Performance Assessment: The State of the Art

Suzanne Lane
University of Pittsburgh

scope
Stanford Center for
Opportunity Policy in Education

This study was conducted by the Stanford Center for Opportunity Policy in Education (SCOPE) with support from the Ford Foundation and the
Nellie Mae Education Foundation.

The Stanford Center for Opportunity Policy in Education (SCOPE) supports cross-disciplinary research, policy analysis, and practice that address issues of educational opportunity, access, equity, and diversity in the United States and internationally.

Citation: Lane, S. (2010). *Performance assessment: The state of the art.* (SCOPE Student Performance Assessment Series). Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stanford Center for Opportunity Policy in Education
Barnum Center, 505 Lasuen Mall
Stanford, California 94305
Phone: 650.725.8600
scope@stanford.edu
http://edpolicy.stanford.edu

scope
Stanford Center for
Opportunity Policy in Education

Abstract

Performance assessments have been an integral part of educational systems in numerous countries however they have not been fully integrated in assessment systems in this country. Research has shown that the format of the assessment affects the type of thinking and reasoning skills that are used by students, with performance assessments being better suited to assessing high level, complex thinking skills. Recent advances in the design and scoring of performance assessments, including computer-based task simulations and automated scoring systems, support their increased use in large-scale assessment programs. There are also promising technical advances that support their use. The educational benefit of using performance assessments has been demonstrated by a number of researchers. When students are given the opportunity to work on meaningful, real world tasks in instruction, students have demonstrated improved performance on performance assessments. Sound educational practice calls for the alignment among curriculum, instruction and assessment, and there is ample evidence to support the use of performance assessments in both instruction and assessment to improve student learning for all students.

Table of Contents

Preface and Acknowledgements

This paper was supported by the project, Investigating Teaching and Learning Systems Around the World, funded by the Ford Foundation. The opinions expressed in this paper do not necessarily reflect the position of the Ford Foundation. This paper extends the work presented in two documents:

1) Lane & Stone, C.A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational Measurement*. American Council on Education & Praeger: Westport, CT. and
2) Lane, S. (in press). Issues in the design and scoring of performance assessments that sssess complex thinking skills. In G. Schraw (Ed.). *Assessment of Higher Order Thinking Skills*.

This paper is one of eight written through a Stanford University project aimed at summarizing research and lessons learned regarding the development, implementation, consequences, and costs of performance assessments. The project was led by Linda Darling-Hammond, Charles E. Ducommun Professor of Education at Stanford University, with assistance from Frank Adamson and Susan Shultz at Stanford. It was funded by the Ford Foundation and the Nellie Mae Education Foundation and guided by an advisory board of education researchers, practitioners, and policy analysts, ably chaired by Richard Shavelson, one of the nation's leading experts on performance assessment. The board shaped the specifications for commissioned papers and reviewed these papers upon their completion. Members of the advisory board include:

Eva Baker, Professor, UCLA, and Director of the Center for Research on Evaluation, Standards, and Student Testing

Christopher Cross, Chairman, Cross & Joftus, LLC

Nicholas Donahue, President and CEO, Nellie Mae Education Foundation, and former State Superintendent, New Hampshire

Michael Feuer, Executive Director, Division of Behavioral and Social Sciences and Education in the National Research Council (NRC) of the National Academies

Edward Haertel, Jacks Family Professor of Education, Stanford University

Jack Jennings, President and CEO, Center on Education Policy

Peter McWalters, Strategic Initiative Director, Education Workforce, Council of Chief States School Officers (CCSSO) and former State Superintendent, Rhode Island

Richard Shavelson, Margaret Jacks Professor of Education and Psychology, Stanford University

Lorrie Shepard, Dean, School of Education, University of Colorado at Boulder

Guillermo Solano-Flores, Professor of Education, University of Colorado at Boulder

Brenda Welburn, Executive Director, National Association of State Boards of Education

Gene Wilhoit, Executive Director, Council of Chief States School Officers

The papers listed below examine experiences with and lessons from large-scale performance assessment in the United States and abroad, including technical advances, feasibility issues, policy implications, uses with English language learners, and costs.

~ Jamal Abedi, *Performance Assessments for English Language Learners*.

~ Linda Darling-Hammond, with Laura Wentworth, *Benchmarking Learning Systems: Student Performance Assessment in International Context*.

~ Suzanne Lane, *Performance Assessment: The State of the Art*.

~ Raymond Pecheone and Stuart Kahl, *Developing Performance Assessments: Lessons from the United States*.

~ Lawrence Picus, Will Montague, Frank Adamson, and Maggie Owens, *A New Conceptual Framework for Analyzing the Costs of Performance Assessment*.

~ Brian Stecher, *Performance Assessment in an Era of Standards-Based Educational Accountability.*

~ Barry Topol, John Olson, and Edward Roeber, *The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments*.

An overview of all these papers has also been written and is available in electronic and print format:

~ Linda Darling-Hammond and Frank Adamson, *Beyond Basic skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning*.

All reports can be downloaded from http://edpolicy.stanford.edu.

We are grateful to the funders, the Advisory Board, and these authors for their careful analyses and wisdom. These papers were ably ushered into production by Barbara McKenna. Without their efforts, this project would not have come to come to fruition.

Introduction

Educational reform in the 1980s was based on the premise that too many students knew how to repeat facts and concepts, but were unable to apply those facts and concepts to solve realistic problems that require complex thinking and reasoning skills. Assessments need to better reflect students' competencies in applying their knowledge and cognitive skills to solve substantive, meaningful tasks. Promising advances in the study of both cognition and learning in content domains and of psychometrics also prompted individuals to think differently about how students process and reason with information and how assessments can be designed to capture meaningful aspects of student learning. Performance assessments that assess complex cognitive skills were also considered to be valuable tools for educational reform by policy makers and advocates for curriculum reform (Linn, 1993; Resnick & Resnick, 1982). They were thought of as vehicles that could help shape sound instructional practice by modeling to teachers what is important to teach and to students what is important to learn. Carefully crafted performance assessments that measure complex thinking and reasoning skills can serve as exemplars of assessments that stimulate and enrich learning rather than just serve as indicators of learning (Bennett & Gitomer, in press; Black & William, 1998). Performance assessments are needed to assess the types of thinking and reasoning skills that are valued by educators, and cannot be assessed by other item formats such as multiple-choice items.

The use of performance tasks in large-scale assessments has declined with the requirements of the No Child Left Behind (NCLB) Act of 2001 (U.S. Department of Education, 2005). Under the NCLB Act, states are required to test all students from grades 3 through 8 annually in reading and mathematics, and students in high school at one grade level. Students also need to be tested in science at one grade level in elementary, middle, and high school. An example of a successful performance-assessment program prior to NCLB was the Maryland School Performance Assessment Program (MSPAP) that was designed for grades 3, 5, and 8 to measure school-level performance and provides information for school accountability and improvement (Maryland State Board of Education, 1995). These assessments were designed to promote performance-based instruction and classroom assessment in reading, writing, mathematics, science, and social studies. The performance assessment tasks were interdisciplinary, required students to produce both short and extended written responses, and some required hands-on activities and collaboration with peers. Maryland's performance-based assessment was no longer tenable given the constraints imposed by the NCLB Act.

This chapter addresses design, scoring, and psychometric advances in performance assessment that allow for the assessment of 2first-century skills. Although performance assessments for classroom purposes are discussed, the focus of this chapter is on the use of performance assessments in large-scale assessment programs. Advances in the integration of cognitive theories of learning and measurement models as they apply to the design of performance assessments are considered. The chapter begins with a discussion on the advances in the design of performance assessments, including a description of the important learning outcomes that can be assessed by performance assessments, and not by other assessment formats. The second section discusses advances in the scoring of performance assessments, including both the technical and substantive advances in automated scoring methods that allow for timely scoring of student performances to innovative item types. The third section addresses issues related to the validity and fairness of the use and interpretation of scores derived from performance assessments. The type of evidence needed to support the validity of score interpretations and use, such as content representation, cognitive complexity, fairness, generalizability, and consequential evidence, is discussed. It should be noted, however, that validity and fairness are addressed throughout the chapter. The last section briefly addresses additional psychometric advances in performance assessments, including advances in measurement models used to capture student performance and rater inconsistencies as well as advances in linking performance assessments.

Design of Performance Assessments

In the design of any assessment, the type of score inferences one wants to make should first be delineated. This includes deciding on whether one wants to generalize to the larger construct domain of interest, or to provide evidence of a particular accomplishment or performance. The former requires sampling tasks from the domain to ensure content representativeness which will contribute to the validity of the score generalizations. This approach is typically used in the design of large-scale assessments, and the challenges of ensuring valid generalizations to the content domain using scores derived from performance assessments will be addressed later in this chapter. The latter approach requires the specification of a performance assessment that allows for the demonstration of a broader ability or performance which is similar to a "merit badge" approach. This approach, performance demonstration, is commonly used for classroom purposes such as a high-school project or paper.

This section focuses on design issues that need to be considered to ensure that performance assessments are capable of eliciting the cognitive processes and skills that they are intended to measure, and to ensure the coherency among curriculum, instruction, and assessment. Advances in the design of computer-based simulation tasks are addressed. The use of computers allows for the modeling of performance tasks that engage students in meaningful problem solving and reasoning skills and for the monitoring and scoring of student performances. Computer simulation tasks are ideal for capturing the multidimensionality of content domains, and scores can be generated for the different dimensions being assessed. Advances in the use of learning progressions in assessment design are also addressed. The use of learning progressions in the design of assessments is invaluable for monitoring an individual student's or group's progress, and for informing instruction and learning. The importance of expert review and field-testing performance assessments is also discussed. Throughout this section, examples of performance assessments are also provided, some of which have been used in large-scale assessment programs.

Description of Performance Assessment

Performance assessments can measure students' cognitive thinking and reasoning skills and their ability to apply knowledge to solve realistic, meaningful problems. They are designed to more closely reflect the performance of interest, allow students to construct or perform an original response, and use predetermined criteria to evaluate student work. The close similarity between the performance that is assessed and the performance of interest is the defining characteristic of a performance assessment as described by Kane, Crooks, and Cohen (1999). As stated by the *Standards for Educational and Psychological Testing*, performance assessments attempt to “emulate the context or conditions in which the intended knowledge or skills are actually applied” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p.137). As this definition indicates, performance assessments do not have to assess complex reasoning and problem-solving skills. As an example, if the targeted domain is the speed and accuracy at which students can keyboard, a measure that captures the accuracy and speed of students’ keyboarding would be considered a performance assessment. Clearly, keyboarding is not a high-level thinking skill but a learned, automated procedural skill. The focus of this chapter will be on performance assessments that are designed to assess complex reasoning and problem-solving skills in academic disciplines, and can be used for large-scale assessments. The Maryland School Performance Program (MSPAP) was an excellent example of a performance assessment that consisted of interdisciplinary tasks that assessed problem-solving, reasoning, and evaluation skills (Maryland State Board of Education, 1995). As an example for a grade 5 Science MSPAP task, students were asked to investigate how a hydrometer can be used to measure different levels of saltiness (salinity) in various water samples, predict how the hydrometer might float in mixtures of fresh and salt water, and determine how the hydrometer could be used to establish the correct salinity for an aquarium. This hands-on task allowed students to conduct several investigations, make predictions, evaluate their work, and provide explanations for their responses.

Performance assessments require students to perform a task such as conducting a science investigation as described above, or to construct an original product or response such as writing an explanation of one’s solution to a mathematics problem or writing a persuasive essay. Proficiency can be explained by the cognitive processes and skills involved in solving the performance task as well as the strategies chosen for a solution, having the potential to provide rich information for diagnosing strengths as well as gaps in understanding for individual students as well as groups of students. When working on well-designed performance tasks, students may be engaged in applying their knowledge to real world problems, evaluating different approaches to solving problems, and providing reasoning for their solutions. A prevailing assumption underlying performance assessments is that they serve as motivators in improving student achievement and learning, and that they encourage instructional strategies that foster reasoning, problem-solving, and communication (Frederiksen

& Collins, 1989; National Council on Education Standards and Testing, 1992; Resnick & Resnick, 1982). Further, Mislevy (1996) pointed out that they allow for better measurements of change, both quantitatively and qualitatively. A hypothetical example of a quantitative measure of change for a mathematics performance assessment would be that a student used a recursive strategy only 2 out of 10 times during the first administration of a math assessment, but 6 out of 10 times during the second administration. An example of a qualitative evaluation of change would be that the student switched from a less effective strategy to the more sophisticated recursive strategy after instruction.

Performance assessments are contextualized, linking school activities to real world experiences (Darling-Hammond, Ancess, & Falk, 1995), and can include opportunities for self-reflection and collaboration as well as student choice, such as choosing a particular topic for a writing assignment (Baker, O'Neil, & Linn, 1993; Baron, 1991). Collaborative efforts were required on the Maryland State Performance Assessment Program (MSPAP, Maryland State Board of Education, 1995) in that students worked together on conducting science investigations and evaluated each other's essays. Collaboration is required on many performance assessments outside of this country, such as a mathematics assessment in Denmark aimed at 16-year-olds (Black & William, 2007). It can be argued that these types of collaborations on performance assessments better reflect skills required in the 21st century. Performance assessments may also allow for a particular task to yield multiple scores in different content domains, which has practical as well as pedagogical appeal. Tasks that are designed to elicit scores in more than one content domain may not only reflect a more integrated approach in instruction, but also motivate a more integrated approach to learning. As an example, MSPAP tasks were integrated and would yield scores in two or more domains (Maryland State Board of Education, 1995). Practical implications for tasks that afford multiple scores may be reduced time and cost for task development, test administration, and scoring by raters (Goldberg & Roswell, 2001). It is important, however, to provide evidence that each score represents the construct it is designed to measure and does not include construct-irrelevant variance (Messick, 1989).

Performance assessments, in particular, writing assessments, have been included in some large-scale assessment programs in this country for monitoring students' progress towards meeting national or state content standards, promoting educational reform, and holding schools accountable for student learning. High-stakes decisions are typically required as well as an evaluation of changes in performance over time, which requires a level of standardization of the content to be assessed, of the administration of the assessment, and of the scoring of student performances over time. Thus, extended time periods, collaborative work, choice of task, and use of ancillary material may challenge the standardization of the assessment and, consequently, the accuracy of the score interpretations. Large-scale performance assessment programs, such as MSPAP, however, have included these attractive features of performance assessments while ensuring the quality and validity of the score interpretations at the school level.

Another consideration in the design of assessments of complex skills is whether a portfolio approach will be used. The Advanced Placement (AP) Studio Art portfolios provide an excellent example of a large-scale portfolio assessment that has been sustained over time (Myford & Mislevy, 1995). As an example, in the 3-D Design portfolio, students are required to submit a specified series of images of their 3-D artworks and their artworks are evaluated independently according to their quality (demonstration of form, technique, and content), breadth (demonstration of visual principles and material techniques), and concentration (demonstration of depth of investigation and process of discovery). Using a well-delineated scoring rubric for each of these three areas, from three to seven artist-educators evaluate the submitted images of the artwork. The portfolios that are submitted are standardized in that specific instructions are provided to students that specify what type of artwork is appropriate and the students are provided with detailed scoring rubrics that delineate what is expected for each of the dimensions being assessed.

Performance assessments that are aligned with curriculum and instruction can provide valuable information to guide the instructional process. Thus, it is imperative to ensure that both classroom and large-scale assessments are aligned to the curriculum and instruction. A rich, contextualized curriculum-embedded performance assessment that is used for classroom purposes does not require the level of standardization as the typical large-scale performance assessment. These curriculum-embedded assessments allow teachers to more fully understand the ways in which students understand the subject matter, and can help guide day-to-day instruction. Large-scale assessments can also inform instruction, but at a broader level for both individual students and groups of students (e.g., classrooms).

Cognitive Theories in the Design of Performance Assessments

The need for models of cognition and learning and quantitative psychometric models to be used together to develop and interpret achievement measures has been widely recognized (Embretson, 1985; Glaser, Lesgold & Lajoie, 1987; National Research Council, 2001). The deeper the understanding of how individuals acquire and structure knowledge and cognitive skills, and how they perform cognitive tasks, the better able we are to assess students' cognitive thinking and reasoning and obtain information that will lead to improved learning. Substantial theories of knowledge acquisition are needed in order to design assessments that can be used in meaningful ways to guide instruction and monitor student learning. Several early research programs that have had a direct impact on the assessment of achievement studied the difference between experts' and novices' knowledge structures (e.g., Simon & Chase, 1973; Chi, Feltovich, & Glaser, 1981). Chi and her colleagues (1981) demonstrated that an expert's knowledge in physics is organized around central principals of physics, whereas a novice's knowledge is organized around the surface features represented in the problem description. It is important to point out that much of what is known in the development of expertise is based on studies of students' acquisition of knowledge and skills in content domains.

Other early approaches that link cognitive models of learning and psychometrics have drawn upon work in the area of artificial intelligence (e.g., Brown & Burton, 1978). As an example, Brown and Burton (1978) represented the complex procedures underlying addition and subtraction as a set of component procedural skills, and described proficiency in terms of these procedural skills. Using artificial intelligence, they were able to uncover procedural errors or bugs in students' performance that represented misconceptions in their understanding. Cognitive task analysis using experts' talk alouds (Ericcson & Smith, 1991) has also been used to design performance assessments in the medical domain (Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002). Features of the expert's thinking, knowledge, procedures, and problem- posing are considered to be indicators of developing expertise in the domain (Glaser, Lesgold & Lajoie, 1987), and can be used systematically in the design of assessment tasks. These features can then in turn be used in the design of the scoring rubrics by embedding them in the criteria at each score level. Experts need not be professionals in the field. Instead, in the design of K-12 assessments, experts are typically considered students who have attained competency within the content domain.

While there is the recognition that theories of cognition and learning should serve as the foundation for the design and interpretation of assessments, widespread use of cognitive models of learning in assessment design has not been realized. As summarized by Bennett and Gitomer (in press), there are three primary reasons for this: 1) the disciplines of psychometrics and of cognition and learning that have developed separately are just beginning to merge, 2) theories of the nature of proficiency and learning progressions are not fully developed, and 3) there are both economic and practical constraints. There are promising assessment design efforts, however, that are taking advantage of what has been learned about the acquisition of student proficiency. A systematic approach to designing assessments that reflect theories of cognition and learning is embodied in Mislevy and his colleagues' (Mislevy, Steinberg, and Almond, 2003) evidence-centered design (ECD), in which evidence observed in student performances on complex problem-solving tasks (that have clearly articulated cognitive demands) is used to make inferences about student proficiency. Some of these design efforts will be discussed in this chapter.

Delineation of a Conceptual Framework for Design

A well-designed performance assessments begins with the delineation of the conceptual framework. The extent to which the conceptual framework considers cognitive theories of student proficiency and is closely aligned to the relevant curriculum will affect the validity of score interpretations. The delineation of the conceptual framework includes a description of the construct to be assessed, the purpose of the assessment, and the intended inferences to be drawn from the assessment results (Lane & Stone, 2006). Construct theory as a guide to the development of an assessment provides a rationale basis for specifying features of assessment tasks and scoring rubrics as well as for expecting certain empirical evidence, such as the extent of homogeneity of item responses and the relationship between scores with other measures (Messick, 1994; Mislevy, 1996; National Research Council, 2001). Two general approaches to designing performance assessments has been proposed, a construct-centered approach and a task-centered approach (Messick, 1994). Under the construct-centered approach, the construct is specified by identifying the complex set of knowledge and skills that need to be assessed and are valued in instruction. The performances or responses that should be elicited by the assessment are then identified. By using this design approach, the construct guides the development of the tasks as well as the specification of the scoring criteria and rubrics.

By focusing on the construct in assessment design, the test designer can pay attention to both construct-irrelevant variance and construct underrepresentation which may have an impact on the validity of score inferences (Messick, 1994). Construct underrepresentation occurs when the assessment does not fully capture the targeted construct, and therefore the score inferences may not be generalizable to the larger domain of interest. Construct-irrelevant variance occurs when one or more irrelevant constructs is being assessed in addition to the intended construct. For example, students' writing ability may have an unwanted impact on performance on a mathematics assessment. The section on validity and fairness of assessment later in the chapter provides a discussion on construct-irrelevant variance and construct underrepresentation. Scores derived from a construct-centered approach may be more generalizable across variations in tasks, settings, and examinee groups than scores derived from a task-centered approach because of the attention to reducing construct-irrelevant variance and to increasing the representation of the construct (Messick, 1994).

For large-scale educational assessments, the conceptual framework is typically defined by content standards delineated at the state or national level. The grain at which the content standards are specified impacts on whether narrow bits of information will be assessed or whether broader, more contextualized understanding of the content domain will be assessed. This is because the content standards guide the development of the test specifications that include the content, cognitive processes and skills, and psychometric characteristics of the tasks. Thus, the extent to which an assessment is valued will be dependent on the quality of the content standards.

Test specifications need to clearly articulate the cognitive demands of the tasks, problem-solving skills and strategies that can be employed, and criteria to judge performance. This includes the specification of knowledge and strategies that are not only linked closely to the content domain, but also those that are content-domain independent (Baker, 2007). Carefully crafted and detailed test specifications are even more important for performance assessments than multiple-choice tests because there are fewer performance tasks and typically each is designed to measure something that is relatively unique (Haertel & Linn, 1996). The use of detailed test specifications can also help ensure that the content of the assessment is comparable across years so as to allow for measuring change over time. The performance tasks and scoring rubrics are then developed iteratively based on a well-delineated conceptual framework and test specifications (Lane & Stone, 2006).

The use of conceptual frameworks in designing performance assessments leads to assessments that are linked to educational outcomes and provide meaningful information that can guide curriculum and instructional reform. As an example, a construct-centered approach was taken in the design of a mathematics performance assessment that required students to show their solution processes and explain their reasoning (Lane, 1993; Lane et al., 1995). The conceptual framework that Lane and her colleagues proposed then guided the design of the performance tasks and scoring rubrics. Cognitive theories of student mathematical proficiency provided a foundation for defining the construct domain of mathematics. Four components were specified for the assessment-and-task design and were further delineated: cognitive processes, mathematical content, mode of representation, and task context. To reflect the complex construct domain of mathematical problem solving, reasoning, and communication; for example, a range of cognitive processes were specified including discerning mathematical relations, using and discovering strategies and heuristics, formulating conjectures, and evaluating the reasonableness of answers. Performance tasks were then developed to assess one or more of these skills.

Underlying performance assessments is a continuum that represents different degrees of structure versus open-endedness in the response (Messick, 1996). The degree of structure for the problem posed and the response expected should be considered in the design of performance assessments. Baxter and Glaser (1998) characterized performance assessments along two continuums with respect to their task demands. One continuum represents the task demand for cognitive processes ranging from open to constrained, and the other continuum represents the task demand for content knowledge from rich to lean. A task is process open if it promotes opportunities for students to develop their own procedures and strategies, and a task is content rich if it requires substantial content knowledge for successful performance. These two continuums are crossed to form four quadrants so that tasks can be designed to fit one or more of these quadrants. This allows for clearly articulated cognitive and content targets in task design, and for the evaluation of tasks in terms of their alignment with these targets (Baxter & Glazer, 1998). In the design of performance assessments that assess complex cognitive thinking skills, design efforts can be aimed primarily in the quadrant that reflects tasks that are process open and content rich; however, familiarity with these types of tasks in instruction and the age of the student needs to be considered in design efforts. The two continuums (content knowledge and cognitive processes) could allow for more than four quadrants so as to examine students' progression in understanding within a content area.

