Research Memorandum

Transforming K-12 Assessment: Integrating Accountability Testing, Formative Assessment, and Professional Support

Randy Elliot Bennett

Drew H. Gitomer

July 2008

ETS RM-08-13

Transforming K-12 Assessment:
Integrating Accountability Testing, Formative Assessment, and Professional Support

Randy Elliot Bennett and Drew H. Gitomer
ETS, Princeton, NJ

July 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS' constituents and the field.

ETS Research Memorandums provide limited dissemination of ETS research.

Abstract

This paper presents a brief overview of the status of K-12 accountability testing in the United States. Following that review, we describe an assessment-system model designed to overcome the problems associated with current approaches to accountability testing. In particular, we propose a model in which accountability assessment, formative assessment, and professional support are built on the same conceptual base and work synergistically with one another. We close with a brief discussion of the role of technology and a review of the challenges that must be met if the highly ambitious system we suggest is to be realized.

Key words: Accountability assessment, formative assessment, professional development, comprehensive assessment systems, computer-based testing

Can advances in cognitive science, psychometrics, and technology transform the accountability paradigm that is currently in place in the United States? Of course, asking this question implies problems with the present enactment of the No Child Left Behind Act, a system that requires each state regularly to test students in specified grades and subject areas against a state-imposed proficiency standard. We begin the chapter by describing some of the forces that have led to the heightened emphasis on testing, and then we articulate some of the fundamental problems with the system as currently implemented. We then present an assessment-system model that is designed to overcome some of the inherent weaknesses of the present approach. Specifically, we ask whether we can have an assessment system that goes beyond fulfilling a simple accountability function by (a) documenting what students have achieved (*assessment of learning*), (b) helping identify how to plan instruction (*assessment for learning*), and (c) engaging students and teachers in worthwhile educational experiences in and of themselves (*assessment as learning*).

The system we propose is heavily dependent on new technology. However, simply putting current tests on computer will not lead to substantive change in assessment practice. Instead, the system relies on advances in (a) cognitive science and an understanding of how students learn, (b) psychometric approaches that attempt to provide richer characterizations of student achievement, and (c) technologies that allow for the presentation of richer assessment tasks and for the collection and automated scoring of more complex student responses. We close by putting forth the challenges facing the full development and implementation of an assessment system that is intended to support sound educational practice.

A Brief Overview of the Status of Accountability in the United States

The push for educational accountability has its roots in concerns about the ability of the educational system to prepare citizens to meet successfully the challenges of a global economy. One leg of this argument is that maintaining current living standards depends on keeping high-paying jobs at home. Those jobs are created through business investment, and business investment follows labor pools that are skilled and productive. However, when a nation's labor pool begins to become less skilled and productive relative to the pools of other nations, business investment starts to flow elsewhere, jobs leave, the standard of living drops, and in the worst case national economic stability is threatened.

The second leg of the argument is that the U.S. educational system has not effectively addressed fundamental inequity in access to a quality education. This unequal access has been primarily defined by race, income, and home language. As the proportion of students who are poor, non-White, or nonnative speakers of English continues to increase, the need to improve educational quality for all becomes not only an issue of economic necessity, but also one of moral and democratic principles. Education must be able to engender an informed and self-sufficient citizenry for a stable democracy to survive.

Such arguments are captured in three recent reports: (a) *America's Perfect Storm: Three Forces Changing Our Nation's Future* (Kirsch, Braun, Yamamoto, & Sum, 2007), (b) *Tough Times, Tough Choices: The Report of the New Commission on the Skills of the American Workforce* (National Center on Education and the Economy, 2006), and (c) *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future* (Committee on Prospering in the Global Economy of the 21st Century, 2007).

These reports generally claim that the U.S. education system, which is responsible for producing the skilled and productive labor pools of tomorrow, is in danger of failing to meet that responsibility. According to the Organisation for Economic Co-operation and Development (OECD) Program for International Student Assessment (PISA), U.S. 15-year-olds performed below the OECD average in math literacy, science literacy, and problem solving (i.e., below the average for the industrialized nations with which the United States competes economically; Lemke et al., 2004). Upper secondary graduation rates are also below the OECD average (OECD, 2006). Further, in terms of tertiary educational attainment, meaning the number of years completed beyond secondary school, the United States has slipped from first to seventh of the OECD countries. Finally, U.S. university graduation rates are below the OECD average.

This skill profile is highly related to socioeconomic and language status. *America's Perfect Storm* (Kirsch et al., 2007) makes clear that the fastest growing part of the U.S. population is coming from families in which English is not the first language. Other studies show that social mobility has decreased dramatically in recent years. Students born into poor and less-educated families have lower likelihoods of moving into higher socioeconomic strata than did students of previous generations (Beller & Hout, 2006).

These conditions have raised the call for increased use of assessment as a tool for educational accountability in order to evaluate educational effectiveness and make informed

decisions about how to improve the system. Educators and policy makers need mechanisms to identify the competencies, ages, population groups, schools and even the individuals requiring attention.

Assessments, with stakes attached to them, have been viewed as more than information systems. They have been seen as a primary tool to focus attention on achievement in particular subject areas and on the achievement of selected population groups. In the United States, those population groups have included ethnic minorities, economically disadvantaged students, students with disabilities, and students having limited English proficiency. The focal subject areas have been reading; math; and, more recently, science.

