Halliday, M. A. K. (1994). *An introduction to functional grammar.* (2nd ed.). London, UK: Edward Arnold.

Herman, J. L., Baker, E. L., & Linn, R. L. (2004, spring). Accountability systems in support of student learning: Moving to the next generation. *CRESST Line*. Retrieved from: http://www.cse.ucla.edu/ products/newsletters/CLspring2004.pdf

Martin, K., & Miller, E. (1988). Storytelling and history. *Language Arts*, *65* (3), 255-259.

Miller, K. R., & Levine, J. S. (2006). *Prentice hall history*, (Dragonfly version). New Jersey and Massachusetts: Pearson Prentice Hall.

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

Pontecorvo, C. (1987). Discussing for reasoning: The role of argument in knowledge construction. In E. De Corte, J.G.L.C. Lodewijks, R. Parmentier, & P. Span (Eds.), *Learning and instruction. A publication of the European Association for Research on Learning and Instruction* (pp. 71-82). Oxford, UK: Leuven University Press.

Shavelson, R.J., & Webb, N.M. (1991) *Generalizability theory: A primer.* Thousand Oaks, CA: SAGE.

Appendix A:
Biology ILA (Parts 1, 2, 3)

High School Biology Assessment Part 1

Genetics

Student ID	
Teacher ID	

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

Biology Content

Use what you know from your studies in Biology to answer the following questions. Circle the letter next to the correct response.

A___B___C___D___

1. If the diagram above represents the genes on a chromosome, which genes would have the highest frequency of recombination between them?

a. A and B
b. A and D
c. B and C
d. B and D

2. A process that cannot take place in haploid cells is

a. Mitosis
b. Meiosis
c. Cell division
d. Growth
e. Digestion

3. Mendel hypothesized that reproductive cells have only one factor for each inherited trait. This hypothesis is supported by the observation that:

a. Haploid cells are produced by mitosis
b. Diploid cells are produced by mitosis
c. Haploid cells are produced by meiosis
d. Diploid cells are produced by meiosis

	T	T
t	Tt	Tt
t	Tt	Tt

4. What is the chance that an offspring plant shown in the Punnett Square above will be short? (T = tall; t = short)

a. No chance

b. 1 out of 4

c. 2 out of 4

d. 3 out of 4

e. 4 out of 4

5. Laboratory mice are to be classified based on genes A, B, C. How many genetically different gametes can be formed by a mouse that is genotypically AaBbCc? (Assume that none of these is a lethal gene.)

a. 3

b. 6

c. 8

d. 9

e. 12

6. Tall is dominant over short in a certain plant. A tall plant was crossed with a short plant, and both tall and short offspring were produced. This demonstrates

a. The law of segregation

b. Incomplete dominance

c. Linkage

d. Mutation

e. The law of independent assortment

7. Suppose the sequence of bases along one side of a particular section of DNA is ATGTCAGC. Which of the following is the correct sequence of bases with which this sequence would be paired?

a. CTAGATAT

b. CTAGTGCT

c. TACACTCG

d. TACAGTCG

e. ATGTCAGC

8. All of the following about the structure of DNA are correct EXCEPT

a. DNA is a polymer

b. DNA contains Deoxyribose

c. The two strands are connected by hydrogen bonds

d. Adenine bonds to guanine

e. Nucleotides consist of a sugar, phosphate, and nitrogenous base

9. A clone is an organism that develops from one parent through asexual reproduction, inheriting all the genetic material from that parent. Dolly the sheep was the first animal clone produced by scientists. Which of the following is a clone?

a. A baby born to a mother who was artificially inseminated

b. A colt born to a mare who mated with a racing stallion

c. A coleus plant grown from a stem cutting buried in rooting mixture

d. A pea plant grown from a seed that developed after pollination by another plant

e. A spider hatched from an egg fertilized by the sperm of a male spider

10. Which of the following involves taking DNA from two sources and putting it into one cell?

a. Gel electrophoresis

b. Restriction enzymes

c. Polymerase chain reaction

d. Recombinant DNA

High School Biology
Assessment Part 2

Genetics

Student ID	
Teacher ID	

CRESST
National Center for Research on Evaluation, Standards, and Student Testing

In this assessment you will be asked to complete a biology reading task about using biotechnology to improve food production. This is an assessment of your reading in biology. You will have one period to complete the assessment.

Thinking ahead: In Assessment Part 3, half of the class will go on to write an essay in response to the document in this assessment, while the other half will complete additional reading tasks.

Reading Task Directions

Please carefully read the following documents written about using biotechnology to improve food production*. There are five sections.* As you read, consider the five sections individually, but also think about how they relate to one another.

Show your thinking about the reading by taking *notes in the margins or on the texts. These notes will be scored as part of the assessment of your reading.*

Then, respond to the multiple choice and short answer questions after each section of the document.

Section 1: Read carefully, record your thinking about the reading, and answer the questions that follow the section.

The following section was adapted from an online lesson on crop biotechnology and an online article written on genetically modified organisms.

Using Biotechnology to Improve Food Production

The Scoop on Biotechnology

What is "biotechnology"? Biotechnology can be defined in a number of ways. First, it can be defined as "the use of biotechnical methods to modify the genetic material of living cells so they will produce new substances or perform new functions." Second, it can also refer to genetic engineering technology of the 21st century used to directly manipulate the genes of organisms, such as moving or transferring genetic material between sources. Finally, the broadest definition of biotechnology is the use of living organisms to make a product or conduct a process. This includes using bacteria to make yogurt, cheese, and vinegar as well as the use of plant or animal cross-breeding techniques or genetic engineering to produce food with enhanced qualities. Therefore, methods of biotechnology include both the indirect and direct manipulation of genes, such as in the traditional cross-breeding and selective breeding in plants and animals, as well as in engineered recombinant DNA.

The link between biotechnology and food dates back almost 10,000 years, yet scientific experiments with biotechnology were not recorded in writing until the 1860's. Gregor Mendel was the first to document the results of his experiments in the carefully planned traditional cross-breeding of garden peas. Mendel used mathematics to conclude that each true-breeding pea plant had two identical copies of an allele for a particular trait. During meiosis, only one copy of each allele went into each pollen or egg cell. He referred to this separation of alleles in the first generation (F1) as the principle of segregation.

Since Mendel's time, traditional cross breeding has been used to develop lines of plants with desired qualities, such as orchids with brilliant color. Unfortunately, cross-breeding indirectly transfers many unwanted traits along with the trait of interest, and continued selective breeding is necessary to rid the new plant of these unwanted traits.

Text continues on page 5

Sources:
1) Jennifer Flak and Julie Albrecht, Department of Agronomy and Horticulture at the University of Nebraska.
2) Genetically Modified Organisms, Institute of Food Technologists. Internet publication.

Section 1 questions:

1. What does the following phrase refer to in the section? “During meiosis, only one copy of each allele went into each pollen or egg cell.”

A. The process of genetic engineering

B. The principle of segregation

C. The enhancement of the nutrient content of food

D. The genetic limitations of blueberries

2. In this section “to modify the genetic material of living cells” involves?

A. Adding one organism’s cells into another organism

B. Taking care of living cells as they modify themselves

C. Helping organisms become resistant to bacteria

D. Moving or transfer genetic material between sources

3. According to the section, traditional cross breeding:

A. Has not been used to develop plants with desired qualities

B. Involves the removal and transfer of DNA from one organism into another

C. Transfers many unwanted traits along with the trait of interest

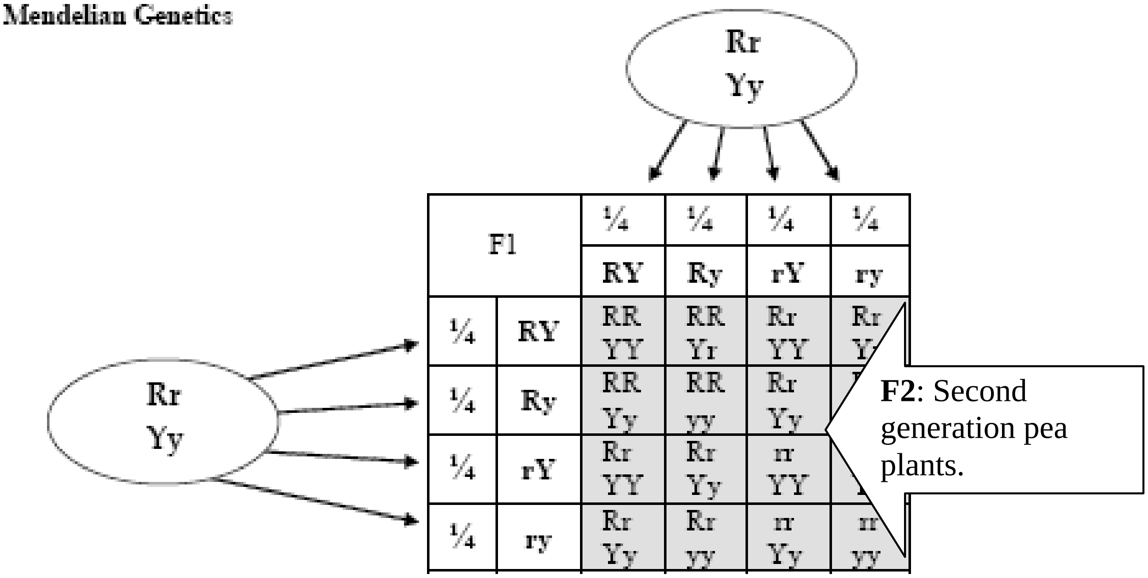
D. Manufactures DNA to create new organisms

Section 2: Read carefully, record your thinking about the reading, and answer the questions that follow the section.

The following section is a figure illustrating traditional cross-breeding and Mendelian genetics, including information about the first and second generation of pea plants and the principles of segregation and independent assortment.

continued from page 3.

Biotechnology: Traditional Cross-breeding and Mendelian Genetics



F1		¼	¼	¼	¼
		RY	**Ry**	**rY**	**ry**
¼	**RY**	RR YY	RR Yr	Rr YY	Rr Y
¼	**Ry**	RR Yy	RR yy	Rr Yy	
¼	**rY**	Rr YY	Rr Yy	rr YY	
¼	**ry**	Rr Yy	Rr yy	rr Yy	rr yy

F2: Second generation pea plants.

Figure. 1. A dihybrid cross illustrates Mendelian principles of segregation and independent assortment. Each pea plant has two alleles for each trait. Round pea (R) is dominant over wrinkled pea (r), and yellow pea (Y) is dominant over green pea (y). This is how Mendel might have illustrated the way that alleles of the same trait segregate from each other and alleles of different traits sort independently, during meiosis.

Text continues on page 7

Source:
BSCS Biology: A Molecular Approach. 8th Edition. Everyday Learning Corp., 2001. P 352-356.

Section 2 questions:

1. The first generation (F1) pea plants are:
   a. Homozygous for shape and homozygous for color with round, yellow seeds
   b. Heterozygous for shape and heterozygous for color with round, yellow seeds
   c. Homozygous for shape and homozygous for color with wrinkled, green seeds
   d. Heterozygous for shape and heterozygous for color with wrinkled, green seeds

2. Of the second generation (F2) pea plants, how many will have phenotype of round peas?
   a. 8
   b. 4
   c. 9
   d. 12

3. Which of the following statements is true for a dihybrid cross?
   a. Each parent pea plant has two genes from which to contribute alleles to their offspring
   b. Each offspring inherits two genes from one parent and one from the other
   c. There are 16 possible genotypes of the offspring
   d. None of the above statements are true of dihybrid crosses

Section 3: Read carefully, record your thinking about the reading, and answer the questions that follow the section.

The following section was adapted from online lessons on crop technology and the genetic modification of organisms. It also contains information from a written publication about genetically modified organisms.

Continued from page 5

The Scoop on Genetic Engineering

The most recent technique in biotechnology is sometimes referred to as genetic engineering. It was developed in 1973 and refers to the ability to directly transfer genetic information between organisms using molecular technology. Genetic engineering physically removes the DNA code for a particular gene from one organism and transfers it into the genome of another organism. A gene holds information that will give an organism a trait. Using this method, a single trait can be added to an organism at a time, making it much more efficient than traditional cross breeding.

Genetic engineering in this sense has been used in many areas related to food and nutrition. A recent focus of genetic engineering techniques has been to enhance the nutrient content of food. This area includes the development of oils with reduced saturated fat content and rice that has been modified to have high carotene levels (a vitamin A precursor). One early experiment attempted to alter tomatoes for the purpose of increasing their cold resistance, thus allowing a longer growing season. While not completely successful, this experiment illustrates the high hopes that scientists have for using biotechnology for the betterment of society. A group of California scientists used genetic engineering techniques to create a synthetic gene based on a specific flounder fish gene (that enables the fish to survive in very cold ocean waters) and inserted it into the DNA of a tomato seed. Today, scientists continue to investigate how genetic engineering can be used to improve quality of life.

Text continues on page 9

Sources:
1) Jennifer Flak and Julie Albrecht, Department of Agronomy and Horticulture at the University of Nebraska.
2) Genetically Modified Organisms, Institute of Food Technologists. Internet publication.
3) Online lesson in genetic modification of organisms. Science Enhancement Programme, UK

Section 3 questions:

According to the section, genetic engineering can be used to:

A. Crossbreed rice to increase carotene

B. Develop oils with reduced saturated fat content

C. Crossbreed tomatoes in a carefully planned experiment

D. Divide alleles for one trait among gametes during meiosis

2. Which statement below most closely expresses the main idea of the section?

A. The principle of segregation developed in the late 1800s was the most recent innovation in biotechnology.

B. Mendel laid the groundwork for scientists to genetically modify organisms.

C. Genetic engineering can efficiently add desired traits to food plants by transferring a single gene from another organism to the food plant.

