CRESST REPORT 727

Alison L. Bailey
Becky H. Huang
Hye Won Shin
Tim Farnsworth
Frances A. Butler

DEVELOPING ACADEMIC ENGLISH LANGUAGE PROFICIENCY PROTOTYPES FOR 5TH GRADE READING: PSYCHOMETRIC AND LINGUISTIC PROFILES OF TASKS

SEPTEMBER 2007

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

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

Developing Academic English Language Proficiency Prototypes for 5th Grade Reading: Psychometric and Linguistic Profiles of Tasks

CSE Technical Report 727

Alison L. Bailey, Becky H. Huang, Hye Won Shin, and Tim Farnsworth
CRESST/University of California, Los Angeles

Frances A. Butler
Language Testing Consultant

September, 2007

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
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002-02, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

DEVELOPING ACADEMIC ENGLISH LANGUAGE PROFICIENCY PROTOTYPES FOR 5TH GRADE READING: PSYCHOMETRIC AND LINGUISTIC PROFILES OF TASKS[1]

Alison L. Bailey, Becky H. Huang, Hye Won Shin, and Tim Farnsworth
CRESST/University of California, Los Angeles

Frances A. Butler
Language Testing Consultant

Abstract

Within an evidentiary framework for operationally defining academic English language proficiency (AELP), linguistic analyses of standards, classroom discourse, and textbooks have led to specifications for assessment of AELP. The test development process described here is novel due to the emphasis on using linguistic profiles to inform the creation of test specifications and guide the writing of draft tasks. In this report, we outline the test development process we have adopted and provide the results of studies designed to turn the drafted tasks into illustrative prototypes (i.e., tried out tasks) of AELP for the 5th grade. The tasks use the reading modality; however, they were drafted to measure the academic language construct and not reading comprehension per se. That is, the tasks isolate specific language features (e.g., vocabulary, grammar, language functions) occurring in different content areas (e.g., mathematics, science and social studies texts). Taken together these features are necessary for reading comprehension in the content areas. Indeed, students will need to control all these features in order to comprehend information presented in their textbooks. By focusing on the individual language features, rather than the subject matter or overall meaning of a text, the AELP tasks are designed to help determine whether a student has sufficient antecedent knowledge of English language features to be able to comprehend the content of a text.

The work reported here is the third and final stage of an iterative test development process. In previous CRESST work, we conducted a series of studies to develop specifications and create tasks of AELP. Specifically, we first specified the construct by synthesizing evidence from linguistic analyses of ELD and content standards, textbooks (mathematics, science, and social studies), and teacher talk in classrooms, resulting in

[1] We would like to thank the following for their role in the preparation of this report: Christine Ong for thorough research assistance, Amy Dray for data collection assistance, Joan Herman for insightful comments on an earlier draft, Erna Aridzanyan for administrative assistance, and Yasmin Damshenas for final editing. Conference participants at the 28th Annual Language Testing Research Colloquium at the University of Melbourne, Australia (June 28 - July 1, 2006) and at the 2006 Midwest Association of Language Testers Conference at the University of Illinois (Sept. 29-30, 2006) provided valuable feedback on the project. Finally, to all the students, teachers, and administrators at the schools in Southern California who made the tryouts and pilot study possible—a very special thank you.

language demand profiles for the 5th grade. After determining task format by frequency of assessment types in textbooks, we then created draft tasks aligned with the language profiles.

The goals of the current effort were to take these previously drafted tasks and create prototypes by trying out the tasks for the first time with 224 students from English language learner (ELL) and native English backgrounds. Students across the 4th-6th grades, as well as native-English students are included in the studies because native speakers and adjacent grades provide critical information about the targeted language abilities of mainstream students at the 5th grade. Phase 1 (n=96) involved various pre-pilot tryouts of 101 draft tasks to estimate duration of administration, clarity of directions, whole-class administration procedures, and an opportunity to administer verbal protocols to provide further information about task accessibility and characteristics. Phase 2, the pilot stage, involved administration of 40 retained tasks (35 of which were modified as a result of Phase 1) to students in whole-class settings (n=128). Analyses included item difficulty and item discrimination. The rationale for retaining or rejecting tasks is presented along with psychometric/linguistic profiles documenting the evolution of example effective and ineffective prototype tasks. The final chapter in this report reflects on the lessons learned from the test development process we adopted and makes suggestions for further advances in this area.

CHAPTER 1:
INTRODUCTION AND LITERATURE REVIEW

Overview and Outline of the Report

The work reported here is the culmination of several years of research at the National Center for Research in Evaluation, Standards, and Student Testing (CRESST) that focused initially on articulation of the academic English construct in school settings and finally on the use of that information for the development of prototype reading tasks of academic English. Specifically, the report presents findings from a series of small-scale try-outs and a pilot study with reading tasks designed to assess 5th grade academic English language proficiency (AELP). In Chapter 1, we first summarize the prior research at CRESST which provides the background and context for the AELP task development. The specific goals of the task development effort are then outlined. In Chapter 2, we describe the procedures and instrumentation of each of the two phases of administering and revising the AELP tasks.

Chapter 3 presents analyses of the data collected during in the pre-pilot phase and the subsequent pilot phase. In Chapter 4, we provide six task profiles that provide information about how tasks were refined in light of feedback from verbal protocols with students and psychometric information on item-level performance. Tasks based on reading passages from mathematics, science, and social studies content areas are used to illustrate in considerable depth the decision-making process for how tasks could be retained without modification, modified and retained for piloting, or rejected as unsuitable for further development. Finally, Chapter 5 presents recommendations for refinement of the research and standards-informed test development process, and implications for further research in this area.

Context and Stages of AELP Test Development

The impetus for this long-term initiative grew out of the need to assure access for all students in evaluation of their academic progress. In the mid to late 1990's, the validity of large-scale (standardized) assessments with English language learner (ELL) students came into question (August & Hakuta, 1997; Butler & Stevens, 1997, 2001; LaCelle-Peterson & Rivers, 1994). This concern led to further issues including the use of test accommodations with ELL students (Abedi, 1997; Abedi, Lord, & Plummer, 1995; Butler & Stevens, 1997) and the effectiveness of existing language proficiency tests for evaluating their English language skills (Bailey & Butler, 2002/2003; Butler & Stevens, 2001; Stevens, Butler, & Castellon-Wellington, 2000). CRESST research was showing that existing language tests were not good predictors of performance on standardized content tests (Butler & Castellon-Wellington, 2000/2005). There was a mismatch between the language tested on language proficiency tests (every-day vocabulary and simple structures) and the language used on

content tests and in the classroom (more precise uses of vocabulary and complex structures; Stevens, Butler, & Castellon-Wellington, 2000). The distinctions between the two are typically characterized as social versus academic English, although the distinctions are not always easy to articulate. Since both are critical to the student's English language development, educators began to recognize the need for expanding the content domain of K-12 English language proficiency tests to include academic English.

The *No Child Left Behind Act of 2001*, which required that ELL students show measurable yearly progress in English language development (ELD), brought the language proficiency assessment of ELL students to the forefront of the national educational discussion. The need for language tests that focused on academic English or at least included features of academic English in the test content, rapidly became apparent because of the high stakes decisions (e.g., redesignation of ELL students) being made on the basis of student performance on language tests and the accountability of schools and states for student performance. Many existing commercial ELD tests failed to capture the language demands required for academic success, thus motivating the development of AELP assessment tasks.

In order to develop tests with an expanded content domain, the academic English construct had to be defined with sufficient specificity to allow for the production of test specifications that would capture the most critical features of the domain, that is, those features that distinguish abilities for decision-making purposes. While general definitions of academic English and information on narrow sets of features were available in the literature (e.g., Cummins, 1981, 2003; Schleppegrell, 2001; Short, 1994; Solomon & Rhodes, 1995), there had been no attempt to articulate a more complete construct for educational purposes including test development. In response to this need, Bailey and Butler (2002/2003) proposed a framework based on evidence of language use for operationalizing the academic English construct (AEL). Undertaking this framework, researchers at CRESST have been conducting AELP research for the past five years. Table 1 below summarizes the timeline of the AELP project stages, with goals and the CRESST reports and publications associated with each stage also provided. Specifically, Stage I extended from 1997 to 2004. The goals were to operationalize the AEL construct as well as establish the evidentiary bases from a range of data sources. Six CRESST reports were produced during Stage I. Stage II was carried out during the next two years, 2004-2005, and focused on the development of prototype reading passages and tasks. The final stage, Stage III, during 2005 and 2006 has involved a pre-pilot phase of initial tryouts of the drafted AELP tasks and a pilot phase with retained and often refined tasks from the initial tryouts.

Table 1

An Overview Timeline of Academic English Language Proficiency (AELP) Research at CRESST

Stage	Goal	CRESST Report
Stage I (1997-2004)	Multi-year research effort to:	
	• operationalize the construct of Academic English (AE)	*Students' concurrent performance on tests of English language proficiency and academic achievement* (Butler, F. A., & Castellon-Wellington, M., 2000/2005, Tech. Rep. No. 663).
		*Academic English and content assessment: Measuring the progress of ELLs (*Stevens, R. A., Butler, F. A., & Castellon-Wellington, M., 2000, Tech. Rep. No. 552).
		An evidentiary framework for operationalizing academic English for broad application to K-12 education: A design document (Bailey, A. L., & Butler, F. A., 2002/2003, Tech. Rep. No. 611).
		*Towards the characterization of academic English in upper elementary science classrooms (*Bailey, A. L., Butler, F. A., LaFramenta, C., & Ong, C., 2001/2004, Tech. Rep. No. 621).
		An approach to operationalizing academic English for language test development purposes: Evidence from 5th-grade science and mathematics (Butler, F. A., Lord, C., Stevens, R., Borrego, M., & Bailey, A. L., 2003/2004, Tech. Rep. No. 626).
	• establish the evidentiary bases for academic English language (AEL) across a range of contexts	*Academic English in 5th-grade Mathematics, Science, and Social Studies Textbooks.* (Butler, F. A., Bailey A. L., Stevens, R., Huang, B. & Lord, C., 2004, Tech. Rep. No. 642).
Stage II (2004-2005)	Development of prototype reading passages and tasks that meet the following criteria:	*Using Standards and Empirical Evidence to Develop Academic English Proficiency Test Tasks in Reading* (Bailey, A. L., Stevens, R., Butler, F. A., Huang, B., & Miyoshi, J. N., 2005, Tech. Rep. No. 664).
	• aligned to California content standards	
	• correspond to the linguistic profiles established by the Stage 1 empirical studies	
	• free of cultural and gender bias	
Stage III (2005-2006) Phase 1	• Gather student performance data on the prototype tasks to make decisions about retention and revision of the tasks	Current Report
Phase 2 (2006)	• Gather student performance data on the tasks on a larger scale to make final decisions about retention and revision of the tasks	Current Report

Note. All CRESST/CSE reports available at: http://www.cse.ucla.edu/products/reports_set.htm, the National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Table 2 provides details of the Stage III research, which is the focus of the current report. Stage III was divided into a phase of initial tryouts, which provided information that led to retention, refinement, or rejection of the tasks. A second phase involved piloting the smaller number of retained/refined tasks to create prototype tasks of AELP.

Table 2

Breakdown of Phases for Stage III AELP Research at CRESST

Phase	Goal	CRESST Report
Phase 1: Pre-pilot tryouts (2005- 2006)		Current Report
Student informant	• Estimate the completion time for the tasks • Make sure the instructions are clear	
Whole-group Administration (n=77)	• Collect information from 4^{th}-6^{th} graders' performance on the prototype tasks for preliminary analysis and revision of those tasks	
Verbal Protocol (n=18)	• Obtain in-depth student feedback on the tasks	
Phase 2: Pilot (2006)		Current Report
Whole-group Administration (n=128)	• Gather student performance data on the tasks on a larger scale to make final decisions about retention and revision of the tasks	

An articulation of the multi-year AELP research by stage is presented in the following paragraphs:

Establishing the evidentiary bases for academic English language across a range of contexts was the first stage (Stage I) of the academic English language research at CRESST. Given the dearth of research and language assessments that had focused on AEL as the target language use (TLU) domain, Stage I produced the necessary linguistic profiles for mathematics, science, and social studies contexts in which to anchor the development of the prototype AEL tasks.

Stage II involved the identification of texts across content areas to be used in the development of the prototype reading tasks. The outcomes of this stage included mathematics word problems, science passages, and social studies expository passages that were aligned to the California content standards, corresponded to the linguistic profiles established by the Stage I empirical studies, and were free of cultural and gender bias.

Stage III, involved two phases: a phase of try-outs in which we asked an in-coming 5th grader to be an informant by completing the tasks while providing feedback to researchers, and tried out a whole-group administration (n=77) and one-on-one think aloud activities using a verbal protocol (n=18) with predominantly English-only and some ELL students in the 4th -6th grades at a university-affiliated elementary school.

The second phase of Stage III involved pilot testing subsequently retained and largely refined tasks with larger numbers of both English-only and ELL students (n=128). Data on the tasks provided the information needed to make further decisions about retention and revision of the tasks and turning the most effective into prototypes. These prototypes are example tasks accompanied by both linguistic and psychometric information. The long-term undertaking, Stages I - III, has yielded a set of guidelines for task development with accompanying prototype tasks that can be used by teachers and test developers to produce assessments appropriate for their specific needs in evaluating academic English. It is important to remember that while this work focuses specifically on the development of tasks that assess academic English, an operational and comprehensive test of English language proficiency for K-12 would also include social language as part of the TLU to measure a student's ability to use language in all school-related contexts (i.e., both the academic and social situations of school).

The next section defines academic English in this work and is followed by an overview of test development practices.

Academic English Defined

Academic English language has become one of the popular foci in the current field of education and assessment (GAO, 2006). At its simplest, AEL refers to the language used for the purpose of "acquiring new knowledge and skills...imparting new information, describing abstract ideas, and developing students' conceptual understanding" (Chamot & O'Malley 1994, p. 40). AEL is distinct from the social language used in school; it encompasses the vocabulary, syntactic structures, and discourse features that are "necessary for a student to access and engage with their grade-level curriculum" (Bailey & Heritage, forthcoming, 2007).

The precursor of the growing body of literature on AEL can be traced back to the BICS/CALP distinction proposed by Jim Cummins (1981). Cummins (1981, 2003) attempted to distinguish academic English, which he labeled *Cognitive Academic English Proficiency* (CALP), from everyday language, which he labeled *Basic Interpersonal Communicative Skills* (BICS). Both CALP and BICS are influenced by two continua: context and cognitive demands. Cummins argued that CALP is both cognitively demanding and context-reduced, whereas BICS is cognitively undemanding and context-embedded. The BICS/CALP distinction, while seminal and useful, is criticized for equating BICS with linguistic simplicity and CALP with complexity (Bailey, 2006). BICS is not necessarily less cognitively demanding than CALP. The mental effort and cognitive ability necessary to contrive plausible excuses, negotiate and persuade, deceive and win over others in our everyday life (BICS) is not any less than that required to comprehend a paragraph on Civil War in a 5th grade social studies textbook (CALP). Nor does it seem the case that BICS cannot exist within context-reduced settings. Children's make-believe play, for example, involves complex and highly abstract reasoning.

Another approach to defining academic English is through analyzing the linguistic elements that make up the register of schooling. Schleppegrell (2001) provided such an analysis of school-based texts and labeled AEL as "language of schooling." Schleppegrell took on Halliday's definition of "register" as "the constellation of lexical and grammatical features that characterize particular uses of language" (p. 431), and compiled a set of linguistic features that constituted academic English. She further argued features of school-based texts are much less familiar for many children because they are not equally prepared to use language in expected ways.

In contrast to the linguistic feature (i.e., form) perspective that describes AEL in discrete linguistic terms, a different approach to characterizing language in general and AE in particular is the language function perspective (e.g., Bachman, 2003; Short, 1994). A representative example from this camp is the work by Chamot & O'Malley (1994), who argued that AEL consists primarily of "language functions needed for authentic academic content" (p. 40).

Rejecting previous conceptualizations of AEL that focus mainly on linguistic features without examining personal and socio-cultural factors, Scarcella (2003) took a broader view and proposed a three-dimensional framework that encompassed linguistic, cognitive, as well as socio-cultural/psychological aspects of AEL. Although this multi-dimensional model is more comprehensive than other theoretical models, it is still of limited practical value due to the lack of specificities required to operationalize tasks for either instruction or test development purposes.

An Evidentiary Framework for Operationalization of Academic English

Following the National Research Council's call for evidence-based educational research, researchers at CRESST adopted a framework that documents a variety of sources of information for operationalizing the AEL construct in order to develop tasks that measure AEL proficiency (Bailey et al., 2001; Bailey & Butler, 2002/2003; 2006; Butler et al., 2003; Butler, et al., 2004). The information includes language prerequisites assumed in national content standards (e.g., National Science Education Standards of the National Research Council), in state content standards (e.g., California, Florida, New York, and Texas), in English as a second language (ESL) standards, the language demands of standardized achievement tests, teacher expectations of language comprehension and production, and the language input students receive in school such as classroom language and textbooks (see Bailey & Butler, 2002/2003; 2006). These different sources of information were used to generate draft specifications and assessment tasks that use actual texts from mathematics, science, and social studies textbooks (see Bailey et al, 2005). With the addition of psychometric and linguistic information provided by the studies conducted for the current report, these tasks can serve as potential models or prototypes for others who are developing AEL language proficiency tests.

The linguistic profiles in Table 3 are taken from Butler, et al. (2004) and show how texts from the different subject matter areas contain different language features.

Table 3

Content Framework for Developing an Assessment of Academic Language Proficiency

Content Category	Mathematics	Science	Social Studies
Vocabulary			
Clause connectors	√	√	√
Non-academic vocabulary			
Academic vocabulary (AV)			
General AV (high-frequency)	√	√	√
Specialized AV (defined in context)	--	√	√
Measurement words	√	√	--
Proper nouns	--	--	√
Grammar			
Nominalizations	--	√	√
Noun phrases	√	√	√
Participial modifiers	--	√	√
Passive forms	--	√	√
Prepositional phrases	√	√	√
Organization of Text			
Comparison	√	√	√
Definition	--	√	√
Description	√	√	√
Enumeration	√	√	√
Exemplification	--	√	√
Explanation	--	√	√
Labeling	--	√	√
Paraphrase	√	√	√
Scenario	√	--	--
Sequencing	√	√	√

Note. From *Academic English in 5th-grade Mathematics Science, and Social Studies Textbooks* (p. 110), by F. A. Butler, A. L. Bailey, R. Stevens, B. Huang, and C. Lord, 2004, CSE report # 642. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Copyright 2004 by CRESST. Reprinted with permission.

These profiles were used as part of the test specifications for AELP tasks and guided the task creation in Bailey et al. (2005). The information guided the prevalence of these linguistic features in the tasks and also the linguistic characteristics of the text selections to which the tasks are attached. The profiles of effective and individual tasks, which we report in Chapter 4, also include the linguistic profiles unique to each task.

After also determining task format by frequency of assessment types found in the textbooks, Bailey et al. (2005) created 101 draft tasks aligned with the language profiles presented above. These tasks are designed to measure student knowledge of academic English through reading. However, these tasks were drafted to measure the academic language construct and not reading comprehension; that is the tasks isolate specific language features (e.g., vocabulary, grammar, language functions) of the different subject matter areas (e.g., mathematics, science, and social studies). Taken together these features are necessary for reading comprehension in the subject areas; indeed students will need to control all these features in order to comprehend information presented in their textbooks. By focusing on the individual language features, rather than the subject matter content or overall meaning of a text, the AELP tasks are designed to help determine whether a student has sufficient antecedent knowledge of English (i.e., linguistic features such the nominalization of verbs and the complex embedding of clauses within sentences) to be able to comprehend the content of a text.

Test Development Research Practices

Properly executed, language test development (TD) is a complex process that begins with specifying the construct and associated skills to be assessed. Following Bailey and Butler (2002/2003), the process begins with a needs analysis that helps establish parameters for both test content and test use.[2] At this initial stage, the construct is an evolving definition of the abilities the test is intended to tap. As part of the needs analysis, the characteristics of the intended test takers and test users, as well as the situations in real-world language use, or TLU domain, to which the test is intended to generalize are described (Bachman 1996). A design or framework document is then written which provides the information above and specifies what aspects of test quality (e.g., reliability, authenticity, etc.) are to be prioritized and the resources to be devoted to the project (Bachman 1990, 1996; Davidson & Lynch 2002). In addition, the document identifies gaps in information about the construct that may require additional research. Once this work has been completed, operationalization of the test commences.

Traditionally, at this stage, construct definitions are made specific at the level of the features of vocabulary, grammar, reading skill, pronunciation skill, or other aspects of language ability thought to be relevant to the construct and language use situations described in the design document above.[3] This step leads to the development of the test specifications

1 The techniques for needs analysis grew out of work in the area of syllabus design. See McNamara (1996, p. 36) for an historical perspective on the use of needs analyses in language test development. See also Witkins and Altschuld (1995) for needs assessment techniques.

2 Bailey and Butler (2002/2003), drawing on Bachman (1990) and Davidson and Lynch (2002), have suggested elsewhere that the entire process should be documented to help establish the validity

that will then guide task and item development. Test tasks tap into the operational construct, keeping in mind the features of the test takers, TLU domain, available resources, time constraints, and other concerns delineated in the design document. Most commonly, these tests tasks are pilot tested, subjected to analysis, then revised and retested until the finished product meets the clients', developers, and legal and professional standards of approval (e.g., AERA, APA, NCME, 1999).

In our own work, we also include an additional phase of try-outs, small-scale studies of tasks and test directions/formatting conducted before the pilot stage. We suggest that while frequently omitted from the traditional TD process, small-scale tryouts or pre-pilot studies should be an inherent part of any test development effort to help assure that the tasks, and directions to be pilot tested on a larger scale have been refined to a high degree so that data from the pilot are not lost due to problems with the tasks that could have been corrected based on limited pre-pilot input.

foundation for the tasks being developed. The audit trail is a critical piece of the overall documentation procedure.

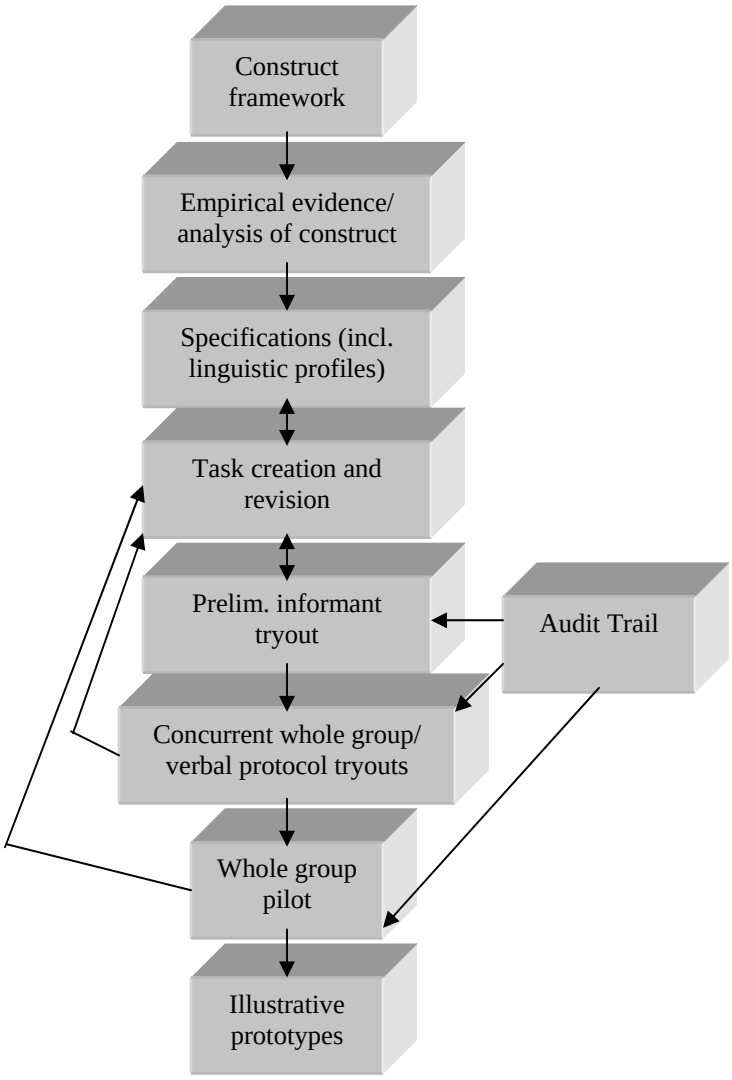


Figure 1. The test development process used with the reported AELP prototypes.

As Figure 1 shows throughout the empirical testing of tasks, accountability is maintained through the documentation of the processes of test task development and modification, and ongoing qualitative and quantitative analysis of test taker performance on the test tasks. This documentation, which we conduct here as an "audit trail" (Davidson et al., 2006), serves as a primary source of evidence for evaluating the overall validity of the test, as well as providing a guide for future item and test development. Starting from test specifications, test developers can use the audit trail to document how tasks change or are eliminated during the test development process due to data from pre-pilot and pilot testing, expert reviews, and revision, thus providing important information for creating a validity argument.

In the work of Davidson and his colleagues, the audit trail is described as a practical method of documenting the myriad changes and revisions in the test development process. In order to establish a sound argument for test validity, Davidson et al. (2006) advocate the practice of using an audit trail for building a strong case of the inferences made from test results. In their words, audits can serve as "lines of defense against attack on the validity" (p. 223). Davidson et al. believe that the audit trail method used to document the changes and decisions made during test development process, should be employed by language test developers to complement the traditional procedures in test development such as piloting and data analysis. To illustrate the audit trail method, Davidson et al. presented the evolution of changes on test specifications from the WIDA assessment project. Test specifications are defined as multi-level generative tools along the spectrum from which test developers can benefit. High-level general guidelines, at one end of the spectrum, are a set of decisions about the entire test, such as the numbers of tasks on A and B target constructs and scoring weighting applications. Low-level guidelines, at the other end, are specific principles for producing actual individual test tasks.

Davidson et al. documented their timeline of WIDA specifications development and the iterative process of the development. Specifically, they demonstrated the process of how they audited their work by using three example areas, i.e., *bias check*, *authenticity*, and *version*. The audit trail illustrates how test specifications and resultant tasks should be considered working documents undergoing a process of revision and improvement as a result of multiphase tryouts before field-testing commences. Our hope is that the audit trail will be used as a template for future test development. This audit trail method is the organizing feature of the current paper.

The following chapters focus on the method and results of Stage III. The goals are to document the test development process we have outlined and take the previously drafted AELP tasks (Bailey et al., 2005) and describe how they evolve (or not) into exemplar prototype tasks with associated psychometric and linguistic information resulting from trying out the tasks for the first time with 224 ELL and native English-speaking students.

CHAPTER 2:
METHOD

Procedures Overview

One hundred and one draft reading tasks were created at Stage II, from test specifications based on linguistic analysis of mathematics, science, and social studies textbooks. A focus group of nine ESL and content-area teachers rated the passages and tasks for linguistic difficulty and item type familiarity. Minor changes were made to the tasks based on feedback from the teachers (see Bailey et al., 2005).

Stage III was designed to include a phase of tryouts that provide additional feedback on the tasks and information on completion time and clarity of the directions, and a pilot phase. Phase 1 included (a) initial tryout with an in-coming 5th-grader, (b) administration of the tasks to 77 students in whole-class settings, and (c) a verbal protocol version of the tasks administered individually with 18 additional students.

Phase 2 was the formal pilot study designed to elevate the draft tasks into formal prototype tasks with accompanying psychometric and linguistic information on each task. The pilot was conducted with the 40 retained and predominantly modified tasks (as will be discussed, 35 of the 40 were modified as a result of the Phase 1 pre-pilot studies). One hundred and twenty-eight students were administered the draft tasks in whole-class settings. Throughout Stage III, an audit-trail method was used to document the tryout and pilot processes and to reveal the shortcomings of the items themselves. In the discussion that follows, we first describe the audit trail method and its adaptation for this work. We then discuss the phases of Stage III.