In the United States, these assessments are being used to evaluate not only students, but also schools and teachers. Schools can be sanctioned, to the point of closing, if performance criteria are not satisfied. States and districts are introducing teacher pay-for-performance systems based on student test scores. In reaction to these highly consequential assessments, educational practices are changing, in intended and unintended ways. While there is significant debate about the efficacy of the current assessment system to meet the intended goals of increasing accountability and improving teaching and learning, there is no reason to believe that the emphasis on accountability testing will abate any time soon.

However, we believe there is a fundamental problem with the system as currently implemented. In the United States, the problem is that the above set of circumstances has fashioned an accountability assessment system with at least two salient characteristics. The first characteristic is that there are now significant consequences for students, teachers, school administrators, and policy makers. The second characteristic is, paradoxically, very limited educational value. This limited value stems from the fact that our accountability assessments typically reflect a shallow view of proficiency defined in terms of the skills needed to succeed on relatively short and, too often, quite artificial test items (i.e., with little direct connection to real-world contexts).

The enactment of the No Child Left Behind Act has resulted in an unprecedented and very direct connection between highly consequential assessments and instructional practice. Historically, the disassociation between large-scale assessments and classroom practice has been decried, but the current irony is that the influence these tests now have on educational practice has raised even stronger concerns (e.g., Abrams, Pedulla, & Madaus, 2003), stemming from a

general narrowing of the curriculum, both in terms of subject areas and in terms of the kinds of skills and understandings that are taught. The cognitive models underlying these assessments are long out of date (Shepard, 2000); evidence is still collected primarily through multiple-choice items; and students are characterized too often on only a single proficiency, when the nature of domain performance is arguably more complex.

Many experts in assessment—as well as instruction—claim that we unintentionally have created a system of accountability assessment grounded in an outdated scientific model for conceptualizing proficiency, teaching it, and measuring it. Further, an entire continuum of supporting products has been developed, including interim (or benchmark) assessments, so-called formative assessments, and teacher professional development that are emulating—and worse, reinforcing—the less desirable characteristics of those accountability tests.

In essence, the end goal for too many teachers, students, and school administrators has become improving performance on the accountability assessment without enough attention to whether students actually learn the deeper curriculum standards those tests are intended to represent.

Designing an Alternative System

The question we are asking at ETS is this: Given the press for accountability testing, could we do better? Could we design a comprehensive system of assessment that:

- Is based on modern scientific conceptions of domain proficiency and that therefore causes teachers to think differently about the nature of proficiency, how to teach it, and how to assess it?
- Shifts the end goal from improving performance on an unavoidably shallow accountability measure toward developing the deeper skills we would like students to master?
- Capitalizes on new technology to make assessment more relevant, effective, and efficient?
- Primarily uses extended, open-ended tasks?
- Measures frequently?

- Provides not only formative and interim-progress information, but also accountability information, thereby reducing dependence on the one-time test?

Developing large-scale assessment systems that can support decision making for state and local policy makers, teachers, parents, and students has proven to be an elusive goal. Yet, the idea that educational assessment ought to better reflect student learning and afford opportunities to inform instructional practice can be traced back at least 50 years, to Cronbach's (1957) seminal article, "The Two Disciplines of Scientific Psychology." These ideas continued to evolve with Glaser's (1976) conceptualization of an *instructional psychology* that would adapt instruction to students' individual knowledge states. Further developments in aligning cognitive theory and psychometric modeling approaches have been summarized by Glaser and Silver (1994); Pellegrino, Baxter, and Glaser (1999); Pellegrino, Chudowsky, and Glaser (2001); the Committee on Programs for Advanced Study of Mathematics and Science in American High Schools and National Research Council (2002); and Wilson (2004).

We are proposing a system that needs to be coherent in two ways (Gitomer & Duschl, 2007). First, assessment systems are *externally coherent* when they are consistent with accepted theories of learning and valued learning outcomes. Second, assessment systems can be considered *internally coherent* to the extent that different components of the assessment system, particularly large-scale and classroom components, share the same underlying views of learners' academic development. The challenge is to design assessment systems that are both internally and externally coherent. Realizing such a system is not straightforward and requires a long-term research and development effort. Yet, if successful, we believe the benefits to students, teachers, schools, and the entire educational system would be profound.

There are undoubtedly many different ways one could conceptualize a comprehensive system of assessment to improve on current practice. We offer one potential solution that we are pursuing, not because we think it is the sole solution, but because we believe it contains certain core elements that would be integral to any system that endeavored faithfully to assess important learning objectives summatively at the same time as it encouraged and facilitated good instructional practice. Our vision entails three closely related systems built upon the same conceptual base: (a) accountability assessment, (b) formative assessment, and (c) professional support.

The Common Conceptual Base

The foundation for all three systems is a common conceptual base that combines curriculum standards with principles from cognitive-scientific research. By cognitive-scientific research, we refer broadly to the multiple fields of inquiry concerned with how students learn (e.g., Bransford, Brown, & Cocking, 1999). (See the Appendix A and B for brief descriptions of cognitive science and psychometric science, respectively.) Of course, calls for assessments driven by theories of learning are not new, so the question is why have such calls not been heeded?

For one, the sciences of educational measurement, and of learning and cognition, evolved separately from one another. Attempts to bring the two fields together are relatively recent and have not yet been incorporated into accountability assessment in any significant way. Second, cognitive-scientific research has produced only partial knowledge about the nature of proficiency in specific domains, and we do not yet know how to create practical assessment systems that use this partial knowledge effectively. Third, practical and economic constraints have inhibited the development and deployment of such systems. However, sufficient progress has been made on a number of relevant fronts to make the pursuit of a more ambitious vision of assessment a worthwhile endeavor.