D. Scientists prefer to transfer large quantities of genetic information from one organism to another, giving them a wide selection of genes from which to choose when genetically modifying an organism**.**

Section 4: Read carefully, record your thinking about the reading, and answer the questions that follow the section.

This section is a figure illustrating a form of biotechnology called genetic engineering using recombinant DNA. Included is information on the general step by step process used to genetically modify a tomato in an attempt to make it more resistant to cold temperatures.

Continued from page 7

Biotechnology: Genetic engineering and recombinant DNA

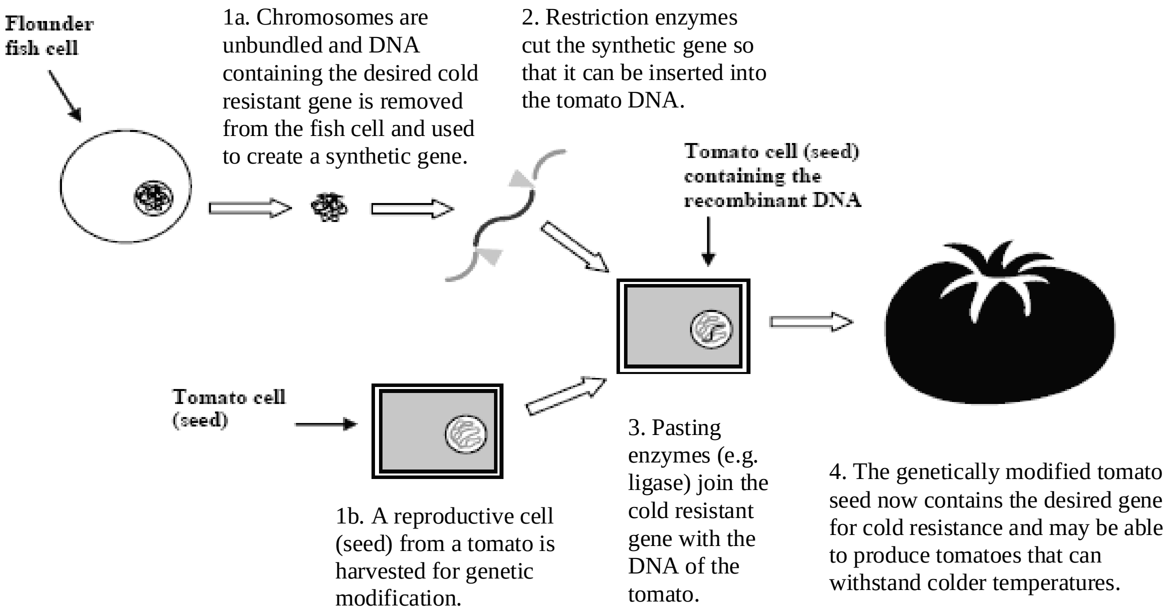

Flounder fish cell
1a. Chromosomes are unbundled and DNA containing the desired cold resistant gene is removed from the fish cell and used to create a synthetic gene.
2. Restriction enzymes cut the synthetic gene so that it can be inserted into the tomato DNA.
Tomato cell (seed) containing the recombinant DNA
Tomato cell (seed)
1b. A reproductive cell (seed) from a tomato is harvested for genetic modification.
3. Pasting enzymes (e.g. ligase) join the cold resistant gene with the DNA of the tomato.
4. The genetically modified tomato seed now contains the desired gene for cold resistance and may be able to produce tomatoes that can withstand colder temperatures.

Figure. 2. The illustration above shows the process of splicing (joining) DNA containing the antifreeze gene and DNA from a tomato for the purpose of increasing the tomato’s resistance to cold. First, the gene for cold resistance is extracted from the genome of the flounder fish and used to create a synthetic antifreeze gene. Using restriction enzymes and pasting enzymes, the antifreeze gene is cut and pasted with another piece of DNA called a plasmid. This hybrid DNA, which joins DNA from two different sources, is called recombinant DNA. The recombinant DNA is inserted into a bacterium that infects a tomato cell, transferring the hybrid DNA and integrating the cold resistance gene into the tomato DNA. The tomato cell’s genome now contains the integrated gene for cold resistance and can be encouraged to grow into a tomato plant.

Text continues on page 11

1) Online lesson in genetic modification of organisms. Science Enhancement Programme, UK
2) J Dale and M von Schantz. From Genes to Genomes: Concepts and Applications of DNA Technology*.*

Section 4 questions:

1. In order for the tomato to acquire the flounder fish's cold resistant trait, a scientist needs to add the following to the tomato's DNA:

a. The genome of a flounder fish

b. Unbundled flounder chromosomes

c. A specific antifreeze gene

d. Pasting enzymes

2. Which of the following statements is true about this process?

a. The tomato will taste like flounder

b. The tomato seed can be used to produce tomatoes that may survive colder temperatures than before

c. The flounder fish egg will produce a fish that requires warmer water

d. Restriction enzymes join the antifreeze gene with the tomato's DNA

3. Parts of this section were complex. What did you do as you were reading to improve your understanding? Please be as detailed as possible.

____

____

____

____

____

____

CRESST

WestEd

High School Biology
Assessment Part 3

Genetics

Student ID	
Teacher ID	

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

Assessment Overview

This is an assessment of your writing in biology as well as the biology content you have learned. For this writing task you will write an essay using information from the sections you read in Assessment Part 2 and from your studies in Biology. <u>You will need the texts from Assessment Part 2 from your teacher so that you can refer to the text sections while working on your essay</u>.

Writing Task Directions

Imagine that you are a biologist. A potato farmer has come to you for advice on how to protect her crop from being destroyed by the Colorado potato beetle. The farmer has just read the text sections on biotechnology (from Assessment Part 2) and wants to modify her potatoes to include a gene that will make them resistant to the beetle using genetic engineering (recombinant DNA biotechnology). The gene for *crown gall disease* is found in an organism called *agrobacterium*. When expressed in potato plants, the gene is harmless to the plants, but deadly to beetles that eat the potatoes. Write an essay that explains to the farmer:

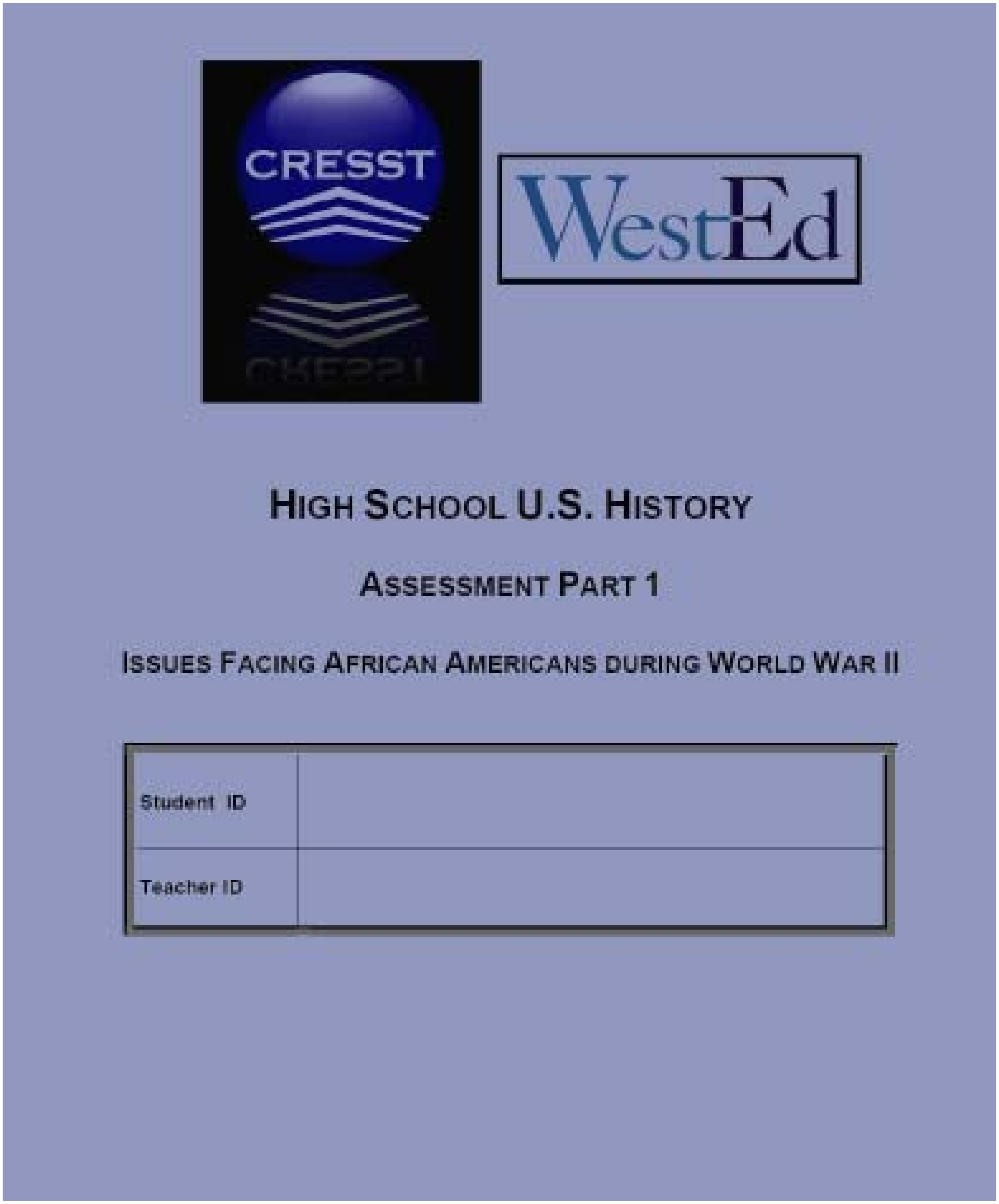
- *how scientists might use recombinant DNA biotechnology to genetically modify potatoes for the purpose of making them more resistant to beetles*; and

- *includes a discussion on (1) the general structure and function of DNA and genes, (2) the benefit of using recombinant DNA compared to traditional biotechnology, and (3) why modifying an organism's genome results in the modification of the entire organism.*

Your task is to write a **science-based essay** that:

1. addresses both of the above bullets;
2. incorporates information from **at least two of the text sections**;
3. includes **relevant knowledge that you have learned from biology class**; and
4. uses your own words, whenever possible

Appendix B:
History ILA (Parts 1, 2, 3)