Adaptation of the Audit Trail Approach

Following Davidson et al.'s (2006) footsteps, we incorporated the audit trail method in our current study with slight modification of the approach. Specifically, instead of focusing primarily on changes to a set of test specifications, we kept detailed notes regarding modifications made to the test format and particular task over the course of the project. In addition, we recorded observations made during whole class administration of the test (i.e., student questions), verbal protocol sessions, and in scoring. At the beginning of Stage III, entries in our audit trail database were originally organized according to type—broadly speaking, changes in format or task, and observational notes. Information regarding the item number, type, and date of entry were also noted. In the second phase of Stage III, entries were collapsed into a single Excel database to make the searches and organization of the entries more effective. More formal definitions for entry type were also created at this time (See Table 4 for definitions and examples of entries by category, and see Appendix A for

excerpts from the audit trail database). In addition, a column identifying the primary information source was added to the database, namely: verbal protocol, whole class administration, data entry, data analysis, and internal review.

Table 4

Audit Trail Organizing Entry Categories

Entry Category	Definition	Example
Directions	Modifications made to passage directions	"Change the instruction. OLD: Read the problem. Then fill in the blanks in the table. NEW: Fill in the blanks for questions in the table below." (10/14/05)
Format	Modifications made to passage or tasks related to format (e.g., spacing, size of font, use of titles)	"…align the four choices, and delete "circle the best answer" from the question stem." (10/14/05)
Item content	Modifications made to the content of the stimulus and/or response	Add the phrase "according to the passage" in the stem to prevent self-invented responses like "weather man" or "weather women." (11/28/05)
Comment	Researcher observations or comments related to students' questions during whole class administration, potential patterns in responses, etc.	

Phase 1: Pre-Piloting for Draft Task Refinement Using Small-Scale Tryouts

Pre-Pilot Samples and Procedures

After internal review and teacher review of the draft tasks (see Bailey et al. 2005), all tasks were additionally completed by an in-coming 5th grader for us to determine any test booklet formatting errors and to make a rough estimate of testing time.[4] To economize on

[4] The student was male and had successfully completed Grade 4, scoring at least at the proficient level on the California Standards Tests in English language arts and mathematics. He was transitioning into Grade 5 and thus represented the youngest of the target group for the draft tasks. The 100 draft tasks were provided in a single booklet format. The administration took place during two informal sessions with a researcher making note of feedback. The informant was directed to attempt all tasks and to verbally comment and make notes about tasks that were confusing in terms of directions or format. The combined sessions totaled approximately 80 minutes. The informant answered 80 tasks correctly. He commented on 17; of those, he answered six incorrectly. Errors were distributed across content areas and item types. Seventeen changes were made in the items based on feedback from the

time during the pre-pilot and pilot administrations, the decision was made after this initial tryout to distribute the 101 draft tasks across two booklets of 66 and 62 tasks each with 27 overlapping tasks across the two booklets (See Appendix B for Forms A & B). The pre-piloting phase itself involved administration of draft tasks with two groups of students; those completing the tasks in whole group settings (n=77) and those completing the tasks one-on-one with a researcher using a verbal protocol to garner in-depth information about the tasks and any continuing issues with directions and formatting of tasks (n=18).

Description of pre-pilot sample. Ninety-five students across Grades 4-6 participated in the pre-pilot. Following UCLA Internal Review Board (IRB) procedures, students were recruited from a university elementary laboratory school situated in a large urban area of Southern California. Enrollment at the school is designed to represent the ethnic diversity of the greater metropolitan area. A number of the students were enrolled in a bilingual program that offers primary language literacy in Spanish. One 6^{th} grade class and two $4^{th}/5^{th}$ combined classes participated in the study. Students across the 4^{th}-6^{th} grades were included in the studies because students in adjacent grades provide critical information about the targeted language abilities of students at the 5^{th} grade. The tasks must be harder for the students in the 5^{th} grade than in the 6^{th} grade, but easier than for students in the 4^{th} grade. Also, the items must be tried out with native speakers of English to make sure they are neither exceptionally easy nor difficult for this population of students who are assumed to have the level of academic English proficiency, which the ELL population is expected to move toward. At this stage in the process native-speaker feedback is critical. Appropriately, the majority of students in the pre-pilot were native speakers of English. Seventy-nine percent of the students in the pre-pilot indicated English as their home language.

Table 5 provides demographic information by grade gathered from a student background survey at the start of the task administrations (See Appendix B). Information about student reading ability (reading group level) was gathered from the classroom teachers for each student.

---

informant. Ideally, future studies would include more student informants at this stage to obtain a range of comments and approximate testing durations.

Table 5

Student Demographic Information for Phase 1 Pre-Pilot Tryouts, Raw number and Percentage (n=95)

Demographic	Grade 4 (n=34)	Grade 5 (n=34)	Grade 6 (n=27)	Grade Total (n=95)
Gender				
girl	17(.50)	16(.47)	17(.63)	50(.53)
boy	17(.50)	18(.53)	10(.37)	45(.47)
Home Language				
English	26(.77)	26(.77)	23(.85)	75(.79)
Spanish	7(.21)	5(.15)	2(.07)	14(.15)
Other[a]	1(.03)	3(.09)	2(.07)	6(.06)
Birthplace				
USA	33(.97)	32(.94)	26(.96)	91(.96)
Other[b]	1(.03)	2(.06)	1(.04)	4(.04)
Grade Entered School in the U.S.				
Pre-Kindergarten	32(.94)	33(.97)	25(.93)	90(.95)
Kindergarten	0	0	2(.07)	2(.02)
Grade 1	2(.06)	1 (.03)	0	3(.03)
Reading Level				
Low	5 (.15)	10 (.29)	3 (.11)	18(.19)
Mid	14 (.41)	13(.38)	13(.48)	40(.42)
High	15(.44)	11(.32)	11 (.41)	37(.39)

[a]Other home languages include English & Spanish (n=3), Arabic & English (n=1), Korean (n=1) and Mandarin (n=1).
[b]Other birthplaces include Argentina (n=1), Australia (n=1), Ireland (n=1) and Mexico (n=1).

In Grades 4 and 5, there was an even to nearly even ratio of girls to boys; in the 6th grade, the girls outnumbered the boys. About 80% of the students reported English as the primary language they spoke at home. Nearly all of the students (96%) reported being born in the United States and entering school in the U.S. (95%). Students were rated on their reading level by their primary classroom teacher. The teachers reported a roughly even mix of middle and high level readers, with some students (about 20%) being rated as low level readers. In summary, the students tested in this portion of the project represented a somewhat diverse population with a range of English language reading abilities.

Whole group administration procedure. Two or more researchers administered the two forms of the test booklet. Thirty-five students received Form A and 42 students received Form B. These were randomly administered to students within or across classrooms if there was more than one classroom at a grade level. The test booklet also contained a short demographic survey about student language use, years in the U.S., etc. (See Appendix B for

survey). Administrations lasted between 30 and 50 minutes and followed a typical formal testing situation with students working on their own in silence with one exception. Students were instructed to attempt each task and to raise their hand to ask about any directions or formatting that were unclear to them (See Appendix C for protocol of directions for whole group administration). The questions were documented and entered into the audit trail database. Requests for help with answering the AELP tasks were responded to with encouragement to complete the task but not with the correct answer to the task. Students were asked to record their finish time from a large clock provided for this purpose.

Verbal protocol individual administration procedure. In the current study, our purpose for conducting the verbal protocols (VP) was two-fold: (a) to gain more in-depth information, which is unavailable from whole-class administration, about the reading tasks, and (b) we also took this opportunity to preliminarily explore the correspondences between the academic English construct and the mental processes of test-takers. In the first instance, to acquire additional information about the tasks, we probed specifically for information about the clarity of the direction and instruction, the difficulty level of the prototype tasks, and the potentially problematic linguistic components such as discrete vocabulary words. In the second instance, we asked whether test takers when given an item that measures academic English vocabulary for instance, actually process and comprehend the item stem to reach the correct answer, or if not, did they arrive at the correct response without any analytical and judgmental activity? Test-takers might get the right answer for the wrong reason, or the wrong answer for the right reason.

Description of the verbal protocol subsample. The data for the verbal protocol analysis come from 18 of the laboratory school students included in the first phase. These students were selected by their teachers such that students from each grade (i.e., Grades 4-6) and different language backgrounds and reading levels are represented. The descriptive statistics of these 18 students are shown in Table 6. Slightly higher ratios of girls to boys are represented in the sample where the number of girls in the 5^{th} and 6^{th} grade are greater than the number of boys. With the exception of one student, all students stated being born in the United States and entering school in the US. Five of the students reported Spanish as at least one language spoken in the home. The students' range of reading ability (indicated as low, mid or high by their teachers) was distributed similarly across the three levels.

Table 6

Student Demographic Information for the Verbal Protocol Subsample, Raw Number and Percentage (n=18)

Demographic	Grade 4 (n=8)	Grade 5 (n=5)	Grade 6 (n=5)	Total
Gender				
Girl	3(.38)	3(.60)	4(.80)	10(.56)
Boy	5(.63)	2(.40)	1(.20)	8(.44)
Home Language				
English	6(.75)	4(.80)	3(.60)	13(.72)
English & Spanish	1(.13)	1(.20)	0(.00)	2(.11)
Spanish	1(.13)	0(.00)	2(.40)	3(.17)
Birthplace[a]				
USA	8(1.0)	5(1.0)	4(.80)	17(.94)
Other	0	0	1(.20)	1(.06)
Grade Entered School in the U.S.				
Pre-Kindergarten	8(1.0)	5(1.0)	4(.80)	17(.94)
Kindergarten	0	0	1(.20)	1(.06)
Reading Level				
low	3(.38)	0(.00)	2(.40)	5(.28)
mid	2(.25)	3(.60)	2(.40)	7(.39)
high	3(.38)	2(.40)	1(.20)	6(.33)

[a]Other birthplace is Mexico (n=1)

Specifically, these students were asked to read the passages aloud, underline any unfamiliar/challenging words or phrases and then "think out loud" as they were answering each item (See Appendix D for the verbal protocol administration directions). Before starting the exercise, students were asked to share the test taking strategies they had learned from their teachers in order to gauge students' previous experience taking standardized tests. They were also asked to select the easiest and most difficult passages (and reasons why) after completing the exercises. Verbal protocol sessions ranged from approximately 25 to 60 minutes in total, with some students completing the exercises across multiple sessions.

Analytic Plan for Quantitative Analysis of Pre-Pilot Data

The following psychometric analyses were conducted on the data collected in the whole group and verbal protocol tryouts. First, descriptive statistics were calculated, specifically means, standard deviations, and range of performance for the overall sample. This information was further subdivided to provide descriptive details by grade level. Second, item difficulty was calculated for each item as the proportion of correct answers over total answers for dichotomously scored tasks, and as the simple mean score, expressed as a

number between 0 and 1, for partial credit scored tasks. The item difficulty was subdivided into the three grades in the study (4^{th}, 5^{th}, and 6^{th}).

Finally, item discrimination was conducted on the tasks for which sufficient data were available. This analysis was conducted to determine the extent to which test tasks distinguish between masters of the content and non-masters of the content, and was operationalized as the difference in mean scores of students whose teachers classified them as proficient readers and students not classified by their teachers as proficient readers. To summarize, the sequence of analyses were:

I. **Distribution of scores:** overall and by grade level

II. **Percent Correct (Item Difficulty):** for each item sample

III. **Discriminant Item Function Analysis:** using teacher reported reading levels

Analytic Plan for Qualitative Analysis of Pre-Pilot Data

Researchers took a selective versus extensive approach to transcribing and analyzing the verbal protocol data collected from the 18 additional students. First, summary questions regarding students' thoughts about the easiest and most difficult passages and questions as well as student general feedback on the format (i.e., directions, layout) of entire test booklet were transcribed. Next, nine tasks (six retained tasks and three tasks earmarked for rejection at this stage in the test development effort, see Chapter 4 for details) were transcribed in their entirety and coded for six further aspects of task characteristics commented on by the students: *format* (e.g., font size, familiarity with charts), *directions* (e.g., student interpretation of what a question is asking them to do), *word level comments* (e.g., familiarity of vocabulary, use of abbreviations), *item level comments* (e.g., what response students felt they should give to a item), *test taking/reading comprehension strategies* (e.g., how students find the right answer within a reading passage), and *use of background information/prior knowledge* (e.g., prior study of a topic).

Reliability of coding student comments for these six categories was conducted on approximately 50% the data (182 coding decisions) by combinations of two of the three researchers who coded the transcribed student comments. Simple inter-rater agreement (agreements divided by agreements plus disagreements) ranged from .80 to .91. Disagreements were resolved by consensus.

Refinements to Tasks

The initial informant commented on 17 tasks, namely feedback pertaining to formatting issues and typographical errors (n=13) as well as content (n=4) (see Appendix A, audit trail database, for further information). One task, based on the informant's feedback, was later

split into two tasks to total 101 tasks for the pre-pilot and pilot studies. An example of a substantive comment was to change the ordering of two noun phrases in a gap-fill sentence. Specifically, test-takers were requested to fill in the blank using words from a science passage. The relevant words in the passage were "much drier," so the order of the two geographical areas needed to be reversed so the target words could be used and the sentence completed in a meaningful manner otherwise the correct answer would have required students to introduce antonyms (e.g., "much wetter") of the target vocabulary in the passage.

Original:

Northern California is _____ than Southern California.

Modified:

The southern part of California is _____ than Northern California.

Before moving to the Phase 2 pilot, tasks also under went modification as necessary based on the results of the pre-pilot tryouts. Where possible, we refined tasks in terms of continued formatting errors or confusions, clarity of directions, and ambiguity in the tasks. In some cases, tasks were rejected if they could not be easily modified. In some cases where tasks showed no discrimination across masters and non-master, they were also rejected (See Chapters 3 and 4 below for further details and discussion of these processes).

Phase 2: Pilot Administration of Retained Draft Tasks for Prototype Creation

Pilot Sample and Procedures

Description of pilot sample. Following UCLA IRB procedures, two urban elementary schools in Southern California were recruited for the pilot study. School one consisted of predominantly Caucasian students (66.4%). The other major ethnic groups were Hispanic (16.1%) and Asian (8.4%). There were 6.2% English learners, and 9.3% of the student body qualified for free/reduced-price meals. In contrast, the majority of the student population in the school two was Hispanic (84.4%). More than half of the student body was designated as English learners (59.2%), and a high proportion of the students qualified for free/reduced-price meals (81.8%). The average Academic Performance Index (API) score (calculated with standardized and standards-based statewide assessments) for California in 2006 was 720, and the API scores for the two schools were 887 and 661, respectively (Source: http:// dq.cde.ca.gov/dataquest).

A total of 128 students across Grades 4 though 6 participated. Table 6 provides details about the students' backgrounds obtained from the background survey presented in the front of the pilot test booklet (See Appendix B). Tests scores for the California Standards Test in

English Language Arts (CST-ELA) were also available for 121 of the students. For 73 students who were designated ELL students by their districts, scores on the California English Language Development Test (CELDT) were also available. The Phase 2 student sample included native-English speaking students for the reasons mentioned above, but a larger number of ELL students was included as well to ensure information about the revised tasks from the target population.

Sixty percent of the students were in 5th grade which was deliberate over-sampling because this is the target grade for the 5th grade standards-based tasks we had developed. There were slightly more girls as participants in grades four and six than boys. There were comparable numbers of boys and girls in the 5th grade. About 41% of the students reported English as the main language they spoke at home, whereas 56% reported Spanish as their main language. The majority of students (78%) reported being born in the United States and entering school in the US (86%), as either a preschooler or Kindergartner. Table 6 also shows that the average test scores on the CST-ELA and the CELDT are higher for the 5th grade than for other grades. This was likely the effect of the higher API school which contributed only 5th graders to the study. However, even the CELDT scores of the ELL students are higher for the 5th grade than for the other grades and most of these ELL students came from the lower API school.

Table 6

Student Demographic Information for Phase 2 Pilot, Raw Numbers and Percentages (n=128)

Demographic	Grade 4 (n=20)	Grade 5 (n=77)	Grade 6 (n=31)	Total
Gender				
Girl	12 (.60)	38 (.49)	19 (.61)	69 (.54)
Boy	8 (.40)	39 (.51)	12 (.39)	59 (.46)
Home Language				
English	6 (.30)	40 (.52)	6 (.19)	52 (.41)
Spanish	10 (.50)	27 (.35)	21 (.68)	58 (.45)
Other[a]	4 (.20)	10 (.13)	4 (.13)	18 (.14)
Birthplace				
USA	18 (.90)	59 (.77)	21 (.68)	98 (.77)
Other[b]	2 (.10)	15 (.19)	10 (.32)	27 (.21)
Missing	0 (0)	3 (.04)	0 (0)	3 (.02)
Grade Entered School in the U.S.				
Pre-Kindergarten	4 (.20)	36 (.47)	7 (.22)	47 (.37)
Kindergarten	15 (.75)	31 (.40)	17 (.55)	63 (.49)
Grade 1 or later	1 (.05)	10 (.13)	7 (.22)	18 (.14)
CELDT Standardized Score [SD]	483.77 [35.41] (n=13)	517.40 [44.31] (n=35)	499.84 [76.42] (n=25)	505.40 [57.00] (n=73)
CST-ELA Standardized Score [SD]	295.26 [43.42] (n=19)	362.59 [62.39] (n=74)	310.96 [33.44] (n=28)	340.07 [61.08] (n=121)

[a]Other home languages include English & Spanish (n=14), Vietnamese & English (n=1), Filipino & English (n=1), Malaysia (n=1), and Urdu (n=1).
[b]Other birth places include Mexico (n=22), Kenya (n=1), Salvado (n=1), Colombia (n=1), Phillipines (n=1), and South America (n=1).

Analytic Plan for Quantitative Analysis of Pilot Data

The following psychometric analyses were conducted on the data collected in the pilot phase. First, descriptive statistics were calculated, and the mean, standard deviation, and range of scores for all tasks were determined both for the entire sample and for subgroups such as grade level, gender, and ELL status. Correlations between percent of items correct and the standards-based test scores were calculated. Next, item difficulty was again calculated as the proportion correct, or as the mean score on partial credit scored tasks expressed as a number from 0 to 1.

Item discrimination analysis was then conducted, using test takers' designation on the California Standards Test of English Language Arts (CST ELA) as the determinant of Master/non-Master status. Although the target construct of the AELP tasks is academic

English language rather than English language arts, we used the CST ELA scores to create the mastery categories rather than the CELDT Reading subtest scores because more students had CST ELA scores available (i.e., both English-only and ELL students totaling 121 students). However, we are reasonably confident that the CST ELA also reflected the ELL status of the students due to the very high correlation between CELDT Reading scores and CST ELA scores for the ELL students (r (69) =.754, p < .0001). Students designated as Proficient or better at the 5th-grade level (regardless of students' actual grade) were judged to be masters of the content, while students not meeting this standard were judged to be non-masters. Item discrimination function was calculated as the difference in mean score between these two groups.

The focus of the project was on the creation of task templates or prototypes, and not on the creation of a test. Moreover, it was not possible to conduct reliability with the pilot data for two reasons: First, the total test scores from the two schools in our study resulted in a bimodal distribution from which we could not calculate alpha values or other inferential statistics. Second, while splitting the dataset into two separate populations by school would have eliminated the first problem, another would have been created. Sample sizes for the two schools, once separated and after the removal of any students who did not complete the test in its entirety, would have proven too small for meaningful analysis. To summarize, the sequence of analyses for this phase was:

I. **Distribution of scores:** overall and by grade level

II. **Correlations:** using available CST- ELA and CELDT scores

III. **Percent Correct (Item Difficulty):** for each item sample

IV **Discriminant Item Function Analysis:** using CST ELA scores

The following chapter provides the results of the pre-pilot and pilot phases of the project.

CHAPTER 3:
RESULTS OF THE PRE-PILOT AND PILOT STUDIES

Results of Quantitative Analysis of Phase 1 Pre-pilot Data

Distribution of Scores

While the focus of the analyses for this report is item-level rather than student-level performance, for completeness we present the average raw scores and proportion correct for the pre-pilot sample in Table 7. Overall, the average raw score was roughly half of tasks correct (test Forms A and B contained 66 and 62 items, respectively). The trend of student performance being better for the 6th grade than for the 5th grade and being worse for the 4th grade than for the 5th grade was observed. The amount of variation in student performance relative to the mean was similar across the three grades. Student performance is also reported as the proportion correct of all attempted tasks to take account of different amounts of time taken or allowed for the completion of the task booklets. As Table 7 shows similar trends to raw scores were also obtained.[5]

Table 7

Mean Raw Score and Proportion Correct by Grade and Overall (n=95)

	Grade 4 (n=34)	Grade 5 (n=34)	Grade 6 (n=27)	Total (n=95)
Raw Score (SD)	20.44 (8.99)	26.47 (10.30)	38.96 (14.72)	27.86 (13.48)
Proportion Correct (SD)	.67 (.14)	.73 (.14)	.78 (.11)	.72 (.14)

Item Difficulty

Item difficulty (P values) was calculated for all tasks in the total pre-pilot dataset (n=95). However, some tasks had a low number of students attempting the item, making interpretation of item difficulty problematic for some tasks. Therefore, a 95% confidence interval was created around the P values (the proportion of test takers who got the item correct) to allow us to determine how sample size would affect interpretations. Appendix E contains P values with 95% confidence intervals for all tasks trialed in this stage.

Despite some large confidence intervals stemming from the sometimes small samples available, this item difficulty index was useful at this early stage. First, the majority of tasks had difficulty estimates in the middle range, indicating that most tasks were neither

[5] Further analyses were conducted at the item level to examine any trends that suggest items distinguished between gender and home language backgrounds of students (See Chapter 4 task profiles).

exceptionally easy nor difficult for this population. For these data, P values ranged from 0 to 1. Figure 2 shows the range of P values observed for all 100 tasks, with the X axis representing tasks ordered by increasing difficulty and the Y axis representing the P values, or proportion correct (0 =no students answered correctly, 1 =all students answered correctly).

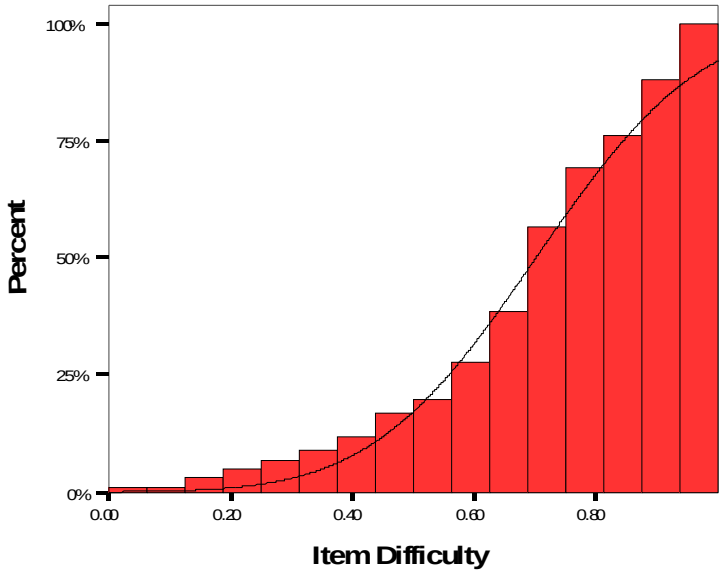


Figure 2. Cumulative Histogram of Item Difficulty of the Pre-pilot Data.

Second, for tasks which had moderately sized confidence intervals, inferences were made about whether the item was too easy or too difficult, and either modified, deleted, or flagged for further examination. If an item had a large confidence interval, indicating that judgments about item difficulty were likely to be inaccurate at this stage, it was not deleted due to excessive difficulty or easiness at this stage.

Item Discriminant Functioning Analysis

This portion of the analysis dealt with calculation of an item discrimination index for further analysis of item performance. Essentially item discrimination is the difference in mean score on an item between groups of study participants. In this case items need to discriminate between a high performing or "masters" group and a low performing or "non-masters" group; it is a measure of the extent to which the content of the item is harder for the non-masters group than for the masters group (Bachman, 2002). It can indicate whether the item is in fact measuring something that differs between the two groups, in other words if the item is measuring the right construct. However, an item will of course have low discrimination if it is very easy or very difficult for most test takers, thus interpreting this

statistic must be done in combination with other indicators of item performance such as item difficulty.

The first procedure was to identify the two groups: masters and non-masters. Two sources of information were collected: grade and reading level as determined by teacher. Since there was likely to be a degree of overlap between grades, and reading level is relative to grade, the two pieces of information were combined in the following way to create the following simplified "reading group" variable:

Table 8

Reading Group "Master/Non-Masters" Variable Creation

grade	Teacher Rating	Reading group
4th	Below	1
	At	1
	Above	2
5th	Below	1
	At	2
	Above	3
6th	Below	2
	At	3
	Above	3

Thus, the three groups created were:

1 =4th graders reading at or below grade level, and 5th graders reading below grade level

2 =4th graders reading above grade level, 5th graders at grade level, and 6th graders below grade level

3 =5th graders above grade level, 6th graders at or above grade level

The Level 2 group scores were then eliminated from the data, to set up a clear distinction between the "masters" (Reading Group 3) and "nonmasters" (Reading Group 1). In other words, it was reasonable to expect that everyone in Group 3 did well on this test, and that everyone in Group 1 should have struggled. This also had the advantage of combining groups (from nine categories to three) and increasing the sample sizes of the groups.

Unfortunately, available data were limited especially because there were few students in Reading Group 1; many of these students answered relatively few questions. A cut-off size of 10 in each group (which actually included several tasks with 9 non-masters) was set in order to reduce the number of potentially misleading conclusions that could be drawn from

such small sample sizes. Forty-two tasks were then eligible for item discrimination analysis, as the table in Appendix F shows.

In general, psychometricians recommend (e.g., Purpura, personal communication, 2006; Bachman, 2002) that all tasks have a discrimination value of greater than .3, which would mean 12 of 42 tasks examined were acceptable according to this criterion. A further 8 items had marginal discrimination with values between .2 - .29. These items are in need of improvement. The remaining 22 items had values below .2 and these are in need of rejection or extensive modification and improvement before being retried.

Refinement of Tasks

This pre-pilot phase was not viewed as "validating" any of the tasks, but merely as identifying as many problematic tasks as possible given the data available. Given that the pre-pilot data did not allow for definitive quantitative conclusions about the effectiveness of every draft task (i.e., some items were answered by too few students), we focused on tasks which did have sufficient responses and seemed not to be working well according to item difficulty or item discrimination functions. These tasks were then later subject to extra scrutiny for possible refinement or even rejection (See Chapter 4).

Tables 9 and 10 show the test form, subject matter, content of the reading passage upon which a task is based, and the question number of the task that was identified as problematic (either exceptionally easy or hard) according to the item difficulty analysis. In instances where the tasks overlapped across Forms A and B, the form designation in the table is C. Only one task was found to be extremely difficult, whereas nine tasks were found to be extremely easy.

Table 9

Exceptionally Difficult Task (P < .2)

Form	Area	Passage	Question No.
C	Soc Stu	George Washington	6

Table 10

Exceptionally Easy Tasks (P> .9)

Form	Area	Passage	Question No.
C	Math	Stilt Walker	4
C	Math	Stilt Walker	2
C	Science	Water Cycle	6
C	Science	Water Diagram	1
C	Math	Stilt Walker	3
A	Soc Stu	Colonial Women	8
B	Math	Organizing Books	5
C	Math	Lemonade	1
C	Math	Lemonade	2

Table 11 shows comparable information about tasks that were identified as problematic according to the item discrimination analyses. Twenty-two tasks were identified as having poor item discrimination.

Table 11

Tasks with Poor Discrimination (D < .2)

Form	Area	Passage	Question No.
A	Math	Camping	1
A	Math	Camping	2
A	Soc Stu	Colonial Women	1
A	Soc Stu	Colonial Women	4
A	Soc Stu	Colonial Women	5
A	Soc Stu	Colonial Women	6
A	Soc Stu	Colonial Women	7
A	Soc Stu[a]	Colonial Women	8[a]
A	Soc Stu	Colonial Women	9
A	Soc Stu	Colonial Women	10
A	Soc Stu	Colonial Women	11
B	Math	Traffic Light	2
C	Math[a]	Lemonade[a]	1[a]
C	Math[a]	Lemonade[a]	2[a]
C	Math	Lemonade	*4*
C	Math	Stilt Walker	1
C	Math[a]	Stilt Walker[a]	3[a]
C	Science[a]	Water Cycle[a]	6[a]
C	Science	Water Cycle	*7*
C	Science[a]	Water Diagram[a]	1[a]
C	Soc Stu	George Washington	*4*
C	Soc Stu	George Washington	6

[a]Tasks that were also flagged for high or low difficulty.

Although the focus of these analyses is the individual tasks, some passages had multiple problematic tasks, raising questions about the suitability of some of the passages themselves. In particular, the mathematics “Lemonade” passage was notable for its multiple issues, and the social studies “Colonial Women” passage was notable for its poor discrimination function in general.