Templates for Task Design

It is beneficial to develop templates for task design to ensure that the cognitive skills that are of interest are assessed. Templates can be developed for performance tasks that allow for tasks to be designed that assess the same cognitive processes and skills, and a scoring rubric can then be designed for the tasks that can be generated from a particular template. The use of templates for task design allows for an explicit delineation of the cognitive skills to be assessed, and can improve the generalizability of the score inferences. A model-based assessment approach that uses task templates has been proposed by Baker (2007). The major components of the model are the cognitive demands of the task, criteria to judge performance derived by competent performance, and a content map that describes the subject matter, including the interrelationships among concepts and the most salient features of the content. The cognitive demands of the tasks can then be represented in terms of families of tasks (or task templates) such as reasoning, problem solving, and knowledge representation tasks (Baker, 2007). As an example, the explanation task template asks students to read one or more texts that require some prior knowledge of the subject domain, including concepts, principles, and declarative knowledge, in order to understand them, and to evaluate and explain important issues introduced in the text (Niemi, Baker, & Sylvester, 2007). A task from the explanation family that was developed for assessing student proficiency in Hawaii is provided below (Niemi, Baker, & Sylvester, 2007).

> Imagine you are in a class that has been studying Hawaiian history. One of your friends, who is a new student in the class, has missed all the classes. Recently, your class began studying the Bayonet Constitution. Your friend is very interested in this topic and asks you to write an essay to explain everything that you have learned about it.
>
> Write an essay explaining the most important ideas you want your friend to understand. Include what you have already learned in class about Hawaiian history and what you have learned from the texts you have just read. While you write, think about what Thurston and Liliuokalani said about the Bayonet Constitution, and what is shown in the other materials.
>
> Your essay should be based on two major sources:
> 1. The general concepts and specific facts you know about Hawaiian history, and especially what you know about the period of Bayonet Constitution
> 2. What you have learned from the readings yesterday.

Prior to receiving this task, students were required to read the primary source documents that were referred to in the prompt. This task requires students to not only make sense of the material from multiple sources, but to integrate material from these multiple sources in their explanations. This provides just one example of a task that can be generated from the explanation task template. Task templates can also be used to design computer-based simulation tasks.

Design of Computer-Based Simulation Tasks

Computer-based simulations have made it possible to assess complex thinking skills that cannot be measured well by more traditional assessment methods. Using extended, integrated tasks, a large problem-solving space with various levels of complexity can be provided in an assessment (Vendlinski, Baker, & Niemi, 2008). Computer-based simulation tasks can assess student competency in formulating, testing, and evaluating hypotheses; selecting an appropriate solution strategy; and when necessary, adapting strategies based on the degree of success to a??? solution. An attractive feature of computer-based simulation tasks is that they can include some form of immediate feedback to the student according to the course of actions taken by the student. Other important features of computer-based simulations include the variety of the types of interactions that a student has with tools in the problem-solving space, and the monitoring and recording of how a student uses these tools (Vendlinski et al., 2008). Technology used in computer-based simulations allow assessments to provide more meaningful information by capturing students' processes and strategies, as well as their products. Information on how a student arrived at an answer or conclusion can be valuable in guiding instruction and monitoring the progression of student learning (Bennett, Persky, Weiss, & Jenkins, 2007). The use of automated scoring procedures for evaluating student performances to computer-based simulation tasks addresses the cost and time demands of human scoring.

Several issues need to be considered in the design of computer-based simulations such as the examinee's familiarity with the navigation rules and controls imposed by the computer interface and testing network requirements, the potential requirement of examinees' to record their answers in an unusual manner, and the large amount of data that needs to be summarized in a meaningful way (Bennett, et al., 2007; DeVore, 2002). Like all assessments, computer-based tasks have the potential to measure factors which are irrelevant to the construct that is intended to be assessed, and therefore the validity of the score interpretations can be hindered. It is important to ensure that the computer interface is one in which examinees are familiar with, and students have had the opportunity to practice with the computer interface and navigation system. It is also important to ensure that the range of cognitive skills and knowledge assessed are not narrowed to those that are more easily assessed using computer technology. Further, the automated scoring procedures need to reflect important features of proficiency so as to ensure that the generated scores provide accurate interpretations (Bennett, 2006; Bennett & Gitomer, in press). The use of test specifications that delineate the cognitive skills and knowledge that are intended to be assessed by the computer-based simulations will help ensure representation of the assessed content domain in both the tasks and scoring procedures so as to allow for valid score interpretations. Further, task templates can be used to ensure that the tasks and scoring rubrics embody the important cognitive demands.

The advancements of computer technology have made it possible to use performance-based simulations, which assess problem-solving and reasoning skills in large-scale, high-stakes assessment programs. The most prominent large-scale assessments that use computer-based simulations are licensure examinations in medicine, architecture, and accountancy. As an example, computer-based case simulations have been designed to measure physicians' patient-management skills, providing a dynamic interaction simulation of the patient-care environment (Clyman, Melnick, & Clauser, 1995). In this assessment, the examinee is first presented with a description of the patient and then the examinee must manage the patient case by selecting history and physical examination options or by making entries into the patient's chart to request tests, treatments, and/or consultations. The condition of the patient changes in real time based on the patient's disease and the examinee's course of actions. The computer-based system generates a report that displays each action taken by the examinee and the time that the action was ordered. The examinee performance is then scored by a computer-automated scoring system according to the appropriateness of the sequence of the ordered actions. It is apparent that this licensure examination captures some essential and relevant problem-solving, judgment, and decision- making skills that are required of physicians.

A research project that used the architecture computer-based exam demonstrated how the format of a task can affect the problem-solving and reasoning skills that are used by examinees (Martinez & Katz, 1996). Differences in the cognitive skills assessed by computer-based figural responses items as compared to multiple-choice items in the architecture exam were observed. As an example, for one figural response item, a building site, which is surrounded by icons that represent a parking lot, playground, and library, are presented on the computer screen. The examinee is asked to select a tool, such as one that rotates or moves an icon, and, through a series of mouse movements and clicks, is then asked to arrange the icons to meet particular criteria. Other figural response items require students to draw lines or arrows or attach labels to parts of a diagram. The results of their study suggest that on items that required students to use their own strategies, the skills used to solve the tasks differed, dependent on whether it was a figural response item or a multiple-choice item. For the figural response items, students devised a strategy, generated a response, and evaluated it based on the criteria, whereas on the multiple-choice items, students just examined each alternative with respect to the criteria. The cognitive demands of the item formats were clearly different, with the skills engaged by students on the figural response items being better aligned to the skills of interest than those used on the multiple-choice items.

Using evidence-centered design (Mislevy, Steinberg, and Almond, 2003), computer-simulation tasks in the physics domain were developed in the context of a NAEP research project (Bennett, Persky, Weiss, & Jenkins, 2007). The project goal was to examine the feasibility of including computer-based simulations on the NAEP science assessment. The computer-simulation tasks were designed to represent exploration features of real-world problem solving, and incorporated "what-if" tools that students used to uncover underlying scientific relationships. To assess scientific-inquiry skills, students were required to design and conduct experiments, interpret results, and formulate conclusions. As part of the simulations, students selected values for independent variables and made predictions as they designed their experiments. To interpret their results, students needed to develop tables, graphs, and formulate conclusions. In addition to these scientific-inquiry tasks, tasks were developed to assess students' search capabilities on a computer. One eighth grade inquiry, computer-based simulation task required students to investigate why scientists use helium gas balloons to explore out of space and the atmosphere (Bennett et al., 2007). An example of an item within this task that required students to search a simulated World Wide Web is provided below (Bennett et al., 2007, p. 41).

> Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from a spacecraft near other planets.
>
> Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.
>
> Base your answer on more than one web page or site. Be sure to write your answer in your own words.

This task assesses students' research skills using a computer, which is typical of what is expected in their instructional experiences. An example of a related scientific-inquiry task that required students to evaluate their work, form conclusions, and provide rationales after designing and conducting a scientific investigation is provided below (Bennett et al., 2007, p. 46).

> How do different amounts of helium affect the altitude of a helium balloon?
>
> Support your answer with what you saw when you experimented.

These simulation tasks were based on models of student cognition and learning and allowed for the assessment of problem-solving, reasoning, and evaluation skills that are valued within the scientific discipline. It should be noted that for the 2009 science NAEP, a sample of the students were administered these types of computer-based simulation tasks, requiring them to engage in the processes of scientific inquiry by working on a simulated experiment, recording data, and critiquing a hypothesis.

Computer-based simulation tasks in the reading, mathematics, and writing domains are being designed and evaluated for their potential inclusion in an integrated accountability and formative assessment system (Bennett & Gitomer, in press; O'Reilly & Sheehan, in press). In the reading domain, a cognitive model of reading competency serves at the basis for both assessing learning and advancing learning. Three assessment design features that are aimed at assessing deeper processing by requiring students to actively construct meaning from text, and are based on a cognitive model of reading are described by O'Reilly and Sheehan (in press). First, in the assessment, a scenario is provided that describes the purpose of reading. Because students engage in the reading process in meaningfully different ways dependent on the purpose of reading, the purpose of reading is clearly articulated. Second, students are required to read multiple texts so as to encourage students to integrate and synthesize information across texts. Lastly, to assess students' evaluation skills, texts of varying quality are provided.

One of the four important components assessed in their reading competency model is the student's ability to extract discourse structure (the other three are understanding vocabulary, drawing necessary inferences, and identifying important details). As O'Reilly and Sheehan (in press) pointed out, requiring students to construct a lengthy written summary may be more appropriate in the assessment of writing and not reading since the quality of students' response to a reading task can be affected by their writing ability. Instead, they use graphical representations for students to map out the structure of the text, including graphic hierarchical organizers and construct maps. The use of graphical representations instead of written summaries helps ensure that a student's writing ability does not unduly affect their performance on the reading tasks. This may help in minimizing construct-irrelevant variance. Further, the use of graphical representations will more easily allow for computer-automated scoring procedures to be used in the scoring of students' competency in organizing and summarizing information that they have read from one or more texts. This research program draws on models of cognition and learning and advances in technology and measurement to design assessments that capture students' complex thinking skills; it, therefore, has the capacity to provide meaningful information to guide instruction. As the researchers have indicated, there are a number of things that are being addressed in the design of these computer-based simulation tasks so as to ensure the validity of the score interpretations. Response formats are being chosen to minimize the extent to which writing is affecting the scores on reading and mathematics tasks, and that allow for automated scoring. Also, careful attention is being paid to representing the content and cognitive skills across the tasks so as to ensure the validity of the score generalizations.

Design of Assessments that Measure Learning Progressions

Assessments that reflect learning progressions are capable of identifying where students are on the learning progression and the skills and knowledge they need to acquire to become more competent. There have been some recent advances in assessment design efforts that reflect learning progressions or sometimes referred to as construct maps. Learning progressions indicate what it means to acquire understanding within a construct domain, and they identify where a student is on the continuum of the underlying construct. More specifically, they have been defined as "descriptions of successively more sophisticated ways of reasoning within a content domain based on research syntheses and conceptual analyses" (Smith, Wiser, Anderson, & Krajcik, 2006, p. 1), and should be organized around central concepts or big ideas within a content domain. Empirically validated models of cognition and learning can be used to design assessments that monitor students' learning as they develop understanding and competency in the content domain. These models of student cognition and learning across grade levels can be reflected in a coherent set of content standards across grade levels. This will help ensure the continuity of the assessment of students across grades, and will allow for monitoring student understanding and competency and for informing instruction and learning.

An issue in the design of learning progressions is that there may be multiple paths to proficiency; however, some paths typically are followed by students more often than others (Bennett & Gitomer, 2006). The use of these common paths to define learning progressions and the ways in which students gain a deep understanding of the content domain can be used as the foundation for designing assessments that monitor student achievement and learning. Learning progressions that are based on cognitive models of learning and are supplemented by teacher knowledge of student learning within content domains can inform the design of assessments that will elicit evidence to support inferences about student achievement at different points along the learning progression (NRC, 2006). Further, they have the potential to lead to more meaningful scaling of assessments that span grade levels, and thus more valid score interpretations regarding student growth.

Wilson and his colleagues have designed an assessment system that incorporates information from learning progressions and advances in both technology and measurement referred to as the BEAR Assessment System (Wilson, 2005; Wilson & Sloane, 2000). One application of this assessment system is for measuring a student's progression for one of the three "big ideas" in the domain of chemistry, namely *matter* which is concerned with describing molecular and atomic views of matter (Wilson, 2005). The two other "big ideas are *change* and *stability,* the former is concerned with kinetic views of change and the conservation of matter during chemical change, and the latter is concerned with the system of relationships in conservation of energy. Figure 1 illustrates the construct map for the *matter* big idea for two of its substrands, visualizing and measuring.

Figure 1: BEAR Assessment System Construct Map for the Matter Strand in Chemistry

Levels of Success	Matter Substrands	
	Visualizing Matter: Atomic and Molecular Views	Measuring Matter: Measurement and Model Refinement
5 - Integrating	bonding and relative reactivity	models and evidence
4 - Predicting	phase and composition	limitations of models
3 - Relating	properties and atomic views	measured amounts of models
2 - Representing	matter with chemical symbols	mass with a particulate view
1 - Describing	properties of matter	amounts of matter

Source: Adapted from Wilson (2005)

Level 1 in the table is the lowest level of proficiency and reflects students' lack of understanding of atomic views of matter, reflecting only their ability to describe some characteristics of matter, such as differentiating between a solid and a gas (Wilson, 2005). At Level 2, students begin to use a definition or simple representation to interpret chemical phenomena, and at Level 3 students begin to combine and relate patterns to account for chemical phenomena. Items are designed to reflect the differing achievement levels of the learning progression, or construct map, and empirical evidence is then collected to validate the construct map. A task designed to assess the lower levels of the construct map depicted in Figure 1 asks students to explain why two solutions with the same molecular formula have two very different smells. The task presents students with the two solutions, butyric acid and ethyl acetate; their common molecular formula, $C_4H_8O_4$; and a pictorial representation depicting that one smells good and the other bad. The students are required to respond in writing to the following prompt (Wilson, 2005, p. 11):

> Both of the solutions have the same molecular formulas, but butyric acid smells bad and putrid while ethyl acetate smells good and sweet. Explain why these two solutions smell differently.

By delineating the learning progressions within each of the "big ideas" of chemistry based on models of cognition and learning, assessments can be designed so as to provide evidence to support inferences about student competency at different achievement levels along the learning progressions. Performance assessments are well-suited for capturing student understanding and thinking along these learning progressions. Smith and her colleagues (Smith et al., 2006) proposed a learning progression around three key questions and six big ideas within the scientific topic of matter and atomic-molecular theory and provided examples of performance tasks that can assess different points along the continuum of understanding and inquiry within this domain.

Learning progressions are also considered in the BioKIDS project that is based on the Principled Assessment Designs for Inquiry (PADI) system. Within this system, three main design patterns for assessing scientific inquiry were identified, including formulating scientific explanations from evidence, interpreting data, and making hypotheses and predictions (Gotwals & Songer, 2006). Tasks based on a specific design pattern have many features in common. As an example, the design pattern, Formulating Scientific Explanations from Evidence, has two dimensions that are crossed: the level of inquiry skill required for the task and the level of content knowledge required for the task. This allows for the design of assessment tasks in nine cells, each cell representing a task template. There are three inquiry skill steps, from Step 1 to Step 3: "students match relevant evidence to a given claim, students choose a relevant claim and construct a simple explanation based on given evidence (construction is scaffolded), students construct a claim and explanation that justifies the claim using relevant evidence (construction is unscaffolded)" (Gotwals & Songer, 2006, p. 13). The level of content knowledge required for the task is classified as simple, moderate, or complex, requiring minimal content knowledge and no interpretation to applying extra content knowledge and interpretation of evidence. This is similar to Baxter and Glaser's (1998) conceptualization of four quadrants that differ in terms of content-richness and level of inquiry skills, but further divides these two dimensions into nine quadrants. To better reflect scientific inquiry, Gotwals and Songer (2006) have proposed a matrix for each of the three design patterns (formulating scientific explanations from evidence, interpreting data, and making hypotheses and predictions).

In designing performance tasks, the amount of scaffolding needs to be considered. Scaffolding is a task feature that is manipulated explicitly in the BioKIDS design patterns (Gotwals & Songer, 2006; Mislevy & Haertel, 2006). For example, Figure 2 (following page) presents two scientific inquiry assessment tasks that require scientific explanations from the BioKIDS project. The first task requires more scaffolding than the second task. The amount of scaffolding built into a task depends on the age of the students and the extent to which students have had the opportunity in instruction to solve tasks that require explanations and complex reasoning skills. The first task in the figure represents the second step in the level of inquiry skills and the second level of content knowledge (moderate) in that evidence is provided (pictures of invertebrates that must be grouped together based on their characteristics), but the student needs to choose a claim and construct the explanation, and they must interpret evidence and/or apply additional content knowledge (need to know which characteristics are relevant for classifying animals). The second task is at Step 3 of the level of inquiry skill and the third level of content knowledge (complex) in that the student needs to construct a claim and an explanation that requires the interpretation of evidence and application of additional content knowledge (Gotwals & Songer, 2006). More specifically, in this second task, "Students are provided a scenario, and they must construct (rather than choose) a claim and then, using their knowledge of food web interactions, provide evidence to back up their claim" (Gotwals & Songer, 2006, p. 16).

Additional Examples of Performance Tasks

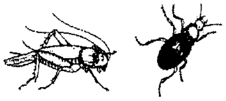
This section provides additional examples of performance tasks that draw on cognitive theories of student thinking and learning. The National Assessment of Educational Progress (NAEP) has included hands-on performance tasks in their science assessment (U.S. Department of Education, 2005) so as to better measure complex problem solving and reasoning skills. These tasks require students to engage in scientific inquiry, and to record their observations and conclusions by answering both multiple-choice and constructed-response items. As an example, a public release fourth-grade task, Floating Pencil, provides students with a set of materials, including bottles of freshwater, salt water, and "mystery" water. Students are required to perform a series of investigations to determine the properties of salt and freshwater, and to determine whether the bottle of mystery water is salt water or freshwater. After responding to a number of questions throughout their investigation, the students are asked (U.S. Department of Education, 2005, p. 10):

- Is the mystery water fresh water or is it salt water?
- How can you tell what the mystery water is?
- When people are swimming, is it easier for them to stay afloat in the ocean or in a freshwater lake?
- Explain your answer.

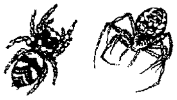
Figure 2: BioKids Assessment Tasks for "Formulating Scientific Explanations Using Evidence"

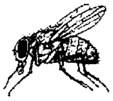
Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:

Group A:



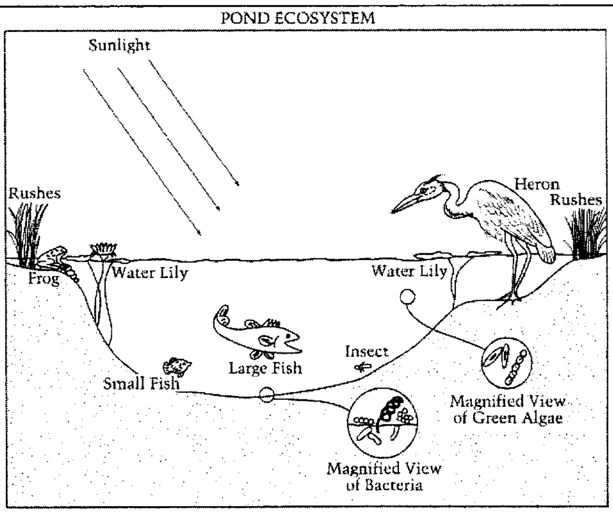
Group B:



They want to place this fly  in either Group A or Group B. Where should this fly be placed?

A fly should be in Group A /Group B
(Circle one)

Name two physical characteristics that you used when you decided to place the fly in this group:
(a)
(b)



...If all of the small fish in the pond system died one year from a disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so.

What would happen to the large fish? Explain why you think so.

The use of these hands-on performance tasks and constructed-response items allow NAEP to better assess scientific inquiry skills. As previously discussed, NAEP is currently examining the use of computer-based simulations to assess scientific inquiry.

The most commonly used large-scale performance assessments in this country are writing assessments. Writing assessments may consist of stand-alone writing prompts or text-based writing prompts. Stand-alone writing prompts require students to produce a written response to a given brief topic or prompt; whereas, text-based writing prompts reflect the reading and writing connection, in that students are asked to read about a topic from one or more sources, analyze it from a particular perspective, and then write a response (Nelson & Calfee, 1998). It has been argued that text-based writing assessments are more aligned to the writing that occurs in most classrooms in grades 6 through 12, higher education, and the workplace. An example of a writing assessment that includes both stand-alone and text-based writing prompts is the Delaware Student Testing Program (DSTP; Delaware Department of Education, 2000). A text-based writing task in the Delaware state assessment is linked to a passage in the reading assessment, and student responses to the task are scored twice, once for reading and once for writing. Below is an example eighth-grade, text-based persuasive writing prompt from the DSTP which requires students to read an article prior to writing:

> The article you have just read describes some problems and possible solutions for dealing with grease. Do you think grease should be classified or labeled as a pollutant?
>
> Write a letter to the Environmental Protection Agency explaining whether or not grease should be classified as a pollutant. Use information from this article to support your position (Delaware Department of Education, 2005).

This task is aligned to the reading and writing connection that occurs in instruction in Delaware classrooms. Students are first asked to read about a topic and then to use the information that they have read to support their position in their written product.

Another example of an assessment program that reflected the reading and writing connection was the Maryland School Performance Assessment Program (MSPAP; Maryland State Board of Education, 1995). It consisted of integrated performance tasks in writing, reading, social studies and science. For example, a writing task may have required the student to read one or more text in the social studies or science domains, analyze the texts from a particular perspective, and then write an essay. A complex integrated performance task from MSPAP that addressed issues related to child labor is provided in the Appendix. This is an integrated task that assesses reading, writing, language usage, and social studies. For this task, students were required to read “A Letter to Hannah” and “Mill Children” (only the first two pages of the two texts are included), and were asked to respond to a series of questions based on these readings, some of the questions required students to integrate their understanding of the two texts. They were also provided with two maps and were asked to respond to questions based on their understanding of the texts, maps, and general social studies content knowledge. A persuasive writing prompt was also given in which students needed to use information from the texts to support their views on child labor. The writing portion of the task also required students to engage in the different stages of the writing process (prewriting, writing, review and editing, and final version) as well as peer review. As indicated previously, MSPAP was the only completely state performance-based assessment program in multiple content areas that sustained success over a number of years.

Review and Field-Testing Performance Assessments

Performance assessments need to be appraised with regard to the quality and comprehensiveness of the content and processes being assessed and with regard to potential issues of bias in task content, language, and context. The review process is an iterative process in that when tasks are developed they may be reviewed by experts and modified a number of times prior to and after being field-tested. This involves logical analyses of the tasks to help evaluate whether they are assessing the intended content and processes, worded clearly and concisely, and free from anticipated sources of bias. The development process also includes field-testing the tasks and scoring rubrics to ensure they elicit the processes and skills intended.

It is important to field-test items individually as well as in a large-scale administration. For example, protocol analysis in which students are asked to think aloud while solving a task or to describe retrospectively the way in which they solved the task can be conducted to examine whether the intended cognitive processes are elicited by the task (Chi, Glaser, & Farr, 1988; Ericsson & Simon, 1984). These individual pilots afford rich information from a relatively small number of students regarding the degree to which the tasks evoke the content knowledge and complex thinking processes that they were intended to evoke, and allows for additional probing regarding the processes underlying student performance. The individual piloting of tasks also provides an opportunity for the examiner to pose questions to students regarding their understanding of task wording and directions, and to evaluate their appropriateness for different subgroups of students, such as students whose first language is not English.

A large-scale, field-testing provides additional information regarding the quality of the tasks including the psychometric characteristics of the items. Student work from constructed-response items or essays can also be analyzed to ensure that the tasks evoke the content knowledge and cognitive processes that they are intended to evoke, and the directions and wording are as clear as possible. Multiple variants of tasks can also be field-tested to further examine the best way to phrase and format tasks to ensure that all students have the opportunity to display their reasoning and thinking. Any one of these analyses may point to needed modifications to the tasks.