The first advance has been in the depth and breadth of our understanding of learning and performance in academic domains. Depending upon the content domain, research offers us the following: cognitive-scientific principles, competency models, and developmental models.

Principles present an important contrast to the outcomes that often characterize curriculum standards. Cognitive-scientific principles describe the processes, strategies, and knowledge structures important for achieving curriculum standards, and the features of tasks—or more generally, of situations—that call upon those processes, strategies, and knowledge structures.

For example, cognitive principles suggest working with multiple representations because information does not come in only one form. Indeed, Sigel (1993) and others have made a compelling case that conceptual competence is, at its core, the ability to understand and navigate between multiple representations. For example, the child who learns to read moves from the direct experience of an object to a picture representation, to a word (e.g., *cat*), to increasingly abstract descriptions, all signifying the same concept. Across domains, students need to

understand and use representational forms that may include written text, oral description, diagrams, and specialized symbol systems, moving easily and flexibly among these different representations.

Cognitive principles also suggest embedding tasks in meaningful contexts, since meaningful contextualization can engage students and help them link solution strategies to the conditions under which those strategies might be best employed.

Cognitive principles suggest integrating component skills, because real-world tasks often call for the execution of components in a highly coordinated fashion, and achieving that coordination requires the components to be practiced, and assessed, in an integrated manner.

Fourth, cognitive principles suggest developing component skills to automaticity (Perfetti, 1985). If low-level components—like the ability to decode words—are not automatic, attention must be devoted to them, drawing limited cognitive resources away from higher-level processes, like making meaning from text.

Finally, cognitive principles suggest designing assessment so that it supports—or at least does not conflict with—the social processes integral to learning and performance. At one level, the *sociocultural and situative* perspective focuses on the nature of social interactions and how these interactions influence learning. From this perspective, learning involves the adoption of sociocultural practices, including the practices within particular academic domains. Students of science for example, not only learn the content of science, but also develop an intellective identity (Greeno, 2002) as scientists, by becoming acculturated to the tools, practices, and discourse of science as a discipline (Bazerman, 1988; Gee, 1999; Hogan, 2007; Lave & Wenger, 1991; Rogoff, 1990; Roseberry, Warren, & Contant, 1992). Similarly, students learn to engage in the practices of writers or mathematicians as they become more accomplished in a domain. This perspective grows out of the work of Vygotsky (1978) and others and posits that learning and disciplinary practice develop out of social interaction. The second social dimension that needs to be attended to in an assessment design that produces meaningful results is the accommodation of students with a wide range of cultural, linguistic, and other characteristics

Competency models define, from a cognitive perspective, what it means to be skilled in a domain. Ideally, these models can tell us not only the processes, strategies, and knowledge structures important for achievement and the features of tasks that call upon those processes, strategies, and knowledge structures, but also how the components of domain proficiency might

be organized and how those components work together to facilitate skilled performance. For example in our work on writing, the competency model is shaped around the interaction of (a) the use of language and literacy skills (skills involved in speaking, reading, and writing standard English), (b) the use of strategies to manage the writing process (e.g., planning, drafting evaluating and revising), and (c) the use of critical-thinking skills (reasoning about content, reasoning about social context). Assessment is then designed to assess the interplay of these skills using tasks that reflect legitimate writing activity.

Developmental models define, from a cognitive perspective, what it means to progress in a domain. In addition to providing principles and a proposed domain organization, these models tell us how proficiency develops over time, including how that development is affected by the diverse cultural and linguistic backgrounds that students bring to school.

Together, these cognitive-scientific principles and models help us determine:

- the components of proficiency critical to achieving curriculum standards that should, therefore, be assessed;
- the features of test questions to manipulate to distinguish better among students at different proficiency levels, to give diagnostic information, or to give targeted instructional practice;
- how to anchor score scales so that test performance can be described in terms that more effectively communicate what students know and can do;
- the components of proficiency that should be instructional targets;
- how teachers might arrange instruction for maximum effect; and
- how to better account for cultural and linguistic diversity in assessment.

It is important to note that the nature of most curriculum standards is such that they are not particularly helpful in making these decisions. Current standards are not helpful because they are often list-like, rather than coherently grouped; may be overly general, so that specifically what to teach may be unclear; or, at the other extreme, are too molecular, encouraging a piecemeal approach to instruction that neglects meaningful integration of components. Thus, in principle, having a modern cognitive-scientific basis should help us build better assessments in the same way as having an understanding of physics helps engineers build better bridges.

The Accountability System

For purposes of this paper, *accountability assessment* is defined as a standardized, summative examination, or program of examinations, used to hold an entity formally or informally responsible for achievement. That entity could be a learner, as when a school-leaving examination is used to determine if a student can graduate; a school, as when league tables are compiled; or the education system as a whole, as when the achievement of different countries is compared.

Our conception for an accountability system begins with the strong conceptual base described above. Foundational tasks are administered periodically with information aggregated over time to dynamically update proficiency estimates. Timely reports are produced that are customized for particular audiences. Each of these features is described in more detail.

Foundational tasks. Assessments composed of foundational tasks are built upon the conceptual base so that they are demonstrably aligned to curriculum standards and to cognitive principles or models. That is, these tasks should be written to target processes, strategies, and knowledge structures central to achieving curriculum standards and to proficient performance in the domain. The foundational tasks are the central (but not exclusive) means of measuring student competency. These foundational tasks generally are intended to do the following:

1. Require the integration of multiple skills or curriculum standards.
2. Be extended, offering many opportunities to observe student behavior.
3. Be meaningfully contextualized.
4. Call upon problem-solving skills.
5. Utilize constructed-response formats.
6. Be regarded by teachers as learning events worth teaching toward.