Qualitative Results of Verbal Protocols

Review of Verbal Protocols

The verbal protocol originated from the field of cognitive psychology as a research technique to gain information about human cognitive processes (Ericsson & Simon, 1993). Undertaking an information processing theory that assumes information is stored in short

term memory and thus available for retrieval, Ericsson and Simon (1993) proposed the use of verbal protocols to elicit such information. Although they were initiated and established in cognitive psychology, they had been widely adapted in other disciplines to gather information about learners' cognitive processes. Despite frequent criticism on the reliability and validity of verbal protocol methodology and the challenge to the assumption that a verbal protocol is a solely cognitive instead of socially constructed activity (Pressley & Afflerbach, 1995; Smagorinsky, 2001), the use of verbal protocols remains unabated because the methodology yields data on the cognitive processes that otherwise would have to be studied only indirectly.

Specifically, in the past few decades, verbal protocols have become an increasingly popular research methodology for examining the strategies employed in taking language tests (Cohen, 1998, 2000). In the extant literature, the focuses of test taking strategy research that utilizes verbal protocols range from validating the language tests (e.g., Anderson, Bachman, Perkins, & Cohen, 1991; Storey, 1997), understanding tests of specific skills such as listening and reading comprehension (e.g., Anderson, 1991 & Nevo, 1989 for reading comprehension tests), to the investigation of strategies related to learner characteristics such as gender and language proficiency level (e.g., Purpura, 1997, for language proficiency level).

Cohen (2000) categorized verbal protocols into three major data sources: *self-report*, *self-observation, and self-revelation*. *Self-report* refers to participants' generalized statements about their test-taking strategies (e.g., "If there is a word I don't understand, I read it and if it sounds like a word before that I've actually heard of, I'll try connect it and see if it sounds similar" reported a 6th grader from the study), and is usually elicited via questionnaires or prompts that request general description of participants' use of strategies. In contrast, *self-observation* involves inquiries into specific, rather than generalized, language behavior or responses to test tasks. Self-observation data can be collected via introspection, which can be *retrospective introspection,* which entails references to actual instances of strategy utilization on specific test tasks (e.g., "I was trying to see if there is any clue. Then I saw it [the answer] in the passage" as one 5th grader reported).

On the other hand, *self-revelation*, also more commonly known as a "think-aloud," attempts to capture participants' stream-of-consciousness thought processes while they are attending to the information (e.g., "I can't find anything that would go after *was* from this passage [to put in the blank]", according to another study 5th grader). Although introspective self-observation also involves online processing, self-revelation distinguishes itself from introspective self-observation by its basic assumption of narrative description of thoughts; if the participant consciously analyzes his/her thoughts rather than simply describes them, the piece of data immediately falls into the introspective self-observation category instead.

Verbal protocols of students completing test tasks usually use some combination of these strategies (Cohen, 2000). In the case of this study, we used the more familiar "think-aloud" or *self-revelation* technique and *self-observation* specifically *retrospective introspection* to understand student responses to the draft tasks.

Results of "Think-aloud" Self-Revelation Data

The students' comments, as they thought aloud, revealed the dynamics of comprehension difficulties. As mentioned above, six types of information on the draft tasks could be categorized and emerged from the students' spontaneous comments as they completed the tasks. These categories were labeled: *format, directions, word level, item level, strategy, and background information*. To reduce subjectivity in coding the data, three raters independently read and scored two types of responses. Each rater scored one additional kind of response to check for inter-rater reliability. The few discrepancies, which occurred in coding, were resolved through consensus. The six categories are discussed with examples below:

Format

There were only three verbal protocol comments made by three different students on the format of the passage and tasks, all of which pertained to the Stilt Walker passage. Two of the instances were students' spontaneous comments on the format of the passage (*The period after the second sentence "threw me off because they told me to stop."* Comment from a 6th grade student) and the item (Referring to a specific word in the item prompt: "*What does it mean 'from'?"* Comment from a 4th grade student). However, since the passage was taken from a California-approved Grade 5 mathematics textbook without any modification, no revision was made in order to maintain the authenticity of the passage. The student immediately understood how to carry out the task and did not seem to have trouble with the stem. Consequently, no modification was made to the stem. The other instance was from the researcher's observation regarding the limited space provided for student response. (Researcher comment: *One student wrote outside the blank provided. The blank space seemed to be limited.*) The blank space was then increased based on the researcher's comment.

Directions

The directions are seen as an interpretation of what the item is asking the student to do. The reactions that students exhibited in interpreting what the directions required of them are diverse, although there were only eight comments of this type made by six of the students. The transcription of a student's response found below shows that the student had not understood the directions precisely.

Example 1

A student asked if she should write 'stilt walker' or 'the man.' The interviewer clarifies by saying, "either stilt walker 1 or stilt walker 2."

Word Level

One hundred and eighteen comments during the verbal protocol related to word-level issues. All 18 students participating in the verbal protocol made at least one comment on word level aspects of the tasks as they completed them. In general, it appears that unfamiliar vocabulary interfered with reading fluency and comprehension, as well as the completion of task tasks. We categorized those instances into three sources of difficulties at the word level: novel words, word form variation, and morphological variation. We also asked students to highlight which words they did not know throughout the test booklet. We report on these words at the end of this section on word level issues. The three categories of unfamiliar vocabulary identified during the students' self-revelation comments, however, are described below.

1. Novel words. Novel words, in our definition, were words and phrases for which students had not acquired the full meanings. As described in the Highlighted Unfamiliar Vocabulary section below, the majority of novel words identified by the students were either academic vocabulary (general and specialized), measurement words, or proper nouns. To illustrate, many students had difficulties with specialized academic vocabulary such as *evaporation*, *condensation*, *precipitation* (science-based words), and *surveyor*, *colonist*, *plantation* (social studies-based words). They paused, mispronounced, and struggled with those words. Those specialized academic words appeared to have caused a great amount of frustration (e.g., Researcher comment #1: *The student stumbled on many words and had to either pause or repeat a few times at novel vocabulary. He was very frustrated with the passage.* Researcher comment #2: *The student had to try many times before she got the correct pronunciation for "evaporation" and one or two other words*).

A few general academic words were also identified as challenging by many students. For example, some students had only learned the word *factor* as a mathematical term. They were thus confused about its meaning when it was used to denote the cause of a situation in an item stem (Item Stem: According to the passage, what is the third factor used in describing the weather?). Most students understood what *factor* meant in the specific context after we explained the additional meaning to them. However, not all of the students had answered the item correctly. (Researcher comment*: The researcher asked whether the student knew what "factor" meant. The student said "no" and talked about the meaning of "factor" in mathematical equations. The researcher then explained to him what "factor" meant in this context*).

Another obstacle pertained to proper nouns. Students appeared to have great difficulties with unfamiliar names of people and places. To illustrate, many students had to repeat the Spanish name *Carlotta* a few times and commented that they for example *"didn't recognize the name*" (comment from a 4th grader). Another example was the word *Westmoreland* in the social science passage. (Researcher comment: *The student said that Westmoreland was kind of difficult for him because he didn't know where it is*). Other examples included *Burgesses, Hessian, Custis, Vernon.*

Students had employed different strategies to understand novel words. For example, some students would ask for help from the researcher (*What is a surveyor?* Question from a 4th grader), while others speculated on the meaning based on contextual cues (*I didn't know how to pronounce it [militia] and didn't know what it was, but it made me think of the military and also the next sentence was about military*. Comment from a 6th grader).

2. Word form variation. Most students found the abbreviation of "miles" into "mi" in one of the mathematics word problems to be unfamiliar or confusing (*The student stumbled upon the word "mi."* Researcher comment). Only a few students said that it is not a problem either because they had learned about abbreviations before (*We learned all the abbreviations last year in 5th grade.* Comment from a 6th grader student) or because they were able to derived the meaning of the abbreviation from the context (*I didn't know the abbreviation "mi." I thought that was probably miles, I'm guessing.* Comment from a 5th grader student). One student suggested that the abbreviation be spelled out for easier comprehension (*It would be easier for some students if it is spelled out.* Comment from a 6th grader student).

3. Morphological variation. Morphological variation appeared to also constitute one major source of challenges. Our analysis revealed that students had frequently struggled with verb forms, particularly in answering the tasks. Take the following social studies-based item for example:

George Washington's wife was _____ on the battlefield.

In the spontaneous comments on her decision-making process, a 4th grader described how she came to choose "supportive" rather than "supported": "*Because I saw that she sew socks and cooked soup for the solider*." Other students had also expressed confusion with the verb forms (*Surprised? Surprising?* Comment from a 4th grader).

An interesting finding worth mentioning was the observation that one single novel word should have easily interfered with students' reading comprehension or solving the tasks. Regardless of the category of the word, students appeared to easily become "stuck" on

a single, unfamiliar word even though the word itself might be insignificant (in the passage). For example, unfamiliarity with the Spanish name *Carlotta* in the mathematics word problem scenario should not have been critical to understanding the passage and the associated items. However, clearly unfamiliar proper nouns caused the students taking these tasks a great amount of frustration.

Highlighted Unfamiliar Vocabulary

During the verbal protocol session, students were asked to highlight difficult or unfamiliar words in the passage and tasks that pose challenges for them in reading comprehension or answering the tasks (see Appendix G for the list of identified vocabulary by passage). For the analysis, we categorized the vocabulary across the six passages based on the former coding scheme we had created for textbook analysis (Butler et al., 2004). The identified vocabulary fell into the following categories: general academic vocabulary, specialized academic vocabulary, measurement words, and proper nouns. Both the type and token of identified vocabulary vary with the content area and specific passage and associated tasks. The discussion of the findings presented below is thus structured by content area and by passage. Note that the list of vocabulary students highlighted did not include all of the words that were challenging for students. We found from a comparison between the list and students' spontaneous comments that many words were left unidentified by the highlighting process. However, the list is nonetheless informative and provides us with a general picture of the difficulties students encounter at the word level.

Not many words in the three math-based passages (Camping Trip, Lemonade, and Stilt Walker) were highlighted. The majority of those that were highlighted were proper nouns, such as names of place and people. Specifically, Appalachian from the Camping Trip passage was identified by 11 out of the 18 students as a challenging word and was also the only word highlighted in the passage. Additional examples of identified proper nouns include Carlotta (n=4 out of 18) from the Lemonade tasks and Bowen Moscow (n=3 out of 15) from the Stilt Walker tasks. The other two identified words were the general academic vocabulary word profit in one of the Lemonade tasks (n=3 out of 18), and the abbreviation of the measurement word mile/mi in both the Stilt Walker passage and item (n=2 out of 15).

Students identified several difficult words the two science-based passages (Water Cycle and Water Diagram), most of which fell into the specialized academic vocabulary category. Specifically, specialized words identified in the Water Cycle passage include precipitation (n=8 out of 18), cirrus (n=10), cumulus (n=11 out of 18), evaporation (n=4), condensation/condense (n=9), meteorologists (n=7), stratus (n=9), atmosphere/atmospheric (n=5), and humidity (n=4). There were also three general academic words identified as challenging, including condition (n=1), predict (n=1), and factor (n=3).

Because of the similarity in topical area, several of the specialized academic words in the Water Cycle passage, such as precipitation (n=5 out of 15), condensation (n=6), and evaporation (n=3), were also selected in the Water Diagram passage as challenging words. There was also one general academic word, entire, identified for the Water Cycle passage.

Finally, students highlighted a variety of academic words in the social studies-based passage, George Washington, including both general and specialized academic vocabulary and proper nouns. As in the math-based passages, proper nouns such as Virginia House of Burgesses (n=13 out of 17), Potomac (n=2), Custis (n=1), Continental (n=4), Hessian (n=4), Westmoreland (n=3), Vernon (n=1), were all frequently selected as sources of challenge. The other major category of difficulty was specialized academic words, which included militia (n=10), plantation (n=5), surveyor (n=10), colonist/colonial (n=5), and patriot (n=1). Additionally, a few instances of general academic vocabulary were also identified as obstacles in the reading task, including temperament (n=2), wealthiest (n=1), restored (n=1), and discouraged (n=1).

Item Level

Item level responses identified students' interpretations of the meaning of a text and/or the identification of correct responses for tasks. In all, twenty-two item level responses were identified in the comments of 15 students. The following example is a student debating whether a question refers to physical or temporal distance – i.e., how 'far' or 'how long.'

Example 2

> Student: How far or long is almost the same thing, so it was hard to decide…I decided on 'b' because it said 'how many miles.' Then the student said 'how long' is like maybe how many hours.

Difficulties in determining coding also arose when more than one strategy seemed to overlap in a response to the assignment. As the example illustrates below, one student misinterpreted the science-based question in a way which can be classified as an 'item level' response, based on both the context of the response and inferences drawn from researcher comments. Additionally, the student's response shows increasing build-up of both 'strategy' and 'word level' embedded in the response.

Example 3

> Student mispronounced *precipitation* as *participation.* She highlighted this word as well as *condensation*. For the item, student seemed to have misunderstood the question because of the *factor*. She verbally counted the paragraphs and then looked for the answer in the 'third' paragraph because of the *third factor* phrase in the question. [Researcher comment]

Strategy Comments and Behaviors

The students' comments and behaviors as they completed the think-aloud protocol also revealed the strategies they used in order to formulate their answers to the tasks. Strategy comments, for example, included explicit references to how students found the answer to a question in the body of the text selections. Seventy-two strategy comments and behaviors were identified in the selected tasks. Seventeen of the 18 students contributed comments to this category. Most comments and researcher notes about behaviors suggested these students used a strategy of rereading the text to search for and find a specific word or phrase, or to narrow down the location of a sentence or paragraph that contained the language feature(s) being assessed by the task. One student explicitly stated that he simply "guessed" the answer to one of the tasks. Example 3, above, shows the error that another student made in interpreting the meaning of "third factor" as the third paragraph when asked to state the third factor involved in the water cycle.

Background Knowledge Comments

Comments were also made about background or prior knowledge as students completed the tasks. Such comments include references to information in the text that students were familiar with, such as the names of the towns in a mathematics word problem-based task (Los Angeles being an obvious one for this Southern California-based sample). Students would also comment on a lack of prior knowledge or familiarity with information in the texts. For example, four students said they didn't recognize the proper names used in a mathematics word problem-based task. There were just 23 background or prior knowledge comments made by 12 of the 18 students. The following two examples are representative of most of the comments made about using prior knowledge in order to complete a task.

Example 4

Student: Last year we studied the water cycle a lot so I know a lot of the words.

Example 5

[The student] tried to think from prior knowledge (i.e., previous diagram on water cycle) to answer this question. [Researcher comment]

Results of Self-Observation Retrospective Introspection

These data were generated after the students had completed the tasks during the "think aloud" session, by then asking them to comment on which reading passage they found easiest and which the hardest and which draft task was the easiest and which the hardest.

Analyses of the data revealed that passage length played a role in determining whether or not students found the text of the draft tasks easy or difficult. As shown in the examples from the transcribed data below, participants often considered passages that were short to be

easy (comments made by six of the 14 students who answered this question) [6] while determining the longer passages (n =5/14) to be difficult. In the following example the students refer to a mathematical word problem passage, consisting of three sentences describing how much Carlotta profited from selling lemonade drinks. The students identified the passage as an easy passage.

Example 6

Subject: The lemonade one because it was a short one and they just had a few questions. It was kind of easy because when it's long you just couldn't remember.

Similarly, the next example is commentary on the longest reading passage found in the test. Indeed, our previous textbook analyses (Butler et al., 2004) showed that social studies passages such as this one consisted of more complex structures and a variety of organizational features resulting in long passages.

Example 7

Subject: Well, I would say that this one was the hardest cause it's the longest.

Interviewer: George Washington [reading passage]? Did it seem like a passage that was easy to read, just long, or…?

Subject: It was just long.

Prior knowledge also affected the identification of easy and/or challenging passages. One student was able to remember the answers for a fill-in-the-blank question on the Water Cycle diagram because s/he had previously studied the subject. Likewise, for both of the examples below, students activated their background information to relate to the content of the text. Consistent with other observations found in the data, students exhibited more difficulty in processing the text when they were not familiar with the subject.

Example 8

Subject: George Washington. Because I have not studied George Washington before.

Example 9

Subject: George Washington probably.

Interviewer: Why do you think it's difficult?

Subject: Well, I think it's not um difficult…because I really didn't know that much about him. And there were a lot about, it was all like…back back in the old environment…there were all these different things going on. And I didn't really get most of it like…it was all kind of…oh I don't know.

[6] The denominator of the ratios is different for each subsection because of time constraints or technical difficulty (i.e., lost response on faulty tape recording). Therefore, the denominator corresponds to the number of respondents who answered a particular question rather than the total VP subsample of 18).

Among students who found a reading passage easy, prior background knowledge proved an important variable (n=1/2). Similarly, of the seven students who identified a text as being difficult, five attributed their difficulties to a lack of background knowledge. Thus, a closer look at the data reveals students extracting information from their background knowledge to complete the task.

Students' ability to recognize and comprehend vocabulary suggested which passages might be easy (n=2/14) and/or difficult (n=7/14). As is clear from the examples illustrated below, both students encountered an unknown word or phrase. The student in Example 10 below was aware of his comprehension difficulty, he did not attempt to resolve it, while in Example 11, the student attempted to backtrack to the text to figure out the unfamiliar words.

Example 10

Subject: George Washington. Because there were a lot of words in it. A lot of things like "militia," stuff like that. First I didn't know what all that was. Stuff like that, or "Hessian", "revolution".

Example 11

Subject: George Washington one. There are a few words that I didn't understand. Like that word and that word (point to the words in the booklet).

Interviewer: So the vocabulary in the passage?

Subject: Yeah.

Moreover, task specifications for this passage show that students must use vocabulary words present in the passage in order to complete the task. For these test tasks, both students provided incorrect answers to the George Washington passage when they were not able to successfully understand parts of the passage.

All students (n=11/11) who mentioned directions in their evaluations of the tasks reported directions to be clear and easily understandable. They were also readily responsive to this particular closed-ended question. Consequently, these students did not feel compelled to report any further information on the clarity of directions.

Summary

The analysis of the data generated by the verbal protocol was not without limitations. First, the ability to think aloud was probably not equal for all students. Some may have felt self-conscious about the activity itself, which was indicated by their preference for reading the passages to themselves, a preference often articulated during testing. Another concern is the lack of training received by the learners prior to actual data collection. Students who were chosen for the analysis could have been trained more intensively in order to assist them in providing a more complete description of their behaviors while performing the tasks. The

lack of training resulted in interviewers occasionally reminding the students to think aloud. Next, in coding of the responses during analysis of the verbal protocols, student responses overlapped in terms of the type of responses given. For example, students' single responses contained different response types embedded within the sentences. Another limitation of the protocol was the variation in student responses that were sometimes spontaneous and at other times elicited, given certain degrees of inconsistency in researchers' styles when interviewing the students. Despite these limitations, the verbal protocol analysis provided the researchers with insights into students' processing of the AELP tasks.

To conclude, the findings in the retrospective data identified patterns of easy and difficult passages and draft tasks as determined by text length, prior knowledge, and familiarity of the vocabulary for the students who participated. In sum, analysis of the verbal protocol findings provided information for the refinement of the draft tasks as illustrated in the task profiles in Chapter 4.

Results of the Pilot Phase

Distribution of Scores

Again, while the focus of the pilot study is item-level analysis, for completeness, we present the average raw scores and proportion correct for the pilot sample in Table 12. Overall, the average raw score was roughly half of tasks correct (the pilot form contained 40 items). As expected, the 4th-grade student performance was worse than 5th- and 6th-grade student performances. However, the 5th grade did slightly better than the 6th grade which is possibly explained by the larger percentage of students with Spanish as the dominant home language in the 6th grade (68%) compared with the 5th grade (35%) and the fact that among the ELL students, those in the 5th grade performed better on the CELDT on average than those in the 6th grade (See Table 6). The amount of variation in student performance relative to the mean was similar in Grades 5 and 6, but there was somewhat higher variation (relative to the mean) in the 4th grade. Student performance, reported as the proportion of answers correct of all tasks a student attempted, had similar trends as the findings for the raw scores.[7]

[7] Further analyses were conducted at the item level to examine any trends that suggest items distinguished between gender and home language backgrounds of students (See Chapter 4 task profiles).

Table 12

Mean Raw Score and Proportion Correct by Grade and Overall (n=128)

Category	Grade 4 (n=20)	Grade 5 (n=77)	Grade 6 (n=31)	Total (n=128)
Raw Score (SD)	8.61 (6.56)	21.75 (11.30)	18.71 (8.99)	18.96 (11.11)
Proportion Correct (SD)	.30 (.21)	.60 (.24)	.51 (.21)	.53 (.25)

Correlations with State Test Scores

Using these proportion correct scores for the students it is interesting to see whether student performance on the AELP tasks is related to their performance on concurrent state assessments of their English reading, and in the case of the ELL students, the standardized measure of their English language development. Correlations between percent correct on the AELP tasks and these additional assessments are very high. The correlation with the CST ELA assessment was r (121) =.707 ($p < .0001$). The correlation between the AELP percent correct and total CELDT score (a measure of listening, speaking, reading, and writing) was r (73) =.643 ($p < .0001$). The CELDT Reading subtest which is closest to the AELP tasks in both construct definition and content was not surprisingly even higher (r (70) =.725, $p < .0001$).

Item Difficulty

Item difficulty was calculated for all 40 tasks in the total pilot dataset (n=128). This statistic is the proportion of test takers who got an item correct (0 =no students answered correctly, 1 =all students answered correctly). Some tasks had slightly lower numbers of students attempting the item, but none fell below 98 students attempting an item.

The item difficulty index is displayed in Figure 3. The majority of tasks had difficulty estimates in the .50-60 range, indicating that most tasks were neither exceptionally easy nor difficult for this sample.

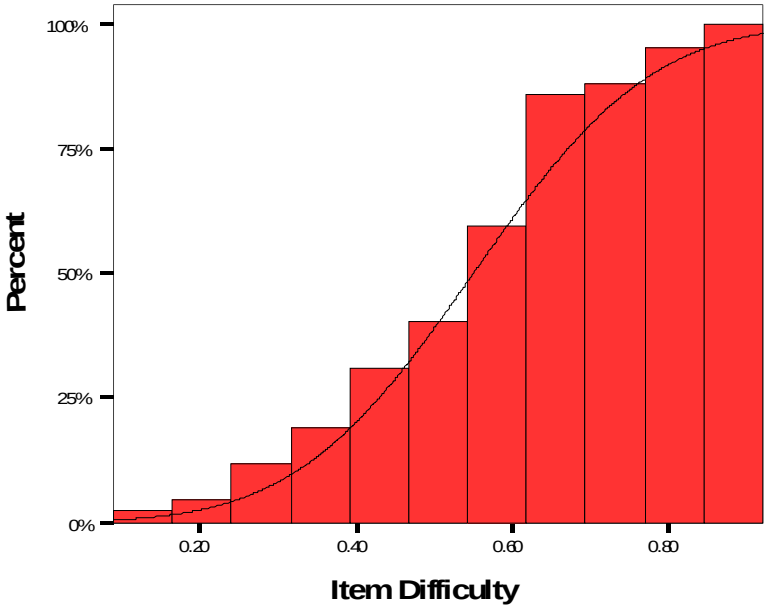


Figure 3. Cumulative Histogram of Item Difficulty for the Pilot Dataset.

Table 13 below shows the range of P values observed in this dataset for all 40 tasks. The clusters of items Q17-Q20 and items Q22-Q25 each constituted two non-independent forced choice tasks and consequently we also analyzed these two clusters as two tasks. These tasks are arrayed from the easiest item (Q31), with 92% of the students attempting this item correctly answering it, to the most difficult item (Q17-20 cluster) with just 9% of students answering it correctly.

Table 13

Average Proportion of Correct Answers by Item for the Pilot Data

Question	N	Missing	Mean	SD	Min	Max
Q31	98	30	0.92	0.22	0	1
Q1	125	3	0.87	0.27	0	1
Q11	111	17	0.82	0.32	0	1
Q2	125	3	0.78	0.34	0	1
Q38	124	4	0.77	0.36	0	1
Q5	119	9	0.75	0.37	0	1
Q4	125	3	0.69	0.39	0	1
Q12	111	17	0.69	0.41	0	1
Q33	123	5	0.68	0.42	0	1
Q13	111	17	0.66	0.42	0	1
Q10	111	17	0.66	0.43	0	1
Q39	116	12	0.66	0.43	0	1
Q14	111	17	0.65	0.44	0	1
Q16	103	25	0.63	0.47	0	1
Q35	125	3	0.62	0.47	0	1
Q29	99	29	0.62	0.47	0	1
Q22	102	26	0.62	0.47	0	1
Q6	116	12	0.60	0.47	0	1
Q9	112	16	0.60	0.48	0	1
Q27	99	29	0.57	0.48	0	1
Q30	99	29	0.57	0.48	0	1
Q23	102	26	0.56	0.48	0	1
Q34	125	3	0.56	0.48	0	1
Q36	125	3	0.56	0.48	0	1
Q24	102	26	0.55	0.49	0	1
Q19	102	26	0.54	0.49	0	1
Q28	99	29	0.53	0.49	0	1
Q20	102	26	0.50	0.49	0	1
Q15	109	19	0.49	0.49	0	1
Q32	100	28	0.45	0.50	0	1
Q3	125	3	0.45	0.50	0	1
Q26	98	30	0.44	0.50	0	1
Q37	125	3	0.42	0.50	0	1
Q7	114	14	0.42	0.50	0	1
Q17	102	26	0.37	0.50	0	1
Q21	102	26	0.37	0.50	0	1
Q8	113	15	0.33	0.50	0	1
Q25	102	26	0.31	0.50	0	1
Q40	115	13	0.26	0.50	0	1
Q18	102	26	0.25	0.50	0	1
Q22-25	102	26	0.23	1.62	0	4
Q17-20	102	26	0.09	1.70	0	4

Item Discriminant Functioning Analysis

This portion of the analysis deals with calculation of an item discrimination index for further analysis of item performance. As with the pre-pilot data analysis, item discrimination is the difference in mean score on an item between “masters” and “non-masters” of the content. In this analysis however, we were able to use student performance on the CST ELA to create the two groups; masters and non-masters. Eleven tasks with dark grey shading in Table 14 below discriminated poorly between the two groups (i.e., $D < .25$). Fourteen tasks with light grey shading discriminated moderately well (i.e., $.25 \leq D < .35$) (including a new cluster item), and the remaining 17 tasks without shading discriminated adequately ($D \geq .35$) (including a second new cluster item).

Table 14

Item Discrimination of the Pilot Data

Q	Non-Masters		Masters		Item
	N	Mean	N	Mean	Discrimination
Q21	52	0.423077	44	0.5	0.08[a]
Q39	66	0.212121	43	0.325581	0.11[a]
Q1	73	0.863014	45	1	0.14[a]
Q31	50	0.03	43	0.174419	0.14[a]
Q4	73	0.753425	45	0.911111	0.16[a]
Q16	53	0.226415	44	0.409091	0.18[a]
Q2	73	0.808219	45	1	0.19[a]
Q33	72	0.194444	44	0.386364	0.19[a]
Q11	61	0.467213	44	0.681818	0.21[a]
Q36	74	0.297297	44	0.522727	0.23[a]
Q9	61	0.557377	44	0.795455	0.24[a]
Q32	51	0.27451	43	0.55814	0.28[b]
Q34	74	0.256757	44	0.545455	0.29[b]
Q8	62	0.435484	44	0.727273	0.29[b]
Q19	52	0.480769	44	0.772727	0.29[b]
Q20	52	0.480769	44	0.772727	0.29[b]
Q38	73	0.34589	44	0.642045	0.30[b]
Q5	67	0.453731	45	0.768889	0.32[b]
Q18	52	0.365385	44	0.681818	0.32[b]
Q25	52	0.365385	44	0.681818	0.32[b]
Q17-20	52	1.711538	44	2.977273	0.32[b]
Q40	65	0.369231	43	0.697674	0.33[b]
Q6	65	0.215385	44	0.545455	0.33[b]
Q24	52	0.461538	44	0.795455	0.33[b]
Q37	74	0.324324	44	0.659091	0.33[b]
Q12	61	0.639344	44	1	0.36[c]
Q22-25	52	1.826923	44	3.272727	0.36[c]
Q3	73	0.30137	45	0.666667	0.37[c]
Q17	52	0.384615	44	0.75	0.37[c]
Q13	61	0.606557	44	0.977273	0.37[c]
Q23	52	0.480769	44	0.863636	0.38[c]
Q29	50	0.5	43	0.895349	0.40[c]
Q22	52	0.519231	44	0.931818	0.41[c]
Q35	74	0.540541	44	0.954545	0.41[c]
Q14	61	0.57377	44	1	0.43[c]
Q10	61	0.04918	44	0.477273	0.43[c]
Q7	63	0.380952	44	0.818182	0.44[c]
Q28	50	0.44	43	0.883721	0.44[c]
Q27	50	0.46	43	0.906977	0.45[c]
Q30	50	0.44	43	0.930233	0.49[c]
Q26	50	0.34	43	0.837209	0.50[c]
Q15	60	0.383333	44	0.909091	0.53[c]

[a]Poor discrimination. [b]Moderate discrimination. [c]Adequate discrimination.