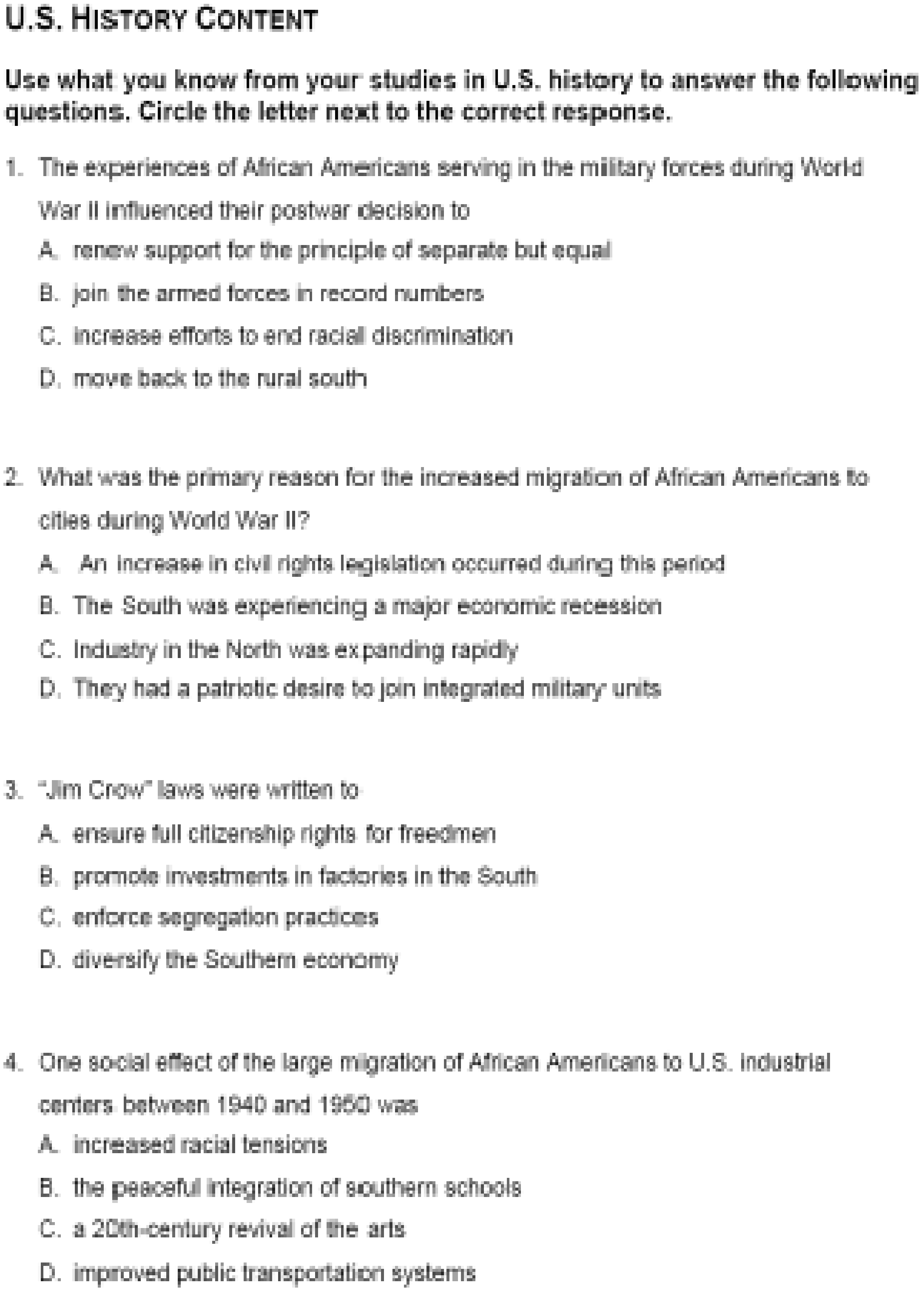


High School U.S. History

Assessment Part 1

Issues Facing African Americans during World War II

Student ID	
Teacher ID	



U.S. History Content

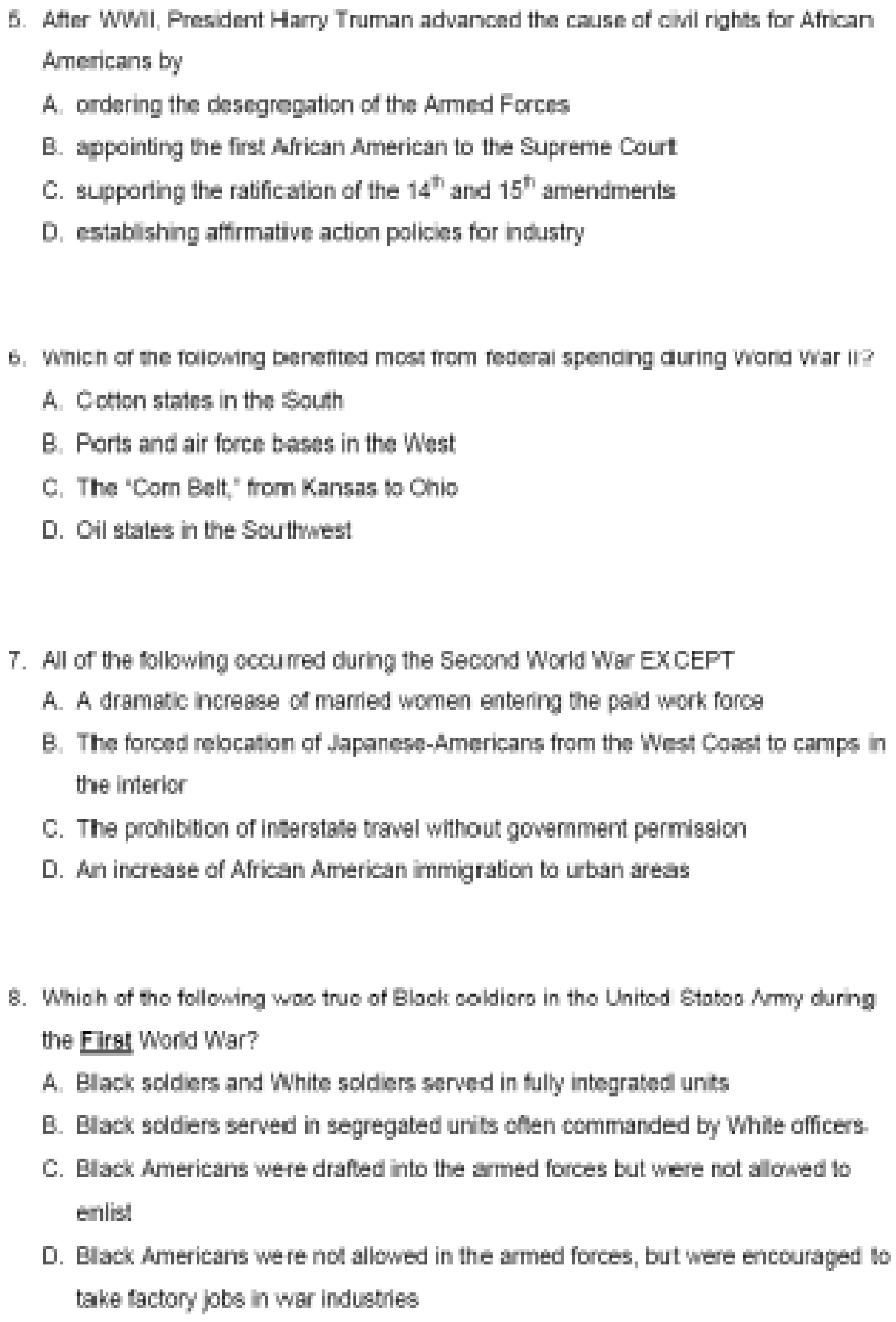
Use what you know from your studies in U.S. history to answer the following questions. Circle the letter next to the correct response.

1. The experiences of African Americans serving in the military forces during World War II influenced their postwar decision to
   A. renew support for the principle of separate but equal
   B. join the armed forces in record numbers
   C. increase efforts to end racial discrimination
   D. move back to the rural south

2. What was the primary reason for the increased migration of African Americans to cities during World War II?
   A. An increase in civil rights legislation occurred during this period
   B. The South was experiencing a major economic recession
   C. Industry in the North was expanding rapidly
   D. They had a patriotic desire to join integrated military units

3. "Jim Crow" laws were written to
   A. ensure full citizenship rights for freedmen
   B. promote investments in factories in the South
   C. enforce segregation practices
   D. diversify the Southern economy

4. One social effect of the large migration of African Americans to U.S. industrial centers between 1940 and 1960 was
   A. increased racial tensions
   B. the peaceful integration of southern schools
   C. a 20th-century revival of the arts
   D. improved public transportation systems

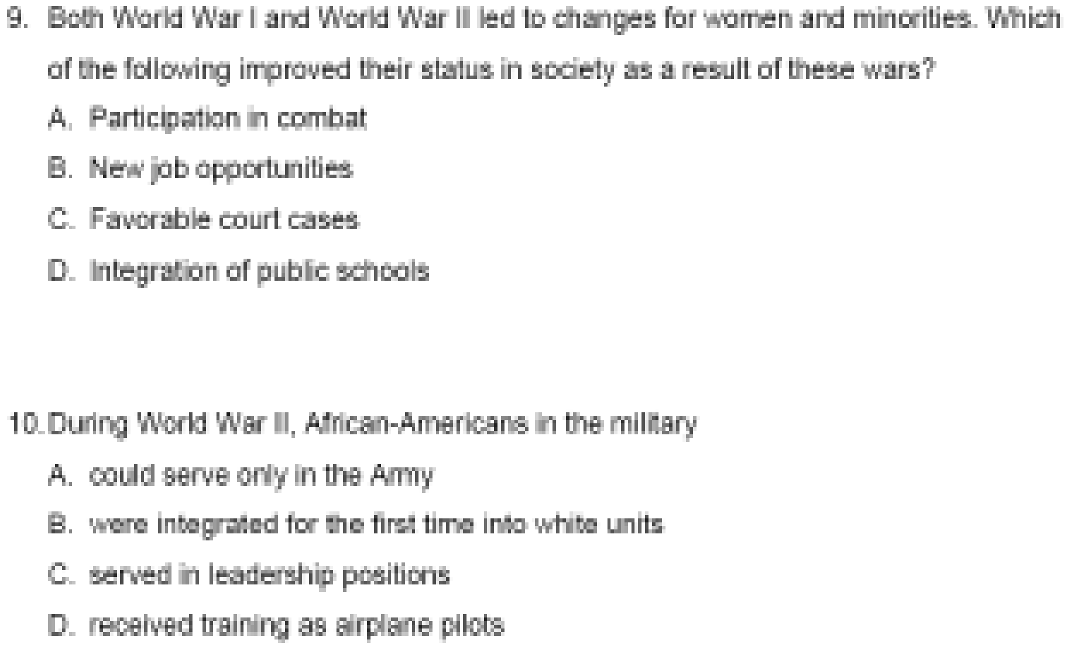


5. After WWII, President Harry Truman advanced the cause of civil rights for African Americans by
   - A. ordering the desegregation of the Armed Forces
   - B. appointing the first African American to the Supreme Court
   - C. supporting the ratification of the 14th and 15th amendments
   - D. establishing affirmative action policies for industry

6. Which of the following benefited most from federal spending during World War II?
   - A. Cotton states in the South
   - B. Ports and air force bases in the West
   - C. The "Corn Belt," from Kansas to Ohio
   - D. Oil states in the Southwest

7. All of the following occurred during the Second World War EXCEPT
   - A. A dramatic increase of married women entering the paid work force
   - B. The forced relocation of Japanese-Americans from the West Coast to camps in the interior
   - C. The prohibition of interstate travel without government permission
   - D. An increase of African American immigration to urban areas

8. Which of the following was true of Black soldiers in the United States Army during the **First** World War?
   - A. Black soldiers and White soldiers served in fully integrated units
   - B. Black soldiers served in segregated units often commanded by White officers.
   - C. Black Americans were drafted into the armed forces but were not allowed to enlist
   - D. Black Americans were not allowed in the armed forces, but were encouraged to take factory jobs in war industries



9. Both World War I and World War II led to changes for women and minorities. Which of the following improved their status in society as a result of these wars?
   A. Participation in combat
   B. New job opportunities
   C. Favorable court cases
   D. Integration of public schools

10. During World War II, African-Americans in the military
   A. could serve only in the Army
   B. were integrated for the first time into white units
   C. served in leadership positions
   D. received training as airplane pilots

CRESST

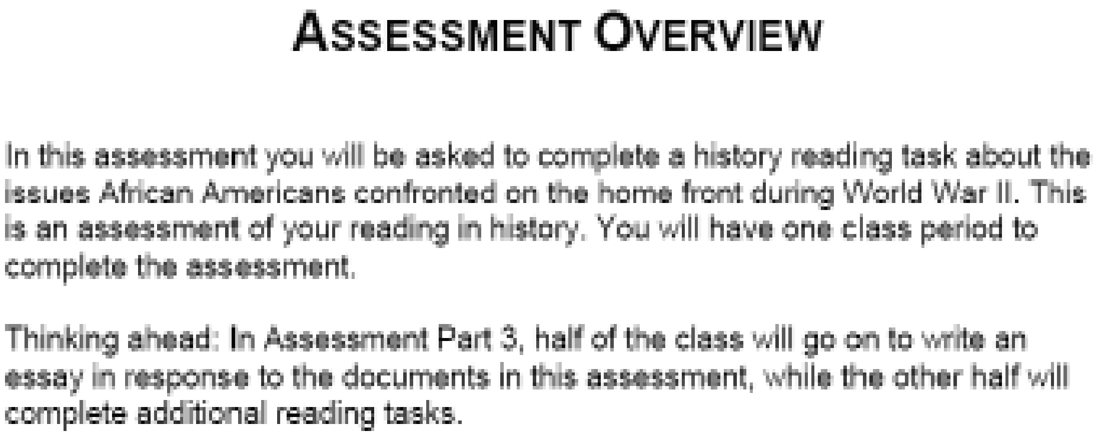
WestEd

High School U.S. History

Assessment Part 2

Issues Facing African Americans during World War II

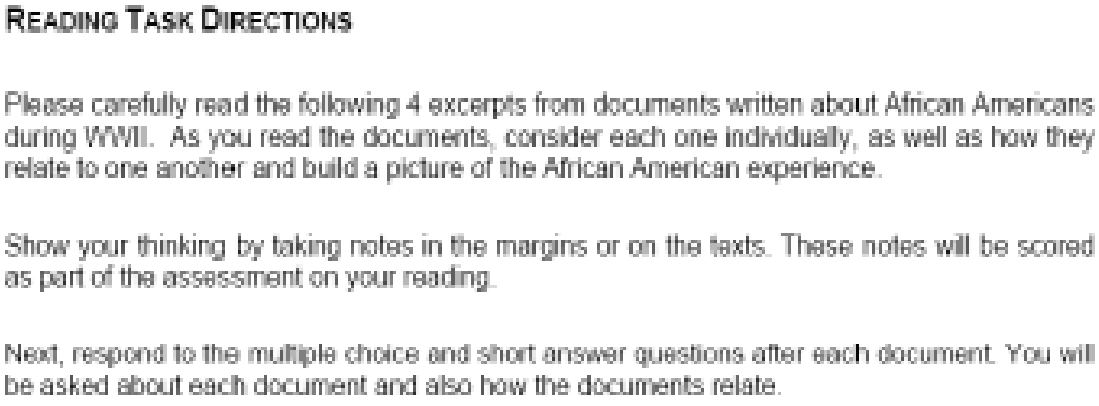
Student ID	
Teacher ID	



Assessment Overview

In this assessment you will be asked to complete a history reading task about the issues African Americans confronted on the home front during World War II. This is an assessment of your reading in history. You will have one class period to complete the assessment.

Thinking ahead: In Assessment Part 3, half of the class will go on to write an essay in response to the documents in this assessment, while the other half will complete additional reading tasks.