Large-scale field-testing of performance tasks poses a risk to security because they tend to be memorable to students. To help ensure the security of performance assessments, some state assessment programs have field-tested new tasks in other states. As an example, the initial field-testing of writing prompts for the Maryland Writing Test (MWT) occurred in states other than Maryland (Ferrara, 1987). However, the state's concern about the comparability of the out-of-state sample with respect to demographics, motivation, and writing instruction led them to an in-state field-test design. While security issues such as students sharing the field-test prompts with other students were considered problematic, the improvement of the field-test data outweighed security concerns (Ferrara, 1987). For example, in 1988, 22 new prompts were field-tested on a sample of representative ninth-grade students in Maryland with each student receiving 2 prompts. The anchor prompts were spiraled with the field-test prompts in the classrooms and each prompt was exposed only to approximately 250 students (Maryland State Department of Education, 1990). Field-test prompts that were comparable to the anchor prompts (e.g., similar means and standard deviations) were selected for future operational administrations (Ferrara, personal communication, July 30, 2009) and sophisticated equating procedures were not used. Enough prompts produced similar mean scores and standard deviations so as to be considered interchangeable.

To help maintain the security of the MWT prompts, a number of procedures were implemented (Ferrara, personal communication, July 30, 2009). First, the number of students who were exposed to any one prompt was small (approximately 250), and the number of teachers involved in the field test was relatively small. Second, the prompts were field-tested two to three years before they were administered on an operational test. Third, there was rigorous enforcement of security regulations. For the field testing of essay topics for the new SAT, a number of steps are implemented to help ensure the security of the prompts (Educational Testing Service, 2004). First, approximately 78 topics are pretested each year to a representative sample of juniors and seniors in high schools across the country. No more than 175 students are involved in the field test in a participating school, and only three prompts are administered in any one school with each student receiving only one prompt. Each prompt is field-tested in approximately 6 schools so that only 300 students are administered a given prompt. Second, prompts are field-tested at least 2 years prior to being on an operational form of the SAT. Third, rigorous security procedures are used for shipping and returning the field-test prompts. Lastly, several security procedures are implemented during the pretest readings such as the requirement of signed confidentiality statements by all prescreened readers who have served on College Board writing committees.

Security issues need to be considered for assessment programs for which the intent is to generalize from the score to the broader content domain. If security is breeched and the assessment tasks are known prior to the administration of the assessment, some scores will be artificially inflated which will have an impact on the validity of the score interpretations. Prior exposure to the task is not a security issue for performance demonstrations such as a high school project that requires students to demonstrate competency within a discipline. However, other issues need to be considered for performance demonstrations, such as ensuring the demonstration reflects the examinee's work and not others' unless a specified amount of collaboration was permitted.

Scoring Performance Assessments

As in the design of performance tasks, the design of scoring rubrics is an iterative process and involves coordination across grades as well as across content domains to ensure a cohesive approach to student assessment (Lane & Stone, 2006). Much has been learned about the design of quality scoring rubrics for performance assessments. First, it is critical to design scoring rubrics that include criteria that are aligned to the processes and skills that are intended to be measured by the assessment tasks. Unfortunately, it is not uncommon for performance assessments to be accompanied by scoring rubrics that focus on lower levels of thinking rather than on the more complex reasoning and thinking skills that the tasks are intended to measure; and, therefore, the benefits of the performance tasks are not fully realized. Typically, scoring rubrics should not be developed to be unique to specific tasks nor generic to the entire construct domain, but should be reflective of the "classes of tasks that the construct empirically generalizes or transfers to" (Messick, 1994, p. 17). Thus, a scoring rubric can be designed for a family of tasks or a particular task template. As previously indicated, the underlying performance on a task is a continuum that represents different degrees of structure versus open-endedness in the response, and this needs to be considered in the design of the scoring rubric and criteria (Messick, 1996).

The design of scoring rubrics requires the specification of the criteria for judging the quality of performances, the choice of a scoring procedure (e.g., analytic or holistic), ways for developing criteria, and procedures used to apply the criteria (Clauser, 2000). The ways for developing criteria include the process used for specifying the criteria and who should be involved in developing the criteria. For large-scale assessments in K-12 education, typically, the scoring criteria are developed by a group of experts as defined by their knowledge of the content domain and experience as educators. Often these experienced educators have been involved in the design of the performance tasks and have knowledge of how students of differing levels of proficiency would perform on the task. There are alternative approaches to specifying the criteria such as analyses of experts' thinking and reasoning when solving tasks. Cognitive task analysis using experts' talk alouds (Ericsson & Smith, 1991) has been used to design performance tasks and scoring criteria in the medical domain (Mislevy, Steinberg, Breyer, Almond, & Johnson, (2002). Features of experts' thinking, knowledge, procedures, and problem posing are considered to be indicators of developing expertise in a domain (Glaser, Lesgold, & Lajoie, 1987), and can be used systematically in the design of assessment tasks and scoring criteria. As mentioned previously, these experts can be students who have demonstrated competency within the domain. Two ways in which the criteria can be applied rely on the use of trained raters and computer-automated scoring procedures (Clauser, 2000). This section discusses the specification of the criteria, different scoring procedures, research on scoring procedures, and computer-automated scoring systems.

Specification of the Criteria

The criteria specified at each score level should be linked to the construct being assessed, and depend on a number of factors including the cognitive demands of the tasks in the assessment, the degree of structure or openness expected in the response, the examinee population, the purpose of the assessment, and its intended score interpretations (Lane & Stone, 2006). Further, the number of scores each performance assessment yields needs to be considered based on how many dimensions are being assessed. Performance assessments are well suited for measuring multiple dimensions within a content domain. For example, a grade 5 mathematics assessment may be designed to yield information on students' strategic knowledge, mathematical communication skills, and computational fluency. Separate criteria would be defined for each of these dimensions and a scoring rubric would then be developed for each dimension.

The number of score levels used depends on the extent to which the criteria across the score levels can distinguish among various levels of knowledge and skills. The knowledge and skills reflected at each score level should differ distinctly from those at other score levels. When cognitive theories of learning have been delineated within a domain, the learning progression can be reflected in the criteria. The criteria specified at each score level are then guided by knowledge of how students acquire understanding and competency within a content domain.

A generic rubric may be designed that reflects the skills and knowledge underlying the defined construct. The development of the generic rubric begins in the early stages of the performance assessment design, and then guides the design of specific rubrics for each family of tasks (task template) or a particular task that captures the cognitive skills and content assessed by the family of tasks or the particular task. An advantage of this approach is that it helps ensure consistency across the specific rubrics and is aligned with a construct-centered approach to test design. Typically, student responses that cover a wide range of competency are then evaluated to determine the extent to which the criteria reflect the components displayed in the student work. The criteria for the generic and/or specific rubrics may then be modified, and/or the task may be redesigned to ensure it assesses the intended content knowledge and processes. This may require several iterations to ensure the linkage among the content domain, tasks, and rubrics.

Scoring Procedures

The design of scoring rubrics has been influenced considerably by efforts in the assessment of writing. There are three major types of scoring procedures for direct writing assessments: holistic, analytic, and primary trait scoring (Huot, 1990; Miller & Crocker, 1990; Mullis, 1984). The choice of a scoring procedure depends on the defined construct, purpose of the assessment, and nature of the intended score interpretations. With holistic scoring, the raters make a single, holistic judgment regarding the quality of the writing and assign one score, using a scoring rubric with criteria and benchmark papers anchored at each score level. With analytic scoring, the rater evaluates the writing according to a number of features, such as content, organization, mechanics, focus, and ideas, and assigns a score indicating level of quality to each one. Some analytic scoring methods weigh the domains, allowing for domains that are assumed to be more pertinent to the construct being measured, such as content and organization, to contribute more to the overall score. As summarized by Mullis (1984), "holistic scoring is designed to describe the overall effect of characteristics working in concert, or the sum of the parts, analytic scoring is designed to describe individual characteristics or parts and total them in a meaningful way to arrive at an overall score." Although the sum of the parts of writing may not be the same as an overall holistic judgment, the analytic method has the potential to provide information regarding potential strengths and weaknesses of the examinee. Evidence, however, is needed to determine the extent to which the domain scores are able to differentiate aspects of students' writing ability.

Primary trait scoring was developed by the National Assessment of Educational Progress (Lloyd-Jones, 1977). The primary trait scoring system is based on the premise that most writing is addressed to an audience with a particular purpose, and levels of success in accomplishing that purpose can be defined concretely (Mullis, 1984). As an example, three common purposes of writing are informational, persuasive, and literary. The specific task determines the exact scoring criteria, although criteria are similar across similar kinds of writing (Mullis, 1984). The design of a primary trait scoring system involves the identification of one or more traits relevant for a specific writing task. For example, features selected for persuasive writing may include clarity of position and support, whereas characteristics for a literary piece may include plot, sequence, and character development. Thus, the primary trait scoring system reflects aspects of a generic rubric as well as task-specific rubrics. By first using a construct-centered approach, the construct, and in this case the type of writing, guides the design of the scoring rubrics and criteria. The development of primary trait rubrics then allows for the general criteria to be tailored to the task allowing for more consistency in raters' application of the criteria to the written response. Thus, in the end there may be one scoring rubric for each writing purpose. This would be analogous of having one scoring rubric for each family of tasks or task template.

Examples of Scoring Rubrics

In the design of performance assessments, Baker and her colleagues (Baker, 2007; Niemi, Baker & Sylvester, 2007) have represented the cognitive demands of the tasks in terms of classes or families of tasks such as reasoning, problem solving, and knowledge representation tasks (Baker, 2007). To ensure a coherent link between the tasks and the score inferences, they have designed a scoring rubric for each of these families of tasks. In the adoption of a construct-driven approach to the design of a mathematics performance assessment, Lane and her colleagues (Lane, 1993; Lane et al., 1995) used this approach in the design of their holistic scoring rubric. They first developed a generic rubric, as shown in Figure 3, that reflects the conceptual framework used in the design of the assessment, including mathematical knowledge, strategic knowledge, and communication (i.e., explanations) as overarching features. These features guided the design of families of tasks: tasks that assessed strategic knowledge, tasks that assessed reasoning, and tasks that assessed both strategic knowledge and reasoning. Mathematical knowledge was assessed across these task families. The generic rubric guided the design of each task-specific rubric that reflected one of these three families. The use of task-specific rubrics helped ensure the consistency in which raters applied the scoring rubric and the generalizability of the score inferences to the broader construct domain of mathematics.

Figure 3: Holistic General Scoring Rubric for Mathematics Constructed-Response Items

Performance Criteria, Level 4

- *Mathematical Knowledge.* Shows understanding of the problem's mathematical concepts and principles; uses appropriate mathematical terminology and notations; executes algorithms completely and correctly.
- *Strategic Knowledge.* Identifies all the important elements of the problem and shows understanding of the relationships among them; reflects an appropriate and systematic strategy for solving the problem; gives clear evidence of a solution process, and solution process is complete and systematic.

- *Communication.* Gives a complete response with a clear, unambiguous explanation and/or description; may include an appropriate and complete diagram; communicates effectively to the identified audience; presents strong supporting arguments which are logically sound and complete; may include examples and counter-examples.

Performance Criteria, Level 3

- *Mathematical Knowledge.* Shows nearly complete understanding of the problem's mathematical concepts and principles; uses nearly correct mathematical terminology and notations; executes algorithms completely; and computations are generally correct but may contain minor errors.
- *Strategic Knowledge.* Identifies the most important elements of the problem and shows general understanding of the relationships among them; and gives clear evidence of a solution process, and solution process is complete or nearly complete, and systematic.
- *Communication.* Gives a fairly complete response with reasonably clear explanations or descriptions; may include a nearly complete, appropriate diagram; generally communicates effectively to the identified audience; presents strong supporting arguments which are logically sound but may contain some minor gaps.

Performance Criteria, Level 2

- *Mathematical Knowledge.* Shows understanding of some of the problem's mathematical concepts and principles; and may contain computational errors.
- *Strategic Knowledge.* Identifies some important elements of the problem but shows only limited understanding of the relationships among them; and gives some evidence of a solution process, but solution process may be incomplete or somewhat unsystematic.
- *Communication.* Makes significant progress towards completion of the problem, but the explanation or description may be somewhat ambiguous or unclear; may include a diagram which is flawed or unclear; communication may be somewhat vague or difficult to interpret; and arguments may be incomplete or may be based on a logically unsound premise.

Performance Criteria, Level 1

- *Mathematical Knowledge.* Shows very limited understanding of some of the problem's mathematical concepts and principles; may misuse or fail to use mathematical terms; and may make major computational errors.
- *Strategic Knowledge.* Fails to identify important elements or places too much emphasis on unimportant elements; may reflect an inappropriate strategy for solving the problem; gives incomplete evidence of a solution process; solution process may be missing, difficult to identify, or completely unsystematic.
- *Communication.* Has some satisfactory elements but may fail to complete or may omit significant parts of the problem; explanation or description may be missing or difficult to follow; may include a diagram, which incorrectly represents the problem situation, or diagram may be unclear and difficult to interpret.

Performance Criteria, Level 0

- Shows no understanding of the problem's mathematical concepts and principles.

Source: Adapted from Lane (1993)

Another important issue in the design of scoring rubrics is that each of the score levels addressed each of the important scoring criteria. As can be seen in Figure 3, at each of the score levels criteria are specified for each of the overarching features: mathematical knowledge, strategic knowledge, and communication.

The scoring rubric for the tasks that assess student learning in the matter strand in the chemistry domain (Wilson, 2005) discussed previously is presented in Figure 4. The scoring rubric is reflective of the construct map, or learning progression, depicted in Figure 1, with students progressing from the lowest level of describe to the highest level of explain. Score Levels 1 (describe) and 2 (represent) in the rubric further differentiate students into 3 levels.

Figure 4: Bear Assessment System Scoring Guide for the Matter Strand in Chemistry

Level	Descriptor	Criteria
0	Irrelevant or Blank Response	Response contains no relevant information
1	Describe the properties of matter	Rely on macroscopic observation and logic skills. No use of atomic model. Uses common sense and no correct chemistry concepts. 1- Makes one or more macroscopic observation and/or lists chemical terms without meaning 1 Uses macroscopic observation AND comparative logic skills to get a classification, BUT shows no indication of using chemistry concepts 1+ Makes simple microscopic observations and provides supporting examples, BUT chemical principle/rule cited incorrectly
2	Represent changes in matter with chemical symbols	Beginning to use definitions of chemistry to describe, label, and represent matter in terms of chemical composition. Use correct chemical symbols and terminology 2- Cites definitions/rules about matter somewhat correctly 2 Cites definition/rules about chemical composition 2+ Cites and uses definitions/rules about chemical composition of matter and its transformation
3	Relate	Relates one concept to another and develops models of explanation
4	Predicts how the properties of matter can be changed	Apply behavioral models of chemistry to predict transformation of matter
5	Explains the interactions between atoms and molecules	Integrates models of chemistry to understand empirical observations of matter

Source: Adapted from Wilson (2005)

A constructed response that reflects Level 2 is (Wilson, 2005):

> They smell differently b/c even though they have the same molecular formula, they have different structural formulas with different arrangements and patterns.

This example response is at the Level 2 because it “appropriately cites the principle that molecules with the same formula can have different arrangements of atoms. But the answer stops short of examining structure-property relationships (a relational, Level 3 characteristic)” (Wilson, 2005, p. 16). A major goal of the assessment system is to be able to estimate, with a certain level of probability, where a student is on the construct map or learning progression. Students and items are located on the same construct map, which allows for student proficiency to have substantive interpretation in terms of what the student knows and can do (Wilson, 2005). The maps can then be used to monitor the progress of an individual student as well as groups of students. Thus, valid interpretations of a student’s learning or progression require a carefully designed assessment system that has well-conceived items and scoring rubrics that represent the various levels of the construct continuum as well as the empirical validation of the construct map, or learning progression. As previously indicated, students do not necessarily follow the same progression in becoming proficient within a subject domain. Consequently, in the design of assessments, considerations should be given to identifying the range of strategies used for solving problems in a content domain, with an emphasis on those strategies that are more typical of the student population (Wilson, 2005). This assessment-design effort provides an interesting example of the integration of models of cognition and learning, and of measurement models in the design of an assessment system that can monitor student learning and inform instruction. Further, a measurement model called the saltus (Latin for leap) model developed by Wilson (1989) can incorporate developmental changes (or conceptual shifts in understanding) as well as the incremental increases in skill in evaluating student achievement and monitoring student learning.

In an effort to assess complex science reasoning in middle and high school, a systematic assessment-design procedure was adopted by Liu, Lee, Hofstetter, and Linn (2008). First, they identified an important construct within scientific inquiry, *science knowledge integration*. A comprehensive, integrated system of inquiry-based science curriculum modules, assessment tasks and scoring rubric were then developed to assess *science knowledge integration*. A scoring rubric was designed so that the different levels captured qualitatively different kinds of scientific cognition and reasoning that focused on elaborated links rather than individual concepts. Their assessment design is similar to the modeling of construct maps, or stages in learning progressions, described by Wilson (2005) and Wilson & Sloan (2000). The knowledge integration scoring rubric is shown in Figure 5.

Figure 5: Knowledge Integration Scoring Rubric

Link Levels	Description
Complex	Elaborate 2 or more scientifically valid links among relevant ideas
Full	Elaborate 1 scientifically valid link between 2 relevant ideas
Partial	State relevant ideas but do not fully elaborate the link between relevant ideas
No	Make invalid ideas or have non-normative ideas

Source: Liu, Lee, Hofstetter, and Linn (2008)

The rubric is applied to all the tasks that represent the task template for *science knowledge integration*, allowing for score comparisons across different items (Liu et al., 2008). As they indicate, having one scoring rubric that can be applied to the set of items that measure knowledge integration makes it more accessible for teachers to use and provides coherency in the score interpretations. The authors also provided validity evidence for the learning progression reflected in the scoring rubric.

Research on Analytic and Holistic Scoring Procedures

The validity of score interpretation and use depends on the fidelity between the constructs being measured and the derived scores (Messick, 1989). Validation of the scoring rubrics includes an evaluation of the match between the rubric and the targeted construct or content domain, how well the criteria at each score level captures the defined construct, and the extent to which the domains specified in analytic scoring schemes each measure some unique aspect of student cognition. Lane and Stone (2006) provide a brief summary of the relative advantages of both analytic and holistic scoring procedures for writing assessments. As an example, Roid (1994) used Oregon's direct-writing assessment to evaluate its analytic scoring rubric in which students' essays were scored on six dimensions. The results suggested that each dimension may not be unique, in that relative strengths and weaknesses for some students were identified for combinations of dimensions. Thus, some of the dimensions could be combined in the scoring system without much loss of information while simplifying the rubric and the scoring process. Other researchers have suggested that analytic and holistic scoring methods for writing assessments may not necessarily provide the same relative standings for examinees. Vacc (1989) reported correlations between the two scoring methods ranging from .56 to .81 for elementary school students' essays. Research that has examined factors that affect rater judgment of writing quality have shown that holistic scores for writing assessments are most influenced by the organization of the text and important ideas or content rather than domains related to mechanics and sentence structure (Breland & Jones, 1982; Huot, 1990; Welch & Harris; 1994). Breland and colleagues (Breland, Danos, Kahn, Kubota, & Bonner, 1994) reported relatively high correlations between holistic scores and scores for overall organization (approximately .73), supporting ideas (approximately .70), and noteworthy ideas (approximately .68).

In the science domain, Klein et al. (1998) compared analytic and holistic scoring of hands-on science performance tasks for grades 5, 8, and 10. The correlations between the total scores obtained for the two scoring methods were relatively high, .71 for grade 5 and .80 for grade 8. The correlations increased to .90 for grade 5 and .96 for grade 8 when disattenuated for the inconsistency among raters within a scoring method. The authors suggested that the scoring method has little unique influence on the raters' assessment of the relative quality of a student's performance. They further suggested that, if school performance is of interest, the use of one scoring method over the other probably has little or no effect on a school's relative standing within a state given the relatively high values of the disattenuated correlations. The time and cost for scoring for both of the methods was also addressed. The analytic method took nearly three times as long as the holistic method to score for a grade 5 response and nearly five times as long to score for a grade 8 response, resulting in higher costs for scoring using the analytic method.

The results of these studies suggest that the impact of the choice of scoring method (e.g., analytic versus holistic) may vary depending on the similarity of the criteria reflected in the scoring methods and for the use of the scores. The more closely the criteria for the analytic method resemble the criteria delineated in the holistic method, the more likely it is that the relative standings for examinees will be similar. The research also suggests that analytic rubrics typically are capable of providing distinct information for only a small number of domains or dimensions (i.e., two or three), and thus providing scores for a small number of domains has the potential for identifying overall strengths and weaknesses in student achievement and for informing instruction. As previously suggested, scores derived from performances on computer-based simulation tasks also allow for addressing different aspects of students' thinking.

Human Scoring

The scoring of student responses to performance assessments may be done by human scorers or automated scoring procedures that have been trained by human scoring. Lane and Stone (2006) provide an overview of the training procedures and methods for human scorers, and discuss rating sessions that may involve raters spending several days together evaluating student work as well as online rating of student work. A consideration in human scoring of performance assessments is rater variability or inconsistency, in particular, with writing assessments. As summarized by Eckes (2008), raters may differ in the extent to which they implement the scoring rubric, the way in which they interpret the scoring criteria, and the extent to which they are severe or lenient in scoring examinee performance; as well as in their understanding and use of scoring categories, and their consistency in rating across examinees, scoring criteria, and tasks (Bachman & Palmer, 1996; McNamara, 1996; Lumley, 2005). Thus, construct representation of the assessment can be jeopardized by the raters' interpretation and implementation of the scoring rubric as well as by features specific to the training session. For example, the pace at which raters are expected to score student responses may affect their ability to use their unique capabilities to accurately evaluate student responses or products (Bejar, Williamson, & Mislevy, 2006). Carefully designed scoring rubrics and training procedures, however, can help alleviate errors in human scoring.

Freedman and Calfee (1983) pointed out the importance of understanding rater cognition and proposed a model of rater cognition for evaluating writing assessments that consisted of three processes: reading the students' text to build a text image, evaluating the text image, and articulating the evaluation. Wolfe (1997) elaborated on Freedman and Calfee's model of rater cognition and proposed a cognitive model of rater cognition for scoring essays, which included a framework for scoring and a framework for writing. He proposed that understanding the process of rating would allow for better design of scoring rubrics and training procedures. The framework of scoring is a "mental representation of the processes through which a text image is created, compared to the scoring criteria, and used as the basis for generating a scoring decision" (Wolfe, 1997, p. 89). The framework for writing, which includes the rater's interpretation of the criteria in the scoring rubrics, emphasized that raters have different interpretations of the scoring rubric and therefore, are not equally proficient at rating student essays (Wolfe, 1997). Through training, however, raters begin to share a common understanding of the scoring rubric so as to apply it consistently. Wolfe (1997) also observed that proficient scorers were better able to withhold judgment as they read an essay and focused their efforts more on the evaluation process than less proficient scorers. This shared common framework for writing and high level of scoring proficiency can lead to a high level of agreement among raters, and has implications for the training of raters (Wolfe, 1997). Thus, raters can be trained to internalize the criteria in a similar manner and to apply it consistently so as to ensure scores that allow for valid interpretations of student achievement.