Summary

Forty draft tasks had been retained or refined after the pre-pilot phase of the project. After piloting with 128 4th -6th grade students, findings show that the majority of these tasks were in the middle range of difficulty. Items that are on the extremes of the difficultly continuum might be avoided as exemplar tasks although an operational test of AELP reading at the 5th grade that had the purpose of measuring student development in AELP would require tasks at both the easy and difficult levels, as well as in the middle range.

Most tasks discriminated good readers from poor readers in the sample moderately to adequately. Those tasks with moderate discrimination are candidates for refinement and subject to further tryouts and re-piloting. While the tasks with adequate discrimination may also require further refinements, these are certainly the most promising tasks from the pilot stage and three are profiled as example effective prototypes in the next chapter.

CHAPTER 4:
EXAMPLE TASK PROFILES

Introduction

In this chapter, we present representative effective and ineffective tasks from three subject areas, mathematics, science, and social studies. Each profile consists of the task specifications, the target features of the AELP construct, the linguistic profiles, relevant audit trail entries, the pre-pilot and pilot results, and relevant verbal protocol excerpts. As mentioned in Chapter 1, the analytical framework for linguistic profiles is adapted from our previous work on textbook analysis (Butler, Bailey, Stevens, Lord, & Huang, 2004; see also Appendix H for the profile template).[8]

For retained/effective tasks, there are two example tasks based on mathematics word-problems and one based on a social studies text. No science task that was retained for the pilot phase had associated verbal protocol data collected at the pre-pilot phase (i.e., all science tasks with verbal protocol data were rejected after evaluation with the pre-pilot findings). Consequently, we profiled an additional mathematics text-based task in order to have relevant verbal protocol data to illustrate task effectiveness. For rejected tasks, there is one example task based on texts from each of the three subject areas, i.e., one mathematics word problem-based task, one science text-based task, and one social studies text-based task.

The task profiles serve as documentation of the evolution of example AELP tasks in our test development process. For each effective task profile, we first introduce the passage and the item. However, for passages with multiple paragraphs, we will only present selective paragraphs that are relevant to the task due to space constraints. We then present the task specifications and the linguistic analysis results, followed by the results from our Phase 1 pre-pilot and Phase 2 pilot studies. Specifically, for the pre-pilot tryout studies we will summarize the results and present the information from the audit trail, followed by our decision on the task after the initial tryouts. For whole group tryouts, we again present the passage and task with modifications highlighted. We then, in order, display the statistical results by different background variables such as grade and home language, give percentages and examples of the range in student responses, show excerpts of verbal protocol analysis and audit trail records, as well as report Phase 2 statistical results. We conclude each profile with a recount of the information to substantiate our argument for its role as AELP prototype

[8] We gratefully acknowledge the following publishers for permission to use textbook excerpts in the CRESST test development process: Harcourt for Math (2002) National Edition, Science (2000) California Edition and Social Studies: Early United States (2002) National Edition; Houghton Mifflin for Mathematics (2002) California Edition; Science (2000) California Edition, Social Studies: America Will Be (1999) National Edition; McGraw-Hill for Math Explorations and Applications (2003) National Edition, Science (2000) California Edition, United States: Adventure in Time and Place (2001) National Edition.

task. We present the comparable information in the same order for three example tasks that we rejected as ineffective AELP tasks.

Profiles for Effective Tasks

Task Profile I – Social Studies-based Task

Original Draft Task

Passage

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

[paragraphs omitted]

Certain of future victories, General Howe decided to rest for the winter in New York City. Washington knew that the British would not try to advance again until the spring. So he planned a surprise attack on the close to 1,400 Hessian troops in Trenton, New Jersey. The password Washington gave his soldiers was "Victory or Death!" After nightfall on Christmas Day, December 25, 1776, Washington and his troops crossed the Delaware River into New Jersey. The next morning, they surprised the Hessians, who quickly surrendered. "This is a glorious day for our country," said Washington.

Fill in the blanks using vocabulary words from the passage.

The Hessian troops were _____(attacked)___ by George Washington.

Task Specifications

Framework category: Vocabulary

- **General description and text type:** *Students will complete a sentence using vocabulary words that are defined in a multi-paragraph expository text.*
- **Task format:** *Sentence completion using words from the passage.*
- **Stimulus attributes:** *A multi-paragraph expository text generally consisting of 3-5 paragraphs.*
- **Response attributes:** *The stimulus is followed by incomplete sentences. Students complete each sentence by filling in the blank with the correct verb from the passage.*
- **Standard addressed:** ELD Standard addressed: *Advanced Vocabulary and Concept Development*; California Content Standard addressed: *Social Studies: 5.5 (4)*
- **Target Academic Language Constructs:** *Specialized and general academic vocabularies are the focal linguistic features. Also measured are academic language functions: "explanation," "description", "provide instruction/guidance" and "reference to text/visual"; simple and complex grammar.*

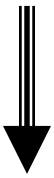
Linguistic Analysis Profile

	Stem/Prompt	*Response*
Descriptive Analysis		
(Mean) no. of words per sentence (range)	10	8
Sum of Words	10	8
Total # of words (token)[a]	10	8
Total # of words (type)[b]	9	8
Lexical Features		
Academic vocabulary - general (token)	3	
Academic vocabulary - general (type)	3	
Academic vocabulary – specialized (token)		3
Academic vocabulary – specialized (type)		3
Low-frequency words (token)	1	2
Low-frequency words (type)	1	2
3-or-more-syllable words (token)	1	1
3-or-more-syllable words (type)	1	1
Avg. % of nominalizations per selection	1	
Sentence Type		
Simple sentences		1
Complex sentences	1	
Grammatical Features		
Noun phrases	3	2
Participial modifiers	1	
Passive voice verb forms		1
Prepositional phrases	1	1
Organizational Features		
Description		1
Explanation	1	
Provide instruction or guidance	1	
Reference to text or visual	1	

a. "Token" refers to the total number of words
b. "Type" refers to the number of different words

PHASE 1: Pre-Pilot Tryouts
Initial Feedback on Task Formatting and Directions

The student informant answered the task correctly. Based on the feedback from phase I tryout, we decided to italicize and bold the phrase "vocabulary words from the passage" in the instructions to make it clear that ***only*** words from the passage are acceptable answers. (See highlighted area in draft task below)



Task Modified; Passage Intact for Phase 1 Pre-pilot

Modified Task

Passage

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

[paragraphs omitted]

Certain of future victories, General Howe decided to rest for the winter in New York City. Washington knew that the British would not try to advance again until the spring. So he planned a surprise attack on the close to 1,400 Hessian troops in Trenton, New Jersey. The password Washington gave his soldiers was "Victory or Death!" After nightfall on Christmas Day, December 25, 1776, Washington and his troops crossed the Delaware River into New Jersey. The next morning, they surprised the Hessians, who quickly surrendered. "This is a glorious day for our country," said Washington.

Fill in the blanks using ***vocabulary words from the passage.***

The Hessian troops were _____(attacked)_____ by George Washington.

Statistical Results from Whole Group Tryout

Item Difficulties (% correct), p = .40, (95% CI = .26-.54)
Item discrimination, D = .456

	n (Total = 45)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4th = 7 5th = 16 6th = 22	4th = 29% (2) 5th = 19% (3) 6th = 64% (14)	Unclear	S
Gender	Girls = 24 Boys = 21	Girls = 38% (9) Boys = 48% (10)	Boys higher	NS
Home Language [a]	English = 38 Non-English = 7	English = 42% (16) Non-English = 43% (3)	Similar	NS

a. In fact, all but 3 students in the Non-English group had Spanish as a home language (one child each for Arabic/English, Korean and Mandarin).

Breakdown of Whole Group Responses

- Correct Answer
  - surprised OR attacked: 42.2% (n=19)
- Incorrect but Meaningful (ICM): 26.7% (n=12)
  - defeated (A 6th grader, Korean as Home Language)
  - killed (A 5th grader, Native-English speaker)
- Incorrect and Irrelevant (ICI): 31.1% (n=14)
  - winners (A 5th grader, Native-English speaker)
  - ordered (A 6th grader, Native-English speaker)

Verbal Protocol Analysis from Tryout

Among the 10 students who answered the item, only half provided spontaneous comments. One 4th grader explained that "*because George Washington planned a surprise attack*", the answer should be "*attacked.*" Another 5th grader found the answer after having gone through the passage a few times and tried different answers (*Were threatened? Surrendered maybe? Surrendered? No, "surprised".*) Generally speaking, although it did take students some time to look for the answer in the passage, those who persisted were able to arrive at the correct answer.

Excerpt from Audit Trail

This item had medium level of difficulty (p = .40) and a very good discrimination index (D = .456). Based on results from whole group tryouts, we revised our scoring rubric and included "*surprised*" as an acceptable answer. Given its reasonable difficulty level and promising discrimination index, we decided to retain this item for Phase 2 pilot.

Passage & Item Intact for Phase 2 Pilot

Statistical Results from Phase 2 Pilot

Item Difficulties (% correct), p = .23, (95% CI = .15-.31)
Item discrimination, D = .43

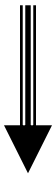
	n (Total = 111)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4^{th} = 15 5^{th} = 66 6^{th} = 30	4^{th} = 0% 5^{th} = 34.8% 6^{th} = 10%	Unclear	S
Gender	Girls = 59 Boys = 52	Girls = 27.1% Boys = 19.2%	Girls higher	NS
Home Language	English = 51 Spanish = 60 Other = 13	English = 33.3% (17) Spanish = 10.6% (5) Other = 30.8% (4)	English group higher than Spanish group	S

Breakdown of Pilot Group Responses

- Correct Answer
  - surprised OR attacked: 23.4% (n=26)
- Incorrect but Meaningful (ICM): 2.7% (n=3)
  - defeated (A 5^{th} grader, Spanish Home Language)
  - killed (A 6^{th} grader, Spanish Home Language)
- Incorrect and Irrelevant (ICI): 73.9% (n=82)
  - winners (A 6^{th} grader, Native-English speaker)
  - against (A 5^{th} grader, Native-English speaker)
  - surrendered (A 4^{th} grader, , Spanish Home Language)

Excerpt from Audit Trail

Compared to the results from Phase I pre-pilot studies, this item yielded item difficulty index in Phase 2 Pilot (p = .23) that suggested it was more difficult for students. However, the discrimination index remained very promising (D = .43), indicating that the item distinguished correctly between good and poor readers. Also, the item significantly distinguished between the performance of students with different language backgrounds. Students from English language backgrounds performed significantly better than those who had Spanish as a home language.



VERDICT: ***Passage & Task Retained as an AELP Prototype***

Summary of Effective Task Profile I

The social studies-based task targeted at students' comprehension of AEL vocabulary and grammar through a sentence completion task. Students first read a multi-paragraph expository text taken from a social studies textbook, and were requested to then identify a verb from the passage to fill in the blank of an incomplete sentence. This item addresses both ELD standard (Advanced Vocabulary and Concept Development) and California Content Standard for Social Studies 5.5 (4).

The linguistic analysis shows that the constructs of both the task stem/prompt and task response pertained predominantly to knowledge of specialized and general academic vocabulary. The task stem also tapped knowledge of simple and complex English grammar, and academic language functions *explanation, provide instruction or guidance*, and *reference to text*. The task response also required knowledge of AEL function *description*.

The student informant suggested that we italicize and bold the phrase "vocabulary words from the passage" in the instructions to make it clear that ***only*** words from the passage are acceptable in the responses. The passage remained intact and the instructions were modified accordingly for the pre-pilot. Statistical results from these tryouts revealed that the task was of medium difficulty level (p = .40), and had a very good discrimination index (D = .456).

We decided to retain this item for the Phase 2 pilot because of its reasonable difficulty level and promising discrimination index. Possibly due to sample background differences, students in the Phase 2 pilot did not perform as well as their counter-parts in the Phase 1 tryouts; item difficulty increased (p = .23). However, the task still discriminated effectively between good and poor readers, as well as distinguished between students with English and Spanish home language backgrounds.

Task Profile II - Math-based

Original Draft Task

Passage

Carlotta bought 9 packages of lemonade for $1.10 each and 2 packages of cups for $1.09 each. She sold 23 cups of lemonade every hour for 4 hours at $0.40 per cup. How much more money did Carlotta earn than she spent on supplies?

What is the word problem asking about?

a) How much Carlotta spent on supplies.
b) How many packages of lemonade she sold.
c) How much profit Carlotta made.*
d) How much lemonade costs.

*correct response

Task Specifications

Framework category: Demonstration of Comprehension (through paraphrase)

- **General description and text type:** *Students will identify the problem statement in a mathematics word problem and select the correct paraphrase from multiple-choice sentence options*
- **Task format:** *'Wh' question with multiple-choice sentence options*
- **Stimulus attributes:** *A mathematics word problem generally of 2-3 sentences in length with a problem question or imperative statement at the end. (empirical evidence) The target academic language function construct is "paraphrase", which requires the processing of the same idea expressed in different words.*
- **Response attributes:** *Circle the correct multiple-choice option from the four options provided.*
- **Standard addressed:** ELD Standard addressed: *Early Advanced Comprehension and Analysis*; California Content Standard addressed: *Math Number Sense 2.0 (2.1)*
- **Target Academic Language Constructs:** *Academic language functions "paraphrase" and "summarize"; and specialized academic vocabulary.*

Linguistic Analysis Profile

	Stem/Prompt	*Response*
Descriptive Analysis		
(Mean) no. of words per sentence(range)	7	5.5 (4-7)
Sum of Words	7	22
Total # of words (token)	7	22
Total # of words (type)	7	15
Lexical Features		
Academic vocabulary - specialized(token)	2	2
Academic vocabulary - specialized(type)	2	2
3-or-more-syllable words(token)		3
3-or-more-syllable words(type)		2
Derived words (token)		2
Derived words (type)		2
Sentence Type		
Simple sentences	1	NA
Other sentence types		4 clauses
Grammatical Features		
Prepositional phrases		1
Organizational Features		
Paraphrase	1	1
Question	1	
Summary	1	1

PHASE 1: Pre-Pilot Tryouts

Initial Feedback on Task Formatting and Directions

The student informant answered the task correctly. Based on feedback from the tryout, a note of caution was added to the end of the reading passage to prevent students from working on the math problem.

Passage Modified; Item Intact for Phase I Pre-pilot

Modified Task

Passage

Carlotta bought 9 packages of lemonade for $1.10 each and 2 packages of cups for $1.09 each. She sold 23 cups of lemonade every hour for 4 hours at $0.40 per cup. How much more money did Carlotta earn than she spent on supplies? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**[a]

What is the word problem asking about?

a) How much Carlotta spent on supplies.
b) How many packages of lemonade she sold.
c) How much profit Carlotta made.*
d) How much lemonade costs.

[a] The highlights of modifications only appear in the example passage and draft task presented above for demonstration purpose. They were removed from the version students receive in pre-pilot and pilot testing.

Statistical Results from Whole Group Tryout

Item Difficulties (% correct) p = .68 (95% CI = .56-.80)
Item discrimination (D) = .332

	n (Total = 75)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4th = 21 5th =27 6th = 27	4th = 54% 5th = 74% 6th = 78%	Positive	NS
Gender	Girls = 36 Boys = 39	Girls = 78% Boys = 62%	Girls higher	NS
Home Language	English = 61 Non-English = 14	English = 74% (45) Non-English = 50% (7)	English group higher	NS

Breakdown of Tryout Group Responses

- 69.3% (n=52) chose the correct answer C: How much profit Carlotta made.
- 13.3% (n=10) chose distractor answer A: How much Carlotta spent on supplies.
- 5.3% (n=4) chose distractor answer B: How many packages of lemonade she sold.
- 1.3% (n=1) chose distractor answer D: How much lemonade costs
- 10.6 % (n=8) given opportunity to work on the item but provided no response

Verbal Protocol Analysis from Tryout

Among the students who had provided comments on this item (n=13), some of them had specifically identified either the name "Carlotta" in the word problem or the word "profit" in the correct answer as difficult words. Although theoretically unfamiliarity with the proper noun "Carlotta" would not impede reading comprehension, some students would pause at the word and make extra efforts to pronounce the word correctly. About half of the students who had answered the question got the correct answer (n=6 out of 13). The strategies those students reported included going back to the passage (*I really didn't exactly understand so I went back up to the passage and read the question that they asked, so then I noticed that profit is basically the same thing earned of...how much she's earned…so it means how much profit…so profit means the same thing.* Comment from a 6th grader) and eliminating answers (*I just eliminated…[unintelligible]. How many packages of lemonade she sold, it doesn't say that there, how much Carlotta spent on supplies, and the problem doesn't really say that. It said how much more money did Carlotta earn than she spent on supplies. So that's different.* Comment from a 4th grader). On the other hand, based on the comments from students who had answered the question incorrectly, it appeared that some of them still treated the question as more of a "math" word problem than a reading comprehension item of their academic English proficiency. For example, a 4th grader chose the wrong answer "a) How much Carlotta spent on supplies" and rationalized his answer as following: *Because it tells you all the prices for sure. It's not how many packages of lemonade she sold…How much profit Carlotta made…It doesn't even tell you that.*

Excerpts from Audit Trail

Although this item had a relatively low difficulty index (p=.68), it reasonably discriminated among good and poor readers (D=.332). In addition, it also distinguished across grade and home language background. Review of the student responses revealed that the distractors were also plausible and effective.

Passage & Item Intact for Phase 2 Pilot

Statistical Results from Phase 2 Pilot

Item Difficulties (% correct) p = .42 (95% CI = .34-.51)
Item discrimination (D) = .37

	n (Total = 125)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4^{th} = 18 5^{th} = 76 6^{th} = 31	4^{th} = 33.37% 5^{th} = 47.4% 6^{th} = 35.5%	Unclear	NS
Gender	Girls = 67 Boys = 58	Girls = 47.8% Boys = 36.2%	Girls higher	NS
Home Language	English = 52 Spanish = 56 Other = 17	English = 55.8% (29) Spanish = 32.1% (18) Other = 35.3% (6)	English group higher than both Spanish and Other	S (English higher than Spanish)

Breakdown of Pilot Group Responses

- 42.4% (n=53) chose the correct answer C: How much profit Carlotta made.
- 38.4% (n=48) chose distractor answer A: How much Carlotta spent on supplies.
- 12% (n=15) chose distractor answer B: How many packages of lemonade she sold.
- 5.6% (n=7) chose distractor answer D: How much lemonade costs
- 1.6 % (n=2) given opportunity to work on the item but provided no response

Excerpts from Audit Trail

The item difficulty level changed from .68 to .42 in the Phase 2 pilot findings, suggesting it was harder for the pilot students. The item discrimination index remained adequate. The task also significantly distinguished between students with English and Spanish home language backgrounds.

VERDICT: ***Passage & Task Retained as an AELP Prototype***

Summary of Effective Task Profile II

This math-based task was created to measure students' knowledge of English grammar and discourse through a multiple-choice task. Students encountered a word problem of 2-3 sentences in length. They are then required to select the correct answer from four options that answered the main-idea question. This required students to understand that "how much more" in the passage is, in this context, equivalent to the word "profit" in the correct response. This item addressed both ELD standard (Early Advanced Comprehension and Analysis) and California Content Standard for Math (Number Sense 2.1).

The linguistic analysis reveals that the task stem/prompt and the response involved knowledge of the *paraphrase* academic language function or organizing feature, as well as the *summary* function. Knowledge of specialized academic vocabulary is also required for both the stem/prompt and response. The student informant suggested that a note be added to the end of the passage to refrain students from working on the math problem. The item remained intact and the passage was modified accordingly for whole group tryout.

Statistical results from pre-pilot revealed that the item had a low difficulty index (p = .68). However, it reasonably discriminated among good and poor readers (D = .332) and distinguished across grade and home language background. The distractors were also shown to be plausible and effective. We thus retained the task for the Phase 2 pilot. Similar to the findings of the first effective AELP task above, students in the Phase 2 pilot (p = .42) performed less well than those in the Phase 1 tryouts. However, the task maintained an adequate discrimination index (D= .37). It also significantly distinguished between students with English and Spanish home language backgrounds.

Task Profile III - Math-based

Original Draft Task

Passage

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster?

Read the problem. Then complete the table.

Person	Year	Distance	Days	From	To
Stilt walker #1	1980		158		Bowen, Kentucky
Stilt walker #2	1891	1830 miles		Paris, France	***Target Item (Moscow, Russia)***

Task Specifications

Framework category: Academic Language Function "Comparison" and Vocabulary

- **General description and text type:** *Students will read the word problem and retrieve appropriate information from the text to fill in the gaps in a table.*
- **Task format:** *Fill in the gaps in a table using information from the text.*
- **Stimulus attributes:** *A mathematics word problem generally of 2-3 sentences in length with a problem question or imperative statement at the end.*
- **Response attributes:** *Fill in the blanks in a table by retrieving requested information from the text.*
- **Standard addressed:** ELD Standard addressed: *Advanced Reading Comprehension*; California Content Standard addressed: *Math Number Sense 2.0 (2.3).*
- **Target Academic Language Constructs:** *Focal academic language function: "comparison," also "scenario," "labeling", "provide instruction/guidance" and "summarize;" and general academic vocabulary.*

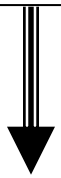
Linguistic Analysis Profile

	Stem/Prompt	*Response*
Descriptive Analysis		
(Mean) no. of words per sentence(range)	3.5 (3-4)	NA
Sum of Words	7	21
Total # of words (token)[a]	28	21
Total # of words (type)	25	17
Lexical Features		
Academic vocabulary - general (token)	4	
Academic vocabulary - general (type)	4	
Low-frequency words (token)	7	1
Low-frequency words (type)	5	1
3-or-more-syllable words(token)		1
3-or-more-syllable words(type)		1
Derived words (token)		3
Derived words (type)		2
Avg. % of nominalizations per selection		1
Sentence Type		
Simple sentences	2	
Grammatical Features		
Noun phrases	3	16
Organizational Features		
Scenario		1
Comparison		1
Labeling		1
Provide instruction or guidance	1	
Reference to text or visual	1	1

[a] This frequency also includes column headings and content of the table in the task.

PHASE 1: Pre-Pilot Tryouts
Initial Feedback on Task Formatting and Directions

The student informant answered the task correctly. Based on the feedback from the student and notes from our internal review meeting, we had made a few changes to the format of the passage and the item. We added a cautionary note at the end of the math word problem to prevent students from treating the item as a mathematical question. Instead of treating the whole table as one item, we separated each blank in the table into individual items. Additionally, we modified the instruction from "Read the problem. Then complete the table" to "Fill in the blanks for questions 1 through 4 in the table below," and accordingly added lines and corresponding numbers for each blank in the table for students to fill in their responses.



Passage & Item Modified for Phase 1 Pre-pilot

Modified Task

Passage

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Fill in the blanks for questions 1 through 4 in the table below.

Person	Year	Distance	Days	From	To
Stilt walker #1	1980	*(1.)* _____	158	*(2.)* _____	Bowen, Kentucky
Stilt walker #2	1891	1,830 miles	*(3.)* _____	Paris, France	*(4.)* _(Moscow, Russia)___

Statistical Results from Whole Group Tryout

Item Difficulties (% correct) p = .90 (95% CI = .82-.98)
Item discrimination (D) = .263

Item discrimination (D) = .263	n (Total = 69)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4th = 17 5th = 27 6th = 25	4th = 88% 5th = 93% 6th = 100%	Positive	NS
Gender	Girls = 37 Boys = 32	Girls = 95% Boys = 94%	Similar	NS
Home Language	English = 56 Non-English = 13	English = 93% (52) Non-English = 100% (13)	Spanish group higher	NS

Breakdown of Tryout Group Responses

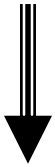
- Correct Answer = Moscow, Russia: 94.2% (n=65)
- Incorrect but Meaningful (ICM): 2.9% (n=2)
  - Russia (A 4th grader, Native-English speaker)
- Incorrect = 2.9% (n=2)
  - Paris, France, Moscow, Russia (A 5th grader, Native-English speaker)
  - Paris, France. (A 5th grader student, Native-English speaker)

Verbal Protocol Analysis from Tryout

All of the fourteen students who worked on the question answered it correctly, and most of them also answered the question promptly. However, only half of them had provided comments on the item either spontaneously or with prompts. Specifically, five of the comments pertained to the format of the item and familiarity with the item type (i.e. tabular form). One 6th grader initially had difficulty understanding the task, but was able to answer the questions after the researcher's explanation. Another 5th grader commented on the limited space for writing down her answer (*I don't think there'd be enough room to write the whole entire name...[referring to place name]*). On the other hand, two students reported using the strategy of going back to the passage to look for the answer (*I first read all the stuff that they said and then I went back to the passage to look what the answers were*. Comment from a 6th grader).

Excerpts from Audit Trail

Although this item turned out to be quite easy for the students (p = .90), it nonetheless had a fair item discrimination index (D = .263). It also distinguished across grades and yielded similar performances for girls and boys. We thus decided to retain this item for the Phase 2 pilot with some modifications to the item. The modifications included adding a title to the table to make the task more transparent and italicizing and reformatting each question number in the table.



Item Modified; Passage Intact for Phase 2 Pilot

Modified Task

Passage

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Fill in the blanks for questions *(1)* through *(4)* in the table below.

Table: Stilt walker Travel by Distance and Days

Person	Year	Distance	Days	From	To
Stilt walker #1	1980	*(1)* _____	158	*(2)* _____	Bowen, Kentucky
Stilt walker #2	1891	1,830 miles	*(3)* _____	Paris, France	*(4)* _____

Statistical Results from Phase 2 Pilot

Item Difficulties (% correct) p = .75 (95% CI = .67-.83)
Item discrimination (D) = .43

	n (Total =111)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4^{th} = 15 5^{th} = 66 6^{th} = 30	4^{th} = 53% 5^{th} = 80% 6^{th} = 73%	Unclear	NS
Gender	Girls = 59 Boys = 52	Girls = 81% Boys = 67%	Girls higher	NS
Home Language	English = 51 Spanish = 47 Other = 13	English = 88% (45) Spanish = 64% (30) Other = 62% (8)	English group higher than Spanish and other group	S

Breakdown of Pilot Group Responses

- Correct Answer = Moscow, Russia: 74.8% (n=83)
- Incorrect but Meaningful (ICM): 1% (n=1)
  - Russia (A 6th grader, Spanish Home Language)
- Incorrect = 24.3% (n=27)
  - Paris, France (A 5th grader, Spanish Home Language)
  - Los Angeles (A 6th grader, Spanish Home Language)
  - 1830 (A 5th grader, Spanish Home Language)

Excerpts from Audit Trail

The item difficulty level changed from .90 in the pre-pilot phase to .75 in the pilot phase, suggesting fewer students answered correctly. The finding was likely due to the differences in sample demographics: there were more ELL students in the pilot. However, the item had a higher discrimination index (D = .43) and distinguished between students from different home language backgrounds.

VERDICT: ***Passage & Task Retained as an AELP Prototype***

Summary of Effective Task Profile III

This math-based item was intended to measure students' knowledge of discourse and vocabulary through a graphic organizer task in a table format. Students first read a word problem of 2-3 sentences in length. They were then required to retrieve information from the word problem to fill in the blanks in a table. This item addressed both ELD standard (Early Advanced Reading Comprehension) and California Content Standard for Math (Number Sense 2.3).

The linguistic analysis reveals that the task stem/prompt construct tapped into knowledge of academic language functions *provide instruction or guidance* and *reference to text*. The task response pertained not only to these functions, but a focal function *comparison*, and additional functions *labeling* and *scenario*, as well as general academic vocabulary.