Reading Task Directions

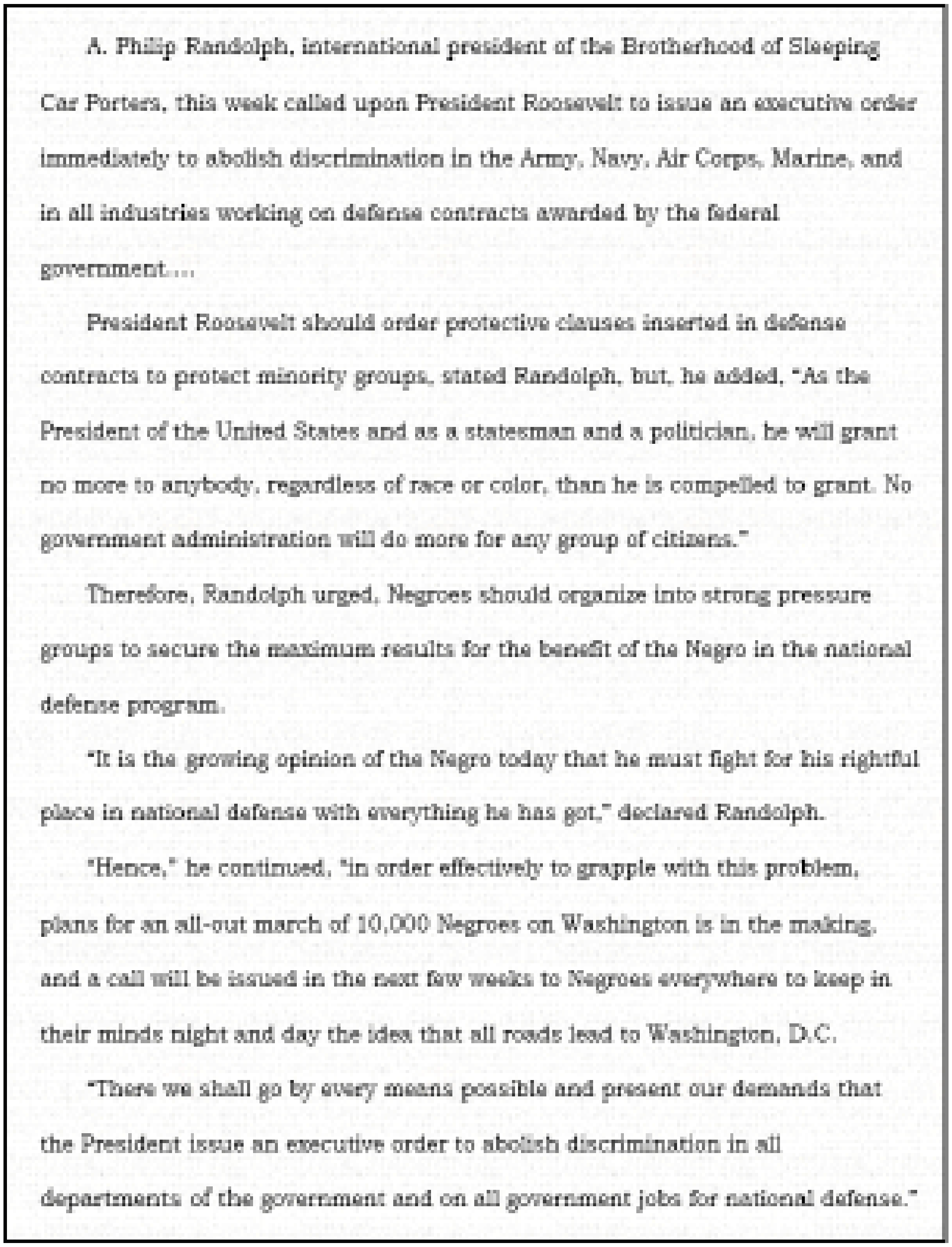
Please carefully read the following 4 excerpts from documents written about African Americans during WWII. As you read the documents, consider each one individually, as well as how they relate to one another and build a picture of the African American experience.

Show your thinking by taking notes in the margins or on the texts. These notes will be scored as part of the assessment on your reading.

Next, respond to the multiple choice and short answer questions after each document. You will be asked about each document and also how the documents relate.

DOCUMENT 1: Read carefully, record your thinking, and answer the questions that follow.

The following document is an excerpt from a newspaper article published less than one year before the United States entered World War II.



A. Philip Randolph, international president of the Brotherhood of Sleeping Car Porters, this week called upon President Roosevelt to issue an executive order immediately to abolish discrimination in the Army, Navy, Air Corps, Marine, and in all industries working on defense contracts awarded by the federal government....

President Roosevelt should order protective clauses inserted in defense contracts to protect minority groups, stated Randolph, but, he added, "As the President of the United States and as a statesman and a politician, he will grant no more to anybody, regardless of race or color, than he is compelled to grant. No government administration will do more for any group of citizens."

Therefore, Randolph urged, Negroes should organize into strong pressure groups to secure the maximum results for the benefit of the Negro in the national defense program.

"It is the growing opinion of the Negro today that he must fight for his rightful place in national defense with everything he has got," declared Randolph.

"Hence," he continued, "in order effectively to grapple with this problem, plans for an all-out march of 10,000 Negroes on Washington is in the making, and a call will be issued in the next few weeks to Negroes everywhere to keep in their minds night and day the idea that all roads lead to Washington, D.C.

"There we shall go by every means possible and present our demands that the President issue an executive order to abolish discrimination in all departments of the government and on all government jobs for national defense."

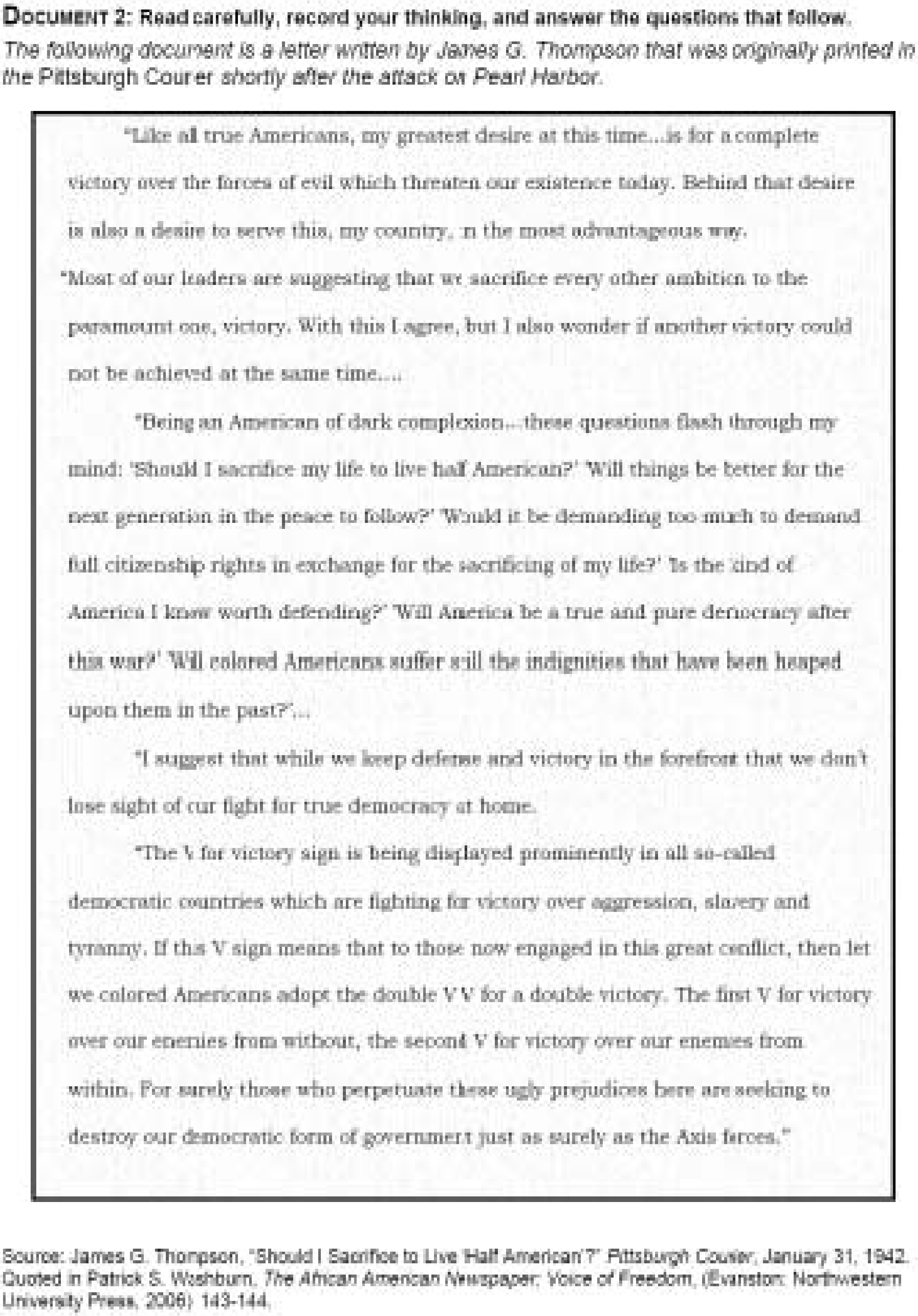
Source: "A.P. Randolph In Appeal To F.D.R. On Bias," *The Chicago Defender (National edition)*, April 12, 1941.

Document 1 Questions:

1. What did Philip Randolph believe President Roosevelt would be likely to do regarding discrimination without being pressured?
   A. Issue an executive order to abolish discrimination in the military
   B. As little as possible
   C. Support a march on Washington
   D. Fight for African Americans' rightful place in national defense

2. How did Randolph view politicians?
   A. As proven allies in the fight against racism
   B. As enemies that must be ignored
   C. As self-interested, but potential agents of change
   D. As immovable

3. Why was 1941 an opportune time for the march?
   A. Because African Americans were already a central part of the U.S. military in Europe
   B. Because of the recent passage of the Civil Rights Act
   C. Because of the need to improve military readiness
   D. Because the African American population was now over 10,000 in Washington



DOCUMENT 2: Read carefully, record your thinking, and answer the questions that follow.

The following document is a letter written by James G. Thompson that was originally printed in the Pittsburgh Courer *shortly after the attack on Pearl Harbor.*

"Like all true Americans, my greatest desire at this time...is for a complete victory over the forces of evil which threaten our existence today. Behind that desire is also a desire to serve this, my country, :n the most advantageous way.

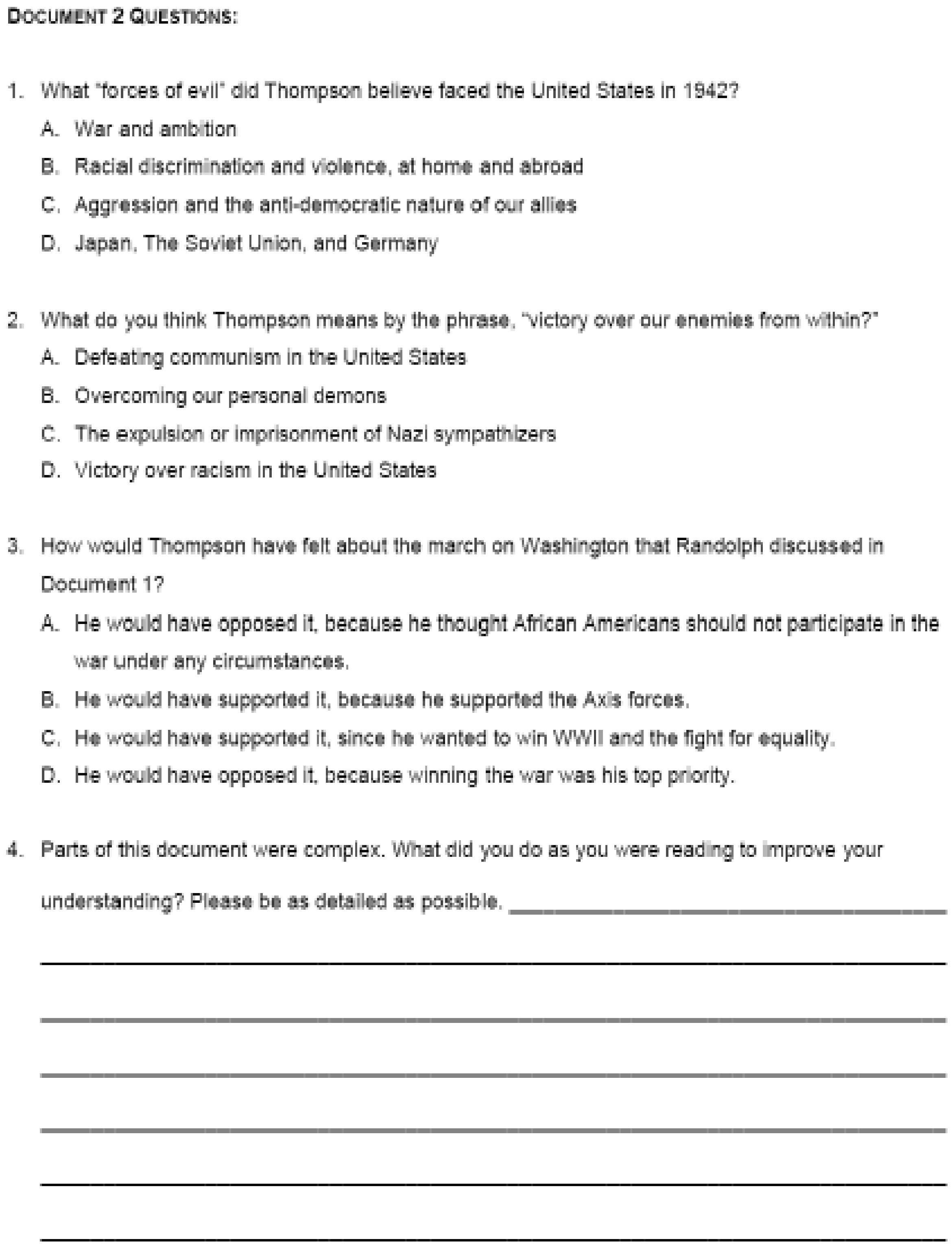
"Most of our leaders are suggesting that we sacrifice every other ambition to the paramount one, victory. With this I agree, but I also wonder if another victory could not be achieved at the same time....