Automated Scoring Systems

Automated scoring systems have supported the use of computer-based performance assessments such as computer-delivered writing assessments and computer-based simulation tasks, as well as paper-and-pencil assessments that are scanned. Automated scoring procedures have a number of attractive features. They apply the scoring rubric consistently, but more importantly they allow for the test designer to control precisely the meaning of scores (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). In order to accomplish this, they "need to elicit the full range of evidence called for by an appropriately broad definition of the construct of interest" (Bejar, Williamson, & Mislevy, 2006, p 52.). Automated scoring procedures allow for an expanded capacity to collect and record many features of student performance from complex assessment tasks that can measure multiple dimensions (Williamson, Bejar, & Mislevy, 2006). A very practical advantage is that they allow for scores to be generated in a timely manner. Automated scoring is defined by Williamson, Bejar and Mislevy (2006) as "any computerized mechanism that evaluates qualities of performances or work products" (p. 2). Automated scoring of complex constructed-response computerized tasks has been proven effective for large-scale assessments as well as for classroom assessment purposes. Project essay grader developed by Page (1994, 2003) in the 1960s paved the way for automated scoring systems for writing assessments, including e-rater (Burstein, 2003), Intelligent Essay Assessor (IEA; Landauer, Foltz, & Laham, 1998; Landauer, Laham, & Folz, 2003), and Intellemetric (Elliot, 2003).

Automated scoring procedures have also been developed to score short constructed-response items. C-rater is an automated scoring method for scoring constructed-response items that elicit verbal responses that range from one sentence to a few paragraphs, have rubrics that explicitly specify the content required in the response, but do not evaluate the mechanics of writing items (Leacock & Chodorow, 2003, 2004). It has been used successfully in Indiana's state end-of-course, grade 11 English assessment, the NAEP Math Online project that required students to provide explanations of their mathematical reasoning, and the NAEP simulation study that required students to use search queries (Bennett et al., 2007; Deane, 2006). C- rater is a paraphrase recognizer in that it can determine when a student's constructed response matches phrases in the scoring rubric regardless of their similarity in word use or grammatical structure (Deane, 2006). In the NAEP study that used physics computer-based simulations, c-rater models were built using student queries and then cross-validated using a sample of queries that were independently hand-scored. The agreement between human raters and c-rater for the cross-validation study was 96 percent.

Automated scoring procedures have also been developed and used successfully for licensure examinations in medicine, architecture, and accountancy. These exams use innovative computer-based simulation tasks which naturally lend themselves to automated scoring. The previously mentioned assessment that uses computer-based case simulations to measure physicians' patient-management skills (Clyman, Melnick, & Clauser, 1995) and the figural response items in the architecture assessment (Martinez & Katz, 1996) are excellent examples of the feasibility in using automated scoring procedures for innovative item types.

Scoring Algorithms for Writing Assessments.

The most widely used automated scoring systems are those that assess students' writing. Typically, the design of the scoring algorithms for automated scoring essay systems requires humans to first rate a set of student essays to a prompt. The student essays and their ratings then serve as calibration data that are used by the software to train it for scoring. The scoring algorithm is designed to analyze specific features of essays, and weights are assigned for each of these features. As summarized by Deane and Gurevich (2008), the fields of computational linguistics, artificial intelligence, and natural language processing have produced a number of methods for investigating the similarity of text content, including latent semantic analysis (LSA) and content vector analysis (CVA). These text-similarity methods have been applied to automated, essay scoring applications. As an example, e-rater, developed by the Educational Testing Service, uses natural language processing techniques and identifies linguistic features of text in the evaluation of the quality of an essay (Burstein, 2003; Attali & Burstein, 2005). The first version of e-rater used over 60 features in the scoring process, whereas the latter versions use only "a small set of meaningful and intuitive features" (Attali & Burstein, 2005) that better captures the qualities of good writing, and thus simplifying the scoring algorithm. The scoring system uses a model-building module to analyze a sample of student essays to determine the weight of the features for assigning scores.

Evaluation of Automated Scoring Procedures

As with any assessment procedure, validation studies are imperative for automated scoring systems so as to provide evidence for appropriate score interpretations. Yang, Buckendahl, Juszkiewicz, and Bhola (2002) identified three categories of validation approaches for automated scoring procedures, including approaches focusing on the relationship among scores given by different scorers (human and computer), approaches focusing on the relationship between test scores and external measures of the construct being assessed, and approaches focusing on the scoring process. Most studies have examined the relationship between human- and computer-generated scores, typically indicating that the relationship between the scores produced by computer and humans is similar to the relationship between the scores produced by two humans, indicating the potential interchangeability of human and automated scoring. There have been few studies, however, that focus on the latter two categories. In particular, validation studies focusing on the scoring process for automated scoring procedures are limited. As Bennett (2006) has argued, automated scoring procedures should be grounded in a theory of domain proficiency, using experts to delineate proficiency in a domain rather than having them as a criterion to be predicted. Both construct irrelevant variance and construct underrepresentation may affect the validity of the scores obtained by automated scoring systems (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). With respect to construct irrelevant variance, automated scoring procedures may be influenced by irrelevant features of students' writing and assign a higher or lower score than deserved. In addition, they may not fully represent the construct of good writing which can affect the score assigned (Powers et al., 2002).

Studies have been conducted that require experts to evaluate the relevance of the computer-generated features of the target construct, identify extraneous and missing features, and evaluate the appropriateness of the weights assigned to the features (Ben-Simon & Bennett, 2007). Ben-Simon and Bennett (2007) found that the dimensions that experts in writing believe are most important in the assessment of writing are not necessarily the same as those obtained by automated scoring procedures that statistically optimize weights of the dimensions. As an example, experts indicated that approximately 65 percent of the essay scores in the study should be based on organization, development and topical analysis, while empirical weights gave approximately 21 percent of the emphasis to these dimensions. The opposite pattern occurred for the dimensions related to grammar, usage, mechanics, style, and essay length, with a much lower emphasis assigned by experts and a higher emphasis given by the automated scoring procedure. As indicated by Ben-Simon and Bennett (2007), the parameters of automated scoring procedures can be adjusted to be more consistent with those that experts believe are features of good writing; however, these adjustments may not be based on the criteria specified in the scoring rubric implemented in the study but rather on the criteria used by the scorers in assigning scores. The authors indicated that the rubric employed in their study was missing key features of good writing, leaving experts to apply some of their own criteria in the scoring process. This result illustrates the importance of linking the cognitive demands of the tasks to the criteria specified in the scoring rubric regardless if responses are to be scored by human raters or automated scoring procedures. The authors further suggested that current theories of writing cognition should be used in assessment design so as to ensure that a more theoretical, coherent model for identifying scoring dimensions and features is reflected in the criteria of the rubrics.

Typically, the agreement between the scores that are assigned by human raters and those assigned by the automated scoring procedure is very high. There is some recent research, however, that indicates scores assigned by human raters and by automatic scoring procedures may differ to some extent depending upon student demographics. Bridgeman, Trapani, and Attali (2009) examined whether there were systematic differences in the performance of subgroups using an automated scoring procedure versus human scoring for an 11th-grade, English state assessment. The prompt required students to support an opinion on a proposed topic within a 45- minute class period. The essays were scored holistically using a 6-point scale. The results indicated that on average, both Asian American and Hispanic students received higher scores from the automated scoring procedure than from human raters, whereas African American students scored similarly across the two scoring methods. Under the assumption that Asian American and Hispanic subgroups have a higher proportion of students with English as a second language, the authors suggested that this finding may not be due to minority status, but instead it may be related to having English as a second language. This may be reasonable given that the African American subgroup performed similarly across the two scoring methods. In their conclusions, they suggest that “although we treat human scores at the gold standard, we are reluctant to label discrepancies from the human score as bias because it is not necessarily the case that the human score is a better indicator of writing ability than the e-rater score (Bennett & Behar, 1997)” (Bridgeman, Trapani, & Attali, 2009, p. 17). As suggested by the authors, additional research needs to examine features that contribute to differential subgroup results for human and automated scores, especially for students for which English is a second language. An understanding of the features of automated scoring systems that led to differential subgroup patterns will inform future designs of these systems.

Automated scoring systems need to be capable of flagging bad faith essays because of the possibility of examinees trying to trick the systems into providing scores that are not warranted. Advances in the design of the more recent versions of automated scoring systems have led to accurate identification of bad faith essays. Bad faith essays include essays that are off topic and are written to a different prompt, essays that repeat the prompt, essays that consist of multiple repeated text, and essays that are a mix of a genuine response and a repetition of the prompt. Studies have been conducted that demonstrate the capability of automated scoring procedures in detecting bad faith essays. In an early study, Powers and his colleagues (2002) examined the extent to which an early version of e-rater could be tricked into assigning either too high or too low of a score. Writing experts were asked to fabricate essays in response to the writing prompts in the Graduate Record Examination (GRE) that would trick e-rater into assigning scores that were either higher or lower than deserved. The writing experts were instructed on how e-rater scores student essays, and were asked to write one essay for which e-rater would score higher than human readers and one essay for which e-rater would score lower than human readers. E-rater scores on these fabricated essays were then compared with the scores of two human readers. The predictions that e-rater would score higher than the human readers were upheld for 87 percent of the cases. Some of the essays that were scored higher by e-rater as compared with the human raters consisted of repeated paragraphs with or without a rewording of the first sentence in each paragraph. E-rater also provided higher scores than human readers for essays that did not provide a critical analysis, but focused on the features that e-rater attends to such as relevant content words and complex sentence structures. An important result is that only 42 percent of the cases were upheld when the predictions were that e-rater would score lower than the human raters (Powers et al., 2002). Thus, the experts were less able to trick e-rater to provide a lower score than human raters. It should be noted that e-rater has been revised substantially and there have been numerous versions of e-rater since this study. Further, to detect off-topic essays, which may occur when students are trying to fool the system, a content vector analysis program is used along with the more recent versions of e- rater (Higgins, Burstein & Attali, 2006).

In an evaluation of IntellicMetric for use with the Graduate Management Test (GMAT), Rudner, Garcia, and Welch (2006) examined its ability to detect common cheating techniques. For three Analysis of an Issue prompts and three Analysis of an Argument prompts, approximately 13 essays for each prompt were fabricated, resulting in 78 fabricated essays. The fabricated essays were evaluated with 500 validation essays for each prompt. Five of the fabricated essays were off topic and written to a response to a different prompt but of the same type (Issues or Arguments prompt type), five essays were off-topic and written to a response to a different prompt of a different type, one essay was a repetition of the entire prompt, one essay consisted of multiple repeated text, and one essay consisted of half a genuine response and have a repetition of the prompt. Their results indicated that the system was successful at identifying fabricated essays that were a copy of the prompt, consisted of multiple repeated text, and consisted of the prompt and partly a genuine response. For each detected essay, the system provided specific warning flags for plagiarism, copying the prompt, and nonsensical writing. The system was not successful at detecting off-topic responses; however, as the authors indicated this version of the system did not include a routine to flag off-topic essays.

The current versions of automated scoring systems for essays have shown strength in not only having high rates of agreement with human raters in assigning scores, but also in detecting bad faith essays. Automated scoring procedures for computerized short constructed-response items and innovative item types have also been used effectively for large-scale assessment programs. Further, various features of students' performances can be captured with automated scoring procedures which is ideal for computerized innovative tasks that reflect multiple dimensions within a content domain. Typically, most of the work and costs in designing automated scoring systems occur prior to the operational administration of the assessments, allowing for timely scoring and reporting of the results.

Evaluating the Validity and Fairness of Performance Assessments

Assessments are used in conjunction with other information to make important inferences about proficiency at the student, school, and state level, and therefore it is essential to obtain evidence about the appropriateness of those inferences and any resulting decisions. In evaluating the worth and quality of any assessment, including performance assessments, evidence to support the validity of the score inferences is at the forefront. Validity pertains to the meaningfulness, appropriateness, and usefulness of test scores (Kane, 2006; Messick, 1989). The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) state that "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). This requires specifying the purposes and the uses of the assessment, designing the assessment to fit these intentions, and providing evidence to support the proposed uses of the assessment and the intended score inferences. As an example, if the purpose of a performance assessment is to assess complex thinking skills so as to make inferences about students' problem solving and reasoning, one of the important validity studies would be to examine the cognitive skills and processes underlying task performance for support of those intended score inferences. The alignment between the content knowledge and cognitive skills underlying task responses and those underlying the targeted construct domain needs to be made explicit because typically the goal is to generalize assessment-score interpretations to the broader construct domain (Messick, 1989). Fundamental to the validation of test use and score interpretation is also the evaluation of both intended and unintended consequences of the use of an assessment (Kane, 2006; Messick, 1989). Because performance assessments are intended to improve teaching and student learning, it is essential to obtain evidence of such positive consequences as well as any evidence of negative consequences (Messick, 1994).

As previously discussed, there are two sources of potential threats to the validity of score inferences—construct underrepresentation and construct-irrelevant variance (Messick, 1989). Construct underrepresentation occurs when the assessment does not fully capture the targeted construct, and therefore the score inferences may not be generalizable to the larger domain of interest. Issues related to whether the content of the assessment is representative of the targeted domain will be discussed later in this section. Construct-irrelevant variance occurs when one or more irrelevant constructs is being assessed in addition to the intended construct. Sources of construct-irrelevant variance for performance assessments may include, but are not limited to, task wording and context, response mode, and raters' attention to irrelevant features of responses or performances. As an example, in designing a performance assessment that measures students' mathematical problem solving and reasoning, tasks should be set in contexts that are familiar to the population of students. If one or more subgroups of students are unfamiliar with a particular problem context and it affects their performance, the validity and fairness of the score interpretations for those students is hindered. Similarly, if a mathematics-performance assessment requires a high level of reading ability and students who have very similar mathematical proficiency perform differently due to differences in their reading ability, the assessment is measuring in part a construct that is not the target, namely, reading proficiency. This is of particular concern for English Language Learners (ELLs). Abedi and his colleagues (Abedi, Lord, & Plummer, 1997; Abedi & Lord, 2001) have identified a number of linguistic features that slow down the reader, increasing the chances of misinterpretation. In one study, they used their linguistic modification approach in that mathematics items were modified to reduce the complexity of sentence structures and unfamiliar vocabulary was replaced with familiar vocabulary (Abedi & Lord, 2001). The mathematics scores of both ELL students and non-ELL students in low- and average-level mathematics classes improved significantly when the linguistic modification approach was used. In another study, Abedi and his colleagues (Abedi, Lord, Hofstetter, & Baker, 2000) found that out of four different accommodation strategies for ELLs, only the linguistically modified English form narrowed the gap between ELLs and other students. Thus, the linguistic modification approach can be used in the design of performance assessments to help ensure a valid and fair assessment of not only ELLs, but other students who may have difficulty with reading.

When students are asked to explain their reasoning on mathematics and science assessments, the writing ability of the student could be a source of construct-irrelevant variance. To help minimize the impact of writing ability on math and science assessments, scoring rubrics need to clearly delineate the relevant criteria. Construct-irrelevant variance may also occur when raters score student responses to performance tasks according to features that do not reflect the scoring criteria and are irrelevant to the construct being assessed (Messick, 1994). This can also be addressed by clearly articulated scoring rubrics and the effective training of the raters.

Validity criteria that have been suggested for examining the quality of performance assessments include, but are not limited to, content representation, cognitive complexity, meaningfulness, transfer and generalizability, fairness, and consequences (Linn, Baker, & Dunbar, 1991; Messick, 1994). The discussion that follows is organized around these validity criteria. These criteria are closely intertwined to some of the sources of validity evidence proposed by the *Standards for Educational and Psychological Measurement*. (AERA, APA, and NCME, 1999): evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing.

Evaluating Content Representativeness

An analysis between the content of the assessment and the construct it is intended to measure provides important validity evidence (AERA, APA, NCME, 1999). Test content refers to the skills, knowledge, and processes that are intended to be assessed by tasks as well as the task formats and scoring procedures. Performance tasks can be designed so as to emulate the skills and processes reflected in the targeted construct. For many large-scale assessment programs, it is important to ensure that the ability to generalize from a student's score on a performance assessment to the broader domain of interest is not limited by having too small of a number of tasks on the performance assessments. Although the performance tasks may be assessing students' understanding of some concepts or set of concepts at a deeper level, the content of the domain may not be well represented by a relatively small subset of performance tasks. This can be addressed by including other item formats that can appropriately assess certain skills, and using performance tasks to assess complex thinking skills that cannot be assessed by the other item formats. For some high-stakes, large-scale assessments, including state assessment and accountability systems, performance tasks are used in conjunction with multiple-choice items to ensure that the assessment represents the content domain and to allow for inferences about individual student performance to the broader domain.

Methods are currently being investigated that will allow for accurate student-level scores derived from mathematics and language arts performance assessments that are administered on different occasions throughout the year (Bennett & Gitomer, in press). This will not only allow for content representation across the performance assessments, but also the assessments can be administered in closer proximity to the relevant instruction, and information from any one administration can be used to inform future instructional efforts. If school level scores are of interest primarily, matrix-sampling procedures can be used to ensure content representation on the performance assessment as was done on the Maryland State Performance Assessment Program (Maryland State Board of Education, 1995).

The coherency and representativeness among the assessment tasks, scoring rubrics and procedures, and the target domain are other aspects of validity evidence for score interpretations. It is important to ensure that the cognitive skills and content of the target domain are systematically represented in the tasks and scoring procedures. The method used to transform performance to a score can provide evidence for the validity of the score interpretation. Both logical and empirical evidence can support the validity of the method used for transforming performance to a score.

For performance demonstrations such as a high school project, we are not interested in generalizing the student performance on the demonstration to the broader domain; so, the content domain does not need to be represented fully. The content and skills being assessed by the performance demonstration should be meaningful and relevant within the content domain. Performance demonstrations provide the opportunity for students to show what they know and can do on a real world task, similar to a driver's license test.

Evaluating Cognitive Complexity

One of the most attractive aspects of performance assessments is that they can be designed to assess complex thinking and problem-solving skills. As Linn and his colleagues (1991) have cautioned, however, it should not be assumed that a performance assessment measures complex thinking skills; evidence is needed to examine the extent to which tasks and scoring rubrics are capturing the intended cognitive skills and processes. The alignment between the cognitive processes underlying task responses and the construct domain needs to be made explicit because typically the goal is to generalize scores interpretations to nonassessment construct-domain interpretations (Messick, 1989). The validity of the score interpretations will be affected by the extent to which the design of performance assessments is guided by cognitive theories of student achievement and learning within academic disciplines. Further, the use of task templates will allow for the explicit delineation of the cognitive skills required to perform particular task types.

Several methods have been used to examine whether tasks are assessing the intended cognitive skills and processes (Messick, 1989), and they are particularly appropriate for performance assessments that are designed to tap complex thinking skills. These methods include *protocol analysis*, *analysis of reasons,* and *analysis of errors*. In *protocol analysis*, students are asked to think aloud as they solve a problem or describe retrospectively how they solve the problem. In the method of *analysis of reasons*, students are asked to provide rationales, typically written, to their responses to the tasks. The method of *analysis of errors* requires an examination of procedures, concepts, or representations of the problems in order to make inferences about students' misconceptions or errors in their understanding. As an example, in the design of a science-performance assessment, Shavelson and Ruiz-Primo (1998) used Baxter and Glaser's 1998 analytic framework, which reflects a content-process space depicting the necessary content knowledge and process skills for successful performance. Using protocol analysis, Shavelson and Ruiz-Primo (1998) compared expert and novice reasoning on the science performance tasks that were content-rich and process-open. Their results from the protocol analysis confirmed some of their hypotheses regarding the different reasoning skills that tasks were intended to elicit from examinees. Further, the results elucidated the complexity of experts' reasoning as compared to the novices and informed the design of the tasks and interpretation of the scores.

Evaluating Meaningfulness and Transparency

An important validity criterion for performance assessments is their meaningfulness (Linn et al., 1991) which refers to the extent to which students, teachers, and other interested parties find value in the tasks at hand. Meaningfulness is inherent in the idea that performance assessments are intended to measure more directly the types of reasoning and problem-solving skills that are valued by educators. A related criteria is transparency (Frederiksen & Collins, 1989), that is, students and teacher need to know what is being assessed, by what methods, the criteria used to evaluate performance, and what constitutes quality performance. It is important to ensure that all students are familiar with the task format and scoring criteria for both large-scale assessments and classroom assessments. Teachers can use performance tasks with their students, and engage them in discussions about what the tasks are assessing and the nature of the criteria used for evaluating student work. Teachers can also engage students in using scoring rubrics to evaluate their own work and the work of their peers.

Evaluating the Generalizability of Score Inferences

For many large-scale assessments, the intent is to draw inferences about student achievement in the domain of interest based on scores derived from the assessment. A potential threat to the validity of score interpretations, therefore, is the extent to which the scores from the performance assessments can be generalized to the broader construct domain (Linn, Baker, & Dunbar, 1991). It should be noted, however, that this is not the intent for performance demonstrations as discussed previously.

Generalizability theory provides both a conceptual and statistical framework to examine the extent to which scores derived from an assessment can be generalized to the domain of interest (Brennan, 1996, 2000, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). It is particularly relevant in evaluating performance assessments that assess complex thinking skills because it examines multiple sources of errors that can limit the generalizability of the scores, such as error due to tasks, raters, and occasions. Error due to tasks occurs because there are only a small number of tasks typically included in a performance assessment. As explained by Haertel and Linn (1996), students' individual reactions to specific tasks tend to average out on multiple-choice tests because of the relatively large number of items, but such individual reactions to specific items have more of an effect on scores from performance assessments that are composed of relatively few items. Thus, it is important to consider the sampling of tasks and by increasing the number of tasks on an assessment, the validity and generalizability of the assessment results is enhanced. Further, this concern with task specificity is consistent with research in cognition and learning that underscores the context-specific nature of problem solving and reasoning in subject matter domains (Greeno, 1989). The use of multiple-item formats, including performance tasks, can improve on the generalizability of the scores.

Error due to raters can also affect the generalizability of the scores in that raters may differ in their evaluation of the quality of students' responses to a particular performance task and across performance tasks. Raters may differ in their leniency (resulting in rater mean differences), or they may differ in their judgments about whether one student's response is better than another student's response (resulting in an interaction between the student and rater facets) (Hieronymus & Hoover, 1987; Lane, Liu, Ankenmann, & Stone, 1996; Shavelson, Baxter, & Gao, 1993). Typically, occasion is an important hidden source of error because performance assessments are only given on one occasion and occasion is not typically considered in generalizability studies (Cronbach, Linn, Brennan, & Haertel, 1997).

Generalizability theory estimates variance components for the object of measurement (e.g., student, class, school) and for the sources of error in measurement such as task and rater. The estimated variance components provide information about the relative contribution of each source of measurement error. The variance estimates are then used to design measurement procedures that allow for more accurate score interpretations. As an example, the researcher can examine the effects of increasing the number of items or number of raters, or both, on the generalizability of the scores. Generalizability coefficients are estimated to examine the extent to which the scores generalize to the larger construct domain for relative or absolute decisions, or both.

Generalizability studies have shown that error due to raters for science hands-on performance tasks (e.g., Shavelson et al., 1993) and mathematics-constructed response items (Lane et al., 1996) tends to be smaller than for writing assessments (Dunbar et al., 1991). To help achieve consistency among raters, attention is needed in the design of well-articulated scoring rubrics, selection and training of raters, and evaluation of rater performance prior to and throughout operational scoring of student responses (Lane & Stone, 2006; Linn, 1993; Mehrens, 1992). Researchers have shown that task-sampling variability as compared to rater-sampling variability in students' scores is a greater source of measurement error in science, mathematics, and writing performance assessments (Baxter, Shavelson, Herman, Brown, & Valdadez, 1993; Gao, Shavelson, & Baxter, 1994; Hieronymus & Hoover, 1987; Lane et al., 1996; Shavelson et al., 1993). In other words, increasing the number of tasks in an assessment has a greater effect on the generalizability of the scores than increasing the number of raters scoring student responses.