A few changes in format were made to the task for the pre-pilot based on student informant feedback and internal review meeting notes. Statistical results from Phase 1 tryouts revealed that the task was easy (p = .90). However, it moderately discriminated among good and poor readers (D = .263) and distinguished across grades. The task was further modified as described above and was retained for the Phase 2 pilot. Comparing findings from the Phase 1 tryouts and the Phase 2 pilot, the item difficulty level changed from .90 to .75, suggesting it was more difficult for the pilot sample possibly due to differences in sample demographics. However, the task adequately discriminated between good and poor readers (D = .43), as well as distinguished between students from different home language backgrounds.

Profiles for Rejected Tasks

Rejected Task Profile I – Math-based Task

Original Draft Task

Passage

On a weekend camping trip, Ken, Eric, and their dad went for a walk on the Appalachian Trail. The first hour, they walked 3/8 mile. The second hour, they walked 4/5 mile. About how many miles did the boys and their dad walk?

What is the math problem in the sample passage asking about?

a) How long the family walked.
b) How far the family walked.*
c) How many people walked.
d) How many hours the boys walked.

Task Specifications

Framework category: Overall Comprehension of a Text

- **General description and text type:** *Students will identify the problem statement in a mathematics word problem and select the correct paraphrase from multiple-choice sentence options*
- **Task format:** *'Wh' question with multiple-choice sentence options*
- **Stimulus attributes:** *A mathematics word problem generally of 2-3 sentences in length with a problem question or imperative statement at the end (empirical evidence).*
- **Response attributes:** *Circle the correct multiple-choice option from the four options provided.*
- **Standard addressed:** ELD Standard addressed: *Early Advanced Comprehension and Analysis*; California Content Standard addressed: *Math, Number Sense 2.0 (2.3)*
- **Target Academic Language Constructs:** *The academic language functions are "paraphrase," and "summarize" which require the processing of the same idea(s) expressed in different words. Also included in the construct is vocabulary knowledge, such as "how far" and "how long."*

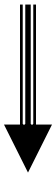
Linguistic Analysis Profile

	Stem/Prompt	*Response*
Descriptive Analysis		
(Mean) no. of words per sentence(range)	11	5(4-6)
Sum of Words	11	20
Total # of words (token)	11	20
Total # of words (type)	10	10
Lexical Features		
Academic vocabulary - general (token)	1	
Academic vocabulary - general (type)	1	
3-or-more-syllable words (token)		2
3-or-more-syllable words (type)		1
Sentence Type		
Simple sentences	1	NA
Other sentence types		4 clauses
Grammatical Features		
Noun phrases	2	7
Organizational Features		
Paraphrase	1	1
Question	1	
Summary	1	1

PHASE 1: Pre-Pilot Tryouts

Initial Feedback on Task Formatting and Directions

While the student informant answered the task correctly, he suggested that we add a note of caution in parenthesis to the end of the reading passage to prevent students from working on the math problem. The task remained intact for the next tryout phase.



Passage Modified; Task Intact for Phase 1 Pre-Pilot

Modified Draft Task

Passage

On a weekend camping trip, Ken, Eric, and their dad went for a walk on the Appalachian Trail. The first hour, they walked 3/8 mile. The second hour, they walked 4/5 mile. About how many miles did the boys and their dad walk? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

What is the math problem in the sample passage asking about?

a) How long the family walked.
b) How far the family walked.*
c) How many people walked.
d) How many hours the boys walked.

Statistical Results from Whole Group Tryout

Item difficulties (% correct) p = .68 (95% CI =.52-.84)
Item discrimination, D = .086

	n (Total =37)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4th =12 5th =13 6th =12	4th =58% (7) 5th =69% (9) 6th =75% (9)	Positive	NS
Gender	Girls =14 Boys =23	Girls =71% (10) Boys =65% (15)	Girls higher	NS
Home Language	English =30 Non-English =7	English =73% (22) Non-English =43% (3)	English group higher	NS

Breakdown of Tryout Group Responses

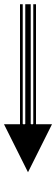
68% (n=25) chose the correct answer B: How far the family walked.
27 % (n=10) chose distractor answer A: How long the family walked.
5% (n=2) chose distractor answer D: How many hours the boys walked

Verbal Protocol Analysis from Tryout

Verbal protocol analysis revealed that students found the word problem to be relatively easy. The two words they had problems with were the proper noun *Appalachian* (n=4 out of 9) and the fraction *4/5* (n=2 out of 9). With regard to the item, some students arrived at the correct answer immediately and appeared to have fully grasped the task. For example, when asked how he could have figured out the answer immediately, one 5th grader explained that the word *miles* referred to distance, so the answer should be (b) *how far the family walked.* Other students, however, had had to take longer to decide among the four options. A few students were confused about *how long* and *how far*. For example, one 4th grader specifically commented "*How far or how long is almost the same thing, so it was hard to decide*."

Excerpts from Audit Trail

This task turned out to have very poor discrimination (D = .08). That is, this item did not discriminate among good and poor readers as expected. Although the item difficulty index (p = .68) is reasonable, the 95% confidence interval range is wide (.52-.84). The item also seemed to slightly privilege girls.



VERDICT: ***Passage & Task Rejected***

Summary of Rejected Task Profile I

This math text-based task was created to measure students' overall comprehension of a text through a "wh-" question that requires students to identify the main idea of the text. Specifically, students first read a word problem of 2-3 sentences in length. They were then required to choose from among four options the one that correctly answers the main idea question. This task addresses both ELD standard Early Advanced Comprehension and Analysis and California Content Standard for Math Number Sense 2.0.

Results from the linguistic analysis reveal that the item stem/prompt and the response involved knowledge of the *paraphrase* academic language function or organizing feature, as well as the *summary* function. Knowledge of various vocabulary and syntactic features are also required for both the stem/prompt and response.

The student informant recommended that a note of caution be added to the end of the passage to prevent future attempts of solving the math problem. The passage was thus modified accordingly while the task remained intact for the pre-pilot. Statistical results from pre-pilot indicated that the task was of medium difficulty level (p = .68), but with a problematic wide confidence interval (95% CI=.52-.84). The task also had low discrimination (D = .086), and seemed to slightly privilege girls. We thus decided to discard this task after the pre-pilot phase.

Rejected Task Profile II – Science-based Task

Original Draft Task

Passage

In addition to temperature and air pressure, humidity, or the amount of water in the air, is an important factor in describing weather. But how does water get into the air?

[paragraphs omitted]

Whether a cloud forms near the ground or high in the atmosphere, it forms in the same way. Water vapor condenses onto dust and other tiny particles in the air when it rises and cools. Another way in which air cools enough for water vapor to condense is by moving from a warm place to a colder place. For example, moist air that moves from over a warm body of water to over cooler land forms clouds or fog…
[passage continues]

Complete the sentence below using verbs from the passage.

Water vapor condenses when it _____(moves)_____ from a warm place to a cooler place.

Task Specifications

Framework category: Comprehension of Academic Vocabulary

- **General description and text type:** *Students will complete a sentence using verbs that are used in context of a multi-paragraph expository text.*
- **Task format:** *Sentence completion using verbs from the passage.*
- **Stimulus attributes:** *A multi-paragraph expository text generally consisting of 3-5 paragraphs.*
- **Response attributes:** *The stimulus is followed by incomplete sentences. Students complete each sentence by filling in the blank with the correct verb from the passage.*
- **Standard addressed:** ELD Standard addressed: *Advanced Vocabulary and Concept Development*; California Content Standard addressed: *Science: Earth Science 3.0 (C)*
- **Target Academic Language Constructs:** *Academic language functions "explanation" and "reference to text/visual"; complex grammar, and focal feature specialized academic vocabulary.*

Linguistic Analysis Profile

	Stem/Prompt	*Response*
Descriptive Analysis		
(Mean) no. of words per sentence(range)	9	14
Sum of Words	9	14
Total # of words (token)	9	14
Total # of words (type)	8	12
Lexical Features		
Academic vocabulary - specialized(token)	2	2
Academic vocabulary - specialized(type)	2	2
Low-frequency words (token)		1
Low-frequency words (type)		1
3-or-more-syllable words (token)	1	1
3-or-more-syllable words (type)	1	1
Derived words (token)	1	1
Derived words (type)	1	1
No. of unique clause connectors	1	
Avg. % of nominalizations per selection	1	
Sentence Type		
Simple sentences		
Complex sentences	1	1
Other sentence types		
Grammatical Features		
Noun phrases		1
Participial modifiers	1	
Passive voice verb forms		
Prepositional phrases	1	2
Dependent clauses	1	
Organizational Features		
Explanation	1	1
Reference to text or visual	1	

PHASE 1: Pre-Pilot Tryouts
Initial Feedback on Task Formatting and Directions

The correct answer for the task is *moves*, and the response from the student informant was *switches*. Although the response was semantically appropriate and grammatically acceptable, the word was not found in the passage. Given that we had made it explicit in the instruction that the answer should be from the passage, the informant failed to meet this criterion, which might indicate lack of comprehension of instructions. We thus decided not to credit the response. Feedback from the student revealed that there could be alternative and acceptable answers to the item. To make it clear to the students that only when their response is a verb "from the passage" will it be considered as an acceptable response, we decided to italicize and bold the phrase in the instructions (modified phrase was highlighted in the following example).

Task Modified; Passage Intact for Phase 1 Pre-pilot

Modified Draft Task

Passage

In addition to temperature and air pressure, humidity, or the amount of water in the air, is an important factor in describing weather. But how does water get into the air?

[paragraphs omitted]

Whether a cloud forms near the ground or high in the atmosphere, it forms in the same way. Water vapor condenses onto dust and other tiny particles in the air when it rises and cools. Another way in which air cools enough for water vapor to condense is by moving from a warm place to a colder place. For example, moist air that moves from over a warm body of water to over cooler land forms clouds or fog... **[passage continues]**

Complete the sentence below using ***verbs from the passage***.

Water vapor condenses when it _____(moves)_____ from a warm place to a cooler place.

Statistical Results from Whole Group Tryout

Item difficulties (% correct) p = .49, (95% CI =.37-.61)
Item discrimination, D = .099

	n (Total =62)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4^{th} =17 5^{th} =20 6^{th} =25	4^{th} =59% (10) 5^{th} =50% (10) 6^{th} =52% (13)	Negative	NS
Gender	Girls =31 Boys =31	Girls =58% (18) Boys =48% (15)	Girls higher	NS
Home Language	English =48 Non-English =14	English =54% (26) Non-English =50%	English group higher	NS

Breakdown of Tryout Group Responses

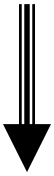
- Correct Answer:
  - *moves* : 53.2% (n=33)
- Incorrect but Meaningful (ICM): 9.7% (n=6)
  - *goes* (A 4^{th} grader, Spanish as Home Language)
- Incorrect and Irrelevant (ICI): 37% (n=23)
  - *rises* (A 5^{th} grader, Native-English Speaker)
  - *condensates* (A 6^{th} grader, Native-English Speaker)

Verbal Protocol Analysis from Tryout

Student responses from verbal protocol data were divided for this task. Some students selected the correct answer immediately without making any comment (n=7 out of 17) while others found this item to be challenging and took some time searching for the answer in the passage (n=8 out of 17). It appeared that unfamiliar words were the sources of challenge for students who had difficulty with this task (*The student read the question, and said that he didn't know what "condenses" mean.* Researcher comment#1; *Cause I kept thinking about these different words and then I remember, then I saw the word "move" somewhere...and I just, oh, that makes sense.* Comment from a 6^{th} grader). Additionally, one student stated that his prior knowledge of this topic helped with comprehension and answering the item (*Last year we studied the water cycle a lot so I know a lot of the words*. Comment from a 6^{th} grader).

Excerpts from Audit Trail

We decided to reject this item for several reasons. The item had poor discrimination (D = .099). It did not distinguish good and poor readers as expected. In addition, this item did not distinguish between students' home language, nor did it distinguish across grades.



VERDICT: Passage & Task Rejected

Summary of Rejected Task Profile II

The science-based task specifically targeted at students' comprehension of general AE vocabulary through a sentence completion task that requires students to fill in a blank using words from a passage previously read. Specifically, students first read a multi-paragraph authentic expository text taken from a state-approved science textbook. They were then required to identify a verb from the passage to fill in the blank of an incomplete sentence. This item addresses both ELD standard (Advanced Vocabulary and Concept Development) and California Content Standard for Science (Earth Science 3.0).

Linguistic analysis of this item revealed that the construct of the item stem/prompt involved knowledge of English grammar and academic language functions, i.e., *explanation* and *reference to the text*, whereas the response construct relates to knowledge of unfamiliar and specialized academic vocabulary, as well as complex syntax.

The student informant provided an answer (i.e., "switches" rather than the correct answer "moves") that was both semantically and grammatically appropriate, yet not present in the passage. We decided not to give credit for the answer given that we had made an explicit statement in the instruction about using words from the passage. However, in order to prevent similar mistakes, we highlighted this requirement by bolding and italicizing the specific phrase (i.e., from the passage). The item was thus modified accordingly while the passage remained intact for whole group tryout.

Statistical results from whole group tryout revealed that the item was of middle to high difficulty level ($p = .49$), but again with a problematic wide range of confidence interval (95% CI = .37-.61). Although statistically non-significant, a reverse trend among grades and a potential bias for better performance among girls were observed. In addition, the item had poor discrimination index ($D = .099$), and was identified as a problematic item according to reliability analysis. All those identified problems led to our verdict of rejecting this item.

Rejected Task Profile III – Social Studies-based Task

Original Draft Task

Passage

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

[paragraphs omitted]

In 1759 Washington retired from military life to manage his lands. By then he had become the most famous American in the military. That same year he married a wealthy widow named Martha Custis. George and Martha Washington moved to Mount Vernon, the plantation he owned on the Potomac River in Virginia. Martha Washington also supported the patriots. During the American Revolution, she helped her husband with his paperwork. She also sewed socks and cooked soup for the soldiers.

Martha Washington often joined George Washington in the field, where things were going badly for the Continental Army at the end of 1776. Washington was discouraged. He wrote, "Such is my situation that if I were to wish the bitterest curse to an enemy on this side of the grave, I should put him in my [place] with my feelings."

[passage continues]

Fill in the blanks using vocabulary words from the passage.

George Washington's wife was _____(often)_____ at home and on the battlefield.

Task Specifications

Framework category: Specialized and Academic Vocabulary

- **General description and text type:** *Students will complete a sentence using vocabulary words that are defined in a multi-paragraph expository text.*
- **Task format:** *Sentence completion using words from the passage.*
- **Stimulus attributes:** *A multi-paragraph expository text generally consisting of 3-5 paragraphs.*
- **Response attributes:** *The stimulus is followed by incomplete sentences. Students complete each sentence by filling in the blank with the correct word from the passage.*
- **Standard addressed:** ELD Standard addressed: *Advanced Vocabulary and Concept Development*; California Content Standard addressed: *Social Studies: 5.5 (4)*
- **Target Academic Language Constructs:** *Academic language functions "description" and "reference to text/visual"; complex grammar, and focal features are specialized and general academic vocabulary.*

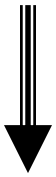
Linguistic Analysis Profile

	Stem/prompt	*Response*
Descriptive Analysis		
(Mean) no. of words per sentence(range)	10	11
Sum of Words	10	11
Total # of words (token)	10	11
Total # of words (type)	9	11
Lexical Features		
Academic vocabulary - general (token)	2	
Academic vocabulary - general (type)	2	
Academic vocabulary - specialized(token)	1	1
Academic vocabulary - specialized(type)	1	1
Low-frequency words (token)		1
Low-frequency words (type)		1
3-or-more-syllable words (token)	1	2
3-or-more-syllable words (type)	1	2
Derived words (token)	1	
Derived words (type)	1	
Sentence Type		
Simple sentences		1
Complex sentences	1	
Grammatical Features		
Noun phrases		1
Prepositional phrases	1	2
Dependent clauses	1	
Organizational Features		
Description	1	1
Reference to text or visual	1	1

PHASE 1: Pre-Pilot Tryouts

Initial Feedback on Task Formatting and Directions

The answer the student informant provided was *there*. Although the word was found in the passage, it was semantically a misfit with the question (It is not possible for someone to be "there" in two places – at home and on the battlefield at the same time). Based on the feedback from phase I tryout, we decided to italicize and bold the phrase "vocabulary words from the passage" in the instruction to make it clear that only words from the passage are acceptable answers.



Item Modified; Passage Intact for Phase 1 Pre-pilot

Modified Task

Passage

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

[paragraphs omitted]

In 1759 Washington retired from military life to manage his lands. By then he had become the most famous American in the military. That same year he married a wealthy widow named Martha Custis. George and Martha Washington moved to Mount Vernon, the plantation he owned on the Potomac River in Virginia. Martha Washington also supported the patriots. During the American Revolution, she helped her husband with his paperwork. She also sewed socks and cooked soup for the soldiers.

Martha Washington often joined George Washington in the field, where things were going badly for the Continental Army at the end of 1776. Washington was discouraged. He wrote, "Such is my situation that if I were to wish the bitterest curse to an enemy on this side of the grave, I should put him in my [place] with my feelings." **[passage continues]**

Fill in the blanks using ***vocabulary words from the passage.***

George Washington's wife was _____often_____ at home and on the battlefield.

Statistical Results from Whole Group Tryout

Item difficulties (% correct), p = .18, (95% CI =.08-.28)
Item discrimination, D = .094

	n (Total =43)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4th =8 5th =14 6th =21	4th =38% (3) 5th =21% (3) 6th =14% (3)	Negative	NS
Gender	Girls =22 Boys =21	Girls =32% (7) Boys =10% (2)	Girls higher	NS
Home Language	English =35 Non-English =8	English =17% (6) Non-English =38% (3)	Other group higher	NS

Breakdown of Tryout Group Responses

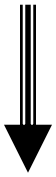
- Correct Answer
  - often : 20.9% (n=9)
- Incorrect but Meaningful (ICM): 44.2% (n=19)
  - there (A 6th grader, Native English speaker)
  - supporting (A 5th grader, Native-English speaker)
- Incorrect and Irrelevant (ICI): 34.9% (n=15)
  - honored (A 5th grader, Native-English speaker)
  - sometimes (A 5th grader, Native-English speaker)

Verbal Protocol Analysis from Tryout

A few students chose to skip this item after having spent some time trying to find the answer in the passage (n=5 out of 11). One 4th grade student had explicitly identified this item to be particularly challenging when asked to identify difficult items among all the tasks in the booklet. It appeared that the minimal clue to the answer posed great obstacles to solving the problem (*She said she just didn't have any clue to this question*. Researcher comment; *I can't really find in the passage, because usually the passage always gives a clue.* Comment from a 6th grader). The wide range of possible answers to the item also threatened its validity (*I think this is pretty like too open. It doesn't really say that much about her on the battlefield. She could be doing anything, she could be not on the battlefield or helping on the battlefield or something.* Comment from a 6th grader).

Excerpt from Audit Trail

This item turned out to be too difficult (item difficulty index p = .18 with 95% CI =.08-.28) and showed negative patterns across grade ($4^{th} > 5^{th} > 6^{th}$) and home languages (Spanish home language > English home language). Also discrimination either (D = .094) was poor. There were additional issues with regard to the wide range of possible answers, which were attested to by the verbal protocol data. Overall, this is a problematic item, thus we rejected it.



VERDICT: ***Passage Retained (for use with other tasks), Task Rejected***

Summary of Rejected Task Profile III

The social studies-based item was designed to measure comprehension of AL vocabulary and grammar through a sentence completion task that requires students to fill in a blank using words from the associated passage. Students were first required to read a multi-paragraph authentic expository text taken from a state-approved social studies textbook. They then identified a word from the passage to fill in the blank of an incomplete sentence. This item addresses both the ELD standard (Advanced Vocabulary and Concept Development) and California Content Standard for Social Studies 5.5 (4).

Linguistic analysis of this item revealed that both the constructs of the item stem/prompt and item response involved knowledge of English grammar and unfamiliar, specialized and general academic vocabulary. The item also tapped knowledge of the academic language functions *description* and *reference to text*.

The student informant provided the answer "George Washington's wife was *there* at home and on the battlefield." While this is grammatically acceptable, his answer is incorrect because it did not use vocabulary from the passage. His response was still meaningful however, because it was relevant to the topic. The passage remained intact but instruction was modified to highlight the requirement of "using vocabulary words from the passage."

Statistical results from whole group tryout revealed that the item was of a high difficulty level (p = .18) and did not have good item discrimination (D = .094). Although statistically non-significant, a reverse trend among grades and home language was observed. Lower graders and the Spanish home language group performed better than higher graders and the English home language group. Additionally, the wide range of possible answers cast doubt on the validity of the item, a problem which was attested to by the verbal protocol data (*She could be doing anything, she could be not on the battlefield or helping on the battlefield or something.* Comment from a 6th grader). We thus decided to reject this item in favor of other more promising ones.

CHAPTER 5:
CONCLUSIONS AND RECOMMENDATIONS

Summary of Results

The work reported here yielded prototype tasks designed to assess academic English at the 5th grade level through the reading modality. The tasks are designed to test a range of language functions and features, not reading comprehension per se and not content knowledge, although the language functions and features being tested are drawn from content material to make the tasks most relevant to language use in the academic context.

The rationale for trying out the tasks with native English-speaking students stemmed from the concern that the tasks not be beyond the ability level of native English-speaking students, nor that we develop too many tasks that are far below the ability levels of native speakers. Information on native English-speaking student performance provides critical information about the targeted language abilities of mainstream students at the 5th grade. This is the level of language demand that ELL students will encounter if they are redesignated fluent English proficient. Determining if the AELP tasks can capture the level of proficiency necessary for participation in mainstream mathematics, science and social science classrooms, etc., is therefore fundamental to the AELP test development effort. The obvious way to determine if this level of prolificacy is met by at least several of the tasks is by piloting them with native English-speaking students.

The pre-pilot phase consisted of group administrations with 77 predominantly English-only 4th-6th graders and verbal protocol data from an additional 18 students distributed across these grades and representative of different reading ability levels and Spanish- and English-dominant language backgrounds. Results suggested that of the original 101 draft tasks, 40 were sufficiently effective for retention in terms of a combination of quantitative and qualitative factors, including item difficulty, item discrimination on reading ability, distinguishing between Spanish-dominant versus English-dominant home language backgrounds, being free of gender biases, and being free from anomalies in directions and formatting ambiguities or at least contained formatting and wording issues that could be refined. Indeed, 35 of these AELP tasks required modifications of some sort (e.g., rewording of directions, etc.) before they were considered acceptable by internal review for the pilot phase.

The intent of verbal protocols at the pre-pilot phase was to gain more in-depth information about the draft tasks for use in making any necessary refinements at the end of Phase 1. The data driven analyses provided feedback on formatting issues, clarity of directions, word-level issues, item-level issues, answer strategies, and use of background

information. Specifically, results suggested that the students found *passage, prior knowledge*, and *familiarity with vocabulary* to be particularly important for comprehension of the academic English texts. This result is not surprising in that a longer passage, for example, could be more difficult for students simply because it contains more language to process. Students also identified prior knowledge as a means of comprehending the informational load of reading passages. Students vary in what they can interpret based on their prior knowledge or experience. Similarly, comprehending a passage involves familiarity with vocabulary, which is again a function of prior knowledge. Students either knew the word or did not, depending on their exposure and familiarity with the language used in the text.

The observed behaviors during the verbal protocol revealed important results on the process; namely, the comprehension and cognitive behavior of the students as they attempted to understand the reading text. Thus, the verbal protocol in this study served as a critical tool utilized in conjunction with the quantitative information to improve the AELP tasks before they were submitted to pilot-testing at Phase 2.

The pilot phase was conducted using group administrations of the 40 retained and largely refined tasks with 128 4^{th}-6^{th} grade English-only and ELL students. After linguistic and psychometric analysis during iterative testing phases, 17 of 101 original draft tasks stand out as sufficiently promising (linguistically and psychometrically) to serve as possible AELP prototypes. One of the 17 is a cluster of four original items that were non-independent forced-choice items. A further 14 tasks can be considered moderately effective. One of these tasks is also a cluster of four original items. This second group of tasks will need greater refinement in any future efforts with these tasks to increase the degree of discrimination they make between good and poor readers. In this report, we feature three of the 17 most promising tasks with all their supporting linguistic and psychometric information reported alongside the example task.

What Makes for Effective AELP Tasks?

What made these tasks effective prototypes from our perspective was the fact that while they targeted academic English in different linguistic domains (lexical, syntactic, discourse), the measurement of specific aspects of academic English was still predominant (e.g., specialized and general academic vocabularies are the focal linguistic features measured by Effective Task Profile I in Chapter 4). The selection of these tasks presents the full range of difficulty from quite difficult (p =. 23) to relatively easy (p =.75), which would be necessary for an operational assessment of AELP to capture if the purpose was to measure progress in academic English language skills. These tasks all distinguished between students who came from Spanish-dominant home language backgrounds and those who came from English-dominant home language backgrounds, although we caution that this supplementary

language variable can only be a rough proxy for proficiency. Additionally, this variable may function as an indicator of cultural differences not only of language differences, and significant differences in task performance may also be a reflection of cultural biases in the tasks. On most tasks in this study, girls outperformed boys. However, for tasks to be considered effective, we required that the difference in performance by gender be slight and always statistically non-significant. Finally, the tasks all had "adequate" discrimination indices suggesting that they distinguished between the good and poor readers who attempted them.

Recommendations

Recommendations for Further Research

We make recommendations for future research in three main areas: First, further research can be conducted with the data collected during this stage of the AELP project. For example, at the item level, further examination of the tasks can be made in terms of difficulty. Specifically, we can investigate further characteristics (e.g., CELDT score, years in the US) of students who incorrectly answered the tasks and those who correctly answered the task, as well as examine correspondences between difficulty and specific linguistic characteristics.

Second, further research on the AELP construct and how tasks are designed to assess the construct is needed at other grade levels and in other modalities. The current study targeted the 5th grade and reading modality only. However, to respond to the needs of students across the K-12 span and to address the demands of academic language in the areas of listening, speaking, and writing, the efforts and processes we have described here will need to be repeated to take account of all grades and the additional modalities. Opportunity to continue with this line of research is possible with new CRESST projects focused on the validity of assessments used with ELL students currently underway.

Third, prior knowledge of the content is not, or should not be, necessary for providing the correct response for a language task; however, the verbal protocol data in this study show that the students who articulated their thought processes were, in fact, strongly influenced in completing tasks by their prior knowledge. Thus the interrelationship between language and content knowledge should not be minimized (e.g., Haladyna & Downing, 2004). Immediate further investigation is warranted to help ensure that interaction between the two does not interfere with assessment of the academic language construct.

Recommendations for the Test Development Process

The goal of the work here was to describe a process for developing tasks that tap academic English and to provide examples or prototypes of such tasks that could be used as

models for similar test development efforts. The strength of the test development process we followed for this AELP project was that we could improve both the tasks *and* the process itself. Specifically, we learned from implementing this process what could be improved, and the changes we recommend are captured in Figure 4. The transparency in the process we followed was achieved through the use of an audit during the pre-pilot and pilot stages. This trail of decision-making served to document the evolution of each task. However, as Figure 4 illustrates, we recommend that the process be expanded to include an audit trail at every stage of test development from initial construct definition all the way through to prototype creation.

The return arrows show how information from the tryouts and pilot administration impacted our task revision (including rejection). However, the bidirectional arrows in Figure 4, which were absent in Figure 1 (the process we followed), are meant to illustrate the suggestion of information flowing back from the various phases of empirical testing of tasks to the specification stage and further back to the construct framework and its formulation so that specifications and the language construct(s) to be measured can be modified as new information comes to light as a result of tryouts and pilots (and by extension, ideally also modified based on information from the field-testing of any preoperational test form).