An example of a framework our colleagues have developed for the design of foundational tasks in writing is described in Figure 1.

Periodic accountability assessment. A second characteristic of the accountability system is to employ a series of periodic administrations instead of the model of assessment as a one-time event. In order to faithfully assess the intent of curriculum standards, in terms of both depth and breadth, as well as to provide models of sound educational practice, it is necessary to construct a

The goal is to help students display their writing skills to best advantage by providing multiple opportunities, guidance, and resources for assessments.
Tests and rubrics emphasize the role of critical thinking in writing proficiency.

Each Periodic Accountability Assessment, or PAA, is a "project"

- Each test is a small-scale project centered on one topic, thereby providing an overall context, purpose, and audience for the set of tasks.
- Each test usually focuses on one genre or mode of discourse and the critical-thinking skills and strategies associated with that mode of discourse.
- Short prewriting/inquiry tasks serve as thematically related but psychometrically independent steps in a sequence leading up to and including a full-length essay or similar document.
- The smaller tasks provide measurement of component skills—especially critical-thinking skills—as well as a structure to help students succeed with the larger, integrated task (essay, letter, etc.).
- Task formats vary widely (mostly constructed-response, with some selected-response), but all tests include "writer's checklists" and glossaries of words used in the test.

The project comes with its own resource materials

- To help address varying levels of background knowledge about the PAA's topic, the tests often include short documents that students are required or encouraged to use.
- This approach permits students to engage in greater depth with more substantive topics and meshes with current curricular emphasis on research skills and use of sources.

Tripartite analytic scoring is based on the three-strand competency model

- Strand I (use language and literacy skills):
  - Instead of using multiple-choice items to measure these skills, the approach is to apply a generic Strand I rubric to all written responses across tasks. This rubric focuses on sentence-level features of the students' writing.
- Strand II (use strategies to manage the writing process):
  - A generic Strand II rubric is applied to all written responses of sufficient length in order to measure document-level skills, including organization, structure, focus, and development.
- Strand III: (use critical-thinking skills):
  - Each constructed-response task includes a task-specific Strand III rubric used to evaluate the quality of ideas and reasoning particular to the task. In addition, most of the selected-response tasks measure critical-thinking skills.

Figure 1. **A framework for the design of periodic accountability assessments in writing.**

Note. This framework was developed by Paul Deane, Nora Odendahl, Mary Fowles, Doug Baldwin, and Tom Quinlan.

relatively long test that consists of integrated, cognitively motivated tasks. However, it is impractical to administer such a test at a single point in time. It is also educationally counterproductive to delay assessment feedback until the end of the school year. Therefore, we divide this hypothetical, long test into multiple parts, with each part including one or more foundational tasks, supplemented by shorter items to test skills that can be appropriately assessed in that latter fashion. Test parts are administered across the school year. Student-status information and formative hypotheses about achievement are returned after each administration. A final accountability result is derived by aggregating performance on the parts. (How best to accomplishment this aggregation is the subject of our ongoing research. However, the magnitude of weights assigned to particular assessment tasks and skills may, in part, be a policy decision determined by the test sponsors, such as state education department staff.)

Periodic administration has multiple benefits. It allows for greater use of tasks worth teaching toward, because there is more time for assessment in the aggregate. Also, the test more effectively can cover curriculum standards, making for a more valid measure. Because the scores can be progressively accumulated, the accumulated scores should gain in reliability as the year advances; the end-of-year scores should be more reliable than scores from a traditional one-time test, thereby giving a truer picture of student competency. There also should be a greater chance of generating instructionally useful profile information, because more information has been systematically assembled than would otherwise be the case. Finally, in contrast to most existing accountability systems, no single performance is determinative. Instead, similar to the way teachers assign course grades, accountability scores come from multiple pieces of information gathered in a standardized fashion throughout the school year. The more pieces of information, the less each counts individually, so no student, teacher, school, or administrator can be held to one unrepresentative performance.

Timely results. Since accountability administration is periodic, student status with respect to curriculum standards can be updated regularly. That regular updating allows targets for formative assessment to be suggested and at-risk students to be identified while there is still time to take instructional action.

Customized reports. Customized reports will be designed that are appropriate to the audience, be it student, parent, teacher, head teacher, local administrator, or national policy

maker. These reports should be available on demand and suggest actions, not only for students, but also for instructional policy and teacher professional development.

The Formative System

The formative system is built on a concept of formative assessment as an ongoing process in which teachers and students use evidence gathered through formal and informal means to make inferences about student competency and, based on those inferences, take actions intended to achieve learning goals. This conception implies that formative assessment encompasses a process aided by some type of instrumentation, formal or not. First, this instrumentation should be fit for use (i.e., suited to instructional decision-making). Not all instruments can be used effectively in a formative assessment process by the typical teacher, because not all instruments are fit for that purpose. Second, the conception depicts formative assessment as a hypothesis-generation-and-testing process, where what we observe students do constitutes evidence for inferences about their competency, which in turn directs instructional action as well as the collection and interpretation of further evidence. Third, the conception attempts to focus formative assessment on an underlying competency model, in contrast to focusing it on classroom activities or assessment tasks. Through the competency model, the formative system is linked to the accountability system, with both systems deriving from the same conceptual base. The intent is to facilitate student growth, not in the shallow way characteristic of many current formative assessments built to improve achievement on multiple-choice or short-answer accountability tests, but in a deeper fashion consistent with cognitive principles and models. Finally, the conception identifies the end purpose of formative assessment as the modification of instruction to facilitate learning of competencies.