Shavelson and his colleagues (1993) reported that task-sampling variability was the major source of measurement error using data from mathematics- and science-performance assessments. The results of their generalizability studies on a math assessment and two science assessments indicated that the person x task variance component accounted for the largest percentage of total score variation, approximately 49 percent, 48 percent, and 82 percent, respectively. This indicates that students were responding differently across the performance tasks. The variance components that included raters (i.e., rater effect, person x rater interaction, and task x rater interaction) were either zero or negligible, indicating that sampling variability due to raters contributed little to no measurement error. They reported that to reach a .80 generalizability coefficient 15 tasks were needed for the math assessment, 8 for the state science assessment, and 23 for the other hands-on science performance assessment. Lane and her colleagues (1996) found similar results with a mathematics-performance assessment that consisted of constructed-response items requiring students to show their solution processes and explain their reasoning. The results indicated that error due to raters was negligible, whereas error due to tasks was more substantial indicating that there was differential student performance across tasks. Generalizability studies for each form of the mathematics assessment indicated that between 42 percent and 62 percent of the total score variation was accounted for by the person x task variance component. Again, persons were responding differently across tasks due to task specificity. The variances due to the rater effect, person x rater interaction, and rater x task interaction were negligible. When the number of tasks was equal to 9, the generalizability coefficients ranged from .71 to .84. They also examined the generalizability of school-level scores for each form. The coefficients for absolute decisions (e.g., standards-based decisions) ranged from .80 to .97 when the number of tasks was equal to 36 using a matrix sampling design, providing evidence that the assessment allowed for accurate generalizability of grade-level scores for the schools.

Shavelson and his colleagues (Shavelson et al., 1993; Shavelson, Ruiz-Primo, and Wiley, 1999) provided evidence that the large task sampling variability in science-performance assessments was due to variability in both the person x task interaction and the person x task x occasion interaction. They conducted a generalizability study using data from a science-performance assessment (Shavelson et al., 1993). The person x task variance component accounted for 32 percent of the total variability, whereas, the person x task x occasion variance component accounted for 59 percent of the total variability. The latter suggests that students performed differently on each task from occasion to occasion. Shavelson and his colleagues (1999) provided additional support for the large effects due to occasion. In their generalizability study, the person x task variance component accounted for 26 percent of the total variability and the person x task x occasion variance component accounted for 31 percent of the total variability, indicating that there was a tendency for students to change their approach to each task from occasion to occasion. The variance component for the person x occasion effect was close to zero. In summary, "even though students approached the tasks differently each time they were tested, the aggregate level of their performance, averaged over the tasks, did not vary from one occasion to another" (Shavelson et al., 1999, pp. 64-65).

In summary, the results from generalizability studies indicate that scoring rubrics and the procedures used to train raters can be designed so as to minimize rater error. Further, the use of well-designed automated scoring systems allows for consistent application of the scoring rubrics in evaluating student work. Also, increasing the number of performance tasks will increase the generalizability of the scores. Likewise, including other item formats on performance assessments will aid in the generalizability of scores to the broader content domain.

Fairness of Assessments

The evaluation of the fairness of an assessment is inherently related to all sources of validity evidence. Bias can be conceptualized "as differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (Cole & Moss, 1989, p. 205). A fair assessment therefore requires evidence to support the meaningfulness, appropriateness, and usefulness of the test score inferences for all relevant subgroups of examinees. Validity evidence for assessments that are intended for students from various cultural, ethnic, and linguistic backgrounds needs to be collected continuously and systematically as the assessment is being developed, administered, and refined. The linguistic demands on items can be simplified to help ensure that ELLs are able to access the task as well as other students. As Abedi and Lord (2001) have demonstrated through their language modification approach, simplifying the linguistic demands on items can narrow the gap between ELLs and other students. The contexts used in mathematics tasks can be evaluated to ensure that they are familiar to various subgroups and will not negatively affect the performance on the task for one or more subgroups. The amount of writing required on mathematics, reading, and science assessments, for example, can be examined to help ensure that writing ability will not unduly influence the ability of the students to demonstrate what they know and can do on these assessments. Scoring rubrics can be designed to ensure that the relevant math, reading, or science skills are the focus, and not students' writing ability. The use of other response formats, such as graphic organizers, on reading assessments may alleviate the concerns of writing ability confounding student performance on reading assessments (O'Reilly & Sheehan, in press).

Some proponents of performance assessments in the 1980s had hoped that subgroup differences that were exhibited on multiple-choice tests would be smaller or alleviated by using performance assessments. However, as stated by Linn and his colleagues (1991), differences among subgroups most likely occur because of differences in learning opportunities, familiarity, and motivation, and are not necessarily due to item format. Research that has examined subgroup differences has focused on both the impact of an assessment on subgroups by examining mean differences or differential group performance on individual items when groups are matched with respect to ability, that is, differential item functioning (Lane & Stone, 2006). Differential item functioning (DIF) methods are commonly used for examining whether individual test items are measuring a similar construct for different groups of examinees (e.g., gender and ethnic groups) of similar ability. Differential item functioning occurs when there is differential performance on an item for subgroups of students of approximately equal ability. The presence of DIF may suggest that inferences based on the test score may be less valid for a particular group or groups. Although researchers have argued that performance assessments offer the potential for more equitable assessments, performance assessments may measure construct-irrelevant features that contribute to DIF. Gender or ethnic bias could be introduced by the typical contextualized nature of performance tasks or the amount of writing and reading required. In addition, the use of raters to score responses to performance assessments could introduce another possible source of differential item functioning (see for example, Gyagenda & Engelhard, in press). Results from DIF studies can be used to inform the design of assessment tasks and scoring rubrics so as to help minimize any potential bias.

Some researchers have supplemented differential item functioning methods with cognitive analyses of student performances designed to uncover reasons why items behave differently across subgroups of students of approximately equal ability. In a study to detect DIF in a mathematics-performance assessment consisting of constructed-response items that required students to show their solution processes and explain their reasoning, using the *analyses of reasons* method, Lane, Wang, and Magone (1996) examined differences in students' solution strategies, mathematical explanations, and mathematical errors as a potential source of differential item functioning. They reported that, for those items that exhibited DIF and favored females, females performed better than their matched males because females tended to provide more comprehensive conceptual explanations and were more complete in displaying their solution strategies. They suggest that increasing the opportunities in instruction for students to provide explanations and show their solution strategies may help alleviate these differences. Ericikan (2002) examined differential item response performances among different language groups. In her research, she conducted linguistic comparisons across different language test versions to identify potential sources of differential item functioning. Her results suggest that care is needed in the writing of items so as to minimize linguistic demands of items. As Wilson (2005) has suggested, the inclusion of DIF parameters into measurement models would allow for a direct measurement of different construct effects such as using different solution strategies and different types of explanations or to capture linguistic differences.

Some research studies have shown both gender and ethnic mean differences on performance assessments that measure complex thinking skill. As an example, ethnic and gender differences in persuasive writing were observed by Gabrielson, Gordon and Engelhard (1995). Their results indicated that high school female students wrote higher-quality persuasive essays than male students, and white students wrote essays of higher quality than black students. The scores for conventions and sentence formation were more affected by gender and ethnic characteristics than the scores in content, organization, and style — which were consistent with results from Engelhard, Jr., Gordon, Walker, and Gabrielson (1994). These differences may be more reflective of differences in learning opportunities and motivation than true differences in ability, again suggesting the need for instruction to provide similar opportunities for all students.

More recently, studies have used advances in statistical models to examine subgroup differences so as to better control for student demographic variables and school level variables. One study examined the extent to which potentially heavy linguistic demands of a performance assessment might interfere with the performance of students who have English as a second language (Goldschmidt, Martinez, Niemi, and Baker, 2007). The results obtained by Goldschmidt and his colleagues (2007) revealed that subgroup differences on student written essays to a writing prompt were less affected by student background variables than a language arts commercially developed test consisting of multiple-choice items and some constructed-response items. The performance gaps between white students, English-only students, and traditionally disadvantaged students (e.g., ELLs) were smaller for the writing performance assessment than the commercially developed test (Goldschmidt et al., 2007). Thus, the performance of students on the writing assessment used in this study was less affected by student demographic variables than their performance on the commercially developed test. As the authors indicate, although these are promising results, additional research is needed to determine if they can be replicated in other settings and with other subgroups. In particular, students in this study had opportunities in instruction to craft written essays, and such learning opportunities may have led to the results because of the alignment between instructional opportunities and the writing performance assessment.

Consequential Evidence

The evaluation of both intended and unintended consequences of any assessment is fundamental to the validation of score interpretation and use (Messick, 1989). Because a major goal of performance assessments is to improve teaching and student learning, it is essential to obtain evidence of such positive consequences and any potentially negative consequences (Messick, 1994). As Linn (1993) stated, the need to obtain evidence about consequences is "especially compelling for performance-based assessments… because particular intended consequences are an explicit part of the assessment systems' rationale" (p. 6). Further, adverse consequences bearing on issues of fairness are particularly relevant because it should not be assumed that a contextualized performance task is equally appropriate for all students because:

> … contextual features that engage and motivate one student and facilitate his or her effective task performances may alienate and confuse another student and bias or distort task performance may alienate and confuse another student and bias or distort task performance. (Messick, 1994).

This concern can be addressed by a thoughtful design process in which fairness issues are addressed, including expert analyses of the tasks and rubrics as well as analyses of student thinking as they solve performance tasks with special attention to examining potential subgroup differences and features of tasks that may contribute to these differences.

Large-scale performance assessments that measure complex thinking skills have been shown to have a positive impact on instruction and student learning (Lane, Parke, & Stone, 2002; Stecher, Barron, Chun & Ross, 2000; Stein & Lane, 1996; Stone & Lane, 2003). In a study examining the consequences of Washington's state assessment, Stecher and his colleagues (Stecher, et al., 2000) indicated that approximately two thirds of fourth- and seventh-grade teachers reported that the state standards and the state assessment short-answer and extended-response items were influential in promoting better instruction and student learning. An important aspect of consequential evidence for performance assessments is the examination of the relationship between changes in instructional practice and improved student performance on the assessments. A series of studies examined the relationship between changes in instructional practice and improved performance on the MSPAP, which was comprised entirely of performance tasks that were integrated across content domains (Lane, et al., 2002; Parke, Lane, & Stone, 2006; Stone & Lane, 2003). The results revealed that teacher-reported, reform-oriented instructional features accounted for differences in school performance on MSPAP in reading, writing, mathematics, and science; and they accounted for differences in the rate of change in MSPAP school performance in reading and writing. The former suggests that schools in which teachers reported that their instruction over the years reflected more reform-oriented problem types and learning outcomes similar to those assessed by MSPAP had higher levels of school performance on MSPAP than schools in which teachers reported that their instruction reflected less reform-oriented problem types and learning outcomes. The latter suggests that increased reported use of reform-oriented performance tasks in writing and reading and a focus on the reading and writing learning outcomes in instruction was associated with greater rates of change in MSPAP school performance over a 5-year period. Support for these results in the mathematics domain was provided by a study conducted by Linn, Baker, and Betebenner (2002). They demonstrated that the slope of the trend lines for the math assessments on both NAEP and MSAP were similar, suggesting that the performance gains in Maryland are not specific to the content and format of either test, but, rather, are due to deepened mathematical understanding on the part of the students.

When using test scores to make inferences regarding the quality of education, contextual information is needed to inform the inferences and actions (Haertel, 1999). Stone and Lane (1993) indicated that a school contextual variable, percent free or reduced lunch (which is typically used as a proxy for SES), was significantly related to school-level performance on MSPAP in mathematics, reading, writing, science, and social studies. That is, schools with a higher percentage of free or reduced lunch tended to perform poorer on MSPAP. There was no significant relationship, however, between percent free or reduced lunch and growth on MSPAP at the school-level in four of the five subject areas—mathematics, writing, science and social studies. This result indicates that improved school performance on performance assessments like MSPAP is not affected by contextual variables such as SES (as measured by percent free or reduced lunch). In other words, school level growth on the science, math, writing and social studies performance assessment was *not* related to the percentage of students who were eligible for free or reduced lunches within the school.

Instructional Sensitivity

An assessment concept that can help inform the consequential aspect of validity is instructional sensitivity. Instructional sensitivity refers to the degree to which tasks are sensitive to improvements in instruction (Popham, 2003; Black & William, 2007). Performance assessments are considered to be vehicles that can help shape sound instructional practice by modeling to teachers what is important to teach, and to students what is important to learn. In this regard, it is important to evaluate the extent to which improved performance on an assessment is linked with improved instructional practices. To accomplish this, the assessments need to be sensitive to improvements of instruction. Assessments that may not be sensitive to well-designed instruction may be measuring something outside of instruction such as irrelevant constructs or learning that may occur outside of the school.

Two methods have been used to examine whether assessments are instructional sensitive: studies have either examined whether students have had the opportunity to learn (OTL) the material, or they have examined the extent to which differences in instruction affect performance on the assessment. In a study using a model-based approach to assessment design (Baker, 2007), it was found that student performance on a language-arts performance assessment was sensitive to different types of language instruction and was able to capture improvement in instruction (Niemi, Wang, Steinberg, Baker, & Wang, 2007). This study examined the effects of three different types of instruction (literary analysis, organization of writing, and teacher-selected instruction) on student responses to an essay about conflict in literary work. The results indicated that students who received instruction on literary analysis were significantly more able to analyze and describe conflict in literature that students in the other two instructional groups, and students who had direct instruction on organization of writing performed significantly better on measures of writing coherency and organization. These results provide evidence that performance assessments can be instructional-sensitive with respect to different types of instruction, and suggest the need to ensure alignment and coherency among curriculum, instruction, and assessment.

Evaluation of Additional Psychometric Characteristics of Performance Assessments

This section briefly discusses additional psychometric issues in the design of performance assessments. First, a brief presentation on the measurement models that have been developed for performance assessments and extended constructed-response items will be provided. Measurement models that account for rater effects will also be introduced. These types of models have been used successfully in large-scale assessment programs to account for rater error in the scores obtained when evaluating performance assessments, allowing for more valid score interpretations. This will be followed with a brief discussion on issues related to linking performance assessments.

Measurement Models and Performance Assessments

Item Response Theory (IRT) models are typically used to scale assessments that consist of performance tasks only and assessments that consist of both performance tasks and multiple-choice items. IRT involves a class of mathematical models that are used to estimate test performance based on characteristics of the items and characteristics of the examinees that are presumed to underlie performance. The models use one or more ability parameters and various item parameters to predict item responses (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). These parameters in conjunction with a mathematical function are used to model the probability of a score response as a function of ability.

The more commonly applied models assume one underlying ability dimension determines item performance (Allen & Yen, 1979), and accommodate ordinal response scales that are typical of performance assessments. They include the graded-response model (Samejima, 1969; 1996), the partial-credit model (Masters, 1982), and the generalized partial-credit model (Muraki, 1992). As an example, Lane, Stone, Ankenmann, and Liu (1995) demonstrated the use of the graded- response model with a mathematics performance assessment and Allen, Johnson, Mislevy, and Thomas (1994) discussed the application of the generalized, partial-credit model to NAEP, which consists of multiple-choice items and constructed-response items.

Performance assessment data may be best modeled by multidimensional item response theory

(MIRT), which allows for the estimation of student proficiency on more than one skill area (Reckase, 1997). The application of MIRT models to assessments that are intended to measure student proficiency on multiple skills can provide a set of scores that profile student proficiency across the skills. These scores can then be used to guide the instructional process so as to narrow gaps in student understanding. Profiles can be developed at the student level or the group level (e.g., class) to inform instruction and student learning. MIRT models are particularly relevant for performance assessments because these assessments are able to capture complex performances that draw on multiple dimensions within the content domain, such as procedural skills, conceptual skills, and reasoning skills.

Modeling Rater Effects Using IRT Models

As previously discussed, performance assessments require either human scorers or automated scoring procedures in evaluating student work. When human scorers are used, performance assessments are considered to be "rater-mediated' since they do not provide direct information about the domain of interest but, rather, mediated information through interpretations by raters (Engelhard, 2002). Engelhard (2002) provided a conceptual model for performance assessments in which the obtained score is not only dependent on the domain of interest (e.g., writing ability), but also on rater severity, difficulty of the task, and the structure of the rating scale (e.g., analytic versus holistic, number of score levels). Test developers exert control over the task difficulty and the nature of the rating scale; however, a number of potential sources of construct-irrelevant variance are introduced into the rating process by the raters, including differential interpretation of score scales, differential assignment of ratings to males and females, halo effects, and bias in rater interpretation of task difficulty (Engelhard, 2002). These sources of construct-irrelevant variance can affect the validity and fairness of score interpretations.

Models have been developed that account for rater variability in scoring performance assessments. As an example, Patz and his colleagues (Patz, 1996; Patz, Junker, & Johnson, 2002; and Patz, Junker, Johnson, & Mariano, 2002) developed a hierarchical rating model to account for the dependencies between rater judgments. A parameter was introduced into the model that could be considered an "ideal rating" or expected score for an individual, and raters could vary with respect to how close their rating is to this ideal rating. This variability reflects random error (e.g., lack of consistency) as well as systematic error (e.g., rater tendencies such as leniency). As discussed by Bejar, Williamson, and Mislevy (2006), this modeling of rater variability may reflect an accurate modeling of rater cognition in that, under operational scoring conditions, raters may try to predict the score an expert rater would assign based on the scoring rubric and benchmark papers. In addition, covariates can be introduced into the model to predict rater behaviors such as rater features (e.g., hours of scoring) and item features. The modeling of rater variability is a way to account for error in the scores obtained when evaluating performance assessments, and allows for more valid interpretations of the scores.

Equating and Linking Issues

Equating helps ensure comparability of interpretations of assessment results from assessment forms administered at one time or over time; however, equating an assessment that consists of only performance tasks is complex (Kolen & Brennan, 2004). One way to equate forms so that they can be used interchangeably is to have a common set of items, typically called anchor items, on each of the forms. The anchor items are then used to adjust for any differences in difficulty across forms. An important issue that needs to be addressed in using performance tasks as anchor items in the equating procedure is that rater teams could change their scoring standards over time and the application of standard equating practices would lead to bias in the equating process and consequently, inaccurate scores (Bock, 1995; Kim, Walker, & McHale, 2008a; Tate, 1999). As a solution to this problem, Tate (1999, 2000) suggested an initial linking study in which any changes in rater severity and discrimination across years could be identified. This would allow for an accurate assessment of across- rater team ability differences and equating of the tests. To accomplish the equating, a large representative sample of anchor item papers (i.e., trend papers) from Year 1 are rescored by raters in Year 2. These raters in Year 2 are the same raters who score the new constructed responses in Year 2. These trend papers now have a set of scores from the old raters in Year 1 and a set of scores from the new raters in Year 2. This allows for examining the extent to which the two rater teams across years differ in severity in assigning scores, and then adjustments can be made to ensure the two tests are on the same scale. Tate contends that, instead of having item parameters, there are item/rating team parameters that reflect the notion that, if the rating team changes across years, any differences due to the change in rating teams will be reflected in the item parameters. Another way to conceptualize this is that the item parameters are confounded by rater-team effects so the rating team needs to be considered in the equating.

The effectiveness of this IRT linking method using trend score papers was established by Tate and his colleague (Tate, 2003; Kamata & Tate, 2005). The use of trend score papers in non-IRT equating methods has also proven effective by Kim and colleagues (2008a; 2008b). They compared the effectiveness of equating for a design that required anchor items and a design that did not require anchor items with and without trend score papers. The design that does not incorporate anchor items alleviates the concern of content representativeness of anchor items. Their results indicated that both designs using trend score papers were more effective in equating the scores as compared to those designs that did not use the trend score papers. More importantly, their results indicate that changes in rater severity can be examined and the equating of test forms across years should adjust for differences in rater severity if the trend scoring indicates that a rater shift has occurred (Kim et al., 2008a, 2008b). Trend scoring should be implemented for any assessment program that uses constructed-response items for equating across years to control for equating bias caused by a scoring shift over years. Kim and colleagues (2008b) point out that the trend-scoring method requires additional rating of student papers which, in turn, increases cost; and it may be a bit cumbersome to implement. The use of image and online scoring methods, however, can ease the complexities of the implementation of the trend scoring method.

Conclusion

Performance assessments have been an integral part of educational systems in many countries; however, they have not been fully utilized in this country. There is evidence that the format of the assessment affects the type of thinking and reasoning skills that are used by students, with performance assessments being better suited to assessing high-level, complex thinking skills (e.g., Martinez & Katz, 1996). Recent advances in the design and scoring of performance assessments support their increased use in large-scale assessment programs. In particular, computer simulations allow for the design of meaningful, real world tasks that require students to problem solve and reason. Scores can be generated for computer-simulation tasks across a number of dimensions and reported in a timely manner given the advances in automated scoring systems. Automated scoring systems have also proven effective in the evaluation of student essays to writing prompts and short constructed-response items. Advances in the design of automated scoring systems will continue to support the increased use of computers for assessment design and delivery, allowing for a more integrative, comprehensive approach to assessment design.

Well-specified content standards that reflect high-level thinking and reasoning skills can guide the design of performance assessments so as to ensure the alignment among curriculum, instruction and assessment. Various task design strategies have proven useful in helping ensure the validity and fairness of performance-assessment results. The language-modification approach used by Abedi and Lord (2001) that minimizes the complexity of linguistic demands on mathematics items has led to improved performance for ELLs as well as other students. The use of computer-based reading comprehension items, where students use graphic displays to demonstrate their understanding, will help minimize the extent to which students' writing ability affects their scores on a reading comprehension assessment (O'Reilly & Sheehan, in press). Task templates can be designed so as to ensure tasks embody the intended cognitive demands, and are not measuring one or more irrelevant constructs. Task templates also have the potential to increase the production of tasks, especially for computer-based simulation tasks.

When human raters are used, well-articulated scoring rubrics and rigorous training procedures for raters will minimize error introduced in the scores due to inconsistency within and among raters. Measurement models and procedures have been designed to model rater errors and inconsistencies so as to control them in the estimation of student scores on performance assessments. Procedures and measurement models have also been developed to account for changes in rater performance over years and to adjust student scores due to these rater changes. This is important because any changes in student performance over the years should reflect actual changes in proficiency and learning, and not be a result of inconsistent scoring on the part of the raters. Further, if raters become more stringent over years, this could mask any improved student performance on the assessment.

The educational benefit of using performance assessments has been demonstrated by a number of researchers (Lane et al., 2002; Niemi et al., 2007; Stecher et al., 2000). When students are given the opportunity to work on meaningful, real world tasks in instruction, students have demonstrated improved performance on performance assessments. Moreover, research has shown that growth on performance assessments at the school level is not related to SES variables. Sound educational practice begs for the alignment among curriculum, instruction, and assessment, and there is ample evidence to support the use of performance assessments in both instruction and assessment to improve student learning for all students.

References

Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Abedi, Lord, Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

Abedi, J., Lord, C., & Plummer, J. (1997). *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.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

Allen, N. L., Johnson, E. G., Mislevy, R. J., & Thomas, N. (1994). Scaling procedures. In N. J. Allen, D. L. Kline & C. A. Zelenak (Eds.), *The NAEP 1994 technical report* (pp. 247-266). Washington, D. C. : U. S. Department of Education.

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

Attali, Y., & Burstein, J. (2005). *Automated Essay Scoring with E-Rater v. 2. 0.* (ETS Research Report RR-04-45), Princeton, NJ: Educational Testing Service.

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

Baker, E. L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment*, *12*(3&4), 179-194.

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*(12), 1210-1218.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305-318.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*(3), 37-45.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A. & Valdadez J. R., (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education*, *24*, 190-216.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human Scoring (pp. 49-82). In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing*. Hillside, NJ: Erlbaum.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful essay scoring. *Journal of Technology, Learning and Assessment 6*(1). Retrieved July 15, 2009, from http://escholarship.bc.edu/jtla/.