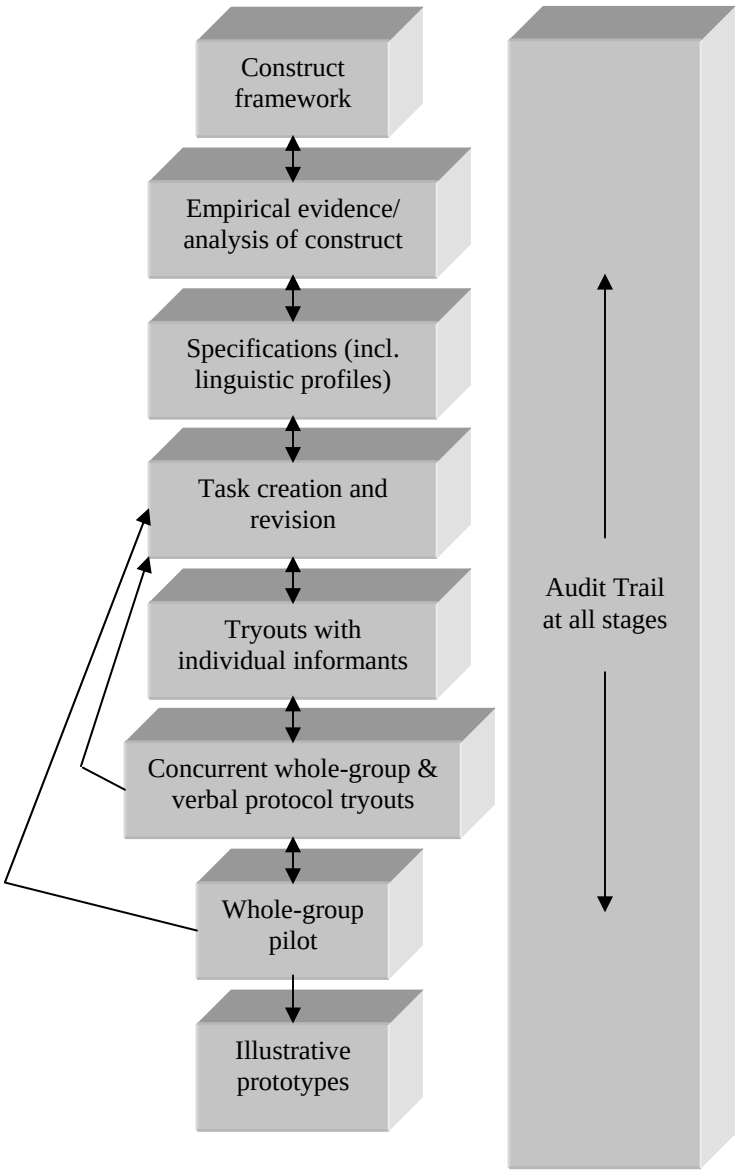


Figure 4. Proposed extensions/modifications to the test development process.

The 40 tasks taken together are not intended to be used as a test because they were not developed as part of a specific assessment plan for a particular purpose such as redesignation. Thus they do not cover the full range of language necessary for a comprehensive evaluation of AEL in reading for such a purpose. Nevertheless, the step-by-step process described here for each task illustrates the complex and iterative nature of task development.

The work here should be viewed within the context of specific test development efforts. That is, a test being developed for a specific purpose such as redesignation or diagnosis would have a set of content requirements and specifications that operationalize the construct to be tested. The range of AEL functions and features to be assessed for a given purpose would be clearly articulated. For a grade or grade span, the appropriate functions would be identified and then the features of the vocabulary and syntactic structure associated with those functions would be specified using academic standards and empirical evidence of classroom talk and texts.

For a test that would be part of redesignation decisions, the construct would be more broadly defined than for a classroom test in which a teacher may be focusing on one or two aspects of language. When decisions have been made about what specifically is to be tested, approaches for measuring the functions and features should be considered, specifications drafted, and tasks prepared for small-scale tryouts. The process described in this report takes potential tasks through tryout, modification, and piloting stages and produces an audit trail for each task.

In addition to following the evolution of each task, the broad picture of the full test must be kept in mind to assure adequate sampling of each content point being assessed. Initially, a target plan for full-test content should be prepared with the number of tasks for each function and features reflecting their importance (empirically established) within the construct. In addition, time limitations and other operational constraints should be noted. In other words, test parameters must be established. In a large-scale assessment several functions might be represented equally, whereas in a classroom test, one or two functions may receive the most emphasis due to coverage in the curriculum during the time prior to testing. Having a full-test plan, which may be modified throughout the process, nevertheless provides a structure for guiding task development.

After small-scale tryouts, revisions, and piloting as described in this report have been completed with a sufficient number of tasks to allow adequate content coverage according to the test design, test assembly for field-testing effort begins. The field-test data provide further evidence about the quality of the tasks as well as whole-test information on the reliability and validity of the instrument. Field-testing provides the first evidence of how well the tasks taken as a whole function to achieve the purpose of the instrument. The more thorough the early stages of test development as we have operationalized them in this report, the fewer tasks need replacing at the later field-testing stage in the test development process.

To conclude, the goal of the CRESST academic English research effort, has been to illustrate a process that would lead to valid and reliable instruments for assessing the English language skills of ELL students K-12. We have tried to show the importance of each step in

the process and along the way have stressed the role of empirical evidence as the foundation for developing instruments of high technical quality. The process is systematic yet flexible by allowing data to continually inform the effectiveness of tasks. Its iterative nature is the key to assuring quality assessments that are revised periodically through a feedback system. Documentation at every stage helps establish the validity argument of the assessment. Only by following a rigorous development path (with of course, ongoing monitoring once an assessment is operational), can we ensure that students' language skills are being accurately and fairly evaluated. We hope that the work reported here contributes to that goal.

References

Abedi, J. (1997). *Dimensionality of NAEP Subscale Scores in Mathematics* (CSE Tech. Rep. No. 428). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory.* Prospect Heights, Ill: Waveland Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standard for educational and psychological testing*. Washington, DC: Author.

Anderson, N. (1991). Individual differences in strategy use in second language reading and testing. *Modern Language Journal, 75,* 460-472.

Anderson, N., Bachman, L., Perkins, K. & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8,* 41-66.

August, D, & Hakuta, K. (1997). *Improving schooling for language-minority children: A research agenda.* Washington DC: National Academy Press.

Bachman, L. F. (2003, March). *Issues in Assessing Language Proficiency and Academic Achievement.* Paper presented at the East Coast Organization of Language Testers Conference (ECOLT), Washington, D.C.

Bachman, L. F. (2002). *Statistical analyses for language assessment.* Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bailey A. L. (2006). From Lambie to Lambaste: *The conceptualization, operationalization and use of academic English in the assessment of ELL students*. In K. Rolstad (Ed.), Rethinking Academic English. Mahwah, NJ: LEA.

Bailey, A.L., & Butler, F.A. (2002/2003). *An evidentiary framework for operationalizing academic English for broad application to K-12 education: A design document.* (Final Deliverable to OERI/OBEMLA Contract No. R305B960002) (CSE Tech. Rep. No. 611). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A.L., & Butler, F. A. (2006). A conceptual framework of academic English language for broad application to education. In A. L. Bailey, (Ed.). *Language*

Demands of School: Putting academic English to the test. New Haven CT: Yale University Press.

Bailey, A.L., Butler, F.A., LaFramenta, C., & Ong, C. (2001/2004). *Towards the characterization of academic English in upper elementary science classrooms*. (Final Deliverable to OERI Contract No. R305B960002) (CSE Tech. Rep. No. 621). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A.L., & Heritage, M., (Forthcoming, 2007). *Formative Assessment for Literacy Learning: Developing reading and academic English proficiency together, K-6.* Thousand Oaks, CA: Sage-Corwin Press.

Bailey, A. L., Stevens, R., Butler, F. A., Huang, B., & Miyoshi, J. N. (2005). *Using standards and empirical evidence to develop academic english proficiency test tasks in reading* (CSE Tech. Rep. No. 664). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., & Castellon-Wellington, M. (2000/2005). *Students' concurrent performance on tests of English language proficiency and academic achievement* (Final Deliverable to OERI, Contract No. R30B60002). (In CSE Tech. Rep. No. 663). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., Bailey, A. L., Stevens, R., Huang, B., & Lord, C. (2004). *Academic English in Fifth-Grade Mathematics, Science and Social Studies Textbooks* (CSE Tech. Rep. No. 642). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F.A., Lord, C., Stevens, R., Borrego, M., & Bailey, A.L. (2003/2004). *An approach to operationalizing academic English for language test development purposes: Evidence from fifth-grade science and math* (CSE Tech. Rep. No. 626). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F.A. & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: current trends and old dilemmas. *Language Testing, 18* (4), 409-427.

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic English learning approach*. Reading, MA: Addison-Wesley Publishing Company.

Cohen, A. (1998). *Strategies in learning and using a second language.* London: Longman.

Cohen, A. (2000). Exploring strategies in test-taking: Fine-tuning verbal reports from respondents. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment In ESL.* Mahwah NJ: Lawrence Erlbaum Associates.

Cummins, J. (1981). *The role of primary language development in promoting educational success for language minority students. In Schooling and language minority students: A theoretical framework.* Office of Bilingual Bicultural Education.

Cummins, J. (2003). BICS and CALP: Origins and rationale for the distinction. In C. B. Paulston, & R. Tucker (Eds.), *Sociolinguistics: The essential readings*, (pp. 322-328). Oxford, UK: Blackwell.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. Newhaven, CT: Yale University Press.

Davidson, F., Kim, J.T., Lee, H. J., Li, J., & Lopez, A. (2006). Using and auditing test specifications in language test development. In A. L. Bailey (Ed). *The Language Demands of School: Putting academic English to the test.* New Haven CT: Yale University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis. Verbal reports as data (2nd*

ed.). Cambridge, MA: The MIT Press.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high stakes testing. *Educational Measurement: Issues and practices, 23,* 17-27.

LaCelle-Peterson, M., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64,* 55-75.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *System, 6*(2), 199–215.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading*. Mahwah, NJ:

Lawrence Erlbaum Associates.

Purpura, J. E. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning, 47,* 289-325.

Scarcella, R. (2003). *Academic English: A conceptual framework* (Tech. Rep. No. 2003-1). Santa Barbara, CA: University of California Linguistic Minority Research Institute.

Schleppegrell, M. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12*(4), 431-459.

Short, D. (1994). Expanding middle school horizons: Integrating language, culture, and social studies. *TESOL Quarterly, 28,* 581-608.

Solomon, J., & Rhodes, N. (1995). *Conceptualizing academic English* (Research Rep. No. 15). Santa Cruz: University of California, National Center for Research on Cultural Diversity and Second Language Learning.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic English and content assessment: Measuring the progress of ELLs* (CSE Tech. Rep. No. 552). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing, 14,* 214–231.

United States Government Accountability Office (2006). No Child Left Behind Act: *Assistance from Education Could Help States Better Measure Progress of Students with Limited English Proficiency* (GAO Publication No. GAO-06-815). Washington, DC.

Witkin, B. R., & Altschuld, J. W. (1995). *Planning and conducting needs assessments*: *A practical guide*. Thousand Oaks: SAGE Publications.

Appendix A
Excerpts from the AELP Audit Trail

Item type	Note	Changes/Decision	Entry type	Source
	Informant Feedback	Lemonade: What did Carlotta buy for her sale? --> What "two things" did Carlotta buy for her sale? Also, split the answers into 2 tasks	Item content	
	Informant Feedback	Stiltwalker: each blank in the table was numbered to show it was an individual item	Format	
	Informant Feedback	Water supply: modify the instruction: -> fill in the missing words in the following passage using the words in the list below	Directions	
	Informant Feedback	Earth & moon: Q51: modify the instruction: the first one is done for you --> the first answer is given	Directions	
3		Align the four choices, and delete "circle the best answer" from the question stem	Format/ Directions	Internal review
10		Change the instruction for the table. Add "for questions 42 through 45" to the end of the instruction. Add lines following the question number in the table for students to fill in their responses	Directions	Internal review
2		Align the four choices, and delete "circle the best answer" from the question stem	Format/ Directions	Internal review
1		Italicize the example answer (earthquake faults). Parallel questions 53-55 with questions 50-52	Format	Internal review
10		Italicize the example answer (e). Align column B choices (a-f)	Format	Internal review
Passage		1) Change the instruction for the table. OLD: Read the problem. Then fill in the blanks in the table. NEW: Fill in the blanks for questions in the table below. 2) Add lines following the question number in the table for students to fill in their responses	Directions	Internal review

6		1) Delete the blank at the end of the question stem and add in a semi-colon (OLD: to organize is to _____. NEW: To organize is to: 2) Align the four choices.	Format	Internal review
7		Modify the instruction. Insert the following phrase "for questions 6 through 13" after the first 5 words.	Directions	Internal review
2		Modify the four choices. Delete the word "that" from all four responses and capitalize the following word.	Item content/ Format	Internal review
10		1) Modify the instruction. Insert the following phrase "for questions 20 through 24" after the first 3 words. 2) Add lines after the question #s in the table for students to fill in their responses.	Format/ Directions	Internal review

1	Two students asked whether question looking for proper names for people who predict the weather.		Comment	Whole class admin.
10	Students seem to be confused by the blanks in chart	Italicized the numbers to make more of an indication what needs to be completed.	Comment	Whole class admin.
1	Students had trouble pronouncing name	None (the name "Carlotta" was an unfamiliar proper noun and difficult to pronounce for many VP students, but we will keep the name because it's probably the publisher's intention to broaden the pool of proper nouns included in the text for cultural diversity)	Comment	Verbal protocol
Passage	Might have to shorten the George Washington passage because it is too long for students		Comment	Verbal protocol
1	Students answering without using passage vocabulary	Add the phrase "according to the passage" in the stem to prevent self-invented responses like "weather man" or "weather women"	Item content	Data entry
8		Add number 1 and 2 as examples to make the instructions clear. Also, option 1 was switched from the first position to the	Item content/ Format	Internal review

		second.		
8		Change the position of option 1	Format	Internal review
1	Students interpreting question: "Why did Washington do well" differently than anticipated. (Focusing on the need for his services or why he did well financially)	Changed stem to read "what made G. Washington a good surveyor"	Item content	Data entry/ verbal protocol
10		Add title for table (all tables in test), e.g., Table: Book Type and Number	Format	Internal review
Demographic Information	Students have difficulty answering; do not include question about gender	Delete 2 questions: 1. Do you speak a language other than English? 2. Did you start school in that country? Add one question about GENDER	Directions	Data entry
Entire Test	Want to mirror classroom texts more closely	Increase the font size from 12 to 13	Format	Internal review
Multiple Passages		*Revised* directions (took out language (1being the first event). Changed to the first event is given to you. Gave students first even (highlighted in bold, moved to middle of selections).	Directions	Internal review
Multiple Passages		Inserted title for table, changed orientation of number *(by rows)*	*Format*	Internal review

Appendix B
Pre-Pilot and Pilot Test Booklets

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

UCLA/CRESST

Academic English Language Proficiency

Reading – Form A [PRE-PILOT VERSION]

Created by the Academic English Language Proficiency research team at UCLA/CRESST

RESEARCH EDITION

This test is for research purpose only, not for general distribution, reproduction, or sale.

Form A

Before you start, please provide us with the following information.

Your Name: _____ (First, Last)

Your School: _____

Your Grade: _____

Your Teacher's Name: _____

Today's Date: _____

Do you speak a language other than English? ☐ Yes ☐ No

If yes, what language(s) do you speak? _____

What language do you speak most of the time at home?

_____

What country were you born in? _____

Did you start school in that country? ☐ Yes ☐ No

When did you start school in the United States?

☐ Preschool ☐ Kindergarten ☐ 1st ☐ 2nd ☐ 3rd

☐ 4th ☐ 5th ☐ 6th

☺ Thank you ☺

DIRECTIONS
Read the following passage. Then study sample questions 1 and 2.

Sample Passage

On a weekend camping trip, Ken, Eric, and their dad went for a walk on the Appalachian Trail. The first hour, they walked 3/8 mile. The second hour, they walked 4/5 mile. About how many miles did the boys and their dad walk? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Sample Questions

1. What two days of the week are weekend days?

_____

2. How many hours did the family walk on the camping trip? [Circle the correct answer.]

a) They walked for at least 2 hours.
b) They walked for an hour.
c) They walked just under 2 hours.
d) They walked for more than 2 hours.

DIRECTIONS
Now answer questions 1 and 2 on your own.

1. What is the math problem in the sample passage asking about?

a) How long the family walked.
b) How far the family walked.
c) How many people walked.
d) How many hours the boys walked.

2. Over which days did the camping trip occur?

_____

DIRECTIONS
Read the following passage and then answer questions 3 through 6.

Passage 1

Carlotta bought 9 packages of lemonade for $1.10 each and 2 packages of cups for $1.09 each. She sold 23 cups of lemonade every hour for 4 hours at $0.40 per cup. How much more money did Carlotta earn than she spent on supplies? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

What two things did Carlotta buy for her sale?

3. _____

4. _____

5. What is the word problem asking about?

e) How much Carlotta spent on supplies.
f) How many packages of lemonade she sold.
g) How much profit Carlotta made.
h) How much lemonade costs.

6. How long did Carlotta sell lemonade? _____

DIRECTIONS

Read the following passage and then answer questions 7 through 13.

Passage 2

In addition to temperature and air pressure, humidity, or the amount of water in the air, is an important factor in describing weather. But how does water get into the air?

Earth's oceans are the biggest source of water. As the sun heats the oceans, liquid water changes into an invisible gas called water vapor, which rises into the air. The process of liquid water changing to water vapor is called evaporation. High up in the atmosphere, where the air is cooler, water vapor turns back into liquid drops of water, forming clouds. This process is called condensation.

When cloud drops come together, gravity returns the water to the Earth's surface as precipitation—usually rain. If the temperature in the clouds is below freezing, the precipitation is sleet, hail, or snow. This transferring of water from the Earth's surface to the atmosphere and back is called the water cycle.

On clear nights, when the surface of the Earth cools quickly, water vapor may condense to form a cloud near the ground. This low cloud is called fog. If you have ever walked through fog, you know what the inside of a cloud is like.

Whether a cloud forms near the ground or high in the atmosphere, it forms in the same way. Water vapor condenses onto dust and other tiny particles in the air when it rises and cools. Another way in which air cools enough for water vapor to condense is by moving from a warm place to a colder place. For example, moist air that moves from over a warm body of water to over cooler land forms clouds or fog.

Even though all clouds form by condensation, different atmospheric conditions produce different types of clouds. Weather scientists, or meteorologists, give clouds three basic names—cirrus, cumulus, and stratus. Along with other information, the types of clouds in the atmosphere can be used to help predict weather changes.

7. According to the passage, what is the third factor used in describing the weather?

_____

What are two names for people who predict the weather?

8. _____

9. _____

10. According to the passage, all clouds are:
    a) formed the same way.
    b) the same as one another.
    c) located close to the ground.
    d) made of invisible gas.

Complete the sentences below using ***verbs from the passage***.

11. This passage explains how water _____ into the air.

12. If you _____ through fog, you will know what it's like to walk inside a cloud.

13. Water vapor condenses when it _____ from a warm place to a cooler place.

DIRECTIONS
Read the following passage and then answer questions 14 through 20.

Passage 3

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

In 1752 the young Washington joined the Virginia militia. Washington hoped a military career would bring him honor. He became angry when he learned that soldiers from the colonies were paid less to fight for the British than soldiers in the regular British army. Then, during the French and Indian War, the British lowered Colonel Washington's rank because they did not want colonists to rise above captain. Washington left the militia in protest. He later returned when the governor of Virginia restored his original rank.

In 1758, while still in the military, Washington was elected to the Virginia House of Burgesses. There he met Thomas Jefferson and Patrick Henry, and later joined colonial protests against the British.

In 1759 Washington retired from military life to manage his lands. By then he had become the most famous American in the military. That same year he married a wealthy widow named Martha Custis. George and Martha Washington moved to Mount Vernon, the plantation he owned on the Potomac River in Virginia. Martha Washington also supported the patriots. During the American Revolution, she helped her husband with his paperwork. She also sewed socks and cooked soup for the soldiers.

Martha Washington often joined George Washington in the field, where things were going badly for the Continental Army at the end of 1776. Washington was discouraged. He wrote, "Such is my situation that if I were to wish the bitterest curse to an enemy on this side of the grave, I should put him in my [place] with my feelings."

Certain of future victories, General Howe decided to rest for the winter in New York City. Washington knew that the British would not try to advance again until the spring. So he planned a surprise attack on the close to 1,400 Hessian troops in Trenton, New Jersey. The password Washington gave his soldiers was "Victory or Death!" After nightfall on Christmas Day, December 25, 1776, Washington and his troops crossed the Delaware River into New Jersey. The next morning, they surprised the Hessians, who quickly surrendered. "This is a glorious day for our country," said Washington.

14. Put the six sentences in the order in which the events occurred.

_____ George Washington was born.
_____ His troops won an important battle.
_____ He became an elected official.
_____ He married his wife.
_____ He joined the military.
_____ He worked as a surveyor.

15. According to the passage, why did George Washington do well at his first job?

_____

_____

16. How were the colonial soldiers and British soldiers treated differently?

a) The British were paid more than the colonists.
b) The colonists had higher ranks than the British.
c) The British and colonists were treated the same.
d) Both types of soldier had socks and soup.

17. The passage says: "Such is my situation that if I were to wish the bitterest curse to an enemy on this side of the grave, I should put him in my [place] with my feelings." Which of the following statements is true? The quote is used to:

a) describe George Washington's temperament.
b) show how George Washington felt at the time.
c) explain why George Washington was happy.
d) prove that George Washington was a good soldier.

Fill in the blanks using ***vocabulary words from the passage***.

18. George Washington and his friends _____ against the British.

19. George Washington's wife was _____ on the battlefield.

20. The Hessian troops were _____ by George Washington.

DIRECTIONS
Read the following passage and then answer questions 21 through 25.

Passage 4

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Fill in the blanks for questions 21 through 24 in the table below.

Person	**Year**	**Distance**	**Days**	**From**	**To**
Stilt walker #1	1980	(21.) _____	158	(23.) _____	Bowen, Kentucky
Stilt walker #2	1891	1,830 miles	(22.) _____	Paris, France	(24.) _____

25. Which walker traveled the longest distance?
_____

DIRECTIONS
Label question 26 through 29 in the diagram with words from the list.

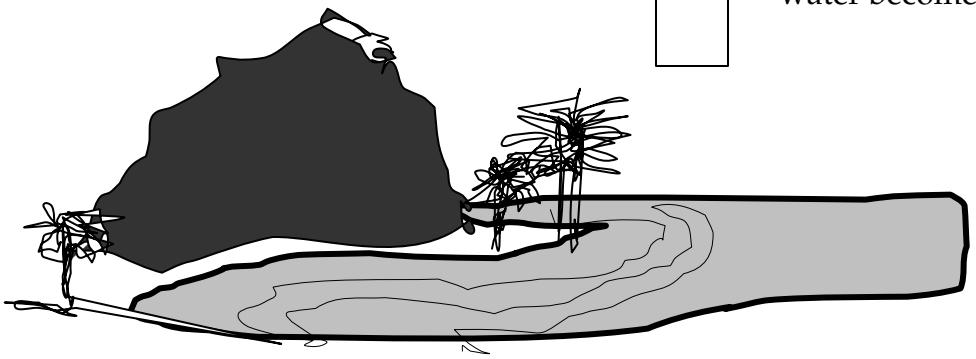
evaporation condensation water cycle precipitation

26. _____
name of the entire process

27. _____
water vapor forms into clouds

28. _____
gravity returns water to Earth

29. _____
water becomes an invisible gas



DIRECTIONS
Read the following passage and then answer questions 30 through 40.

Passage 5

Many women in the colonies took part in the movement, or effort by many people, to gain freedom. When Patriot leaders asked the colonists to boycott British-made goods, women in Boston and other colonial towns banded together to make their own goods. Many women worked for independence in other ways.

Some women formed groups to raise money for the war and collect clothing for the soldiers. In Lancaster, Pennsylvania, women formed a group that was called the Unmarried Ladies of America. Its members promised that they would never give their hand in marriage to any gentleman until he had proved himself a Patriot.

Women also took part in the fighting. Like some women, Mary Ludwig Hays traveled with her husband after he joined the Continental army. When her husband fell during the Battle of Monmouth, Hays took over the firing of his cannon. Mary Slocumb rode through thick forests at night to join her husband and other members of the North Carolina militia. She fought with them in the Battle of Moores Creek Bridge.

In Boston, Phillis Wheatley wrote poems that were praised by George Washington. She used her mind to champion the independence movement. So did Mercy Otis Warren. She wrote a play that made fun of the British and supported the Patriots.

Patriot Abigail Adams wanted to be sure that independence would be good for women as well as men. She wrote to her husband, John: "If particular care and attention are not paid to the ladies we are determined to foment [start] a rebellion and will not hold ourselves bound to obey any laws in which we have no voice or representation."

Not all women were Patriots. There were Loyalist women in every colony. Some of them fought for the British. Many others brought the British food and supplies.

30. What is this passage mainly about? Circle the best answer.

a) It explains how women supported themselves.
b) It lists the women who joined the military.
c) It describes the role of women in the war.
d) It labels women as Patriots.

Match the women's names to their activities. Write the correct letter next to each woman's name.

31. Mary Ludwig Hays	_____	a. supported equality for women after independence
32. Mary Slocumb	_____	b. fought in the war with her husband
33. Phillis Wheatley	_____	c. helped fight after her husband died
34. Abigail Adams	_____	d. used her writing to voice her opinions

35. How were Patriot women different from Loyalist women? Circle the best answer.

a) Patriot women wanted independence.
b) Loyalist women brought colonists food.
c) Patriot women and Loyalist women were the same.
d) Both types of women lived in the colonies.

36. Explain why Abigail Adams said to her husband "…we are determined to foment [start] a rebellion and will not hold ourselves bound to obey any laws in which we have no voice or representation."

_____
___

_____
___

Match the words in Column B to the vocabulary words from the passage in Column A. Put the letter of the word in Column B on the line next to the vocabulary word with the same meaning in Column A. The first answer is given.

Column A		Column B
Example: agree	__*e*___	a. gain freedom
37. goods	_____	b. banded together
38. independence	_____	c. took part
39. formed a group	_____	d. food and supplies
40. joined	_____	e. give their hand

DIRECTIONS
Read the following passage and then answer questions 41 through 47.

Passage 6

Darryl and his classmates were planning to sell hot dogs and soda pop at their school's big baseball game. They planned to donate the money they earned to the local hospital. They knew that last year 400 people came to the game and bought 190 cans of soda pop for $0.50 a can and 110 hot dogs at $1.25 each.

They had to decide how many hot dogs and cans of soda pop to buy this year and how much to charge. They learned that the cost of one hot dog and a bun was $0.80. If they bought at least 125 hot dogs, they would get free mustard, relish, and napkins. They could buy cans of cola or lemon-lime soda pop for $0.25 each. They could return any unsold soda pop, but unsold hot dogs could not be returned. How many cans of soda pop should they buy? What other information would be useful for making a decision? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

41. What are they going to do with the money they make?

_____

Fill in the table for questions 42 through 45 with information from the passage.

	Cost	**Last year's Selling price**	**Number sold last year**
Hot dogs	$0.80	(43.) _____	(45.) _____
Soda pop	(42.) _____	(44.) _____	190

46. In contrast to ordering and selling hot dogs, what is a good thing about ordering and selling soda pop?

_____

47. To get free mustard, relish, and napkins, how many hot dogs do Darryl and his classmates have to buy without worrying about unsold hot dogs they can't return? Choose the best answer.

a) 110 hot dogs
b) 124 hot dogs
c) 150 hot dogs
d) 190 hot dogs

GO ON →

DIRECTIONS

Read the following passage and then answer questions 48 through 60.

Passage 7

Earth's moon, our nearest neighbor in space, is a far different place from earth. There is no evidence of earthquake faults as on earth's crust. There are no erupting volcanoes. In fact there is no evidence of any of the kinds of motion that earth's crust has. Without air and water, there can be very little weathering or erosion. The moon has almost no air or water. There are no streams, no glaciers, and no wind. The only weathering and erosion is due to the impact of rocks from space hitting the moon's surface.

These rocks from space that strike a surface are called meteorites. Some craters formed by the impact of meteorites are big enough to be seen from earth. Others are so tiny the entire crater is on a single mineral crystal.

Can meteorites also strike earth's surface and produce craters? Yes! However, earth's atmosphere protects its surface from many such impacts. Rocks from space "burn up" as they pass through earth's atmosphere. The moon has little atmosphere. How does that fact affect the moon's surface?

Meteorite impacts shatter rocks on the moon and create a lot of heat. The heat melts the rock. Pieces of rock may melt together, and droplets and globs of molten rock can splatter outwards. Over time continual meteorite impacts break down the rock. The end result is a mixture of shattered pieces of rock, rock droplets, and melted-together bits of rock.

48. Which of the following is the best title for the passage?

a) How meteorites strike the moon.
b) How the moon is different from earth.
c) How the moon and earth are similar.
d) How rocks shatter on the moon.

49. Read this sentence from paragraph two of the passage:

> The only weathering and erosion is <u>due to</u> the impact of rocks from space hitting the Moon's surface.