"Being an American of dark complexion...these questions flash through my mind: 'Should I sacrifice my life to live half American?' 'Will things be better for the next generation in the peace to follow?' 'Would it be demanding too much to demand full citizenship rights in exchange for the sacrificing of my life?' 'Is the kind of America I know worth defending?' 'Will America be a true and pure democracy after this war?' 'Will colored Americans suffer still the indignities that have been heaped upon them in the past?'...

"I suggest that while we keep defense and victory in the forefront that we don't lose sight of cur fight for true democracy at home.

"The V for victory sign is being displayed prominently in all so-called democratic countries which are fighting for victory over aggression, slavery and tyranny. If this V sign means that to those now engaged in this great conflict, then let we colored Americans adopt the double V V for a double victory. The first V for victory over our enemies from without, the second V for victory over our enemies from within. For surely those who perpetuate these ugly prejudices here are seeking to destroy our democratic form of governmert just as surely as the Axis forces."

Source: James G. Thompson, "Should I Sacrifice to Live 'Half American'?" *Pittsburgh Couier*, January 31, 1942. Quoted in Patrick S. Washburn, *The African American Newspaper: Voice of Freedom*, (Evanston: Northwestern University Press, 2006) 143-144.



DOCUMENT 2 QUESTIONS:

1. What "forces of evil" did Thompson believe faced the United States in 1942?
   A. War and ambition
   B. Racial discrimination and violence, at home and abroad
   C. Aggression and the anti-democratic nature of our allies
   D. Japan, The Soviet Union, and Germany

2. What do you think Thompson means by the phrase, "victory over our enemies from within?"
   A. Defeating communism in the United States
   B. Overcoming our personal demons
   C. The expulsion or imprisonment of Nazi sympathizers
   D. Victory over racism in the United States

3. How would Thompson have felt about the march on Washington that Randolph discussed in Document 1?
   A. He would have opposed it, because he thought African Americans should not participate in the war under any circumstances.
   B. He would have supported it, because he supported the Axis forces.
   C. He would have supported it, since he wanted to win WWII and the fight for equality.
   D. He would have opposed it, because winning the war was his top priority.

4. Parts of this document were complex. What did you do as you were reading to improve your understanding? Please be as detailed as possible. ______________________________

____

____

____

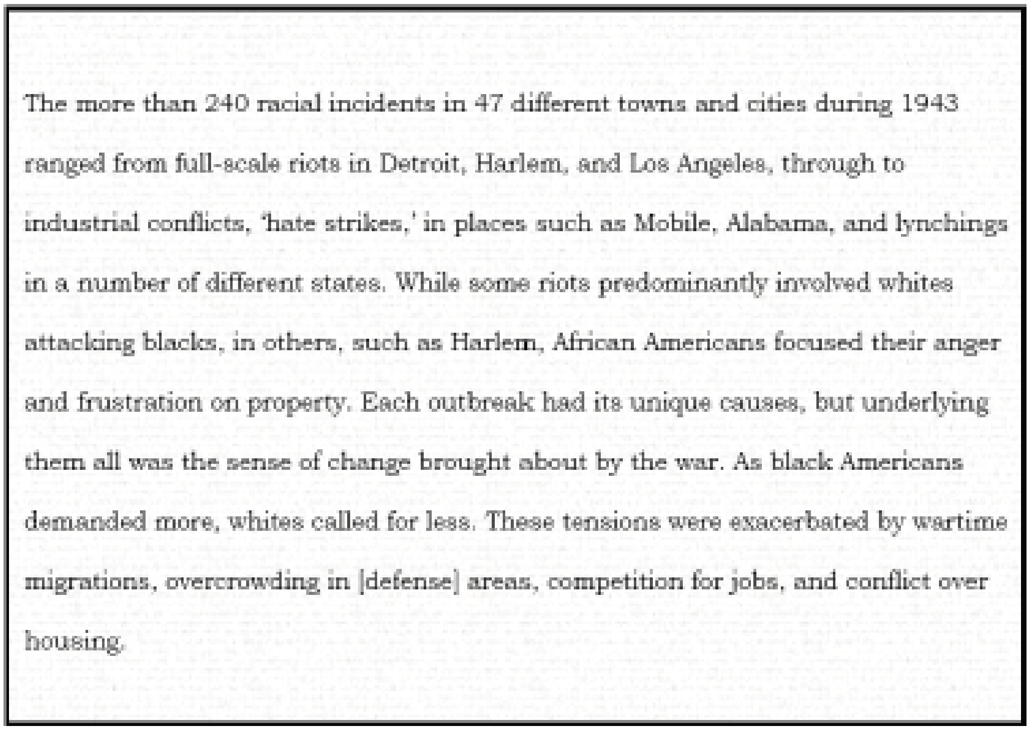
____

____

____

DOCUMENT 3: Read carefully, record your thinking, and answer the questions that follow.

The following document is a secondary source published in the 1990s that documents incidents of racial violence in the United States during 1943.



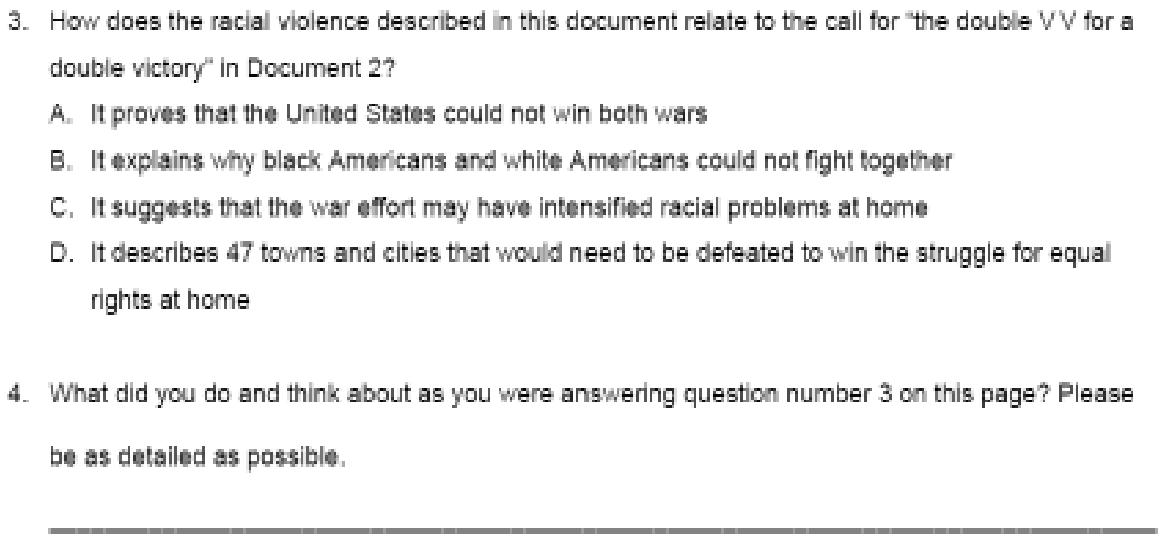
The more than 240 racial incidents in 47 different towns and cities during 1943 ranged from full-scale riots in Detroit, Harlem, and Los Angeles, through to industrial conflicts, 'hate strikes,' in places such as Mobile, Alabama, and lynchings in a number of different states. While some riots predominantly involved whites attacking blacks, in others, such as Harlem, African Americans focused their anger and frustration on property. Each outbreak had its unique causes, but underlying them all was the sense of change brought about by the war. As black Americans demanded more, whites called for less. These tensions were exacerbated by wartime migrations, overcrowding in [defense] areas, competition for jobs, and conflict over housing.

Source: Neil A. Wynn, "The 'Good War:' The Second World War and Postwar American Society," *Journal of Contemporary History* 21.3 (July 1996): 472.

DOCUMENT 3 QUESTIONS:

1. According to the document, what is meant by "wartime migrations?"
   - A. The flight to safety
   - B. Movement of people to cities
   - C. Draft-dodging
   - D. Illegal immigration

2. Racial conflicts took all of the following forms in 1943, EXCEPT
   - A. White people attacking black people
   - B. Emigration to Africa
   - C. Destruction of property
   - D. Housing disputes



3. How does the racial violence described in this document relate to the call for "the double V V for a double victory" in Document 2?

   A. It proves that the United States could not win both wars

   B. It explains why black Americans and white Americans could not fight together

   C. It suggests that the war effort may have intensified racial problems at home

   D. It describes 47 towns and cities that would need to be defeated to win the struggle for equal rights at home

4. What did you do and think about as you were answering question number 3 on this page? Please be as detailed as possible.

____

____

____

____

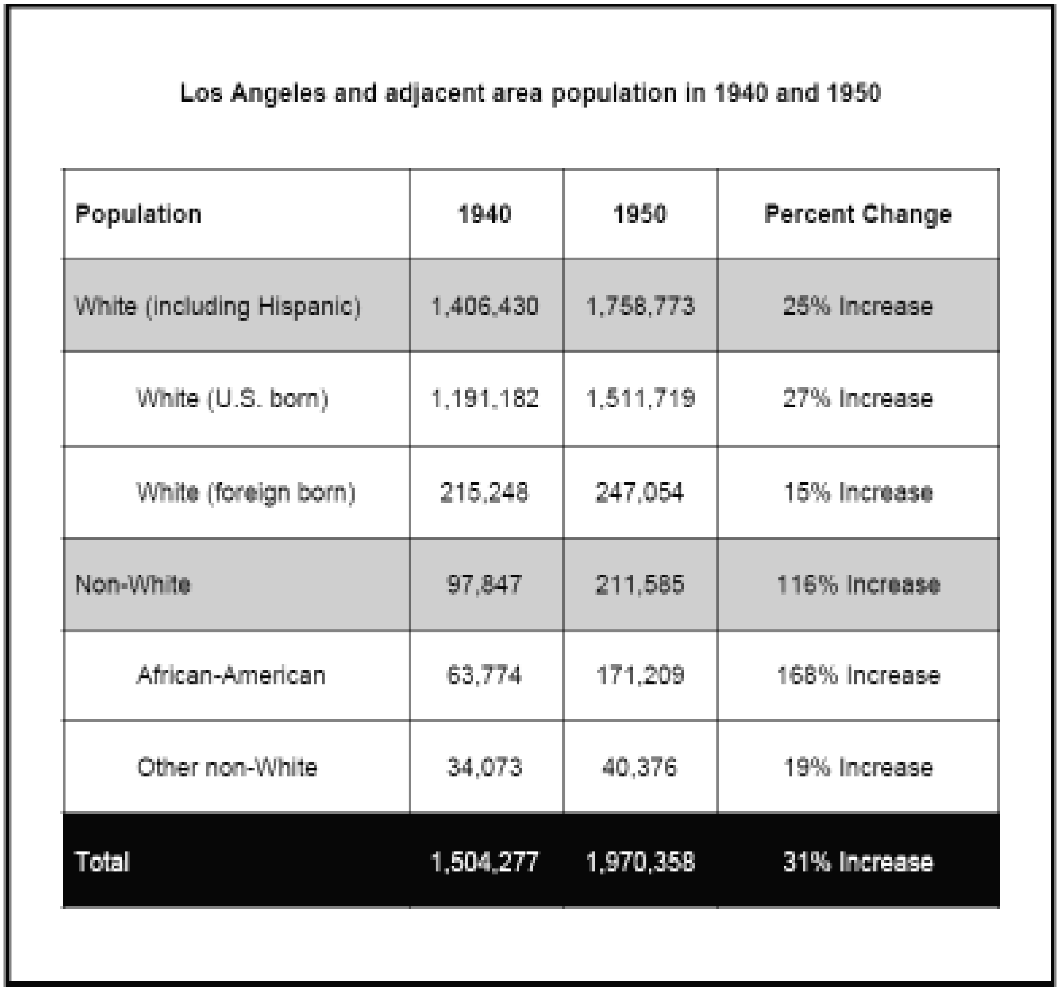
____

____

____

DOCUMENT 4: Read carefully, record your thinking, and answer the questions that follow.

The following document is a table describing the greater Los Angeles area's population data from 1940-1950, including information about race and nationality.