Bennett, R. E., (2006). Moving the field forward: Some thoughts on validity and automated scoring (pp. 403-412). In D. M. Williamson, R. J. Mislevy, & I. I. Behar (Eds.), *Automated scoring of complex tasks in computer-based testing*. Hillside, NJ: Erlbaum.

Bennett, R. E., & Bejar, I. (1997). *Validity and automated scoring: It's not only the scoring.* (ETS RR-97-13). Princeton, NJ: Educational Testing Service.

Bennett, R. E. & Gitomer, D. H. (in press). Transforming K-12 Assessment: Integrating Accountability Testing, Formative Assessment and Professional Support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century.* New York: Springer.

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007-466). Washington, DC: National Center for Education Statistics, U. S. Department of Education. Retrieved May 2, 2009, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466.

Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80,* 139-148.

Black, P. & William, D. (2007). Large scale assessment systems: Design principles drawn from international comparisons. *Measurement: Interdisciplinary Research and Perspectives, 5*(1), 1-53.

Bock, R. D. (1995). Open-ended exercise in large-scale educational assessment. In L. B. Resnick and J. G. Wirt (Eds.), *Linking school and work: Roles for standards and assessment.* (pp. 305-338). San Francisco, CA: Jossey-Bass.

Breland, H., Danos, D., Kahn, H., Kubota, M. & Bonner, M. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement History Examination. *Journal of Educational Measurement, 31*(4), 275-293.

Breland, H. M. & Jones, R. J. (1982). *Perceptions of writing skills* (College Board Report No. 82-4 and ETS Research Report No. 82-47). New York, NY: College Entrance Examination Board.

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment.* Washington, DC: National Center for Education Statistics (NCES 96-802).

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339-353.

Brennan, R. L. (2001). *Generalizability Theory.* New York: Springer-Verlag.

Bridgeman, B., Trapani, C., & Attali, Y. (2009). *Considering fairness and validity in evaluating automated scoring.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Burstein, J. (2003). The E-rater scoring engine: Automated essay scoring with Natural Language Processing. In. M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring* (pp. 113-122). Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155-192.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems of experts and novices. *Cognitive Science*, *5*, 121-152.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.) (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24*(4), 310-324.

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1995). Computer-based case simulations. In E. L. Mancall & P. G. Bashook (Eds.), Assessing clinical reasoning: The oral examination and alternative methods (pp. 139-149). Evanston, IL: American Board of Medical Specialties.

Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement*, (third ed., pp. 201-220). New York: American Council on Education and Macmillan.

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*(3), 373-399.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of school and students at work.* New York City: NY, Teachers' College Press.

Deane, P. (2006). Strategies for evidence identification through linguistic assessment of textual responses. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing.* (pp. 313-362). Mahwah, NJ: Lawrence Erlbaum Associates. Deane, P. & Gurevich, O. (2008). *Applying content similarity metrics to corpus data: Differences between native and non-native speaker response to a TOEFL integrated writing prompt.* (ETS Research Report No. RR-08-5). Princeton, NJ: ETS.

Delaware Department of Education. (2000, November). *Delaware student testing Program Special Writing Study Report.* Retrieved July 5, 2009, from http://www.doe.k12.de.us/aab/report_special_writing%20study.pdf.

Delaware Department of Education. (2005). *Text-based writing item sampler.* Retrieved July 5, 2009, from http://www.doe.k12.de.us/AAB/files/Grade%208%20TBW%20-%20Greaseaters.pdf.

DeVore, R. N. (2002). *Considerations in the development of accounting simulations.* (Technical Report 13). NJ: AICPA.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*(4), 289-304.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Educational Testing Service (2004, July 4). *Pretesting plan for SAT essay topics.* [internal communication].

Elliott, S. (2003). Intellimetric: From Here to Validity. In. M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S. E. (1985). *Test Design: Developments in Psychology and Psychometrics.* Orlando, FL: Academic Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Mahwah, NJ: Erlbaum.

Engelhard, Jr., G., Gordon, B., Walker, E. V., & Gabrielson, S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *Journal of Educational Research, 87,*197-209.

Ericikan, K. (2002) Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing*, 2, 199-215.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis*: *Verbal reports as data.* Cambridge, MA: MIT Press.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 1-38). Cambridge: Cambridge University Press.

Ferrara, S. F. (1987, April). *Practical considerations in equating a direct writing assessment required for high school graduation.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75-98). New York: Longman.

Gabrielson, S., Gordon, B., & Engelhard, G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education, 8*(4), 273-290.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science. Promises and problems. *Applied Measurement in Education*, *7*, 323-334

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. R. Ronning, J. A., Glover, J. C., Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology on testing* (pp. 41-85). Hillsdale, NJ: Erlbaum.

Goldberg, G. L. & Roswell, B. S. (2001). Are multiple measures meaningful?: Lessons learned from a statewide performance assessment. *Applied Measurement in Education, 14*(2), 125-150.

Goldschmidt, P., Martinez, J. F., Niemi, D., & Baker, E. L. (2007) Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment, 12*(3 & 4), 239-266.

Gotwals, A. W., & Songer, N. B. (2006). *Cognitive Predictions: BioKIDS Implementation of the PADI Assessment System.* (PADI Technical Report 10). Menlo Park, CA: SRI International.

Greeno, J. G. (1989). A perspective on thinking. *American Psychologist*, 44, 134-141.

Gyagenda, I. S., & Engelhard, G. (in press). Rater, domain, and gender influences on the assessed quality of student writing. In Garner, M., Engelhard, G., Wilson, & M., Fisher, W. (Eds.). Advances in Rasch Measurement, Volume One. JAM Press.

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan*, *80*(9), 662-667.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical Issues in Large-Scale Performance Assessment* (NCES 96-802). Washington, DC: U. S. Department of Education.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kuwer-Nijhoff.

Hieronymus, A. N., & Hoover, H. D. (1987). *Iowa tests of basic skills: Writing Supplement teacher's guide.* Chicago: Riverside.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering, 12*(2), 145-159.

Huot, B. (1990). The literature of direct writing assessments: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263.

Kane, M. T. (2006). Validation. In B. Brennan (Ed.), *Educational Measurement*. American Council on Education & Praeger: Westport, CT.

Kamata, L., & Tate, R. L. (2005). The performance of a method for the long-term equating of mixed format assessment. *Journal of Educational Measurement, 42,* 193-213.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kim, S., Walker, M. E., & McHale, F. (2008a, May). *Equating of mixed-format tests in large-scale assessments.* (ETS Research Report-08-26). Princeton, NJ: ETS.

Kim, S., Walker, M. E., & McHale, F. (2008b, October). *Comparisons among designs for equating constructed-response items.* (ETS Research Report-08-53). Princeton, NJ: ETS.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., et al. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, *11*(2), 121-137.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices.* (second ed.), New York, NY. Springer.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated Essay Scoring: A cross disciplinary perspective. In M. D. Shermis and J. Burstein (Eds.), *Automated Essay Scoring and annotation of essays with the Intelligent Essay Assessor* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates.

Lane, S. (in press). Issues in the design and scoring of performance assessments that assess complex thinking skills. In G. Schraw (Ed.). *Assessment of Higher Order Thinking Skills.*

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment instrument. *Educational Measurement: Issues and Practice*, *12*(3), 16-23.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *33*(1), 71-92.

Lane, S., Parke, C. S., & Stone, C. A. (2002). The Impact of a State Performance-Based Assessment and Accountability Program on Mathematics Instruction and Student Learning: Evidence from Survey Data and School Performance. *Educational Assessment, 8*(4), 279-315.

Lane, S., Silver, E. A., Ankenmann, R. D., Cai, J., Finseth, C., Liu, M., et al. (1995). *QUASAR Cognitive Assessment Instrument (QCAI)*. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Lane, S., & Stone, C. A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational Measurement*. American Council on Education & Praeger: Westport, CT.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8*, 313-340.

Lane, S., Wang, N., & Magone, M. (1996). Gender related DIF on a middle school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*, 21-27,31.

Leacock, C. & Chodorow, M. (2003). C-rater: Automated scoring of short answer questions. *Computers and humanities, 37*(4), 389-405.

Leacock, C. & Chodorow, M. (2004). *A pilot study of automated scoring of constructed responses.* Paper presented at the 30th Annual International Association of Educational Assessment conference, Philadelphia, PA.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1-16.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Liu, O. L., Lee, H., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Constructs, measures, and evidence. *Educational Assessment, 13*(1), 33-55.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, and judging* (pp. 33-60). Urbana, IN: National Council for Teachers in Education.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt: Lang.

Martinez, M. E., & Katz, I. R. (1996). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. *Educational Assessment, 3*(1), 83-98.

Maryland State Board of Education (1995). *Maryland school performance report*: *State and school systems.* Baltimore, MD: author.

Maryland State Department of Education. (1990). Technical report: Maryland Writing Test, Level II. Baltimore, MD: Author. Retrieved August 1, 2009, from http://www.marces.org/mdarch/htm/M031987.HTM.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

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

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*, 3-9, 20.

Messick, S. (1989) Validity. In R. L. Linn (Ed.), *Educational Measurement*, (third ed., pp. 13-104). New York: American Council on Education and Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (1-18). Washington, DC: National Center for Educational Statistics.

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education, 3*(3), 285-296.

Mislevy, R. J. (1993). Foundations of a new theory. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19-39).

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379-416.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3-62.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*(4), 363-390.

Mullis, I. V. S. (1984). Scoring direct writing assessments: What are the alternatives? *Educational Measurement: Issues and Practice, 3*(1), 16-18.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system*. (Center for Performance Assessment Research Report). Princeton, NJ: Educational Testing Service.

National Council on Education Standards and Testing (1992). *Raising standards for American Education*. Washington, DC: Author.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council (2006). *Systems for State science assessment*. M. R. Wilson & M. W. Bertenthal (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nelson, N., & Calfee, R. C. (1998). The reading-writing connection. In N. Nelson & R. C. Calfee (Eds.), *The reading-writing connection*. University Chicago Press.

Niemi, D., Baker, E. L., & Sylvester, R.M. (2007). Scaling up, scaling down: Seven years of performance assessment devolpment in the nation's second largest school district. *Educational Assessment, 12*(3 & 4), 195-214.

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L., & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment, 12*(3 & 4), 215-238.

O'Reilly, T., & Sheehan, K. M. (in press). Cognitively based assessment of, for and as learning: a framework for assessing reading competency (RR-xx-xx). Princeton, NJ: Educational Testing Service.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*(2), 127-143.

Page, E. B. (2003). Project Essay Grade: PEG. In M. Shermis & J. Burstein, J. (Eds.). *Automated essay scoring: Across-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Erlbaum.

Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation, 12*(3), 239-269.

Patz, R. J. (1996). *Markov Chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress*. Unpublished manuscript, Carnegie Mello University, Pittsburgh PA.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and is application to large-scale educational assessment data. *Journal of educational and Behavioral Statistics, 27*(4), 341-384.

Popham, W. J. (2003). Living (or dying) with your NCLB tests. *School Administrator, 60*(11), 10-14.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Stumpinge-rater: Challenging the validity of automated scoring of essays. *Journal of Educational Computing Research, 26,* 407-425.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Resnick, L. B., & Resnick, D. P. (1982). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Conner (Eds.)., *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37-55), Boston: Kluwer Academics.

Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct-writing assessments. *Applied Measurement in Education, 7*(2), 159-170.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric^{SM} essay scoring system. *The Journal of Technology, Learning, and Assessment, 9*(4), 1-21.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*(No. 17).

Samejima, F. (Ed.). (1996). *The graded response model.* New York: Springer.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232.

Shavelson, R. J., & Ruiz-Primo, M. A. (November 1998). *On the Assessment of Science Achievement Conceptual Underpinnings for the Design of Performance Assessments: Report of Year 2 Activities.* (CSE Technical Report 481). Los Angeles, CA: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability. *Journal of Educational Measurement*, *36*(1), 61-71.

Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Scientist, 61*, 394-403.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement, 14*(1&2), 1-98.

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000, August). *The effects of the Washington state education reform in schools and classrooms* (CSE Tech. Rep. NO. 525). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, *2*(1), 50-80.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1-26.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36,* 336-346.

Tate, R. L., (2000). Performance of a proposed method for the linking of mixed format tests with constructed-response and multiple-choice items. *Journal of Educational Measurement, 36,* 336-346.

Tate, R. L. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and Psychological Measurement, 63*(6), 893-914.

U. S. Department of Education (2005). The Nation's Report Card. Washington, DC: Author Retrieved July, 2009, from http://nationsreportcard.gov/science_2005/s0116. asp.

Vacc, N. N. (1989). Writing evaluation: Examining four teachers' holistic and analytic scores. *The Elementary School Journal,90,* 87-95.

Vendlinski, T. P., Baker, E. L., & Niemi, D. (2008). *Templates and objects in authoring problem-solving assessments.* (CRESST Tech. Rep. No. 735). Los Angeles: University of California, National Center Research on Evaluation, Standards, and Student Testing (CRESST).

Welch, C. J. & Harris, D. J. (1994). A Technical Comparison of Analytic and Holistic Scoring Methods. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Williamson, D. M., Behar, I. I., & Mislevy, R. J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D. M. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.) *Automated scoring of complex tasks in computer-based testing.* (pp. 1-14). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

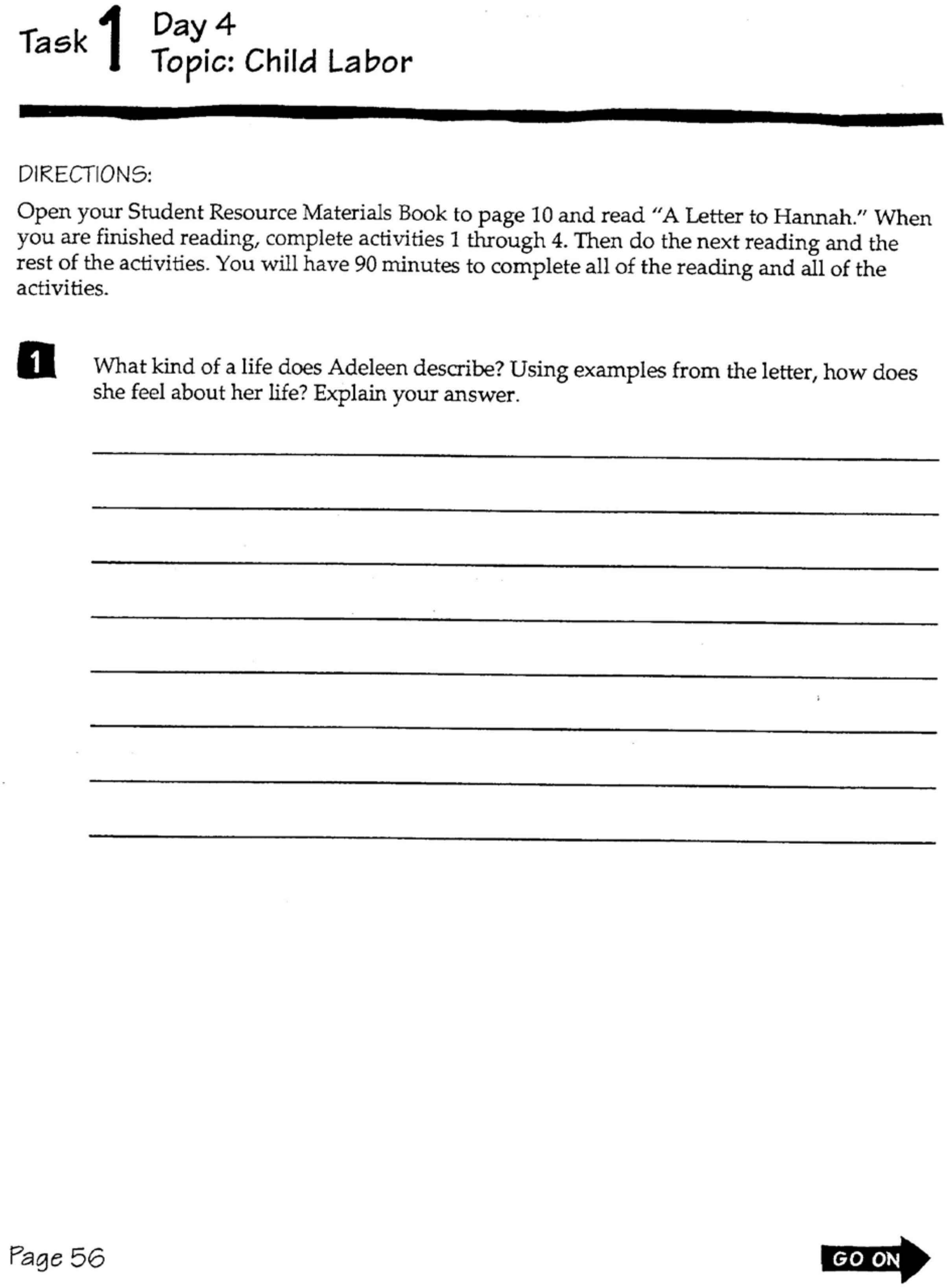
Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181-208.

Wolfe, E. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83-106.

Yang, Y., Buchendahl, C. W., Juszkiewicz, P. J., Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391-412.

Appendix A: MSPAP Performance Task



Task 1 Day 4 Topic: Child Labor

DIRECTIONS:

Open your Student Resource Materials Book to page 10 and read "A Letter to Hannah." When you are finished reading, complete activities 1 through 4. Then do the next reading and the rest of the activities. You will have 90 minutes to complete all of the reading and all of the activities.

1 What kind of a life does Adeleen describe? Using examples from the letter, how does she feel about her life? Explain your answer.

____

____

____

____

____

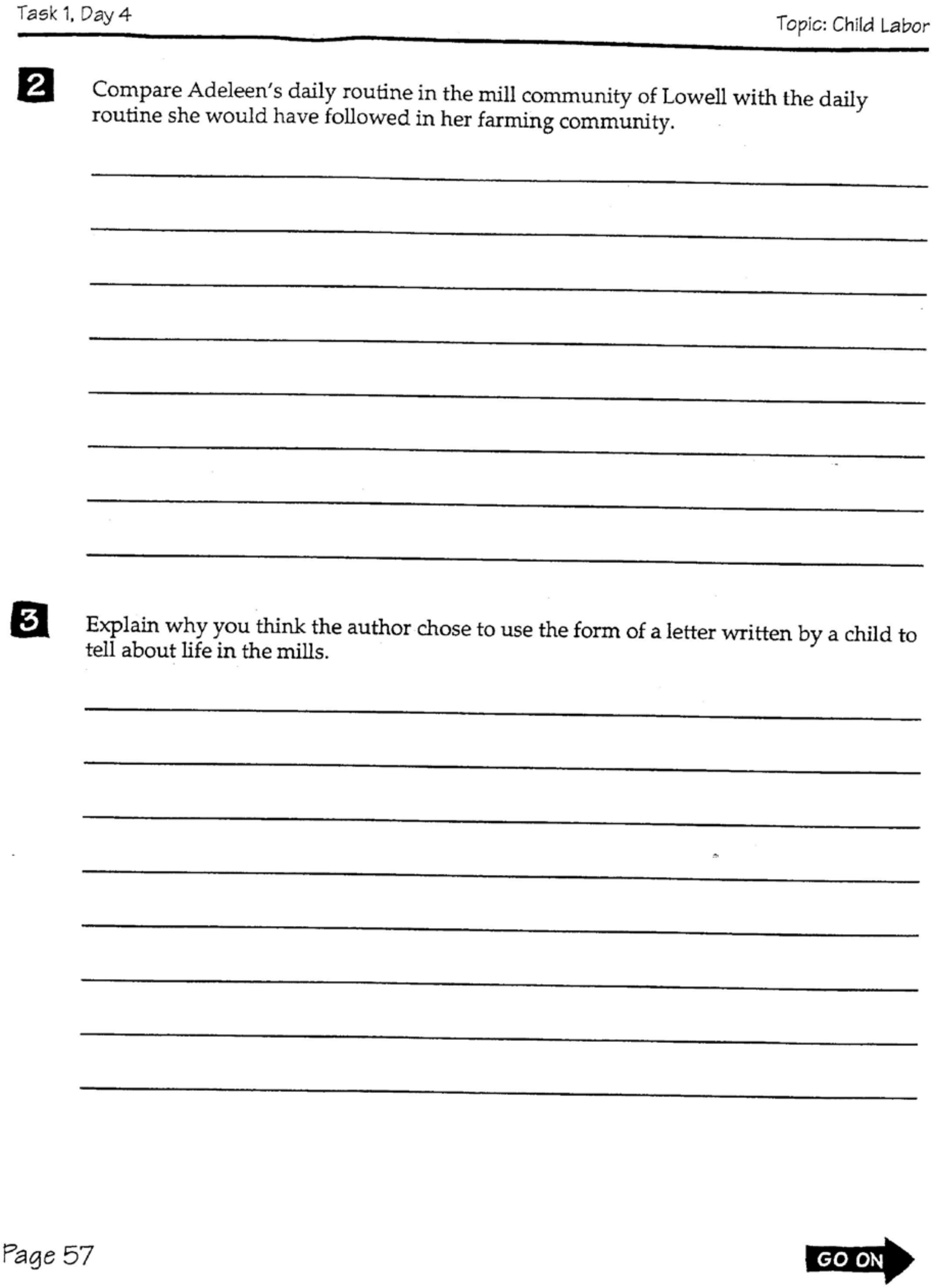
____

____

____

Page 56

GO ON



Task 1, Day 4

Topic: Child Labor

2 Compare Adeleen's daily routine in the mill community of Lowell with the daily routine she would have followed in her farming community.

____

____

____

____

____

____

____

____

3 Explain why you think the author chose to use the form of a letter written by a child to tell about life in the mills.

____

____

____

____

____

____

____

____

Page 57

GO ON

4 Your class is preparing a scrapbook about jobs that children have had. Your task is to write a paragraph telling whether or not you would have been satisfied living as Adeleen did. Use information from the reading and your personal experience to write your paragraph.

Because your paragraph will be read by your classmates and published in the scrapbook, be sure your paragraph is clear and complete. Also, check for correct spelling, grammar, punctuation, and capitalization.