Which words below have the same meaning as the words underlined in the sentence? Circle the best answer.

a) caused by
b) a part of
c) similar to
d) in order to

According to the passage above, what can you find on Earth that is not on the Moon? Fill in all the blanks in the list below. The first answer is given.

Example: *<u>earthquake faults</u>*

50. _____ 53. _____
51. _____ 54. _____
52. _____ 55. _____

Match the words in Column B to the vocabulary words from the passage in Column A. Put the letter of the word in Column B on the line next to the vocabulary word with the same meaning in Column A. The first answer is given.

Column A		Column B
Example: meteorite	__e__	a. unlike
56. impact	_____	b. broken
57. shattered	____	c. closest
58. produce	_____	d. hit
59. different	_____	e. rock
60. nearest	_____	f. make

DIRECTIONS
Read the following passage and then answer questions 61 through 66.

Passage 8

Your digestive system provides the nutrients your cells need to produce energy. To provide nutrients, the digestive system performs two functions. The first is to break food into nutrients. The second is to get the nutrients into the blood. Then the circulatory system transports them to your cells.

Digestion begins as you chew food, breaking it into smaller pieces so that you can swallow it. Glands in your mouth produce saliva. Saliva moistens food and begins to break down starchy foods, such as pasta, into sugars. (If you chew an unsalted cracker for a while, it will begin to taste sweet.)

When you swallow, food passes through the esophagus, a long tube that leads to the stomach. Gastric juice, produced by the stomach, contains acid and chemicals that break down proteins.

After several hours in the stomach, partly digested food moves into the small intestine. Digestion of food into nutrients is completed by chemicals produced in the small intestine. Nutrients diffuse through the villi, projections sticking out of the walls of the small intestine, into the blood. From the small intestine, undigested food passes into the large intestine. There, water and minerals diffuse into the blood, and wastes are removed from the body.

Two other organs have a role in digestion. The liver produces bile, which is stored in the gallbladder until it's needed. Bile breaks down fats into smaller particles that can be more easily digested. The pancreas produces a fluid that neutralizes stomach acid and chemicals that help finish digestion.

Complete each sentence with one of the words in the list. Each word can be used only once.

saliva	chew	gastric juice
bile	pancreas	

61. The _____ is an organ that helps the body complete digestion.

62. Your mouth makes a fluid called _____.

63. A fluid that helps the body break down fats is _____.

64. Your stomach makes _____ to help you digest proteins.

65. Put the following five sentences in the correct order. The first one is done for you.

__*1*__ You chew food into small pieces.
_____ Undigested food passes from the small intestine to the large intestine.
_____ The villi diffuse nutrients into the blood.
_____ Food passes through a long tube to the stomach.
_____ Your glands produce saliva.

66. Explain why it is important to digest food properly.

_____

_____

You are done! Please write down the time when you finish in the space provided below. Then close your booklet and wait for the teacher's instructions. Thank you for your participation. ☺

Finish Time: _____ : _____

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

UCLA/CRESST

Academic English Language Proficiency

Reading – Form B [PRE-PILOT VERSION]

Created by the Academic English Language Proficiency research team at UCLA/CRESST

RESEARCH EDITION

This test is for research purpose only, not for general distribution, reproduction, or sale.

Form B

Before you start, please provide us with the following information.

Your Name: _____ (First, Last)

Your School: _____

Your Grade: _____

Your Teacher's Name: _____ (First, Last)

Today's Date: _____

Do you speak a language other than English? ☐ Yes ☐ No

If yes, what language(s) do you speak? _____

What language do you speak most of the time at home?

_____

What country were you born in? _____

Did you start school in that country? _____

When did you start school in the United States?

☐ PreK ☐ Kindergarten ☐ 1st ☐ 2nd ☐ 3rd

☐ 4th ☐ 5th ☐ 6th

☺ Thank you ☺

DIRECTIONS
Read the following sample passage. Then study sample questions 1 and 2.

Sample Passage

On a weekend camping trip, Ken, Eric, and their dad went for a walk on the Appalachian Trail. The first hour, they walked 3/8 mile. The second hour, they walked 4/5 mile. About how many miles did the boys and their dad walk? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Sample Questions

1. What two days of the week are weekend days?

   _____

2. How many hours did the family walk on the camping trip? [Circle the correct letter.]

   f) They walked for at least 2 hours.
   g) They walked for an hour.
   h) They walked just under 2 hours.
   i) They walked for more than 2 hours.

DIRECTIONS
Read the following passage and then answer questions 1 through 5.

Passage 1

Alanna is organizing her books. There are 21 picture books. There are 15 fewer paperback books than picture books. There are 3 more textbooks than paperback books. There are 7 more textbooks than atlases. How many atlases are there? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Fill in the blanks of question 1 through 3 in the table below.

Type of book	Picture books	(1.) _____	Textbooks	(3.) _____
Number of each type of book	21	15 fewer than picture books	(2.) _____	There are 7 fewer than textbooks.

4. Put the numbers listed below in order from lowest to highest, starting with the lowest number at the top of the list.

21 _____
15 _____
3 _____
7 _____

5. To *organize* is to:

a) count things.
b) separate books.
c) put things in order.
d) summarize.

DIRECTIONS
Read the following passage and then answer questions 6 through 18.

Passage 2

You may have heard the terms supply and demand as they are used to describe the economy. Supply is the amount of a product or service that is available. Demand is the degree to which a product or service is wanted.

The supply of water is fairly steady—it does not change. The demand for this resource, however, has changed. As Earth's human population has increased, so has the demand for water.

In the United States, many areas use more fresh water than they receive from rain and snow. In other words, the demand in these areas is greater than the supply. To make up for this difference in supply and demand, people in such areas need to do two things. First, they have to find ways to use the fresh water they have wisely. Second, they have to find ways to get water from areas that receive more fresh water than those areas use.

The supply of fresh water for a given area is determined by the amount of precipitation—rain, snow, sleet, and hail—the area receives. The demand for fresh water depends mainly on population. Generally, areas around big cities withdraw more fresh water from surface and underground sources than rural areas do. However, less-populated regions with many farms can also use large amounts of fresh water.

Over 33 million people live in California. About 500,000 people move to the state every year. Nearly half of these people settle in southern California. Southern California is a much drier area than the northern part of the state. To meet the great demand for fresh water in southern California, rain and melted snow are collected in reservoirs around the state. This fresh water is then piped through canals, aqueducts, and pipelines to areas where it is needed.

A number of projects have been developed in California to store and deliver fresh water to farms and homes. One of these is the State Water project, begun in 1960. This project is responsible for bringing much of the fresh water to the people of southern California. By 1973 the first facilities were pumping much-needed fresh water in to southern California. Today the project delivers fresh water to two thirds of California's population and provides water for about 1.2 million acres of farmland.

Fill in the missing words for questions 6 through 13 in the following passage using the words in the list below. Use each word once and capitalize the word when necessary.

are	in	the	of
on	were	and	this

Passage 3

Many people (6)_____ beginning to recognize the importance of conservation, the wise use of Earth's natural resources. Water conservation practices include turning off (7)_____ faucet while brushing teeth, taking shorter showers, (8)_____ watering lawns with drip hoses. The activities (9)_____ pages D44-D46 show that everyday activities can be modified to conserve water.

People can conserve water in many creative ways. For example, many lawns (10)_____ California have been replaced with grasses and other plants that are adapted to a dry climate. This practice greatly reduces the amount (11)_____ water needed for watering plants. In the early 1990s, many private toilets in the Los Angeles area (12)_____ replaced with newer models that use much less water. (13)_____ measure saves millions of gallons of fresh water every year.

14. What is the main point of passage 2 and 3?

a) People need to increase the supply of water.
b) People should conserve water.
c) Water conservation is too difficult.
d) Fresh water is used wisely.

Complete the following sentences using words from passage 5 and 6.

15. The southern part of California is _____ than Northern California.

16. Newer toilets use _____ water than older toilets.

17. Rain and snow are two types of _____.

18. You can help _____ water by turning off the faucet while brushing your teeth.

DIRECTIONS
Read the following passage and then answer questions 19 through 33.

Passage 4

During the late 1600s and early 1700s, more and more immigrants traveled to the Americas. England, which later became part of Britain, had settled 13 colonies along the Atlantic coast of North America. In time, some of the early settlements grew into towns and cities. Life in the towns and cities varied from place to place.

Many settlers in the New England colonies lived in towns where, as one settler wrote, "every man...lives in a tidy warm house, has plenty of good food and fuel, with whole clothes from head to foot, [made by] his family." Most New England towns were self-sufficient communities in which the people grew or made most of what they needed.

The earliest New England towns were built on a narrow road. Each of the town's families had a house on this lane. Families had their own gardens and pens for cows, sheep, chicken, or pigs. In the fields near the town, the people grew crops to sell to others and to use for themselves.

A meetinghouse stood at the center of most New England towns. In many places people came to the meetinghouse several times a week to worship together. The meetinghouse was also used for town meetings. At a town meeting male landowners could take part in government.

Two of the most important town workers were the herder and the constable. The herder was the person who took care of the animals on the town's common, an open area where livestock grazed. The constable was a police officer who made sure people obeyed the town's laws. Another important worker was the leader of the town's militia, or volunteer army. Men and boys gathered on the common to train.

Another kind of town developed in many places, especially in the middle colonies. This was the market town. Farmers traveled to market towns to trade their farm produce—grains, fruits, and vegetables—for goods and services.

In most market towns a general store sold imports, or goods brought into the colonies from other countries. The imports included tea, sugar, spices, cloth, shoes, stockings, and buttons. Near the general store was the shop of a cobbler, who made and repaired shoes. There was often a blacksmith's shop, where iron was made into horseshoes, hinges, and nails. Most market towns also had a gristmill, where grain was ground into flour and meal, and a sawmill, where logs were sawed into lumber.

Market towns often had more than one church. A Lutheran church might be a block away from a Quaker meetinghouse or just down the street from a Methodist church.

To carry their produce to the market towns, many Pennsylvania farmers used big covered wagons called Conestogas. Conestogas were much larger than regular wagons. When a visitor from Europe first saw them, he called them "huge moving houses."

When market towns grew along rivers, farmers carried their produce to the towns by boat instead of by wagon. It was easier and cheaper to ship heavy goods by water. After unloading their produce, farmers returned home, their boats filled with goods from the general store.

19. What is the passage mainly about?

a) Trading in small colonial towns.
b) Life styles in different types of colonial towns.
c) The differences between towns and cities.
d) The most important town workers.

Complete the table for questions 20 through 24 following Directions A and B below.

A. Fill in the bottom part of the table with the statements below.

Trading community	Made their own clothes
More than one church	Grew their own food

B. Use the information in the bottom of the table to label the types of towns in the top row.

Type of Town	(20.) _____	**Market town**
Type of community	Self-sufficient community	(23.) _____
Food	(21.) _____	They bought from others
Worship	One meetinghouse	(24.) _____
Clothing	(22.) _____	Imported
Travel	Not necessary	Traveled to trade with others

25. A settler is quoted in the passage as having written: "every man...lives in a tidy warm house, has plenty of good food and fuel, with whole clothes from head to foot, [made by] his family." Choose the best reason this quote is in the passage. The quote is used to:

a) explain how people lived.
b) describe what the settler observed.
c) compare colonists and settlers.
d) label the community.

Match the words in Column A to the definitions in Column B. Put the letter of the word in Column B on the line next to the vocabulary word with the same meaning in Column A. The first answer is given.

Column A		Column B
Example: common	__c___	a. place where lumber is made
26. imports	_____	b. volunteer army
27. militia	_____	c. place where animals ate
28. sawmill	_____	d. product from another country

Match the job titles in Column A to their activities in Column B. Put the letter of the activity in Column B on the line next to the matched job title.

Column A		Column B
29. cobbler	____	a. took care of the animals on the common
30. constable	____	b. turned iron into hardware such as nails
31. herder	____	c. made sure people obeyed laws
32. blacksmith	____	d. made and repaired shoes

33. Explain why the visitor from Europe called Conestogas "huge moving houses".

_____

_____

DIRECTIONS
Read the following passage and then answer questions 34 through 35.

Passage 5

To put a traffic light at an intersection, a city requires an average of 5000 vehicles per day passing through the intersection. In one week the following numbers of vehicles were counted: Sunday, 1812; Monday, 6213; Tuesday, 5935; Wednesday, 6086; Thursday, 6113; Friday, 6184; Saturday, 2593. The city claims the average number of cars is approximately 4991, and there should not be a light. A neighborhood group claims the average number of cars is 6086, and there should be a light. How is each group finding its "average"? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

34. Which sentence is correct?

a) The city and the neighborhood group disagree about the average.
b) An average of 5935 cars passes through the intersection on Tuesdays.
c) The number of cars passing through the intersection each day is unimportant.
d) The average number of cars is reason for a light.

35. Which two words mean the same thing in this math problem? Circle the correct answer.

Car–vehicle Traffic light–intersection Approximately–average

DIRECTIONS

Read the following passage and then answer questions 36 through 40.

Passage 6

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Fill in the blanks for questions 36 through 40 in the table below.

Person	**Year**	**Distance**	**Days**	**From**	**To**
Stilt walker #1	1980	(36.) _____	158	(38.) _____	Bowen, Kentucky
Stilt walker #2	1891	1,830 miles	(37.) _____	Paris, France	(39.) _____

40. Which walker traveled the longest distance? _____

DIRECTIONS

Label question 41 through 44 in the diagram with words from the list.

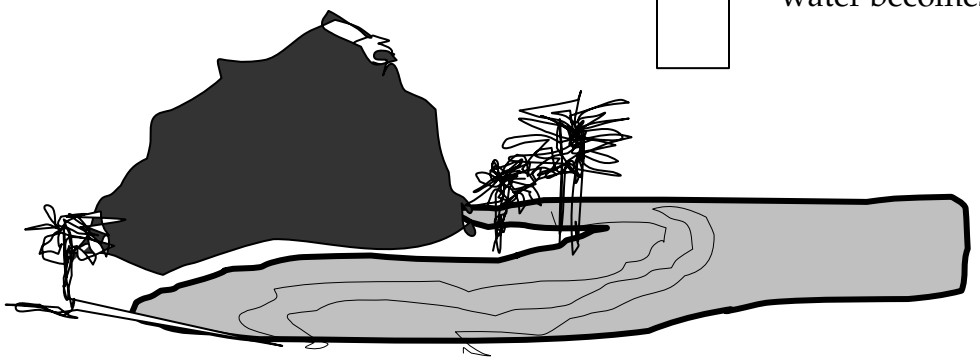
evaporation condensation water cycle precipitation

41. _____
name of the entire process

42. _____
water vapor forms into clouds

43. _____
gravity returns water to Earth

44. _____
water becomes an invisible gas



DIRECTIONS
Read the following passage and then answer questions 45 through 48.

Passage 7

Carlotta bought 9 packages of lemonade for $1.10 each and 2 packages of cups for $1.09 each. She sold 23 cups of lemonade every hour for 4 hours at $0.40 per cup. How much more money did Carlotta earn than she spent on supplies? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

What two things did Carlotta buy for her sale?

45. _____

46. _____

47. What is the word problem asking about?

a) How much Carlotta spent on supplies.
b) How many packages of lemonade she sold.
c) How much profit Carlotta made.
d) How much lemonade costs.

48. How long did Carlotta sell lemonade? _____

GO ON →

DIRECTIONS

Read the following passage and then answer questions 49 through 55.

Passage 8

In addition to temperature and air pressure, humidity, or the amount of water in the air, is an important factor in describing weather. But how does water get into the air?

Earth's oceans are the biggest source of water. As the sun heats the oceans, liquid water changes into an invisible gas called water vapor, which rises into the air. The process of liquid water changing to water vapor is called evaporation. High up in the atmosphere, where the air is cooler, water vapor turns back into liquid drops of water, forming clouds. This process is called condensation.

When cloud drops come together, gravity returns the water to the Earth's surface as precipitation—usually rain. If the temperature in the clouds is below freezing, the precipitation is sleet, hail, or snow. This transferring of water from the Earth's surface to the atmosphere and back is called the water cycle.

On clear nights, when the surface of the Earth cools quickly, water vapor may condense to form a cloud near the ground. This low cloud is called fog. If you have ever walked through fog, you know what the inside of a cloud is like.

Whether a cloud forms near the ground or high in the atmosphere, it forms in the same way. Water vapor condenses onto dust and other tiny particles in the air when it rises and cools. Another way in which air cools enough for water vapor to condense is by moving from a warm place to a colder place. For example, moist air that moves from over a warm body of water to over cooler land forms clouds or fog.

Even though all clouds form by condensation, different atmospheric conditions produce different types of clouds. Weather scientists, or meteorologists, give clouds three basic names—cirrus, cumulus, and stratus. Along with other information, the types of clouds in the atmosphere can be used to help predict weather changes.

49. According to the passage, what is the third factor used in describing the weather?

_____

What are two names for people who predict the weather?

50. _____

51. _____

52. According to the passage, all clouds are :
    e) formed the same way.
    f) the same as one another.
    g) located close to the ground.
    h) made of invisible gas.

Complete the sentences below using ***verbs from the passage***.

53. This passage explains how water _____ into the air.

54. If you _____ through fog, you will know what it's like to walk inside a cloud.

55. Water vapor condenses when it _____ from a warm place to a cooler place.

DIRECTIONS
Read the following passage and then answer questions 56 through 62.

Passage 9

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

In 1752 the young Washington joined the Virginia militia. Washington hoped a military career would bring him honor. He became angry when he learned that soldiers from the colonies were paid less to fight for the British than soldiers in the regular British army. Then, during the French and Indian War, the British lowered Colonel Washington's rank because they did not want colonists to rise above captain. Washington left the militia in protest. He later returned when the governor of Virginia restored his original rank.

In 1758, while still in the military, Washington was elected to the Virginia House of Burgesses. There he met Thomas Jefferson and Patrick Henry, and later joined colonial protests against the British.

In 1759 Washington retired from military life to manage his lands. By then he had become the most famous American in the military. That same year he married a wealthy widow named Martha Custis. George and Martha Washington moved to Mount Vernon, the plantation he owned on the Potomac River in Virginia. Martha Washington also supported the patriots. During the American Revolution, she helped her husband with his paperwork. She also sewed socks and cooked soup for the soldiers.

Martha Washington often joined George Washington in the field, where things were going badly for the Continental Army at the end of 1776. Washington was discouraged. He wrote, "Such is my situation that if I were to wish the bitterest curse to an enemy on this side of the grave, I should put him in my [place] with my feelings."

Certain of future victories, General Howe decided to rest for the winter in New York City. Washington knew that the British would not try to advance again until the spring. So he planned a surprise attack on the close to 1,400 Hessian troops in Trenton, New Jersey. The password Washington gave his soldiers was "Victory or Death!" After nightfall on Christmas Day, December 25, 1776, Washington and his troops crossed the Delaware River into New Jersey. The next morning, they surprised the Hessians, who quickly surrendered. "This is a glorious day for our country," said Washington.

56. Put the six sentences in the order in which the events occurred.

_____ George Washington was born.
_____ His troops won an important battle.
_____ He became an elected official.
_____ He married his wife.
_____ He joined the military.
_____ He worked as a surveyor.

57. According to the passage, why did George Washington do well at his first job?

_____

_____

58. How were the colonial soldiers and British soldiers treated differently?

e) The British were paid more than the colonists.
f) The colonists had higher ranks than the British.
g) The British and colonists were treated the same.
h) Both types of soldier had socks and soup.

59. The passage says: "Such is my situation that if I were to wish the bitterest curse to an enemy on this side of the grave, I should put him in my [place] with my feelings." Which of the following statements is true? The quote is used to:

e) describe George Washington's temperament.
f) show how George Washington felt at the time.
g) explain why George Washington was happy.
h) prove that George Washington was a good soldier.

Fill in the blanks using ***vocabulary words from the passage***.

60. George Washington and his friends _____ against the British.

61. George Washington's wife was _____ on the battlefield.

62. The Hessian troops were _____ by George Washington.

You are done! Please write down the time when you finish in the space provided below. Then close your booklet and wait for the teacher's instructions. Thank you for your participation. ☺

Finish Time: _____:_____

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

UCLA/CRESST

Academic English Language Proficiency

Reading [PILOT VERSION}

Created by the Academic English Language Proficiency research team at UCLA/CRESST

RESEARCH EDITION

This test is for research purpose only, not for general distribution, reproduction, or sale.

Before you start, please provide us with the following information.

Your Name: _____ (First, Last)

Your School: _____

Your Grade: _____ Gender ☐ Boy ☐ Girl

Your Teacher's Name: _____

Today's Date: _____

What language(s) do you speak? _____

What language do you speak most of the time at home?

_____

What country were you born in? _____

Did you start school in that country? ☐ Yes ☐ No

When did you start school in the United States?

☐ Preschool ☐ Kindergarten ☐ 1st ☐ 2nd ☐ 3rd

☐ 4th ☐ 5th ☐ 6th

☺ Thank you ☺

DIRECTIONS
Read the following passage and then study sample questions 1 and 2.

Sample Passage

On a weekend camping trip, Ken, Eric, and their dad went for a walk on the Appalachian Trail. The first hour, they walked 3/8 mile. The second hour, they walked 4/5 mile. About how many miles did the boys and their dad walk? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Sample Questions

1. What two days of the week are weekend days?

_____

2. How many hours did the family walk on the camping trip?

[Circle the correct letter.]

j) They walked for at least 2 hours.
k) They walked for an hour.
l) They walked just under 2 hours.
m) They walked for more than 2 hours.

DIRECTIONS
Read the following passage and then answer questions 1 through 4.

Passage 1

Carlotta bought 9 packages of lemonade for $1.10 each and 2 packages of cups for $1.09 each. She sold 23 cups of lemonade every hour for 4 hours at $0.40 per cup. How much more money did Carlotta earn than she spent on supplies? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

What two things did Carlotta buy for her sale?

1. _____

2. _____

3. What is the word problem asking about?

i) How much Carlotta spent on supplies.
j) How many packages of lemonade she sold.
k) How much profit Carlotta made.
l) How much lemonade costs.

4. How long did Carlotta sell lemonade? _____

DIRECTIONS
Read the following passage and then answer questions 5 through 10.

Passage 2

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

In 1752 the young Washington joined the Virginia militia. Washington hoped a military career would bring him honor. He became angry when he learned that soldiers from the colonies were paid less to fight for the British than soldiers in the regular British army. Then, during the French and Indian War, the British lowered Colonel Washington's rank because they did not want colonists to rise above captain. Washington left the militia in protest. He later returned when the governor of Virginia restored his original rank.

In 1758, while still in the military, Washington was elected to the Virginia House of Burgesses. There he met Thomas Jefferson and Patrick Henry, and later joined colonial protests against the British.

In 1759 Washington retired from military life to manage his lands. By then he had become the most famous American in the military. That same year he married a wealthy widow named Martha Custis. George and Martha Washington moved to Mount Vernon, the plantation he owned on the Potomac River in Virginia. Martha Washington also supported the patriots. During the American Revolution, she helped her husband with his paperwork. She also sewed socks and cooked soup for the soldiers.

Martha Washington often joined George Washington in the field, where things were going badly for the Continental Army at the end of 1776. Washington was discouraged. He wrote, "Such is my situation that if I were to wish the bitterest curse to an enemy on this side of the grave, I should put him in my [place] with my feelings."

Certain of future victories, General Howe decided to rest for the winter in New York City. Washington knew that the British would not try to advance again until the spring. So he planned a surprise attack on the close to 1,400 Hessian troops in Trenton, New Jersey. The password Washington gave his soldiers was "Victory or Death!" After nightfall on Christmas Day, December 25, 1776, Washington and his troops crossed the Delaware River into New Jersey. The next morning, they surprised the Hessians, who quickly surrendered. "This is a glorious day for our country," said Washington.

5. Put the six sentences in the order in which the events occurred by putting a number from ***1-6*** on the line next to each event. The first event is given to you.

_____ His troops won an important battle.
_____ He became an elected official.
1 **George Washington was born.**
_____ He married his wife.
_____ He joined the military.
_____ He worked as a surveyor.

6. According to the passage, what made George Washington a good surveyor?

_____

_____

7. How were the colonial soldiers and British soldiers treated differently?

i) The British were paid more than the colonists.
j) The colonists had higher ranks than the British.
k) The British and colonists were treated the same.
l) Both types of soldier had socks and soup.

8. The passage says: "Such is my situation that if I were to wish the bitterest curse to an enemy on this side of the grave, I should put him in my [place] with my feelings." Which of the following statements is true? The quote is used to:

i) describe George Washington's temperament.
j) show how George Washington felt at the time.
k) explain why George Washington was happy.
l) prove that George Washington was a good soldier.

Fill in the blanks using ***vocabulary words from the passage***.

9. George Washington and his friends _____ against the British.

10. The Hessian troops were _____ by George Washington.

DIRECTIONS
Read the following passage and then answer questions 11 through 15

Passage 3

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Fill in the blanks for questions *(11)* through *(14)* in the table below.

Table: Stilt walker Travel by Distance and Days

Person	**Year**	**Distance**	**Days**	**From**	**To**
Stilt walker #1	1980	*(11)* _____	158	*(12)* _____	Bowen, Kentucky
Stilt walker #2	1891	1,830 miles	*(13)* _____	Paris, France	*(14)* _____

15. Which walker traveled the longest distance? _____

GO ON →

DIRECTIONS
Read the following passage and then answer questions 16 through 25.

Passage 4

Many women in the colonies took part in the movement, or effort by many people, to gain freedom. When Patriot leaders asked the colonists to boycott British-made goods, women in Boston and other colonial towns banded together to make their own goods. Many women worked for independence in other ways.

Some women formed groups to raise money for the war and collect clothing for the soldiers. In Lancaster, Pennsylvania, women formed a group that was called the Unmarried Ladies of America. Its members promised that they would never give their hand in marriage to any gentleman until he had proved himself a Patriot.

Women also took part in the fighting. Like some women, Mary Ludwig Hays traveled with her husband after he joined the Continental army. When her husband fell during the Battle of Monmouth, Hays took over the firing of his cannon. Mary Slocumb rode through thick forests at night to join her husband and other members of the North Carolina militia. She fought with them in the Battle of Moores Creek Bridge.

In Boston, Phillis Wheatley wrote poems that were praised by George Washington. She used her mind to champion the independence movement. So did Mercy Otis Warren. She wrote a play that made fun of the British and supported the Patriots.

Patriot Abigail Adams wanted to be sure that independence would be good for women as well as men. She wrote to her husband, John: "If particular care and attention are not paid to the ladies we are determined to foment [start] a rebellion and will not hold ourselves bound to obey any laws in which we have no voice or representation."

Not all women were Patriots. There were Loyalist women in every colony. Some of them fought for the British. Many others brought the British food and supplies.

16. What is this passage mainly about?

a) It explains how women supported themselves.
b) It lists the women who joined the military.
c) It describes the role of women in the war.
d) It labels women as Patriots.

Match the women's names to their activities. Write the correct letter next to each woman's name.

17. Mary Ludwig Hays	_____	a. supported equality for women after independence
18. Mary Slocumb	_____	b. fought in the war with her husband
19. Phillis Wheatley	_____	c. helped fight after her husband died
20. Abigail Adams	_____	d. used her writing to voice her opinions

21. How were Patriot women different from Loyalist women?

a) Patriot women wanted independence.
b) Loyalist women took part in the war.
c) Patriot women and Loyalist women were the same.
d) Patriot women lived in the colonies.

Match the words in Column B to the vocabulary words from the passage in Column A. Put the letter of the word in Column B on the line next to the vocabulary word with the same meaning in Column A. An example is given.

Column A		Column B
Example: **agree**	__e__	a. gain freedom
22. goods	_____	b. banded together
23. independence	_____	c. took part
24. formed a group	_____	d. food and supplies
25. joined	_____	e. **give their hand**

DIRECTIONS
Read the following passage and then answer questions 26 through 32.

Passage 5

Darryl and his classmates were planning to sell hot dogs and soda pop at their school's big baseball game. They planned to donate the money they earned to the local hospital. They knew that last year 400 people came to the game and bought 190 cans of soda pop for $0.50 a can and 110 hot dogs at $1.25 each.

They had to decide how many hot dogs and cans of soda pop to buy this year and how much to charge. They learned that the cost of one hot dog and a bun was $0.80. If they bought at least 125 hot dogs, they would get free mustard, relish, and napkins. They could buy cans of cola or lemon-lime soda pop for $0.25 each. They could return any unsold soda pop, but unsold hot dogs could not be returned. How many cans of soda pop should they buy? What other information would be useful for making a decision? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

26. What are they going to do with the money they make?

_____

Fill in the blanks in the table below for questions *(27)* through *(30)* by choosing correct words or phrases from the passage.