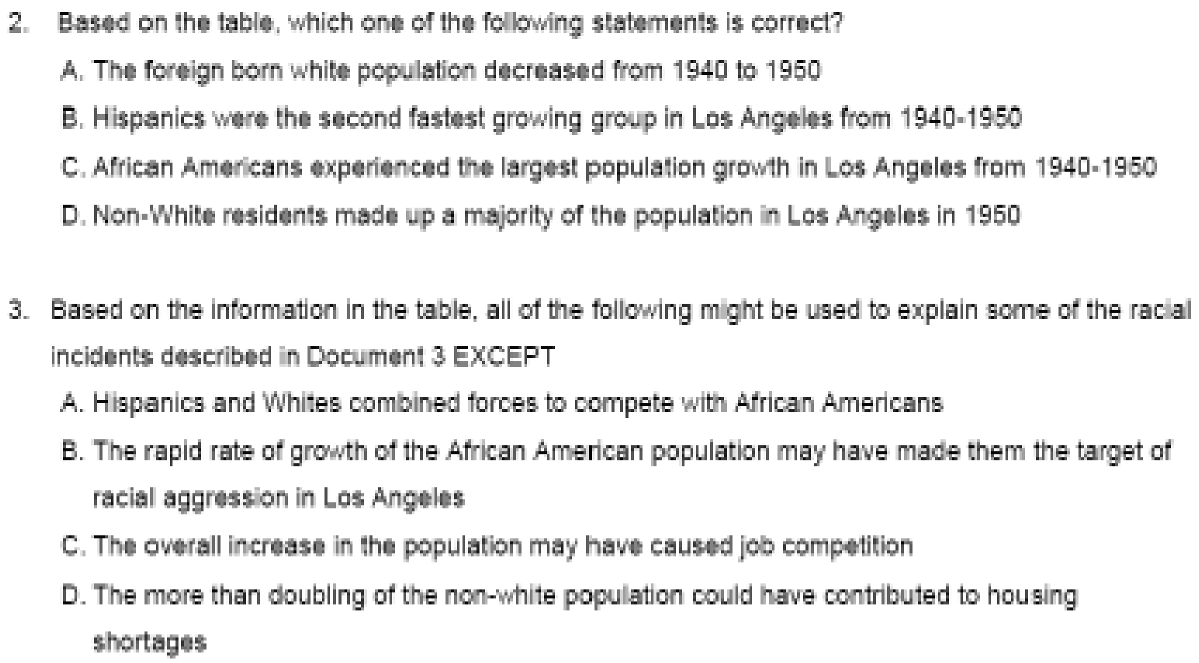
Los Angeles and adjacent area population in 1940 and 1950

Population	1940	1950	Percent Change
White (including Hispanic)	1,406,430	1,758,773	25% Increase
White (U.S. born)	1,191,182	1,511,719	27% Increase
White (foreign born)	215,248	247,054	15% Increase
Non-White	97,847	211,585	116% Increase
African-American	63,774	171,209	168% Increase
Other non-White	34,073	40,376	19% Increase
Total	1,504,277	1,970,358	31% Increase

Source: U.S. Bureau of the Census. *Population and Housing Statistics for Census Tracts*. U.S. Government Printing Office, Washington, D.C, 1942; U.S. Bureau of the Census. *U.S. Census of Population: 1950*. Vol. III. *Census Tract Statistics*. Chap. 28. U.S. Government Printing Office, Washington, D.C. 1952.

DOCUMENT 4 QUESTIONS:

1. Which group made up the largest number of people in Los Angeles in 1940?
   A. White (foreign born)
   B. White (U.S. born)
   C. African-American
   D. Other non-White



2. Based on the table, which one of the following statements is correct?

   A. The foreign born white population decreased from 1940 to 1950

   B. Hispanics were the second fastest growing group in Los Angeles from 1940-1950

   C. African Americans experienced the largest population growth in Los Angeles from 1940-1950

   D. Non-White residents made up a majority of the population in Los Angeles in 1950

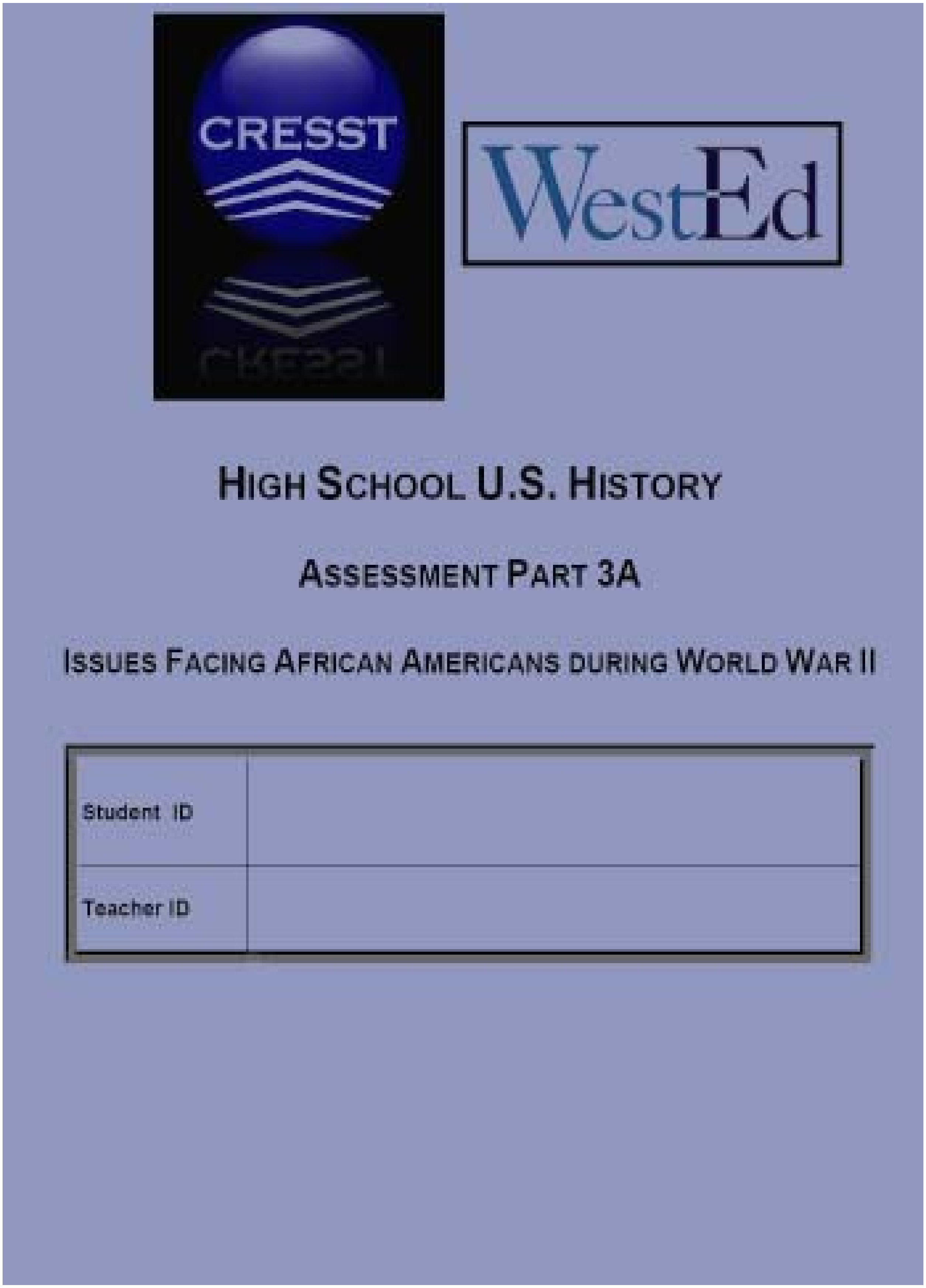
3. Based on the information in the table, all of the following might be used to explain some of the racial incidents described in Document 3 EXCEPT

   A. Hispanics and Whites combined forces to compete with African Americans

   B. The rapid rate of growth of the African American population may have made them the target of racial aggression in Los Angeles

   C. The overall increase in the population may have caused job competition

   D. The more than doubling of the non-white population could have contributed to housing shortages



CRESST

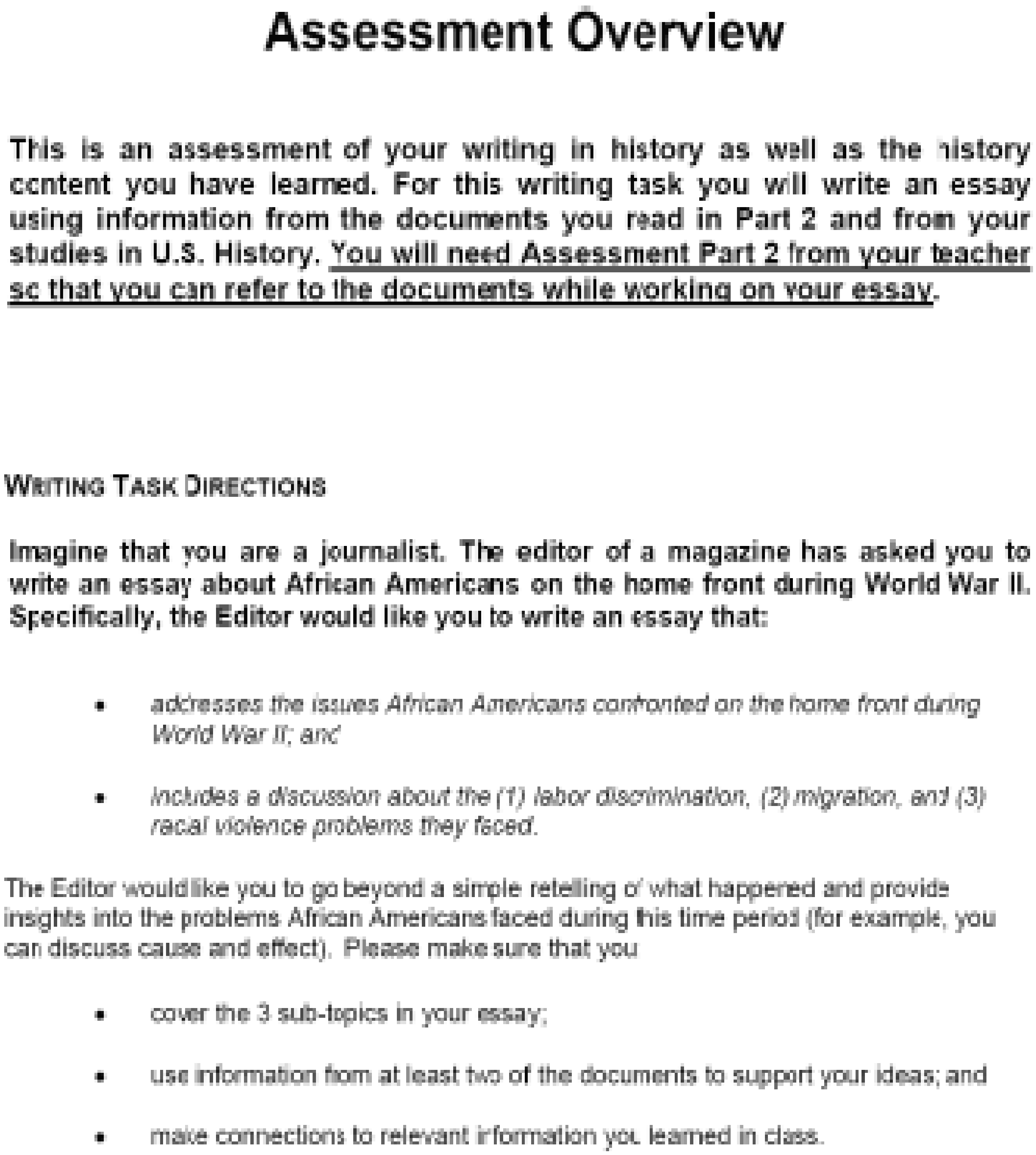
WestEd

HIGH SCHOOL U.S. HISTORY

ASSESSMENT PART 3A

ISSUES FACING AFRICAN AMERICANS DURING WORLD WAR II

Student ID	
Teacher ID	



Assessment Overview

This is an assessment of your writing in history as well as the history content you have learned. For this writing task you will write an essay using information from the documents you read in Part 2 and from your studies in U.S. History. You will need Assessment Part 2 from your teacher so that you can refer to the documents while working on your essay.