An important caveat is that whereas the accountability system may provide information of use to the formative system, the reverse should not occur. That is, performance in the formative system should not be used for accountability purposes. This one-way “firewall” exists for two reasons. First, the formative system is optional and modifiable by design, so students will likely have very different access to formative assessments, making comparability of student results impossible. More importantly, the formative system is for learning, and if students and teachers are to feel comfortable using it for that purpose, they will need to try out problem solutions—and engage in instructional activities—without feeling they are being constantly judged.

The formative system is designed to give students opportunity to develop target competencies through structured instructional practice. Teachers may use formative tasks as part of their lesson designs and also may tailor use on the basis of information from the accountability system. For example, information from the periodic accountability assessments may suggest particular student needs.

The formative system is used at the option of the teacher or school. It is available on demand so that teachers may use it when, and as often as, they need it. The intention behind optional use is the recognition that teachers are dealing with enough mandates already. Our belief is that a formative assessment system is likely to be more effective if teachers choose to use it because they believe it will benefit their practice. The challenge will be in creating a system that can justify such a belief.

The intent underlying the formative system is to give teachers various classroom resources that are instructionally compatible with the accountability system and that they can use in whatever fashion they feel works best. Among these resources would be classroom tasks and focused diagnostic assessment.

Classroom tasks. Classroom tasks are variants of the foundational accountability tasks. They are integrated, extended, problem-solving exercises meant to be learning events worth teaching toward. These tasks should be accessible from an online bank organized by skills required and curriculum level so as to permit out-of-level practice.

Teachers can use these classroom tasks for several purposes. For example, teachers might use them to give practice and feedback to individual students or as the basis for peer interaction (e.g., students might discuss among themselves the different approaches that could be taken to a task). Finally, teachers might use these tasks as the focus of class discussion so that a particular task, and various ways of responding to it, becomes the object of an extended classroom discourse. These uses of the classroom tasks are intended to facilitate not only student achievement of curriculum standards and development of cognitive proficiencies, but also self-reflection and other habits associated with mature practice in a domain. The intention is, as Stiggins has advocated (Stiggins & Chappuis, 2005), to help students develop ownership of their learning processes and investment in the results.

A brief overview of a classroom formative assessment activity is presented in Figure 2 and in Table 1. The activity is designed to help teachers gather evidence about, and facilitate the

development of, persuasive writing skills for middle school students. Included are a sample screenshot that introduces the activity (Figure 2) and a description of the series of classroom tasks that compose the activity (Table 1). Whereas an interactive system can be used to administer the tasks and collect student responses, most of these formative tasks also can be administered outside of a technology-based environment.