Table: Hot Dog and Soda Pop Sales

	Cost	**Last year's selling price**	**Number sold last year**
Hot dogs	$0.80	*(27)* _____	*(28)* _____
Soda pop	*(29)* _____	*(30)* _____	190

31. In contrast to ordering and selling hot dogs, what is a good thing about ordering and selling soda pop?

_____

_____

32. To get free mustard, relish, and napkins, how many hot dogs do Darryl and his classmates have to buy? Choose the best answer.

a) No more than 110 hot dogs
b) Less than 120 hot dogs
c) At least 150 hot dogs
d) Exactly 120 hot dogs

DIRECTIONS
Read the following passage and then answer questions 33 through 38.

Passage 6

Your digestive system provides the nutrients your cells need to produce energy. To provide nutrients, the digestive system performs two functions. The first is to break food into nutrients. The second is to get the nutrients into the blood. Then the circulatory system transports them to your cells.

Digestion begins as you chew food, breaking it into smaller pieces so that you can swallow it. Glands in your mouth produce saliva. Saliva moistens food and begins to break down starchy foods, such as pasta, into sugars. (If you chew an unsalted cracker for a while, it will begin to taste sweet.)

When you swallow, food passes through the esophagus, a long tube that leads to the stomach. Gastric juice, produced by the stomach, contains acid and chemicals that break down proteins.

After several hours in the stomach, partly digested food moves into the small intestine. Digestion of food into nutrients is completed by chemicals produced in the small intestine. Nutrients diffuse through the villi, projections sticking out of the walls of the small intestine, into the blood. From the small intestine, undigested food passes into the large intestine. There, water and minerals diffuse into the blood, and wastes are removed from the body.

Two other organs have a role in digestion. The liver produces bile, which is stored in the gallbladder until it's needed. Bile breaks down fats into smaller particles that can be more easily digested. The pancreas produces a fluid that neutralizes stomach acid and chemicals that help finish digestion.

33. Based on the passage, explain what two things the digestive system does to allow your body to produce energy.

_____

_____

Complete sentences 34-37 with one of the words in the list. Each word can be used only once.

saliva	chew	gastric juice
bile	pancreas	villi

34. The _____ is an organ that helps the body complete digestion.

35. Your mouth makes a fluid called _____.

36. A fluid that helps the body break down fats is _____.

37. Your stomach makes _____ to help you digest proteins.

38. Put the following five sentences in the correct order to show the process of digestion. Put a number for each step ***(1-5)*** on the line next to each sentence. The first step ***(1)*** is given to you.

_____ Undigested food passes from the small intestine to the large intestine.

__*1*__ **You chew food into small pieces**.

_____ The villi diffuse nutrients into the blood.

_____ Food passes through a long tube to the stomach.

_____ Your glands produce saliva.

DIRECTIONS
Read the following passage and then answer questions 39 and 40.

Passage 7

To put a traffic light at an intersection, a city requires an average of 5000 vehicles per day passing through the intersection. In one week the following numbers of vehicles were counted: Sunday, 1812; Monday, 6213; Tuesday, 5935; Wednesday, 6086; Thursday, 6113; Friday, 6184; Saturday, 2593. The city claims the average number of cars is approximately 4991, and there should not be a light. A neighborhood group claims the average number of cars is 6086, and there should be a light. How is each group finding its "average"? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

39. Which sentence is correct?

e) The city and the neighborhood group disagree about the average.
f) An average of 5935 cars passes through the intersection on Tuesdays.
g) The number of cars passing through the intersection each day is unimportant.
h) The average number of cars is reason for a light.

40. Which two words mean the same thing in this math problem? Circle the correct answer.

car–vehicle light–intersection approximately–average

You are done! Please write down the time you finished in the space provided below. Then close your booklet and wait for the teacher's instructions. Thank you for your participation. ☺

Finish Time: _____:_____

Appendix C
Protocol of Directions for Whole-group Administration

Directions (to be read aloud)

Hi, my name is _____ … We are here because we need your help. We are studying what students do when they take tests. You can help us in our research by answering some questions. You aren't expected to know everything. Some questions are hard. Try your best to read the passages and answer the questions. Your answers will help us to create fairer tests for students like you across California. These questions will not affect your grades in school and should take you about thirty minutes to complete.

Please answer the questions in the booklet. Feel free to write any notes or mark the reading passages if you want to. Please try your best to answer the questions. For questions that you are not sure about, make your best guess. If you run into questions that you can't answer, please write "0" as your answer so we know you have still read the passage and the question.

After you are finished, please write the time (*point to the alarm clock in the front of the classroom*) **where it says, "time completed" on the last page** *(point to the time completed blank in the booklet).* **Then close your booklet.**

Do you understand the directions?

Now I'd like for you to turn to page 1. We would like for you to answer a few questions about yourself.

Researcher will determine whether students need every question read aloud or only those pertaining to their home language, etc. (questions 8-11).

1. Write Your Name: _____ (First, Last)
2. Your School: _____
3. Your Grade: _____
4. Your Teacher's Name: _____ (First, Last)
5. Today's Date: _____
6. Do you speak a language other than English? **Check one box** ☐ Yes **OR** ☐ No
7. If yes, what language(s) do you speak? _____
8. What language do you speak most of the time at home? _____
9. What country were you born in? _____
10. Did you start school in that country? _____

11. When did you start school in the United States? **Check one box**

☐ PreK ☐ Kindergarten ☐ 1st ☐ 2nd ☐ 3rd ☐ 4th ☐ 5th ☐ 6th

Thank you, now lets turn to page 2. We'll do some practice questions first. Read the sample passage to yourself. When you are done, we will practice answering some questions.

> On a weekend camping trip, Ken, Eric, and their dad went for a walk on the Appalachian Trail. The first hour, they walked 3/8 mile. The second hour, they walked 4/5 mile. About how many miles did the boys and their dad walk? [DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THI**S PASSAGE.]**

Do you see the word problem at the end of this passage? You do not need to answer it. Instead, let's answer the questions below it together.

1. ***What two days of the week are weekend days?***

Write your answer on the line.

Researcher gives students a few seconds to complete

The two days of the week are *Saturday* and *Sunday*. You should have written *Saturday* and *Sunday* on the line.

Now look at question 2 as I read it.

2. How many hours did the family walk on the camping trip?

You have four answers to choose from.

a) They walked for at least 2 hours.
b) They walked for an hour.
c) They walked just under 2 hours.
d) They walked for more than 2 hours.

Now circle the answer you pick.

Researcher gives students a few seconds to complete

The correct answer is A, you should have circled a) *they walked for at least 2 hours*. We know from the passage that Ken, Eric and their dad walked for two hours. So "*at least*" meaning "*not less than*" two hours is the correct choice. They may have walked longer, but we don't know that from the passage.

Are there any questions?

Now you may begin. Thank you again for your help.

Appendix D
Protocol of Directions for Verbal Protocol Administration

Think Aloud

Introductions (2-3 minutes):
Before starting the exercise, interviewer takes a few minutes to introduce him/herself and chat with the student, making him/her feel more comfortable. This might also help the interviewer gain a sense of the students English Language Proficiency (ELP).

We are studying what students do when they take tests. You can help us in our research by answering some questions. You aren't expected to know everything, but please try your best to read the passages and answer the questions. Your answers will help us to create fairer tests for students like you across California. These questions will not affect your grades in school and should take you about thirty minutes to complete.

Do you understand the directions?

Are there any questions?

Now I'd like for you to turn to page 1. We would like for you to answer a few questions about yourself.

Researcher will judge whether he/she needs help to complete this section accurately. Researcher will either: a) read the following questions out loud for students (leaving then time to write down their answer) or b) allow them to complete the questions on their own depending upon student's English Language proficiency.

1. Write Your Name: _____ (First, Last)
2. Your School: _____
3. Your Grade: _____
4. Your Teacher's Name: _____
5. Today's Date: _____
6. Do you speak a language other than English? **Check one box** ☐ Yes **OR** ☐ No
7. If yes, what language(s) do you speak? _____
8. What language do you speak most of the time at home? _____
9. What country were you born in? _____
10. Did you start school in that country? _____

11. When did you start school in the United States? **Check one box**

☐ PreK ☐ Kindergarten ☐ 1st ☐ 2nd ☐ 3rd ☐ 4th ☐ 5th ☐ 6th

Informal assessment of test taking strategies (5-7 minutes):

Thank you. Are you ready to begin? I'm going to turn on the audiotape recorder now. That way I won't have to write down or try to remember everything you say.

Think of the last time you took an end of the year test (you know, where you read questions in a booklet and answer by filling in little bubbles on a separate sheet?). Think of the ways or strategies you used to complete the test, like rereading a passage, looking for key words, eliminating answers that don't fit, [translating a word or phrase into your first language]

- **What do you usually do when you don't know how to answer a question on an end of the year test?**
- **What do you usually do when you come across a word or sentence you don't know on a test? … when you're reading?**
- **Have your teachers taught you any test taking strategies? Tell me about them.**

When people make something new, they need to talk a lot about it. We would like to hear what you think about the language in this booklet. We'd like for you to think out loud while working on the pages highlighted in blue. What we mean by this is that we would like you to read the passages out loud. Highlight any words or phrases that you don't know. Then, say out loud everything that you are thinking as you answer the questions.

It's OK if you haven't done this before, we'll try a few sample questions before you begin.

Do you understand the directions? Do you have any questions?

Choose a highlighter.

Student selects a highlighter

Practice Think Aloud (3-5 minutes)

OK, lets turn to page 2 and practice thinking out loud. Let's start with sample question #1. Read the passage out loud, highlighting any words or phrases you don't understand. Then tell me everything you're thinking as you answer the questions.

SAMPLE PASSAGE 1 (to be read by student)

On a weekend camping trip, Ken, Eric, and their dad went for a walk on the Appalachian Trail. The first hour, they walked 3/8 mile. The second hour, they walked 4/5 mile. About how many miles did the boys and their dad walk?

1. What two days of the week are weekend days?

2. How many hours did the family walk on the camping trip?
   a) *They walked for at least 2 hours.*
   b) *They walked for an hour.*
   c) *They walked just under 2 hours.*
   d) *d) They walked for more than 2 hours*

[Good job.] The two days of the week are *Saturday* and *Sunday*. You have written *Saturday* and *Sunday* on the line.

The correct answer is A, *they walked for at least 2 hours*. We know from the passage that Ken, Eric and their dad walked for two hours. So "*at least*" meaning "*not less than*" two hours is the correct choice. They may have walked longer, but we don't know that from the passage.

- *If student does not think aloud as answers question, researcher will prompt:*

- **You didn't say very much when you were reading. I'd like to know what was going through your head?** *(give student chance to answer)*

- **What words (or phrases) do you think are difficult?**

- **Why do you think this word (phrase) is difficult for you?**

- *If student needs further scaffolding, interviewer can demonstrate*

Let's do this together. Let's read the passage again. You tell me when we come to a word or phrase you don't understand.

- *While student practices think aloud and during test, interviewer may also remind him/her of directions:*

- **Remember to read the passage out loud.**
- **Highlight the words and phrases you think are difficult.**
- **Talk louder (for audio recorder)**

Do you have any questions?

Test (30 minutes)

Now you're ready to begin. Once you have completed all of the questions on the blue highlighted pages, you may finish the rest on your own. When you are finished, close your booklet and let me know. I will ask you a few short questions before you return to your classroom.

Prompts

- *If student spends more than 15-20 seconds trying to answer a question or says that he/she is unable to answer the question:*

1st Prompt:

Think about some of the ways we talked about answering questions, like rereading passages, looking for key words, eliminating answers that don't fit, [translating a word or phrase into your first language]…

2nd Prompt:

- *If student is not able or does not answer the question after an additional 15-20 seconds have passed, researcher can prompt:*

Let's move on to question X.

Follow-up (5 minutes)

Now that you've finished the exercises, I'd like to know:

- **Which passage do you think was easiest to read and answer the questions?**
- **Why do you think it was easy?**
- **Which passage do you think was most difficult to read and answer the questions?**
- **Why do you think it was difficult?**
- **Did you understand the directions?**

Researcher may also ask about particular passage/question that student had difficulty on, skipped, etc.

Thank you for your time and hard work. You may go back to class.

Appendix E
Item Level Statistics for Pre-pilot Studies

Form	Area	Passage	Q	Type	N	P Value	SE Diff.'w	St.Dev.'w	Range'w	95% Conf. Interval	
A	Math	Camping	1	2	37	0.68	0.08	0.47	0-1	0.52	0.84
A	Math	Camping	2	1	36	0.78	0.07	0.42	0-1	0.64	0.92
A	Soc Stu	Col.Wom.	1	3	32	0.38	0.09	0.49	0-1	0.2	0.56
A	Soc Stu	Col.Wom.	2	9	32	0.69	0.08	0.47	0-1	0.53	0.85
A	Soc Stu	Col.Wom.	3	9	32	0.69	0.08	0.47	0-1	0.53	0.85
A	Soc Stu	Col.Wom.	4	9	32	0.72	0.08	0.46	0-1	0.56	0.88
A	Soc Stu	Col.Wom.	5	9	32	0.72	0.08	0.46	0-1	0.56	0.88
A	Soc Stu	Col.Wom.	6	3	32	0.78	0.07	0.42	0-1	0.64	0.92
A	Soc Stu	Col.Wom.	7	1	31	0.32	0.08	0.46	0-1	0.16	0.48
A	Soc Stu	Col.Wom.	8	9	30	0.97	0.03	0.18	0-1	0.91	1.03
A	Soc Stu	Col.Wom.	9	9	30	0.87	0.06	0.35	0-1	0.75	0.99
A	Soc Stu	Col.Wom.	10	9	30	0.73	0.08	0.45	0-1	0.57	0.89
A	Soc Stu	Col.Wom.	11	1	30	0.6	0.09	0.5	0-1	0.42	0.78
B	Soc Stu	Colonial Towns	1	2	27	0.63	0.09	0.49	0-1	0.45	0.81
B	Soc Stu	Colonial Towns	2	10	27	0.52	0.1	0.51	0-1	0.32	0.72
B	Soc Stu	Colonial Towns	3	10	27	0.59	0.1	0.5	0-1	0.39	0.79
B	Soc Stu	Colonial Towns	4	10	27	0.63	0.09	0.49	0-1	0.45	0.81
B	Soc Stu	Colonial Towns	5	10	27	0.56	0.1	0.51	0-1	0.36	0.76
B	Soc Stu	Colonial Towns	6	10	26	0.65	0.1	0.49	0-1	0.45	0.85
B	Soc Stu	Colonial Towns	7	6	25	0.52	0.1	0.51	0-1	0.32	0.72
B	Soc Stu	Colonial Towns	8	9	25	0.88	0.07	0.33	0-1	0.74	1.02
B	Soc Stu	Colonial Towns	9	9	25	0.84	0.07	0.37	0-1	0.7	0.98
B	Soc Stu	Colonial Towns	10	9	25	0.88	0.07	0.33	0-1	0.74	1.02
B	Soc Stu	Colonial Towns	11	9	24	0.75	0.09	0.44	0-1	0.57	0.93
B	Soc Stu	Colonial Towns	12	9	24	0.79	0.08	0.41	0-1	0.63	0.95
B	Soc Stu	Colonial Towns	13	9	24	0.79	0.08	0.41	0-1	0.63	0.95
B	Soc Stu	Colonial Towns	14	9	24	0.75	0.09	0.44	0-1	0.57	0.93
B	Soc Stu	Colonial Towns	15	1	24	0.21	0.08	0.41	0-1	0.05	0.37
A	Science	Digestive	1	5	15	0.6	0.13	0.51	0-1	0.34	0.86
A	Science	Digestive	2	5	15	1	0	0	1-1	1	1
A	Science	Digestive	3	5	15	0.47	0.13	0.52	0-1	0.21	0.73
A	Science	Digestive	4	5	15	0.73	0.12	0.46	0-1	0.49	0.97
A	Science	Digestive	5	8	13	0.46	0.11	0.38	0-1	0.24	0.68
A	Science	Digestive	6	1	12	0	0	0	0-0	0	0
A	Science	Earth & Moon	1	2	19	0.63	0.11	0.5	0-1	0.41	0.85
A	Science	Earth & Moon	2	2	19	0.9	0.07	0.32	0-1	0.76	1.04
A	Science	Earth & Moon	3	1	18	0.78	0.1	0.43	0-1	0.58	0.98
A	Science	Earth & Moon	4	1	18	0.72	0.11	0.46	0-1	0.5	0.94
A	Science	Earth & Moon	5	1	18	0.78	0.1	0.43	0-1	0.58	0.98
A	Science	Earth & Moon	6	1	18	0.67	0.11	0.49	0-1	0.45	0.89
A	Science	Earth & Moon	7	1	18	0.61	0.12	0.5	0-1	0.37	0.85
A	Science	Earth & Moon	8	1	18	0.5	0.12	0.51	0-1	0.26	0.74
A	Science	Earth & Moon	9	10	18	0.94	0.06	0.24	0-1	0.82	1.06
A	Science	Earth & Moon	10	10	18	1	0	0	1-1	1	1
A	Science	Earth & Moon	11	10	18	1	0	0	1-1	1	1
A	Science	Earth & Moon	12	10	17	1	0	0	1-1	1	1
A	Science	Earth & Moon	13	10	17	1	0	0	1-1	1	1
O	Soc Stu	George	1	8	65	0.79	0.03	0.27	0-1	0.73	0.85

O	Soc Stu	George	2	1	63	0.59	0.06	0.5	0-1	0.47	0.71
O	Soc Stu	George	3	3	63	0.75	0.06	0.44	0-1	0.63	0.87
O	Soc Stu	George	4	2	57	0.68	0.06	0.47	0-1	0.56	0.8
O	Soc Stu	George	5	4	55	0.85	0.05	0.36	0-1	0.75	0.95
O	Soc Stu	George	6	4	51	0.18	0.05	0.39	0-1	0.08	0.28
O	Soc Stu	George	7	4	48	0.4	0.07	0.49	0-1	0.26	0.54
A	Math	Hot Dog	1	1	30	0.77	0.08	0.43	0-1	0.61	0.93
A	Math	Hot Dog	2	10	28	0.71	0.09	0.46	0-1	0.53	0.89
A	Math	Hot Dog	3	10	28	0.75	0.08	0.44	0-1	0.59	0.91
A	Math	Hot Dog	4	10	28	0.61	0.09	0.5	0-1	0.43	0.79
A	Math	Hot Dog	5	10	28	0.79	0.08	0.42	0-1	0.63	0.95
A	Math	Hot Dog	6	1	27	0.17	0.05	0.24	0-.5	0.07	0.27
A	Math	Hot Dog	7	2	27	0.19	0.08	0.4	0-1	0.03	0.35
O	Math	Lemonade	1	1	78	0.97	0.02	0.16	0-1	0.93	1.01
O	Math	Lemonade	2	1	78	0.97	0.02	0.16	0-1	0.93	1.01
O	Math	Lemonade	3	2	77	0.68	0.06	0.47	0-1	0.56	0.8
O	Math	Lemonade	4	1	77	0.88	0.04	0.32	0-1	0.8	0.96
B	Math	Org. Books	1	10	34	0.88	0.06	0.33	0-1	0.76	1
B	Math	Org. Books	2	10	33	0.27	0.08	0.45	0-1	0.11	0.43
B	Math	Org. Books	3	10	33	0.88	0.06	0.33	0-1	0.76	1
B	Math	Org. Books	4	8	32	0.78	0.07	0.42	0-1	0.64	0.92
B	Math	Org. Books	5	6	32	0.97	0.03	0.18	0-1	0.91	1.03
O	Math	Stiltwalker	1	10	74	0.58	0.03	0.25	0-1	0.52	0.64
O	Math	Stiltwalker	2	10	74	0.91	0.03	0.29	0-1	0.85	0.97
O	Math	Stiltwalker	3	10	74	0.95	0.03	0.23	0-1	0.89	1.01
O	Math	Stiltwalker	4	10	72	0.9	0.04	0.3	0-1	0.82	0.98
O	Math	Stiltwalker	5	1	72	0.65	0.06	0.48	0-1	0.53	0.77
B	Math	Traffic Light	1	3	38	0.42	0.08	0.5	0-1	0.26	0.58
B	Math	Traffic Light	2	2	38	0.71	0.07	0.46	0-1	0.57	0.85
B	Science	Wat.Supply	1	7	28	0.82	0.07	0.39	0-1	0.68	0.96
B	Science	Wat.Supply	2	7	28	0.96	0.04	0.19	0-1	0.88	1.04
B	Science	Wat.Supply	3	7	28	0.93	0.05	0.26	0-1	0.83	1.03
B	Science	Wat.Supply	4	7	28	0.82	0.07	0.39	0-1	0.68	0.96
B	Science	Wat.Supply	5	7	28	0.89	0.06	0.31	0-1	0.77	1.01
B	Science	Wat.Supply	6	7	28	0.86	0.07	0.36	0-1	0.72	1
B	Science	Wat.Supply	7	7	28	0.79	0.08	0.42	0-1	0.63	0.95
B	Science	Wat.Supply	8	7	28	0.79	0.08	0.42	0-1	0.63	0.95
B	Science	Wat.Supply	9	2	28	0.68	0.09	0.48	0-1	0.5	0.86
B	Science	Wat.Supply	10	4	28	0.82	0.07	0.39	0-1	0.68	0.96
B	Science	Wat.Supply	11	4	28	0.75	0.08	0.44	0-1	0.59	0.91
B	Science	Wat.Supply	12	4	28	0.36	0.09	0.49	0-1	0.18	0.54
B	Science	Wat.Supply	13	4	28	0.75	0.08	0.44	0-1	0.59	0.91
O	Science	Water Cycle	1	1	72	0.29	0.05	0.46	0-1	0.19	0.39
O	Science	Water Cycle	2	1	71	0.69	0.06	0.47	0-1	0.57	0.81
O	Science	Water Cycle	3	1	71	0.79	0.05	0.41	0-1	0.69	0.89
O	Science	Water Cycle	4	6	70	0.5	0.06	0.5	0-1	0.38	0.62
O	Science	Water Cycle	5	4	70	0.63	0.06	0.49	0-1	0.51	0.75
O	Science	Water Cycle	6	4	70	0.91	0.03	0.28	0-1	0.85	0.97
O	Science	Water Cycle	7	4	68	0.49	0.06	0.5	0-1	0.37	0.61
O	Science	WaterDiagram	1	10	82	0.93	0.03	0.26	0-1	0.87	0.99
O	Science	WaterDiagram	2	10	81	0.6	0.06	0.49	0-1	0.48	0.72
O	Science	WaterDiagram	3	10	81	0.75	0.05	0.43	0-1	0.65	0.85
O	Science	WaterDiagram	4	10	81	0.69	0.05	0.46	0-1	0.59	0.79

Appendix F
Item Discrimination Statistics for Selective 42 Tasks in Pre-pilot Stage

Form	Area	Passage	No.	Type	N for NonMast.	P Value for NM	N for Master	P Value for Master	Discrimination
A	Math	Camping	1	2	11	0.636	16	0.813	0.177
A	Math	Camping	2	1	11	0.727	16	0.813	0.086
A	Soc Stu	Col.Wom.	1	3	10	0.3	14	0.429	0.129
A	Soc Stu	Col.Wom.	2	9	10	0.4	14	0.857	0.457
A	Soc Stu	Col.Wom.	3	9	10	0.4	14	0.857	0.457
A	Soc Stu	Col.Wom.	4	9	10	0.7	14	0.857	0.157
A	Soc Stu	Col.Wom.	5	9	10	0.7	14	0.857	0.157
A	Soc Stu	Col.Wom.	6	3	10	0.7	14	0.857	0.157
A	Soc Stu	Col.Wom.	7	1	9	0.222	14	0.357	0.135
A	Soc Stu	Col.Wom.	8	9	9	0.889	14	1	0.111
A	Soc Stu	Col.Wom.	9	9	9	0.778	14	0.929	0.151
A	Soc Stu	Col.Wom.	10	9	9	0.667	14	0.714	0.047
A	Soc Stu	Col.Wom.	11	1	9	0.556	14	0.643	0.087
B	Math	Traffic Light	1	3	10	0.1	16	0.625	0.525
B	Math	Traffic Light	2	2	10	0.7	16	0.75	0.05
O	Math	Lemonade	1	1	20	0.9	35	1	0.1
O	Math	Lemonade	2	1	20	0.9	35	1	0.1
O	Math	Lemonade	3	2	20	0.55	34	0.882	0.332
O	Math	Lemonade	4	1	20	0.8	34	0.971	0.171
O	Math	Stilt	1	10	21	0.524	33	0.621	0.097
O	Math	Stilt	2	10	21	0.762	33	0.97	0.208
O	Math	Stilt	3	10	21	0.81	33	1	0.19
O	Math	Stilt	4	10	19	0.737	33	1	0.263
O	Math	Stilt	5	1	19	0.526	33	0.727	0.201
O	Science	Water Cycle	1	1	19	0.158	31	0.419	0.261
O	Science	Water Cycle	2	1	19	0.526	31	0.871	0.345
O	Science	Water Cycle	3	1	19	0.631	31	0.903	0.272
O	Science	Water Cycle	4	6	19	0.211	31	0.613	0.402
O	Science	Water Cycle	5	4	19	0.421	31	0.871	0.45
O	Science	Water Cycle	6	4	19	0.79	31	0.968	0.178
O	Science	Water Cycle	7	4	17	0.353	31	0.452	0.099
O	Science	WaterDiagram	1	10	23	0.783	34	0.971	0.188
O	Science	WaterDiagram	2	10	22	0.273	34	0.824	0.551
O	Science	WaterDiagram	3	10	22	0.5	34	0.882	0.382
O	Science	WaterDiagram	4	10	22	0.5	34	0.765	0.265
O	Soc Stu	George	1	8	16	0.675	31	0.886	0.211
O	Soc Stu	George	2	1	16	0.438	30	0.667	0.229
O	Soc Stu	George	3	3	16	0.438	30	0.8	0.362
O	Soc Stu	George	4	2	14	0.571	29	0.759	0.188
O	Soc Stu	George	5	4	13	0.692	28	0.964	0.272
O	Soc Stu	George	6	4	11	0.091	27	0.185	0.094
O	Soc Stu	George	7	4	10	0.1	27	0.556	0.456

Appendix G
List of Identified Vocabulary from Verbal Protocol Analysis

Passage	Total Number of VP Students	Vocabulary (token)
Camping Trip	18	Appalachian (11)
Lemonade	18	Carlotta (4), profit (3)
Water cycle	18	atmosphere/atmospheric(5), cirrus(10), condensation/condense (9), condition(1), cumulus (11), evaporation (4), factor(3), humidity (4), meteorologists (7), precipitation (8), predict (1), stratus (9)
George Washington	17	Captain (1), Colonel (3), colonist/colonial (5), Custis (1), Continental (4), colonies (1), discouraged (1), Hessians (4), landowners (2), mathematics (1), militia (10), patriots (1), Potomac (2), plantation (5), protest, rank, restored (1), surveyor (10), temperament (2), Vernon (1), Virginia House of Burgesses (13), Westmoreland (3), wealthiest (1)
Stilt-walker	15	Bowen Moscow (3), mi (2)
Water Diagram	15	Condensation (6), entire (1), evaporation (3), precipitation (5)

Appendix H
Linguistic Analysis Framework

	Stem/Prompt	*Response*
Construct		
Descriptive Analysis		
Mean no. of words per sentence(range)		
Sum of Words		
Total # of words (token)		
Total # of words (type)		
Lexical Features		
Academic vocabulary - general (token)		
Academic vocabulary - general (type)		
Academic vocabulary - specialized(token)		
Academic vocabulary - specialized(type)		
Low-frequency words (token)		
Low-frequency words (type)		
3-or-more-syllable words(token)		
3-or-more-syllable words(type)		
Derived words (token)		
Derived words (type)		
No. of unique clause connectors		
Avg. % of nominalizations per selection		
Sentence Type		
Simple sentences		
Complex sentences		
Other sentence types		
Grammatical Features		
Noun phrases		
Participial modifiers		
Passive voice verb forms		
Prepositional phrases		
Dependent clauses		
Organizational Features		
Classification		
Description		
Explanation		
Scenario		
Sequencing		
Comparison		
Definition		
Enumeration		
Exemplification		
Labeling		
Paraphrase		

Provide instruction or guidance		
Quotation		
Reference to text or visual		
Question		
Summary		