WRITING TASK DIRECTIONS

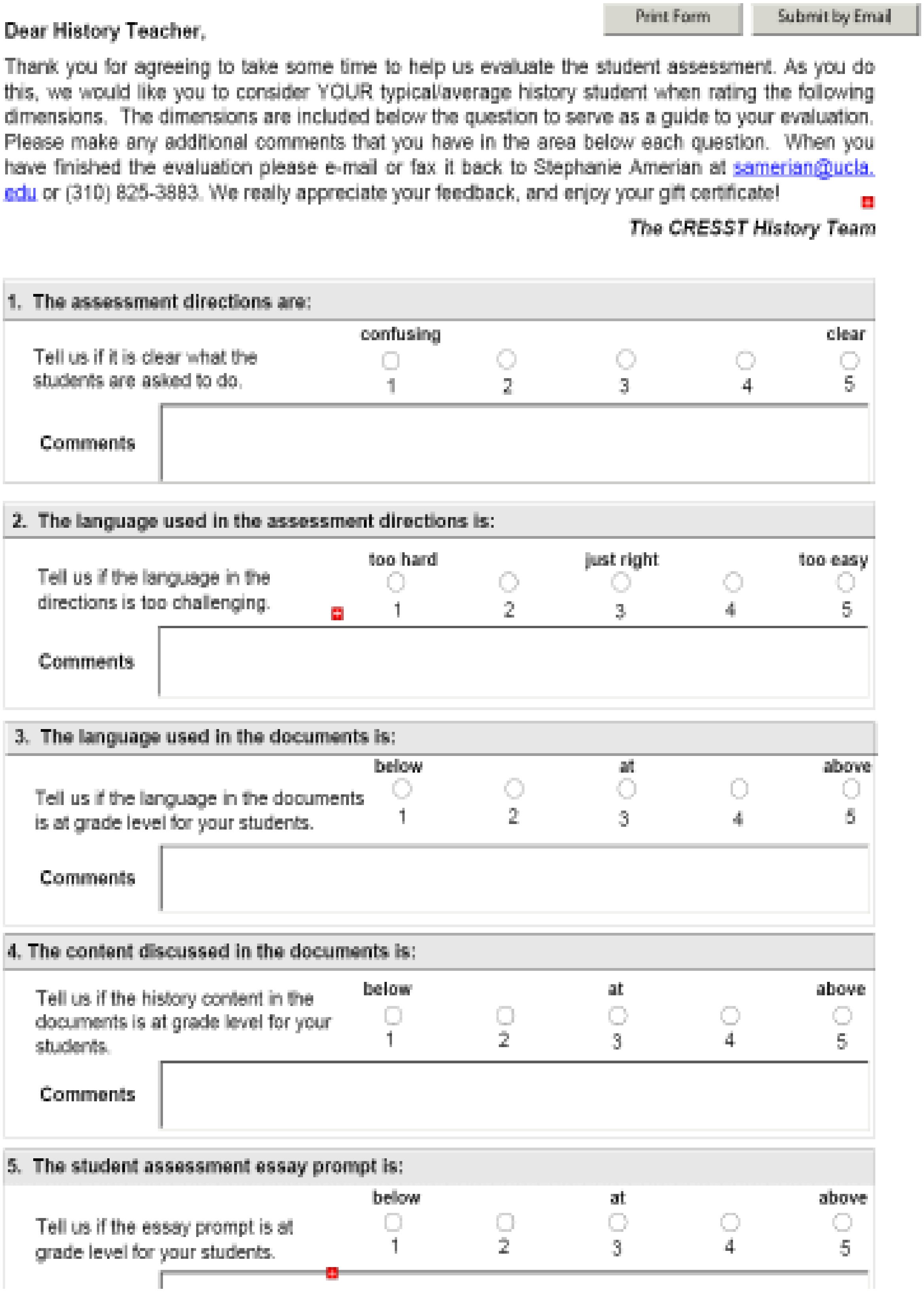
Imagine that you are a journalist. The editor of a magazine has asked you to write an essay about African Americans on the home front during World War II. Specifically, the Editor would like you to write an essay that:

- *addresses the issues African Americans confronted on the home front during World War II; and*
- *includes a discussion about the (1) labor discrimination, (2) migration, and (3) racial violence problems they faced.*

The Editor would like you to go beyond a simple retelling of what happened and provide insights into the problems African Americans faced during this time period (for example, you can discuss cause and effect). Please make sure that you

- cover the 3 sub-topics in your essay;
- use information from at least two of the documents to support your ideas; and
- make connections to relevant information you learned in class.

Appendix C:
ILA Teacher Feedback Survey



Dear History Teacher,

Thank you for agreeing to take some time to help us evaluate the student assessment. As you do this, we would like you to consider YOUR typical/average history student when rating the following dimensions. The dimensions are included below the question to serve as a guide to your evaluation. Please make any additional comments that you have in the area below each question. When you have finished the evaluation please e-mail or fax it back to Stephanie Amerian at samerian@ucla.edu or (310) 825-3883. We really appreciate your feedback, and enjoy your gift certificate!

The CRESST History Team

1. The assessment directions are:

	confusing				clear
Tell us if it is clear what the students are asked to do.	○ 1	○ 2	○ 3	○ 4	○ 5
Comments					

2. The language used in the assessment directions is:

	too hard		just right		too easy
Tell us if the language in the directions is too challenging.	○ 1	○ 2	○ 3	○ 4	○ 5
Comments					

3. The language used in the documents is:

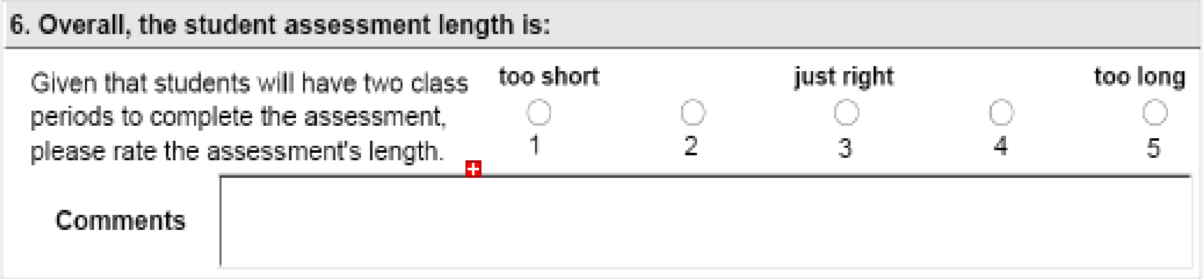
	below		at		above
Tell us if the language in the documents is at grade level for your students.	○ 1	○ 2	○ 3	○ 4	○ 5
Comments					

4. The content discussed in the documents is:

	below		at		above
Tell us if the history content in the documents is at grade level for your students.	○ 1	○ 2	○ 3	○ 4	○ 5
Comments					

5. The student assessment essay prompt is:

	below		at		above
Tell us if the essay prompt is at grade level for your students.	○ 1	○ 2	○ 3	○ 4	○ 5

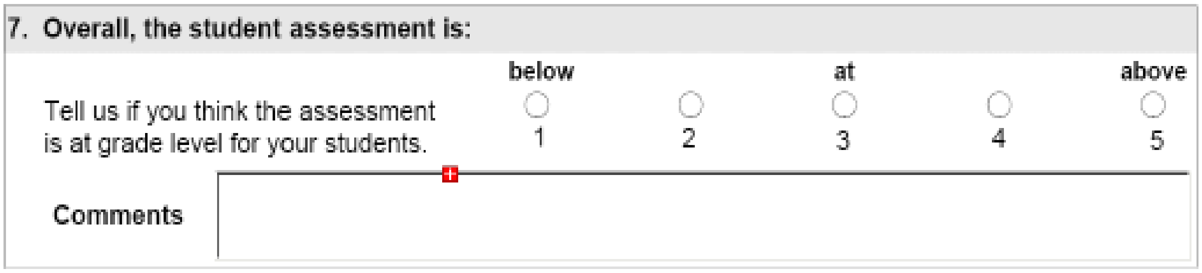


6. Overall, the student assessment length is:

Given that students will have two class periods to complete the assessment, please rate the assessment's length.

too short		just right		too long
○	○	○	○	○
1	2	3	4	5

Comments

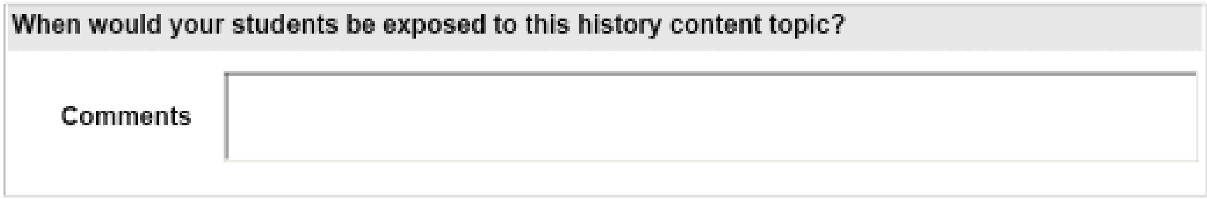


7. Overall, the student assessment is:

Tell us if you think the assessment is at grade level for your students.

below		at		above
○	○	○	○	○
1	2	3	4	5

Comments



When would your students be exposed to this history content topic?

Comments

Appendix D:
Metacognition Scoring Rubric

ILA Scoring Rubric: Metacognition

Score Point	CRITERIA FOR SCORING
4	The response demonstrates **strong** metacognition of ongoing and purposeful interactions with the text and/or its content. This may be evidenced in the following ways: • Engages with complexities in the text or ideas that require attention • Describes multiple thinking processes that occur while reading • Describes more than one approach to how he/she guides his/her thinking about the reading, or gives a sophisticated description of one approach
3	The response demonstrates **adequate** metacognition of purposeful interactions with the text and/or its content. This may be evidenced in the following ways: • Responds to at least one complexity in the text or idea that requires attention • Describes at least one thinking process that occurs while reading • Tells how he/she guides his/her thinking about the reading, albeit with little detail or evidence of thinking processes that occur at multiple points during the reading
2	The response indicates **weak or limited** metacognition. This may be evidenced in the following ways: • Only makes vague reference to complexities in the text or ideas that require attention • Shows limited evidence of thinking processes that occur while reading • Shows little evidence of guiding his/her thinking about the reading
1	The response gives **no evidence of metacognition.** Either there is no response or the student: • Does not identify complexities in the text or ideas that require attention • Gives no indication of thinking processes that occur while reading • Gives no indication of guiding his/her thinking about the reading

Appendix E:
Reading Strategies Scoring Rubric

ILA Scoring Rubric: Use of Reading Strategies

Score Point	CRITERIA FOR SCORING
4	The student text annotations demonstrate **strong** use of **reading strategies**. This may be evidenced in the following ways: • Annotations are seen **frequently** (e.g., seen consistently throughout all five document sections or concentrated in at least two document sections). • Annotations represent a **variety** of reading strategies. **At least 3 reading strategies are used.** • Student utilizes discipline specific reading strategies.
3	The student text annotations demonstrate **adequate** use of **reading strategies.** This may be evidenced in the following ways: • Annotations are seen **somewhat frequently** (e.g., seen to some degree throughout all five document sections or concentrated in at least one document section) • Annotations represent **some variety** of reading strategies. **At least 2 reading strategies are used.** • Student *may* utilize discipline specific reading strategies.
2	The student text annotations demonstrate **weak or limited** use of **reading strategies.** This may be evidenced in the following ways: • Annotations are **sparse** (e.g., annotations appear infrequently in at least two document sections). • Annotations represent **little variety** of reading strategies. **At least 1 reading strategy is used.** • Student utilizes only general reading strategies.
1	The student text annotations demonstrate **no or minimal** use of **reading strategies.** • Annotations are **absent**, **minimal** (e.g., appear in only one document section in a superficial manner), or **indiscriminate** (e.g., large sections of the passage *may* be highlighted or underlined without apparent purpose). • **Only one reading strategy is used, if any.**

Appendix F:
Writing Content Scoring Rubric

ILA Scoring Rubric: Writing Content

To demonstrate understanding of [biological processes and applications/ historical events, changes, and developments]

Score Point	CRITERIA FOR SCORING
4	The response demonstrates a WELL-DEVELOPED understanding and knowledge of the target [biology/history] content. This may be evidenced in the following ways: • The response addresses **all** parts of the essay question. • The response incorporates relevant information from *at least* two document sections. • The response includes **significant** prior knowledge. • The content is **exceptionally** clear, focused, and thoroughly explained. • The response includes strong supportive evidence. • The response relies **very little** on simple (word-for-word) repetition of text.
3	The response demonstrates ADEQUATE understanding and knowledge of the target [biology/history] content. This may be evidenced in the following ways: • The response addresses **most** of the question. • The response incorporates **mostly** relevant information from two document sections. • The response includes **adequate** prior knowledge. • The content is mostly clear and focused. • The response includes **some** supportive evidence. • The response relies **little** on simple (word-for-word) repetition of text.
2	The response demonstrates LOW understanding of the target [biology/history] content. This may be evidenced in the following ways: • The response addresses **some** of the question. • The response includes limited information from the document sections. • The response includes **limited** prior knowledge. • The main idea of the essay is **understandable**, but may be **overly broad** or **simplistic.** • The response includes **insufficient** supportive evidence. • The response may include **some** inaccuracies that detract from the overall essay. • The response may **somewhat** rely on simple (word-for-word) repetition of text.
1	The response represents VERY LOW or NO grasp of the target [biology/history] content. This may be evidenced in the following ways: • The response may address the question **minimally**, or not at all. • The response includes little to no information from the document sections. • The response **does not** include any prior knowledge. • The main idea is not understandable. • The response includes **little or no** supportive evidence to support the main ideas. • The response includes **frequent** inaccuracies that detract from the overall essay. • The response **excessively** relies on simple (word-for-word) repetition of document text.