____

____

____

____

____

____

____

____

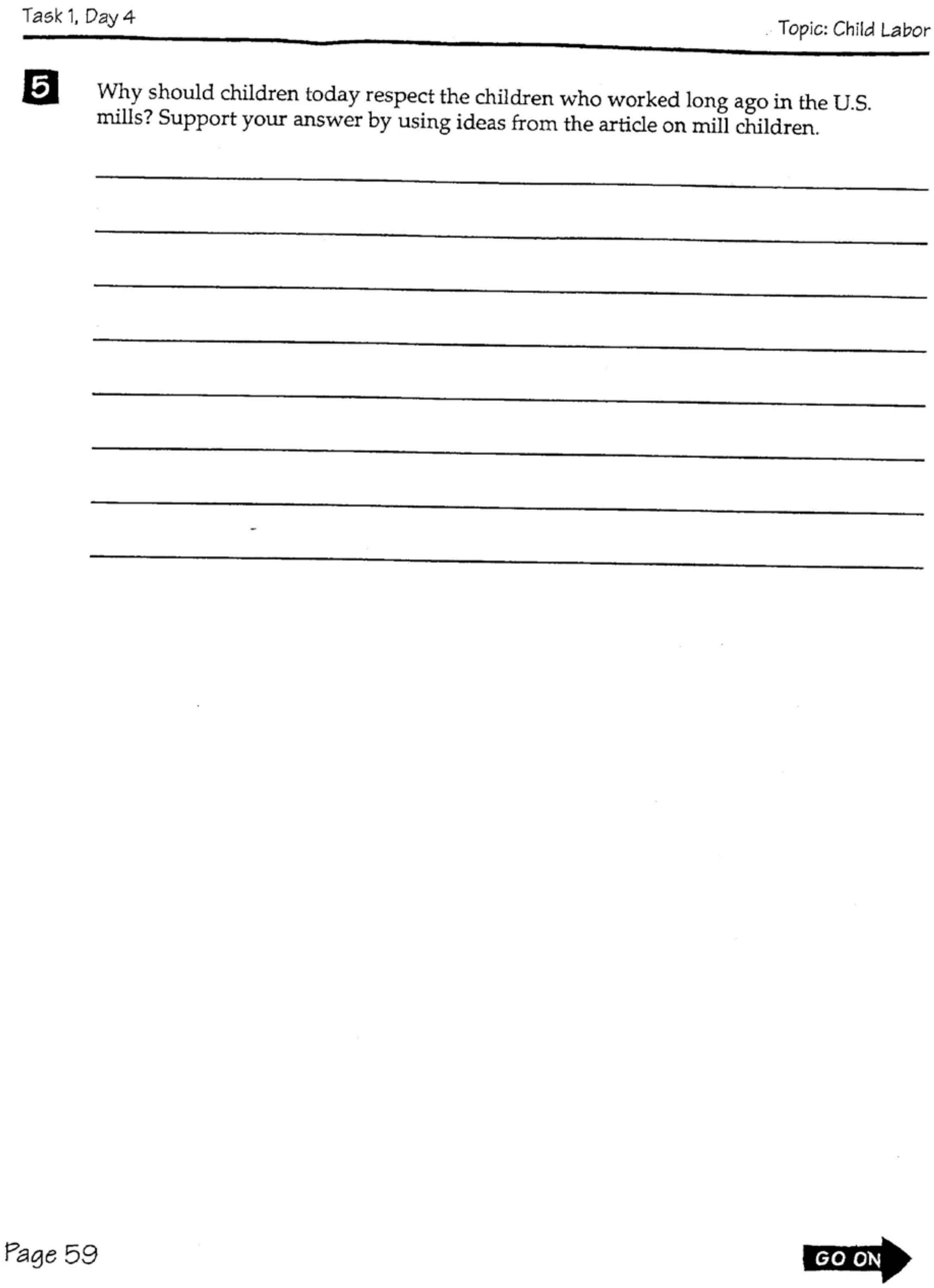
____

____

____

Open your Student Resource Materials Book to page 26 and read the article entitled "Mill Children." Then complete activities 5 through 15.

GO ON



Task 1, Day 4

Topic: Child Labor

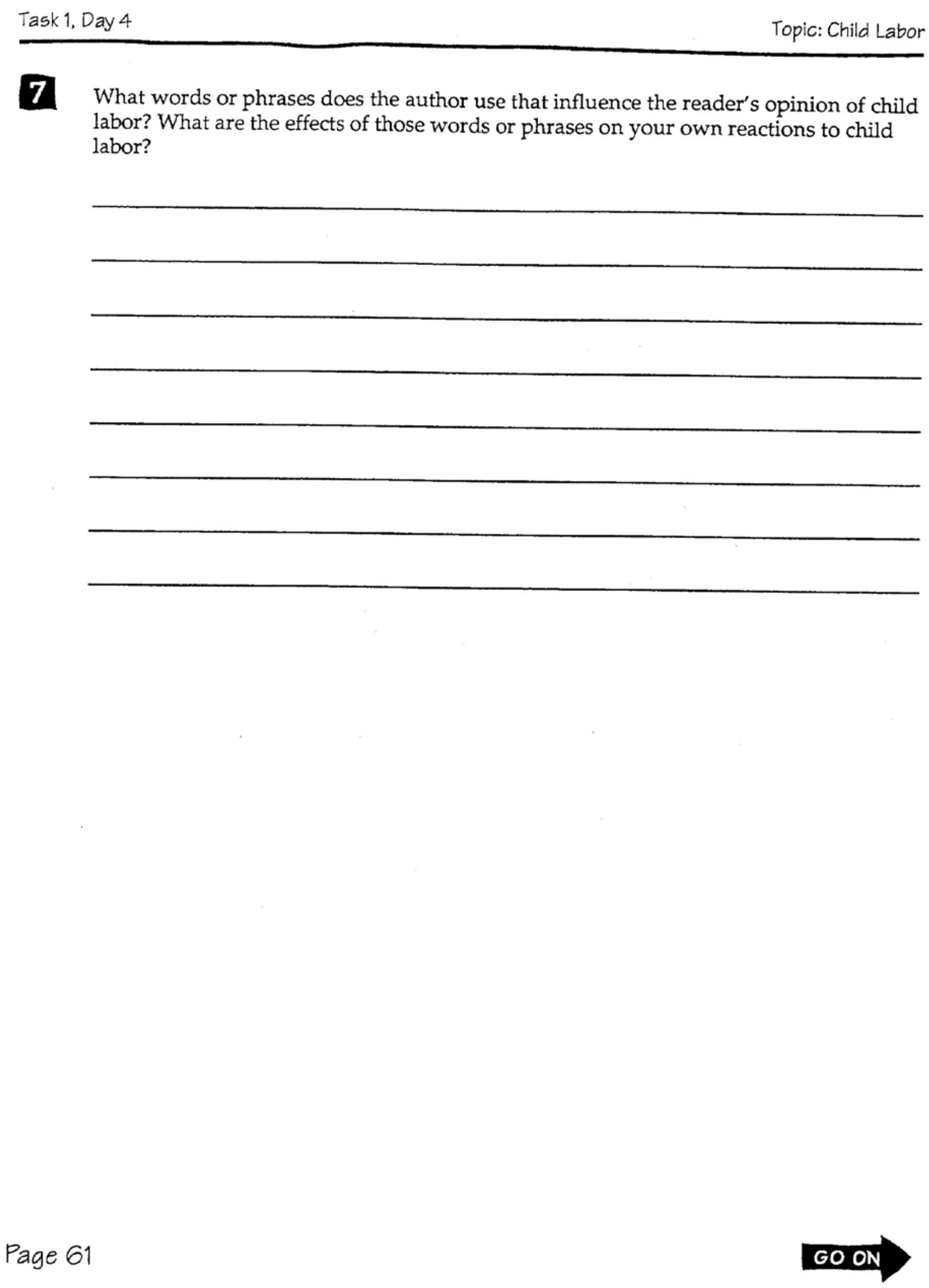
5 Why should children today respect the children who worked long ago in the U.S. mills? Support your answer by using ideas from the article on mill children.

Page 59

GO ON

6 For next year, your teacher is thinking about using only one of these articles to explain life in the mills and children's rights. Which article would you recommend? Write a note to your teacher explaining your recommendation. Be sure to give information from both of the articles you have read that helps your teacher understand why your choice is the best. Because your note will be read by your teacher, be sure your note is clear and complete. Also, check for correct spelling, grammar, punctuation, and capitalization.

GO ON



Task 1, Day 4

Topic: Child Labor

7 What words or phrases does the author use that influence the reader's opinion of child labor? What are the effects of those words or phrases on your own reactions to child labor?

____

____

____

____

____

____

____

____

Page 61

GO ON

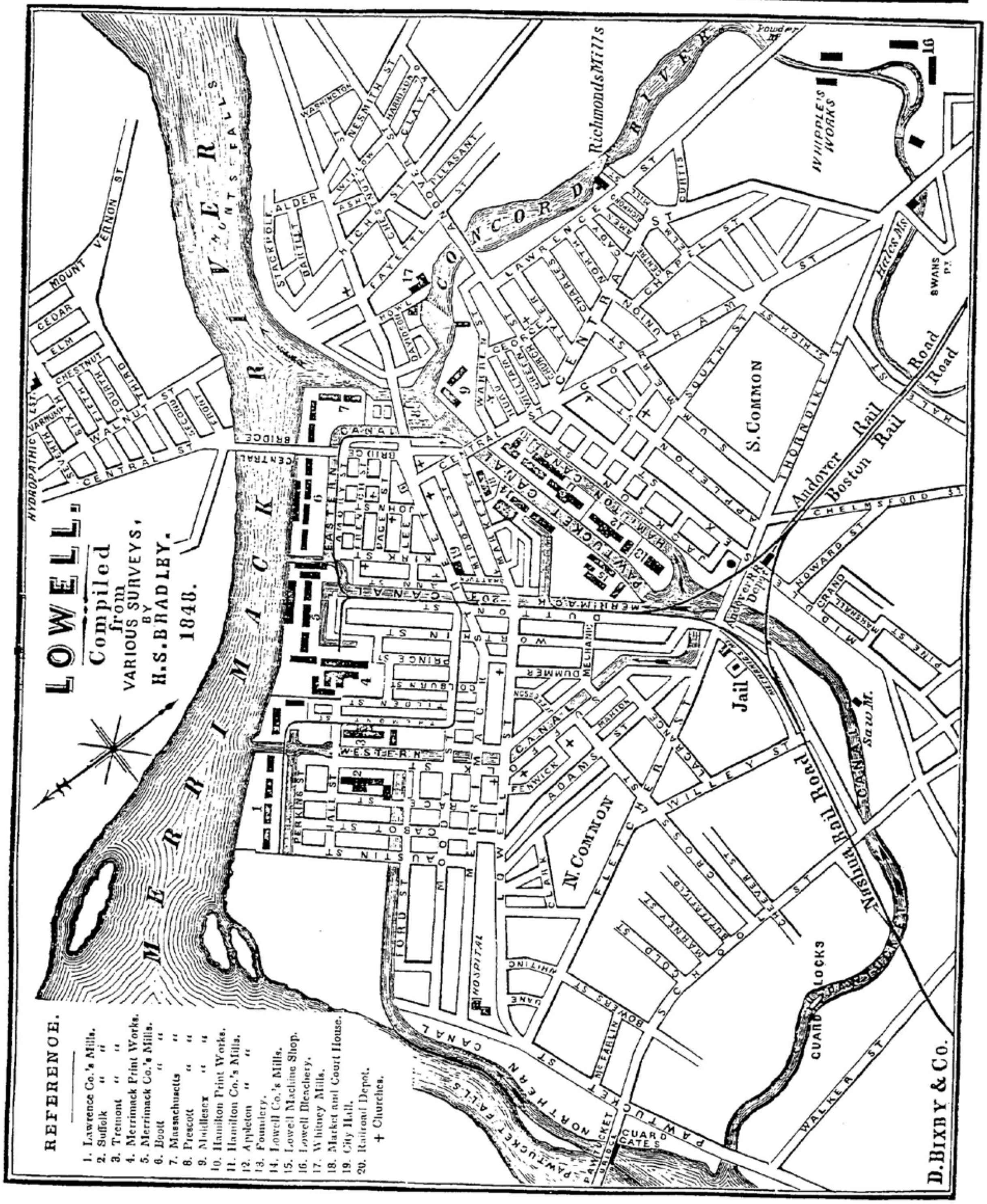
8 Circle the word below which is closest to how you felt after reading "Mill Children."

Angry Confused Fortunate Sad Hopeful

Use information from the text and your personal experience to explain your answer.

9 Now look at the map of Lowell, Massachusetts on the next page, where many textile mills were located. Using this map, locate the Boott Company's mill and mark it with an "X."

GO ON


LOWELL,
Compiled
from
VARIOUS SURVEYS,
BY
H.S.BRADLEY.
1848.
REFERENCE.
1. Lawrence Co.'s Mills.
2. Suffolk " "
3. Tremont " "
4. Merrimack Print Works.
5. Merrimack Co.'s Mills.
6. Boott " "
7. Massachusetts " "
8. Prescott " "
9. Middlesex " "
10. Hamilton Print Works.
11. Hamilton Co.'s Mills.
12. Appleton " "
13. Foundery.
14. Lowell Co.'s Mills.
15. Lowell Machine Shop.
16. Lowell Bleachery.
17. Whitney Mills.
18. Market and Court House.
19. City Hall.
20. Railroad Depot.
+ Churches.
MERRIMACK RIVER
CONCORD RIVER
PAWTUCKET FALLS
S. COMMON
N. COMMON
Jail
Richmonds Mills
WHIPPLE'S WORKS
Saw M.
Powder M.
Boston Rail Road
Andover Rail Road
Nashua Rail Road
PAWTUCKET CANAL
NORTHERN CANAL
WESTERN CANAL
EASTERN CANAL
MERRIMACK CANAL
HAMILTON CANAL
GUARD LOCKS
GUARD GATES
CENTRAL BRIDGE
HOSPITAL
D. BIXBY & Co.

Page 63

GO ON

10 Open your Student Resource Materials Book to page 27 and reread the 5th paragraph of the article "Mill Children."

Then, using the map of Lowell, list at least 2 natural geographic features of this place.

____

____

____

____

11 Tell why these features made this area a good location for a mill town.

____

____

____

____

12 Suppose you were planning to build a textile mill in Maryland. Look at the map of Maryland on the next page. Draw a symbol on the map to show where the mill would be located.

Explain why the mill should be located there.

____

____

____

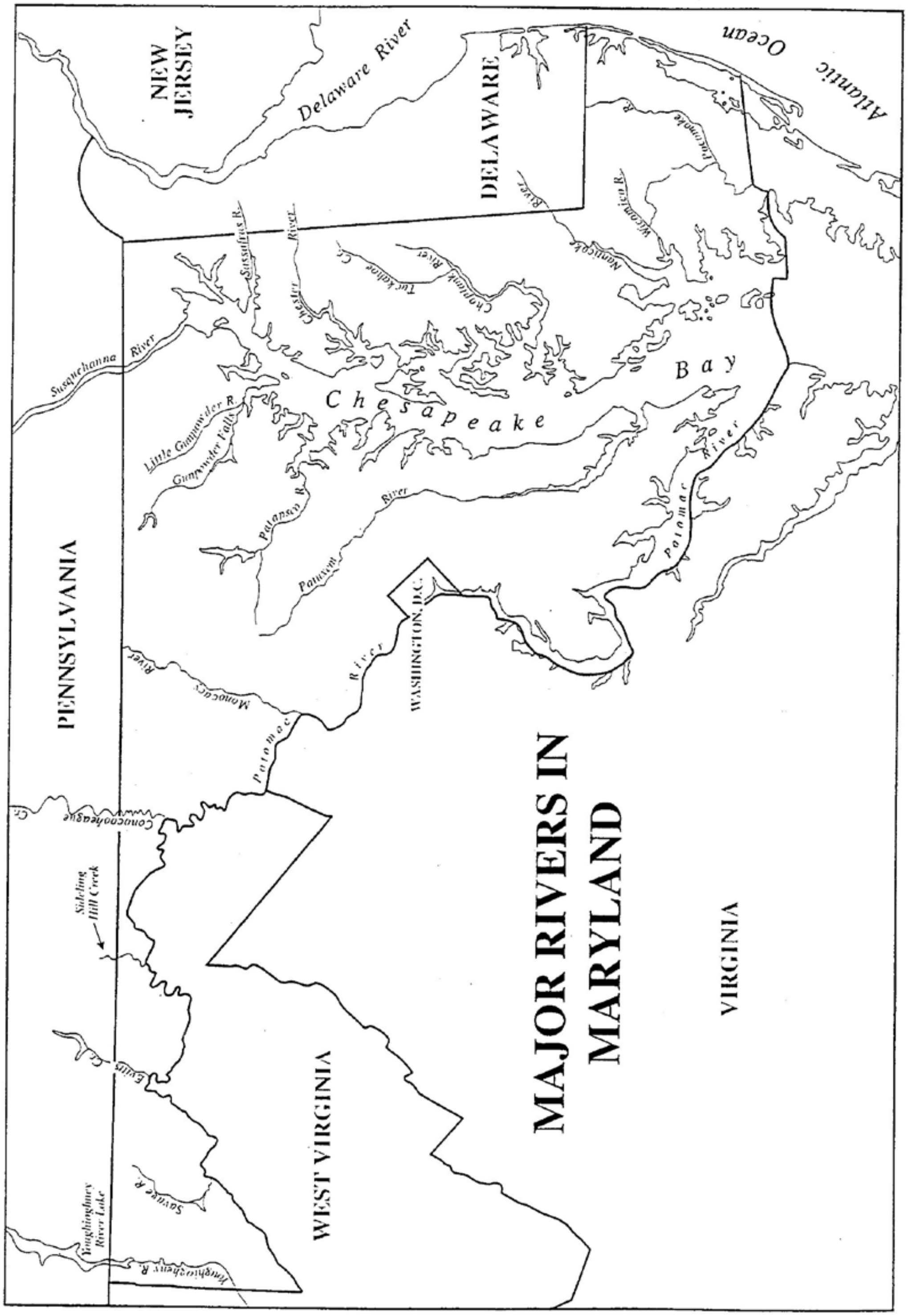
____

____

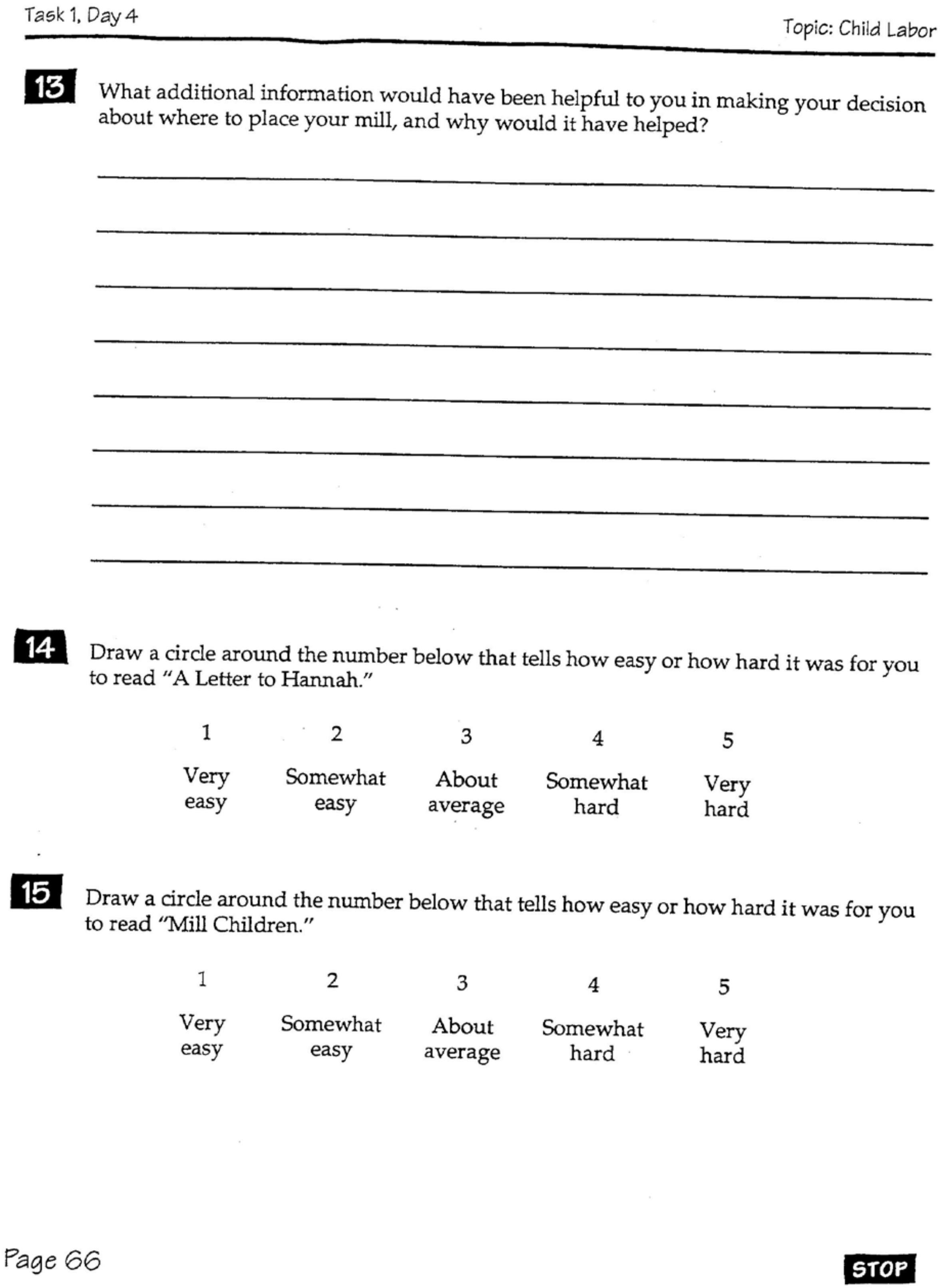
____

____

GO ON


MAJOR RIVERS IN MARYLAND
NEW JERSEY
Delaware River
DELAWARE
Atlantic Ocean
Pocomoke R.
Wicomico R.
Nanticoke R.
River
Choptank River
Tuckahoe Cr.
Chester River
Sassafras R.
Susquehanna River
Chesapeake Bay
Little Gunpowder R.
Gunpowder Falls
Patapsco R.
Patuxent River
Potomac River
WASHINGTON, D.C.
Monocacy River
Potomac River
PENNSYLVANIA
Conococheague Cr.
Sideling Hill Creek
Evitts Cr.
Savage R.
Youghiogheny River Lake
Youghiogheny R.
WEST VIRGINIA
VIRGINIA

GO ON



13 What additional information would have been helpful to you in making your decision about where to place your mill, and why would it have helped?

____

____

____

____

____

____

____

____

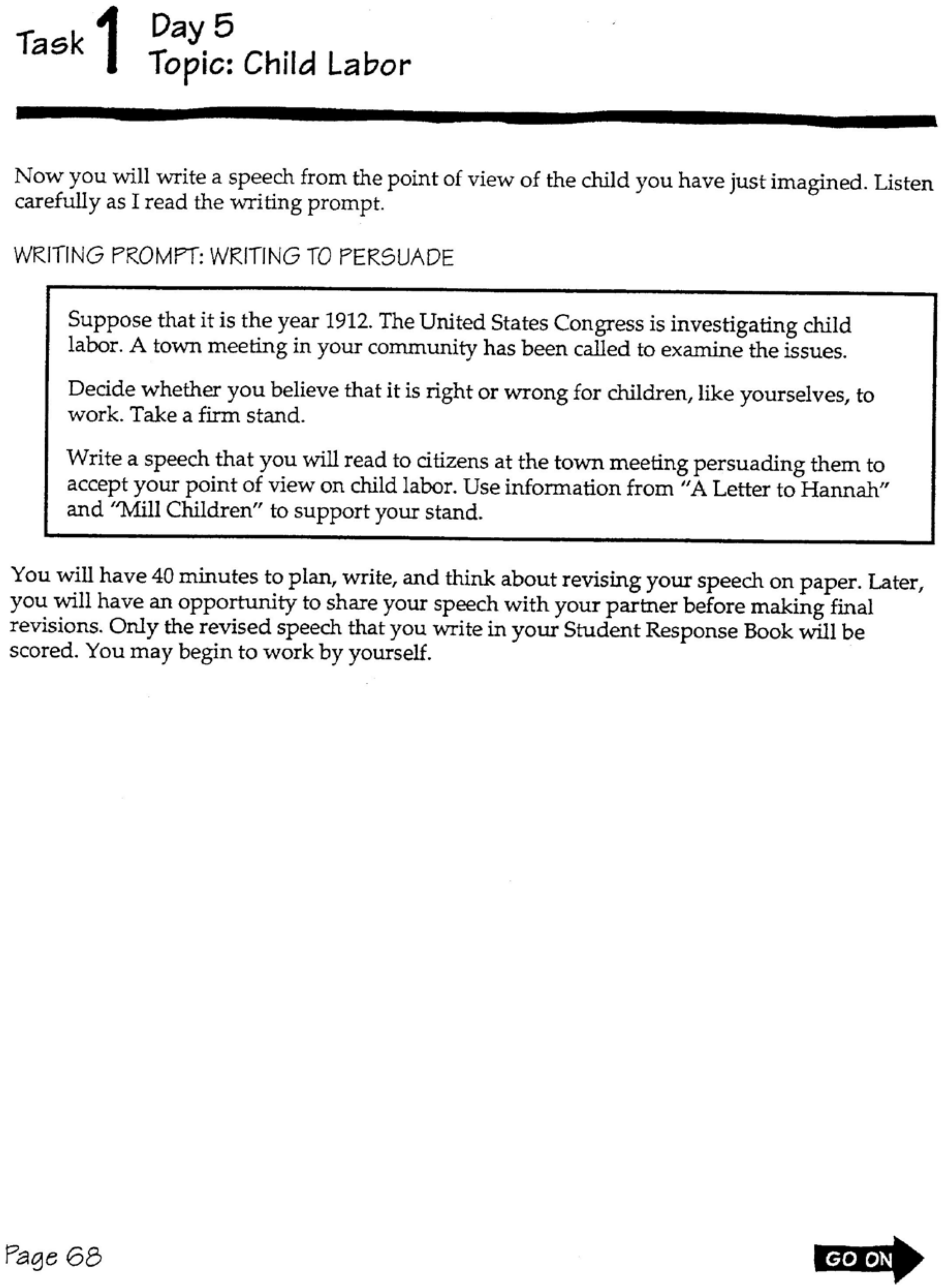
14 Draw a circle around the number below that tells how easy or how hard it was for you to read "A Letter to Hannah."