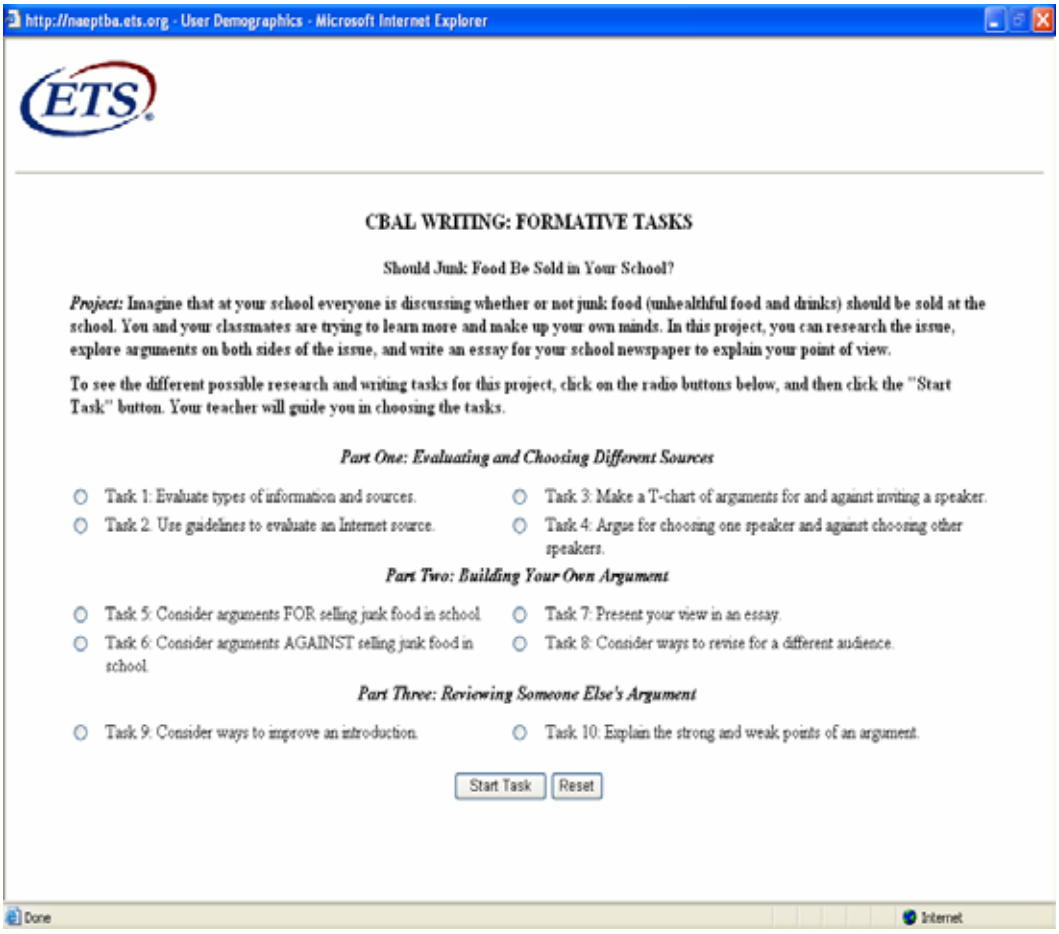

Figure 2. **A formative activity for gathering evidence about, and facilitating the development of, persuasive writing skill.**

Note. This activity was created by Nora Odendahl, Paul Deane, Mary Fowles, and Doug Baldwin.

Diagnostic assessment. The second part of the formative system is diagnostic assessment. Diagnostic assessment is, at the teacher's option, given to students who struggle with certain aspects of performance, either in the accountability system or on classroom tasks. These assessments can be used with students who are at risk of failing or simply with those whom the teacher would like to help advance to the next curriculum level. The diagnostic assessment is

composed of elemental items that test component skills in isolation, something for which multiple-choice or short-answer questions might be used very effectively.

Table 1

Description of Tasks Composing a Formative Activity in Persuasive Writing

Parts	Description of tasks
1	Tasks 1–3 are short exercises that ask students to apply criteria for evaluating various types of research sources. Then, once students have had the opportunity to work with these criteria, they write a persuasive letter arguing in favor of a particular source (in this case, one of three potential speakers). The intent of this group of tasks is to help students develop their ability to judge sources critically and to articulate those judgments. Moreover, the extended writing task (Task 4) gives students an opportunity to write a persuasive piece that is not issue oriented, but instead requires the student to choose from among various alternatives, each with its own pros and cons.
2	Tasks 5 and 6 require the student to read about and consider arguments on each side of the general issue (whether junk food should be sold in schools), before writing an essay presenting his or her own view to a school audience. A follow-up task (Task 8) asks students to consider ways in which they revise the essay for a larger audience outside the school. Thus, this group of tasks takes the student through the stages of persuasive writing—considering arguments on both sides of an issue, formulating and presenting one's own position, and demonstrating awareness of appropriate content and tone for different audiences.
3	Tasks 7 and 8 ask the student to take a given text and apply guidelines for writing an introduction and for presenting an argument. These exercises allow students to work with rubrics and examples of persuasive writing in a very focused way.

Note. This activity was created by Nora Odendahl, Paul Deane, Mary Fowles, and Doug Baldwin.

The diagnostic assessment helps suggest instructional targets by attempting to isolate the causes of inadequate performance on the more integrated foundational tasks comprising the accountability system and classroom assessment. For any student who interacts with the formative system, the reports could provide a dynamic synthesis of evidence, accumulated over time, from the accountability system, the classroom tasks (if administered), and the diagnostic assessment (if administered). Multiple sources of evidence can offer more dependable information about student strengths and weaknesses than any single source alone. For those students who do interact with the system, it should be possible to provide information to the

current teacher, as well as end-of-year formative information to next year's teacher, giving this individual a clearer idea of where to begin instruction than he or she otherwise might have had.

Professional Support

The final component of our vision is professional support. This component has two goals. The first goal is to help teachers and administrators understand how to use the accountability and formative systems effectively. The second goal is to help develop in teachers a fundamentally different conception of what it means to be proficient in a domain, how to help students achieve proficiency, and how to assess it. *Fundamentally different* implies a conception that is based not only on curriculum standards, but also on cognitive research and on recognition of the need to help students develop more positive attitudes toward, and greater investment in, learning and assessment.

To achieve these professional-support goals requires going beyond traditional approaches to teacher in-service training and building more on such ideas as teacher learning communities (McLaughlin & Talbert, 2006). Such communities let interested teachers help one another discover how to use formative assessment best in their own classrooms. We also envision the use of online tools to involve teachers in collaboratively scoring constructed responses to formative system tasks; through scoring, teachers can develop a shared understanding of what it means to be proficient in a domain.

The Role of Technology

The vision presented assumes a heavy presence of technology. For one, technology can help make assessment more relevant, because the computer has become a standard tool in the workplace and in higher education. The ability to use the computer for domain-based work is, therefore, becoming a legitimate part of what should be measured (Bennett, 2002). Second, technology can make assessment more informative since process indicators can be captured, as well as final answers, allowing for the possibility of understanding how a student arrived at a particular result (Bennett, Persky, Weiss, & Jenkins, 2007). Technology can make assessment more efficient because, in principle, moving information electronically is cheaper and faster than shipping paper.

Of great importance is that technology offers a potential long-term solution for the efficient scoring of complex constructed responses. One of the constraints on the widespread use

of constructed-response tasks to date has been the economic expense of human scoring as well as demands on teachers. To the extent that performances can be scored by computer, this limitation will be obviated. Certain kinds of student responses are already reasonably well handled by automated scoring tools (e.g., Shermis & Burstein, 2003; Williamson, Mislevy & Bejar, 2007), whereas other kinds of responses still require long-term research and development efforts.