Appendix G:
Writing Language Scoring Rubric

ILA Scoring Rubric: Writing Language

To communicate ideas clearly with a scholarly scientific writing style

Score Point	CRITERIA FOR SCORING
4	The response is an EXCELLENT [scientific/historical] explanation with very good academic language use. This may be evidenced in the following ways: • **Most** or **all** of the essay's organizational components are strong. • The response includes an introduction with a strong thesis and conclusion that is beyond a restatement of the thesis. • The response demonstrates **very good** text cohesion through the regular use of varied sentence structures and strong links between sentences. • The response demonstrates **consistent** use of precise and varied words, including frequent specific biology terms and expanded noun phrases to describe biology concepts. • The tone is **impersonal** and **authoritative** with no or minimal speech markers. • The response relies **very little** on simple (word-for-word) repetition of text.
3	The response is an ADEQUATE [scientific/historical] explanation with good academic language use. This may be evidenced in the following ways: • The content's organization is **satisfactory**, **generally clear**, **and coherent**. • [History: The response includes a basic introduction and conclusion.] • The response demonstrates a **good level** of text cohesion through the use of sentence structure variety and some marked themes. • The response demonstrates an **adequate** use of precise and varied words, including some specific biology terms and expanded noun phrases to describe biology concepts. • The tone is often **impersonal** and **authoritative**, though the writing may contain some speech markers and personal references. • The response relies **little** on simple (word-for-word) repetition of text.
2	The response is a WEAK [scientific/historical] explanation with only some academic language. This may be evidenced in the following ways: • The content's organization may be **skeletal** and/or **loosely planned**. • [History: The response may lack an introduction and/or conclusion.] • The response demonstrates **some** text cohesion, though the ideas are not linked well with appropriate language features. • The response occasionally demonstrates use of precise and varied words, but generally the vocabulary is **ordinary** and there is **little expansion** of noun phrases. • The tone may be **somewhat informal** with regular uses of speech markers and first or second person references. • The response may **somewhat** rely on simple (word-for-word) repetition of text.
1	The response is a POOR [scientific/historical] explanation with minimal to no academic language use. This may be evidenced in the following ways: • The writing may be **haphazard** and **disjointed**, **with weak organization**. • [History: The response does not include an introduction or conclusion.] • The response demonstrates **minimal to no** text cohesion. • The **word usage is simplistic**, repetitive, inappropriate, or overused with **little to no evidence** of expanded noun phrases. • The tone is usually **informal** and personal with an overuse of speech markers. • The response **excessively** relies on simple (word-for-word) repetition of document text.

Appendix H:
Writing Language Rubric Description

The Language rubric specifically focuses on assessing students' linguistic command of grammatical structures that are directly related to the explanation genre in general and to the history explanation genre in particular and that are aligned with the California Content Standards in writing. The language qualities of history writing that are of interest to us are abstraction, informational density, and technicality.

We related the ideas of abstraction, informational density, and technicality to three systemic functional linguistic concepts. *Field* (the linguistic elements used to communicate them) refers to students' ability to use varied and precise word choice, *Tenor* (the tone of that communication) refers to students' ability to establish a formal, impersonal tone in their writing, and *Mode* (the manner in which ideas are communicated) refers students' ability to create appropriate text cohesion in their writing,.

Varied and precise word choice (Field). The Field of discourse is associated with presentation of ideas, typically involving "content" words such as nominal groups, verbal groups, and adverbial expressions. In history writing in particular, the dimension of Field is characterized by *informational density*, whereby clauses carry a high percentage of content-specific words. These tend to be nouns, verbs, adjectives, and adverbs. Content words are usually clustered into phrases, e.g., expanded noun phrases, which can be used to condense information. This high use of content words and, at times, technical vocabulary leads to another characteristic of history writing, namely, *technicality*. This is realized through the use of noun phrases and verbs that show relationships between them (Fang, 2004). Table H1 below provides additional information about the elements of this dimension.

Table H1

Description of Language Rubric Dimension: Field

Construct: varied and precise word choice	Operationalized in the rubric as:	Specific language features used to realize field	Specific function of language features in history explanations
Information density and technicality	**Word group quality**: Variety and expandedness of word groups **Lexical quality** defined as significant and appropriate use of technical terminology	**Noun groups can consist of:** main noun, adjectives, embedded clauses, prepositional phrases **Verb groups can consist of:** verbs, adverbs, prepositional phrases **Adverbial groups include:** adverbs, subordinate and participial clauses, prepositional phrases **Word Choice** specific to the history domain	**Noun groups:** are often events or happenings instead of personal noun groups. They also name points to be made (e.g., *There are three reasons that…*) **Verb groups:** include frequent action and having/being verbs **Adverbial groups:** rank and condense information through use of subordinate clauses

Formal and impersonal tone (Tenor). In history writing, Tenor reflects a convention of formal, written discourse. That is, personal opinions and stances should be presented in an authoritative and impersonal fashion. This requires the use of interpersonal resources including the declarative mood, modal verbs, and lexical choices that carry an implicit evaluative meaning rather than choices that resort to an emotional appeal (e.g., rhetorical questions) or explicit evaluative meaning (e.g., "I think that" and "I believe that"). In the Language rubric, as shown in Table H2, Tenor is operationalized by considering whether the text has a formal tone and portrays personal opinion implicitly. An author establishes a formal tone by using the linguistic resources of third person and passive voice and by avoiding speech markers ("well", "you know", "like", etc.).

Another consideration for Tenor is the speaker or writer's display of stance (i.e., judgment or interpretation) in the text. The premise is that the speaker or writer expresses his or her personal stance in consideration of the listener or reader. Thus, the display of stance involves various linguistic resources that create the interpersonal meaning. Such interpersonal choices include mood, modality, intonation cues (in spoken discourse), and lexical elements that carry an evaluative and attitudinal meaning. Table H2 provides additional information about the elements of this dimension.

Table H2

Description of Language Rubric Dimension: Tenor

Construct: Formal and impersonal tone	Operationalized in the rubric as:	Specific language features used to realize Tenor	Specific function of language features in history explanations
Tenor	Tone of text	Passive voice	Used to create an impersonal stance
Authoritative quality	Defined as formal and impersonal	Third person	
		Few uses of speech markers	
		Few addresses to oneself or audience	
		Modal verbs and adverbs (e.g., can, will, possibly, perhaps, etc.)	Used to present claims as possibilities
		“It” constructions	
		Precise word choice of nouns, adjectives, verbs, and adverbs	Used to convey evaluation, e.g., responsible

Text cohesion (Mode). The Mode of discourse refers to the way that language is structured in the social context in which it is used. The structure of a text reflects both linguistic and non-linguistic aspects of the social context, such as the availability of feedback between speaker and listener or between writer and reader. Linguistic resources that construe the textual meaning include cohesive devices such as conjunctions and connectors, clause-combining strategies, and thematic organization. In the rubric, we characterized this dimension as text structure in order to reflect the elements of grammar that realize the type and organization of text that serves a specific purpose. When rating the Language dimension, raters considered whether students used a variety of sentence structures, including marked themes (information in front of the subject used to link it to the previous clause). Table H3 provides additional information about the elements of this dimension.

Table H3

Description of Language Rubric Dimension: Mode

Construct: Text Cohesion	Operationalized in the rubric as:	Specific language features used to realize mode in history explanations	Specific function of language features in history explanations
Mode With qualities of: Abstraction and Information density	**Text cohesion:** the flow between clauses and sentences	**Text connectors** (conjunctions, adverbials, verbs) **Marked themes** (information in front of subject) **Thematic progress:** subject of one sentence connected to the predicate of a previous sentence, e.g., nominalization	Text connectors and marked themes link one part of the text to the next with cohesive ties, causal conjunctions and markers of contrast, classification, and logical sequence. They also include grammatical shifts for moving from general to specific and back again **Thematic progress:** Information ranked and condensed through use of clause organization and nominalization

Appendix I:
IRT-Based Analysis of Multiple Choice Tests of Content Knowledge

Item parameter estimates (Est) and their standard errors (*SE*) for confirmatory 3-parameter logistic model with single factor for content knowledge subtest (correlated with reading comprehension subtest factor). Results shown for full-length and reduced-length tests for each subject

Content Knowledge

Biology Content Knowledge (ILA Part 1)

Table I1

Full-Length Test (10 items)

	Slope		Intercept		Guessing	
Item	Est	*SE*	Est	*SE*	Est	*SE*
1	12.23	122.08	-24.64	243.07	.29	.02
2	-2.19	213.26	-12.20	606.84	.07	.02
3	1.36	.40	-1.29	.49	.21	.06
4	1.32	.22	1.04	.18	.16	.07
5	4.70	7.37	-11.00	15.81	.12	.01
6	15.79	425.01	-43.16	1156.00	.19	.01
7	1.58	.27	1.86	.22	.17	.07
8	1.46	.26	-.13	.22	.13	.05
9	2.82	1.30	-2.79	1.27	.20	.04
10	.84	.19	.03	.25	.21	.08

Table I2

Reduced-Length Test[a]

	Slope		Intercept		Guessing	
Item	Est	*SE*	Est	*SE*	Est	*SE*
1						
2						
3	1.30	.40	-1.24	.49	.20	.06
4	1.31	.22	1.04	.18	.16	.07
5						
6						
7	1.59	.26	1.87	.22	.17	.07
8	1.52	.28	-.15	.23	.14	.05
9	2.50	1.05	-2.47	1.03	.19	.04
10	.83	.19	.04	.25	.21	.08

Note.[a]6 items analyzed; items 1, 2, 5, and 6 excluded.

History Content Knowledge (ILA Part 1)

Table I3

Full-Length Test (10 items)

	Slope		Intercept		Guessing	
Item	Est	*SE*	Est	*SE*	Est	*SE*
1	1.23	.21	1.34	.19	.20	.08
2	1.49	.46	-.69	.54	.17	.10
3	2.26	1.34	.20	.46	.29	.10
4	.55	.13	.46	.19	.19	.08
5	1.23	.31	-.88	.34	.15	.06
6	.98	.94	-1.66	1.55	.23	.14
7	1.16	.21	.37	.21	.17	.07
8	.75	.19	-.21	.32	.20	.09
9	.69	.15	1.32	.18	.20	.08
10	.86	.56	-2.20	.96	.16	.06

Table I4

Reduced-Length Test[a]

	Slope		Intercept		Guessing	
Item	Est	*SE*	Est	*SE*	Est	*SE*
1	1.37	.23	1.32	.22	.20	.09
2	1.63	.37	-.74	.28	.17	.05
3	2.03	.49	.33	.28	.24	.07
4	.64	.13	.33	.19	.18	.08
5	1.26	.27	-.86	.28	.14	.05
6	.81	.35	-1.39	.58	.20	.07
7	1.20	.20	.35	.19	.17	.06
8	.87	.20	-.25	.28	.20	.08
9	.80	.16	1.29	.17	.20	.08
10						

Note. [a] 9 items analyzed; item 10 excluded**.**

IRT-Based Analysis of Multiple Choice Tests of Reading Comprehension

Item parameter estimates (Est) and their standard errors (SE) for confirmatory 3-parameter logistic model with single factor for reading comprehension subtest (correlated with content knowledge subtest factor). Results are shown for full-length and reduced-length tests for each subject.

Reading Comprehension

Biology Reading Comprehension (ILA Part 2)

Table I5

Full-Length Test (10 items)

	Slope		Intercept		Guessing	
Item	Est	*SE*	Est	*SE*	Est	*SE*
1	.86	.15	.62	.18	.21	.07
2	.88	.17	-.08	.22	.18	.07
3	1.83	.51	-.96	.48	.16	.06
4	.88	.22	-.42	.31	.24	.08
5	2.07	.48	-.65	.35	.21	.05
6	-.15	4.57	-5.12	17.00	.16	.08
7	1.21	.25	-.49	.28	.20	.06
8	1.26	.20	.81	.20	.16	.07
9	.60	.22	-1.09	.41	.19	.07
10	.76	.15	.42	.19	.18	.07

Table I6[a]

Reduced-Length Test

	Slope		Intercept		Guessing	
Item	Est	*SE*	Est	*SE*	Est	*SE*
1	.89	.16	.62	.18	.18	.07
2	.88	.17	-.08	.22	.16	.07
3	1.73	.47	-.87	.45	.24	.06
4	.85	.21	-.39	.30	.20	.08
5	2.03	.46	-.64	.35	.16	.05
6						
7	1.23	.26	-.48	.28	.16	.06
8	1.28	.20	.81	.20	.19	.07
9	.67	.24	-1.14	.44	.19	.07
10	.79	.15	.42	.19	.18	.07

Note. [a] 9 items analyzed; item 6 excluded.

History Reading Comprehension (ILA Part 2)

Table I7

Full Length Test (12 items)

	Slope		Intercept		Guessing	
Item	Est	*SE*	Est	*SE*	Est	*SE*
1	1.25	.58	-1.94	.88	.18	.06
2	1.13	.32	-.35	.38	.23	.09
3	2.77	1.13	-5.45	1.87	.17	.02
4	.82	.18	-.02	.26	.20	.08
5	1.28	.25	.60	.24	.21	.09
6	1.47	.28	.90	.22	.20	.09
7	.67	.14	.26	.22	.18	.09
8	1.86	.37	1.87	.24	.20	.09
9	.75	.17	-.14	.27	.20	.08
10	1.11	.19	.96	.19	.20	.08
11	1.07	.18	.85	.19	.20	.08
12	1.08	.26	-.97	.33	.15	.05

Table I8

Reduced-Length Test[a]

	Slope		Intercept		Guessing	
Item	Est	*SE*	Est	*SE*	Est	*SE*
1	1.17	.40	-1.88	.56	.18	.08
2	1.14	.31	-.38	.37	.24	.08
3						
4	.89	.19	-.02	.24	.20	.08
5	1.28	.25	.61	.22	.21	.08
6	1.41	.25	.91	.20	.19	.08
7	.66	.14	.27	.20	.18	.08
8	1.79	.34	1.87	.25	.19	.08
9	.79	.18	-.17	.26	.21	.08
10	1.16	.21	.99	.19	.20	.08
11	1.07	.19	.87	.19	.19	.08
12	1.03	.26	-.94	.35	.15	.06

Note.[a] 11 items analyzed; item 3 excluded.