1	2	3	4	5
Very easy	Somewhat easy	About average	Somewhat hard	Very hard

15 Draw a circle around the number below that tells how easy or how hard it was for you to read "Mill Children."

1	2	3	4	5
Very easy	Somewhat easy	About average	Somewhat hard	Very hard

STOP



Task 1 Day 5 Topic: Child Labor

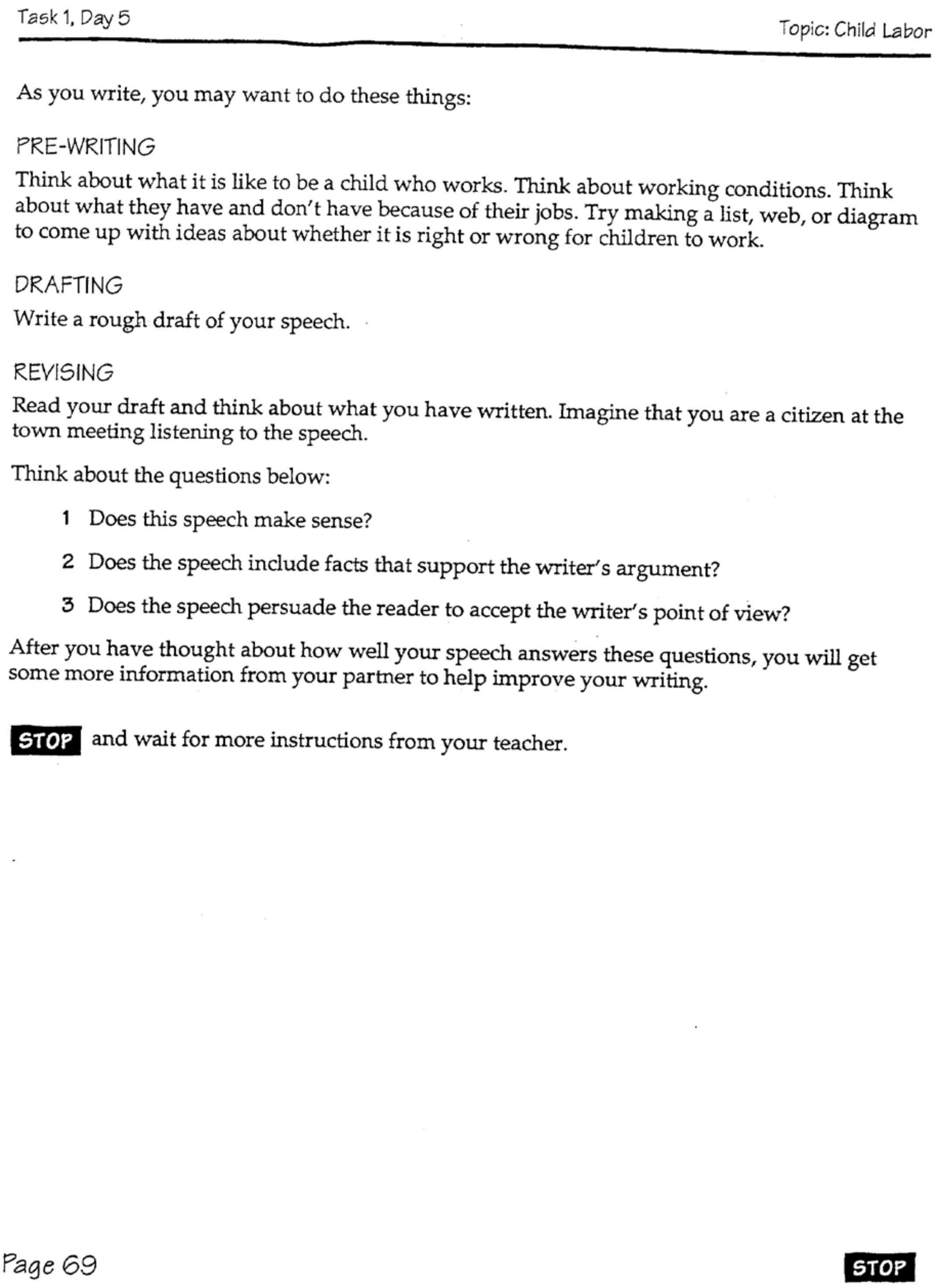
Now you will write a speech from the point of view of the child you have just imagined. Listen carefully as I read the writing prompt.

WRITING PROMPT: WRITING TO PERSUADE

> Suppose that it is the year 1912. The United States Congress is investigating child labor. A town meeting in your community has been called to examine the issues.
>
> Decide whether you believe that it is right or wrong for children, like yourselves, to work. Take a firm stand.
>
> Write a speech that you will read to citizens at the town meeting persuading them to accept your point of view on child labor. Use information from "A Letter to Hannah" and "Mill Children" to support your stand.

You will have 40 minutes to plan, write, and think about revising your speech on paper. Later, you will have an opportunity to share your speech with your partner before making final revisions. Only the revised speech that you write in your Student Response Book will be scored. You may begin to work by yourself.

GO ON



As you write, you may want to do these things:

PRE-WRITING

Think about what it is like to be a child who works. Think about working conditions. Think about what they have and don't have because of their jobs. Try making a list, web, or diagram to come up with ideas about whether it is right or wrong for children to work.

DRAFTING

Write a rough draft of your speech.

REVISING

Read your draft and think about what you have written. Imagine that you are a citizen at the town meeting listening to the speech.

Think about the questions below:

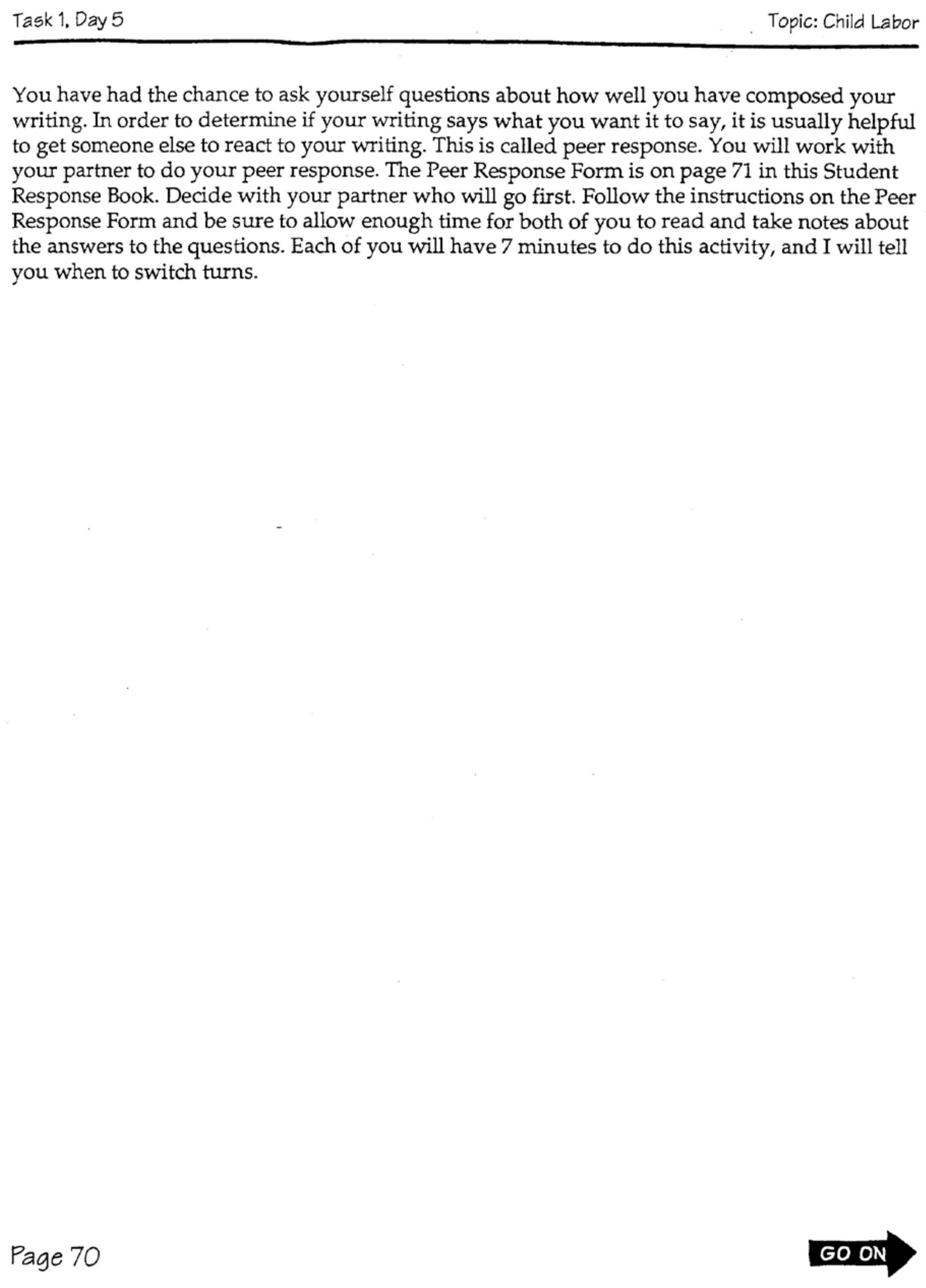
1 Does this speech make sense?

2 Does the speech include facts that support the writer's argument?

3 Does the speech persuade the reader to accept the writer's point of view?

After you have thought about how well your speech answers these questions, you will get some more information from your partner to help improve your writing.

STOP and wait for more instructions from your teacher.

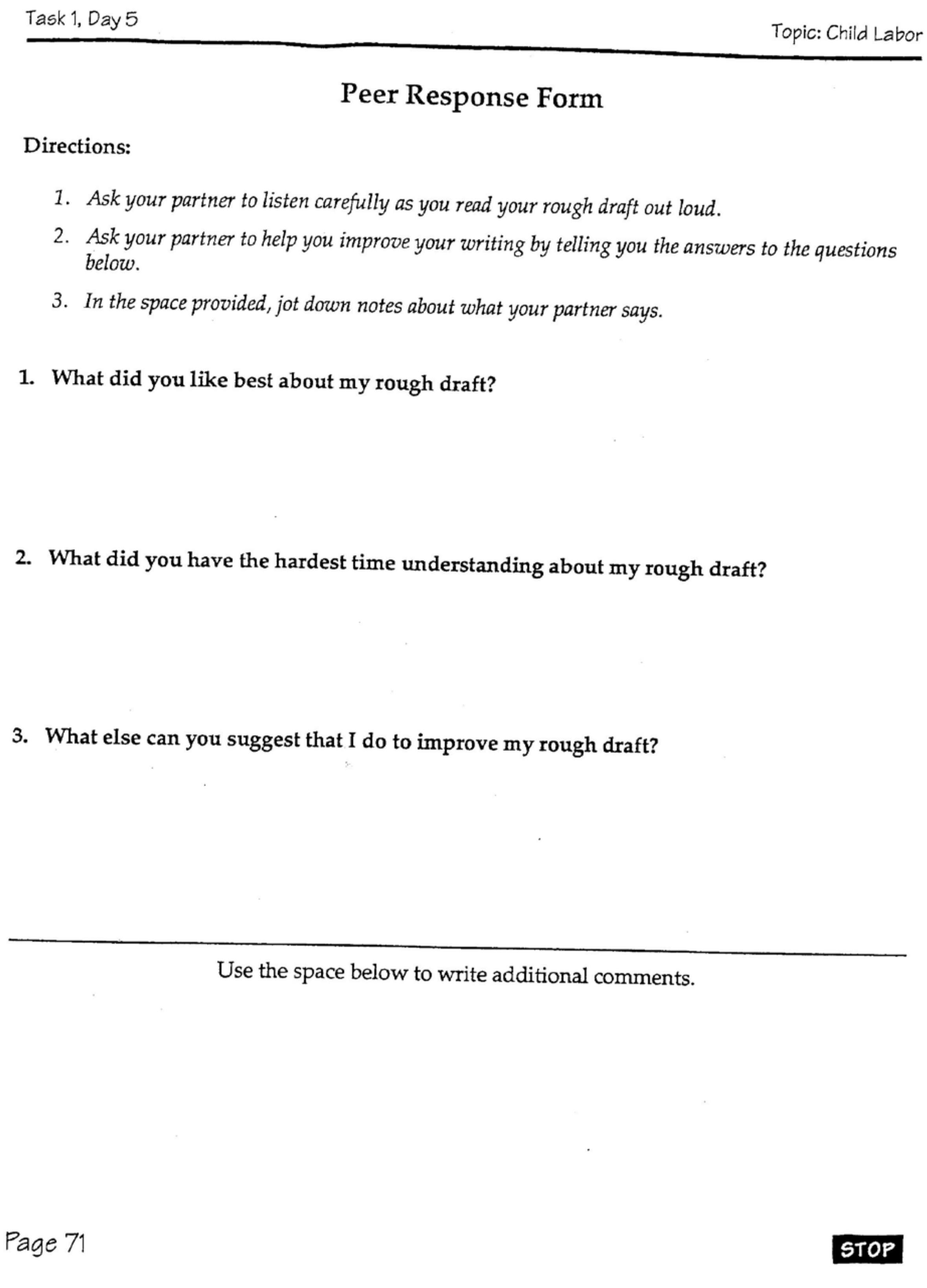


Task 1, Day 5 Topic: Child Labor

You have had the chance to ask yourself questions about how well you have composed your writing. In order to determine if your writing says what you want it to say, it is usually helpful to get someone else to react to your writing. This is called peer response. You will work with your partner to do your peer response. The Peer Response Form is on page 71 in this Student Response Book. Decide with your partner who will go first. Follow the instructions on the Peer Response Form and be sure to allow enough time for both of you to read and take notes about the answers to the questions. Each of you will have 7 minutes to do this activity, and I will tell you when to switch turns.

Page 70

GO ON



Task 1, Day 5

Topic: Child Labor

Peer Response Form

Directions:

1. *Ask your partner to listen carefully as you read your rough draft out loud.*
2. *Ask your partner to help you improve your writing by telling you the answers to the questions below.*
3. *In the space provided, jot down notes about what your partner says.*

1. What did you like best about my rough draft?

2. What did you have the hardest time understanding about my rough draft?

3. What else can you suggest that I do to improve my rough draft?

Use the space below to write additional comments.

Page 71

STOP

WRITING THE REVISED DRAFT

Now that you have had the chance to think about your writing and get information from your partners, it is time to revise your speech. Remember that you are the author, and only you can decide if you want to use your partner's suggestions when you revise your writing. Write your revised draft in this response book on the lines provided below. You will have 35 minutes to revise your draft, to write it in this book, and to do activity 17. Make sure that you get all of your revised draft in the response book in 35 minutes, because only the material that is in the book will be scored.

PROOFREADING

Look over your writing. Because your speech may be printed in the newspaper, be sure your speech is clear and complete. Also, check for correct spelling, punctuation, grammar, and usage. Use the suggestions on the Proofreading Guidesheet to check your work. Turn to page 30 in your resource book to find the Proofreading Guidesheet. Make any necessary corrections on your revised draft.

16 ____

____

____

____

____

____

____

____

____

____

____

____

____

GO ON

A Letter to Hannah

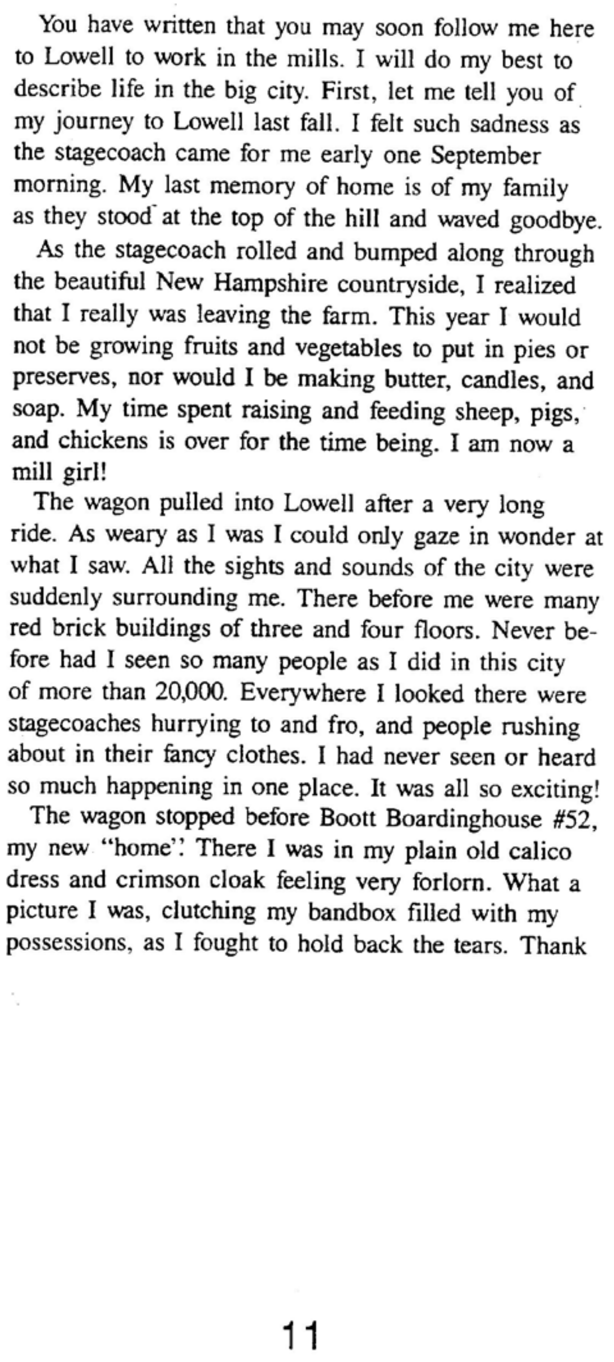
by Shirley Gifford; illustrated by Susan Lippman

March 9, 1840

Dear Cousin Hannah,

Please forgive me for not writing to you since I left home six months ago. Never in my fifteen years has the time seemed to pass so swiftly. Whenever I can, I write to my folks since I am the oldest and the first to leave home to work in the mills.

I feel so proud that I now support myself. I am also able to save money toward my dowry and still have some left for an occasional luxury. I now have a sense of being on my own that I never had on the farm, and as you will see I have learned many things.



You have written that you may soon follow me here to Lowell to work in the mills. I will do my best to describe life in the big city. First, let me tell you of my journey to Lowell last fall. I felt such sadness as the stagecoach came for me early one September morning. My last memory of home is of my family as they stood at the top of the hill and waved goodbye.

As the stagecoach rolled and bumped along through the beautiful New Hampshire countryside, I realized that I really was leaving the farm. This year I would not be growing fruits and vegetables to put in pies or preserves, nor would I be making butter, candles, and soap. My time spent raising and feeding sheep, pigs, and chickens is over for the time being. I am now a mill girl!

The wagon pulled into Lowell after a very long ride. As weary as I was I could only gaze in wonder at what I saw. All the sights and sounds of the city were suddenly surrounding me. There before me were many red brick buildings of three and four floors. Never before had I seen so many people as I did in this city of more than 20,000. Everywhere I looked there were stagecoaches hurrying to and fro, and people rushing about in their fancy clothes. I had never seen or heard so much happening in one place. It was all so exciting!

The wagon stopped before Boott Boardinghouse #52, my new “home”. There I was in my plain old calico dress and crimson cloak feeling very forlorn. What a picture I was, clutching my bandbox filled with my possessions, as I fought to hold back the tears. Thank

11

Mill Children

by Hilda Brucker

In the early 1900s, Lewis Hine photographed these children who tended spinning frames in Fall River, Massachusetts.

At 5:00 A.M., the mill whistle blew, its shrill blast waking the town's children from a deep sleep. There was just time enough for them to wash and dress quickly before hurrying through the darkness to their jobs at the textile mill. Anyone not at work when the whistle sounded again at six would be locked out of the mill and lose much-needed wages. For the next twelve hours, children as young as six years old would toil at the machines, often with only a fifteen-minute break for lunch.

The textile mills of the early 1800s were cold, dark, and noisy. Children labored as bobbin boys who fed yarn into the looms, at the spinning machines, at keeping machines greased and oiled, and at moving supplies from room to room. Many stood all day, their heads throbbing from the continuous roar of the machines. Although their eyes tired quickly, they dared not let their concentration wander, for they had seen others lose fingers in the whirling gears of the machines. Strands of long hair also could get caught in the metal jaws, pulling out not only the hair, but a piece of the scalp as well. Older children tried their best to watch over their younger brothers and sisters, but accidents were common. Any child could easily fall prey to pneumonia or other illnesses due to the dampness of the mill and lack of rest.

For six days out of every week, this nightmare continued. How could parents who loved their children allow them to live with such misery? They saw no other choice.

The grandparents and great-grandparents of these New England mill children had been farmers. Their families had made up a complete work force, with everyone from youngest to oldest pitching in to get the crops planted and harvested. Only the geography would not cooperate. The soil in most of the New England states is stony, difficult to cultivate, and not very fertile. The climate is cool, making the growing season short. Families could raise just enough food to live on, but with nothing left over to sell, they never seemed to get ahead.

Then the Industrial Revolution arrived. The first patented looms, which ran on water power, provided a much quicker method of producing woven cloth, and the textile industry became a profitable one. Businessmen quickly set up mills along the many streams, rivers, and waterfalls that dotted the New England countryside. Men went to work at the mills, leaving the farm chores to women and children. As more and more workers flocked to the mills, towns sprang up around them. Upon the invention of the steam engine, factories were no longer dependent on water power, and even more mills were built.

Machines also became simpler to operate, and that gave mill owners an idea as to how they could increase profits. They began to hire women and children to work at the machines, paying them less money than men because, they reasoned, they were less capable than men. Soon mill owners began firing men and hiring children for all but the most skilled positions. Grown men could not find work, and desperate families, living off only their children's wages, began to send younger and younger children to work in an effort to make ends meet.

A doffer, or bobbin girl, carries bobbins to the spinning frame.

While the situation rapidly worsened, the plight of the mill children did not go unnoticed. Their misery attracted the attention of social reformers who were

Linda Darling-Hammond, Co-Director
Stanford University Charles E. Ducommun Professor of Education

Prudence Carter, Co-Director
Stanford University Associate Professor of Education and (by courtesy) Sociology

Carol Campbell, Executive Director

scope
Stanford Center for
Opportunity Policy in Education

Stanford Center for Opportunity Policy in Education
Barnum Center, 505 Lasuen Mall
Stanford, California 94305
Phone: 650.725.8600
scope@stanford.edu

http://edpolicy.stanford.edu