Technology is not a panacea, however, for it can be a curse as well as blessing. If not used thoughtfully, technology can prevent students from demonstrating skill simply because they do not have enough computer familiarity to respond online effectively (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). Technology can narrow the range of skills measured by encouraging exam developers to use only those tasks most amenable to computer delivery. While such tasks may be quite relevant, they may not cover the full range of skills that should be tested. Technology can distort assessment results when automated scoring neglects important aspects of proficiency (Bennett, 2006). Machines do not do a good job, for example, of evaluating the extent to which a student's essay is appropriate for its intended audience. Finally, technology can encourage students and teachers to focus instructional time on questionable activities like how to write essays that a machine will grade highly, even if the resulting essays are not what an experienced examiner would consider well crafted.

What Are the Challenges?

The successful development and implementation of the aforementioned conception is not a given. Among the challenges that we are working to resolve are:

- The aggregation of results across periodic administrations. For example, should results be weighted according to recency of administration, or some other criterion, so as to account better for growth?
- The problem of missed administrations and missing student-performance data in general.
- The dependence of the system, and interpretation of student results, on specific instructional sequencing within classrooms, schools, and districts.
- Issues of test security related to the memorability of extended tasks.

- Ensuring that generalizable claims about students can be made from assessments composed primarily of extended tasks, which often provide information that is of limited dependability.
- Ensuring the comparability of test forms when different students may be taking different forms and those forms may vary in difficulty.
- Ensuring fairness for special populations.
- Making periodic assessment with extended problems affordable.
- Convincing teachers, administrators, and policy makers to spend *more* time on assessment because the periodic assessments may, in fact, be longer in the aggregate than was the original end-of-year accountability test.
- Making the accountability assessment a worthwhile instructional experience in and of itself.

Indeed, it is only by making the assessment experience educationally worthwhile that we can make a compelling argument for more time and money spent in the process of assessment for accountability.

It is our perception that accountability assessment is unlikely to go away. It is too bound up with the politics of global competition and dissatisfaction with the level of historical accountability by the educational system. However, how we do accountability assessment matters, and it matters a lot, because educational practice (and learning) are influenced considerably by its design, content, and format. Thus, we have a range of choices with respect to how we deal with the influence and, indeed, the permanence of accountability assessment. At one end of this range, we can treat accountability assessment as a necessary evil to be minimized and marginalized as best we can. At the other end, we can attempt to rethink assessment comprehensively from the ground up.

Our work is an invitation to a conversation that needs to start by asking whether we can rethink assessment as a system so that it adequately serves both local learning needs and national policy purposes. That is, can we have an assessment system *of*, *for*, and *as* learning? We do not know the answer, but as assessment professionals, we believe we have a moral obligation to do our best to find out.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice, 42*(1), 8-29.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science.* Madison: University of Wisconsin Press.

Beller, E., & Hout, M. (2006). Intergenerational social mobility: The United States in comparative perspective. *Opportunity in America, 16*(2), 19-37.

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1). Retrieved January 26, 2008, from http://www.bc.edu/research/intasc/jtla/journal/v1n1.shtml

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

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. Retrieved January 26, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466

Bransford, J., Brown, A., & Cocking, R. (Eds.). (1999). *How people learn: Brain, mind, experience and school.* Washington, DC: National Academies Press.

Committee on Programs for Advanced Study of Mathematics and Science in American High Schools & National Research Council. (2002). *Learning and understanding: Improving advanced study in mathematics and science in U.S. high schools*. Washington DC: National Academies Press.

Committee on Prospering in the Global Economy of the 21st Century. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future.* Washington, DC: National Academies Press.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Gee, J. (1999). *An introduction to discourse analysis: Theory and method.* New York: Routledge.

Gitomer, D. H., & Duschl, R. A. (2007). Establishing multilevel coherence in assessment. In P. A. Moss (Ed.), *Evidence and decision making. The 106th yearbook of the National Society for the Study of Education, Part I* (pp. 288-320). Chicago: National Society for the Study of Education.

Glaser, R. (1976). Components of a psychology of instruction: Toward a science of design. *Review of Educational Research, 46*, 1-24.

Glaser, R., & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 20, pp. 393–419). Washington, DC: American Educational Research Association.

Greeno, J. G. (2002). *Students with competence, authority, and accountability: Affording intellective identities in classrooms.* New York: College Board.

Hogan, D. (2007). *Towards "invisible colleges": Conversation, disciplinarity, and pedagogy in Singapore* [Slide presentation]. (Available from Office of Education Research, National Institute of Education, Nanyang Technological University)

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2). Retrieved January 26, 2008, from http://escholarship.bc.edu/jtla/vol5/2/

Kirsch, I., Braun, H., Yamamoto, K., & Sum, A. (2007). *America's perfect storm: Three forces changing our nation's future*. Princeton, NJ: ETS.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, England: Cambridge University Press.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., et al. (2004). *International outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. perspective* (NCES 2005–003). Washington, DC: National Center for Education Statistics.

McLaughlin, M., & Talbert, J. E. (2006). *Building school-based teacher learning communities: Professional strategies to improve student achievement.* New York: Teachers College Press.

National Center on Education and the Economy. (2006). *Tough times, tough choices: The report of the New Commission on the Skills of the American Workforce*. Washington, DC: Author.

Neisser, U. (1967). *Cognitive psychology*. Englewood Cliffs, NJ: Prentice Hall.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Organisation for Economic Co-operation and Development. (2006). *OECD Education at a glance 2006: OECD briefing note for the United States*. Retrieved January 26, 2008, from http://www.oecd.org/dataoecd/51/20/37392850.pdf

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 307–353). Washington, DC: American Educational Research Association.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington DC: National Academies Press.

Perfetti, C. A. (1985). *Reading ability.* New York: Oxford University Press.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context.* New York: Oxford University Press.

Roseberry, A., Warren, B., & Contant, F. (1992). Appropriating scientific discourse: Findings from language minority classrooms. *The Journal of the Learning Sciences, 2*, 61-94.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sigel, I. (1993). The centrality of a distancing model for the development of representation competence. In R. Cocking & K. A. Renninger (Eds.), *The development and meaning of psychological distance* (pp. 141-158). Mahwah, NJ: Lawrence Erlbaum Associates.

Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory Into Practice, 1*(44)**.** Retrieved January 26, 2008, from http://findarticles.com/p/articles/mi_m0NQM/is_1_44/ai_n13807464

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2007). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability*. (NSSE Yearbook, 103 Part 2). Chicago: National Society for the Study of Education.

Appendix A

An Overview of Cognitive Science

Cognitive science comprises the multiple fields concerned with the study of thought and learning. Those fields include psychology, education, anthropology, philosophy, linguistics, computer science, neuroscience, and biology. Because it is an interdisciplinary field, cognitive science has no single genesis. Rather, its roots are found in disparate places.

Cognitive science has supplanted behaviorism as the dominant perspective in the study of thought and learning. Behaviorism grew out of the early 20th-century work of Thorndike, Watson, and Skinner, which rejected the theoretical need for internal mental processes or states. Behaviorism posited that highly complex performance (i.e., behavior) could be decomposed into simpler, discrete units and that such performance could be understood as the aggregation of those units.

The first cognitive science theories, in contrast, highlighted the importance of hypothetical mental processes and states. These theories focused on how individuals processed information from the environment to think, learn, and solve problems. These theories hypothesized specific mental processes as well as how knowledge might be organized in supporting acts of human cognition.

Among the theoretical perspectives commonly identified with cognitive scientific research is information processing. The information processing perspective is commonly traced to the publication in 1967 of Neisser's book, *Cognitive Psychology* as well as to Newell and Simon's 1972 publication of *Human Problem Solving*. This perspective viewed mental activity in terms similar to the way in which a digital computer represents and processes information. Now, with advances in neuroscience, the biological basis for cognitive processes is becoming much more clearly understood.

Alternative perspectives that include activity theory and situated cognition do not view cognition as simply a function of mental processes and knowledge that an individual brings to a task. Rather, in these views, cognition is not separated from context and the interactions in which mental activity and learning occur. Cognition is inherently a social activity, and learning involves increasingly sophisticated participation in the activities of particular social communities. Major contributions to this perspective are attributed to Vygotsky and Wertsch and more recently to Lave and Wegner, Scribner, Cole, and Greeno.

As cognitive science has matured, the field has recognized the importance of both the information-processing and the situated-cognition and activity-theory perspectives. Modern theories of learning, cognition, instruction, and assessment integrate these bodies of work into more unified and complete points of view.

Appendix B

An Overview of Psychometric Science

Psychometrics encompasses the theory and methodology of educational and psychological measurement. Its theory and methods essentially attempt to characterize some unobservable attribute of an individual, in terms of standing on a scale or membership in a category, and the degree of uncertainty associated with that characterization. The characterization may be made in relation to a comparison group (i.e., norm referenced) or it may be made in relation to some performance standard (i.e., criterion referenced).

The emergence of the field is often traced to the late 19th-century and early 20th-century work of such individuals as Wundt and Fechner in Germany; Galton, Spearman, and Pearson in England; and Binet in France. These individuals developed theories of intelligence, methods for quantifying psychological attributes such as the individual intelligence test, and techniques for analyzing the meaning of those quantifications, or scores, like the correlation coefficient and factor analysis. In the United States, the work of Thorndike, Yerkes, Thurstone, and Brigham, among others, led to creation of the group intelligence, aptitude, and achievement tests; the concept of developed ability; and further advances in techniques for analyzing test data.

Because many of the field's pioneers were also psychologists—Thorndike, Yerkes, Thurstone, and Brigham, to name a few—psychometrics was closely associated with, and influenced by, behaviorism, the dominant psychological perspective for most of the 20th century. That perspective is still quite evident in modern psychometrics, where the specifications for test development are commonly stated in terms of lists of behavioral objectives and test scores are transformations of the sum of the items answered correctly. Both practices fit well with the behaviorist notion that complex performance is the aggregation of discrete bits of knowledge.

Among the dominant methodological theories in psychometrics are classical test theory and item-response theory (IRT). Classical test theory is essentially a loose collection of techniques for analyzing test functioning, including but not limited to indices of score reliability, item discrimination, and item difficulty. These techniques include many of those generated in the 19th and 20th centuries by Pearson, Spearman, Thurstone, and others. Classical test theory is built around the idea that the score an individual attains on a test—the observed score—is a function of that individual's "true score" and error.

The second half of the 20th century saw the development of IRT and its widespread application. IRT is a unified framework for solving a wide range of theoretical and practical problems in assessment. Those problems include connecting the item responses made by an individual to inferences about his or her proficiency, summarizing the uncertainty inherent in that characterization at different score levels, putting different forms of a test on a common scale, and evaluating item and test functioning. Most recently, more complex psychometric approaches, including generalizations of IRT, have been created that better capture the multidimensional character typical of cognitive scientific models of cognition and learning.