Journal of Writing Assessment Vol. 3, No.2, pp. 81-98

Issues in Large-Scale Writing Assessment
Perspectives from the National Assessment of Educational Progress*

ARTHUR N. APPLEBEE
University at Albany, SUNY

This article reviews the development of the framework for the 2011 National Assessment of Educational Progress in writing. An issue paper commissioned by the National Assessment Governing Board is used to consider a number of continuing issues in large-scale assessment of writing, including the definition of the domain of writing tasks, which tasks should actually be assessed at which grade levels, the relationship of the assessment to postsecondary demands, the role of commonly available tools such as word processing software in the construct of writing achievement, the specification and measurement of achievement, the development of appropriate topics for writing, the issue of time for writing, and accommodations for English learners, students with disabilities, and low achievers.

During 2006-2007, committees broadly representative of K-12 teachers, school administrators, state departments of education, university specialists in the teaching and assessment of writing, parents, the general public,

*This article is based on an issues paper commissioned by the National Assessment Governing Board as background for the development of a framework for the 2011 NAEP writing assessment (Applebee, 2005).

Arthur Applebee is Distinguished Professor and Chair of the Department of Educational Theory & Practice at the University at Albany, SUNY; he also directs the National Research Center on English Learning and Achievement. He has published widely on the teaching and learning of language and literacy across grade levels, and has been a longtime consultant to the National Assessment of Educational Progress in reading, writing, and literature. His most recent book, *Curriculum as Conversation*, earned the David H. Russell Award for distinguished research from

Direct all correspondence to: Arthur N. Applebee, Department of Educational Theory and Practice, University at Albany, SUNY, 1400 Washington Avenue, Albany, NY 12222.

and the business community worked to develop a new framework for the 2010-2011 writing assessment of the National Assessment of Educational Progress (NAEP). The previous NAEP writing framework dated to 1989-1990 with revisions, primarily to test specifications, in 1995-1996 (National Assessment Governing Board [NAGB], 2002). Much of the substance of the framework went back even further, to the objectives for the 1983-84 assessment (NAEP, 1982. See appendix for a summary of the NAEP objectives from 1969-2011.)

As part of the development process, the NAGB commissioned a background paper to frame issues and debates in writing assessment that could be constructively addressed by the framework committees (Applebee, 2005). The article that follows incorporates that issues paper, adds the recommendations that emerged during the framework development process (NAGB, 2007), and extends the arguments beyond NAEP to concerns that relate more generally to large-scale assessment of writing.

The perspective that frames these issues is a personal and practical one, drawing on recent research and scholarship, changes in policy and practice over the past 20 years, and the collective experience of myself and many others in the development, analysis, and reporting of NAEP assessments.

Underlying all of the specific issues that follow is a larger one: What information about how students write should NAEP and other large-scale assessments provide to interested members of the general public, policymakers, and educators? Although it is a seemingly simple question, buried within it are a variety of difficult issues on which there is currently little consensus, including how to describe the domain of writing tasks; the relationships among component skills, content knowledge, and generalized writing "fluency"; and the relevance of computer-based applications to definitions of writing achievement as well as to assessment techniques.

NAEP itself has a number of constraints and opportunities that set it apart from most other assessments of writing. The opportunities derive from the fact that NAEP does not report scores on an individual level. This allows it to use a matrix sampling design in which different students complete different tasks. It also allows NAEP to use a single rater in evaluating each writing sample (with appropriate checks for interrater reliability). As a result, NAEP assessments are able to include more than 20 writing tasks at a given grade or age level, many more than typically can be administered or scored in other writing assessments. Many states, for example, use a single task, as do the SAT and ACT college entrance examinations.

The constraints on NAEP assessments derive directly from the opportunities: In order to relate student performance across tasks and contexts, NAEP uses a complex balanced-incomplete block design (BIB spiraling) in which all tasks at a given grade level are paired with one another in overlapping subsamples of students. In order to do this, the assessment is organized in blocks of items that take equal time to complete for writing and other subjects being assessed. The result at present is that NAEP writing items are constrained to a maximum of 25 to 30 minutes of testing time. State writing assessments, in contrast, often offer considerably more time for writing and revision.

The Issue: What Types of Writing Should Be Assessed, and How Are They Related to One Another?

Recent research in writing has tended to emphasize the extent to which writing genres are socially situated and context-specific. This is true whether one begins with Miller's (1984) emphasis on genre as social action, or the systemic linguistics approach of the Australian genre theorists (Cope & Kalantzis, 1993; Halliday & Martin, 1993). These perspectives pose a challenge to the traditional emphasis on writing as a generic skill, taught primarily in English language arts or composition classes, and assessable through generic writing tasks detached from particular disciplinary or socially constituted contexts. They suggest that what counts as effective argument and persuasive evidence varies greatly in moving from one context to another, so that what counts as "good writing" is itself socially constructed and context-specific. As Halliday and Martin (1993) demonstrated, for example, science writing has many features such as reliance on technical vocabulary, use of the passive voice, and nominalization (use of verbs and adjectives as nouns) that English teachers would ordinarily find objectionable—although these features have evolved in science writing to serve particular communicative needs.

The current NAEP framework, which will remain in place through the analysis and reporting of results from the 2006-2007 writing assessment, derives from the work of Kinneavy (1980), Britton and colleagues (1975), and Moffett (1968) during the 1960s and 1970s, in interaction with perceptions of typical practice and school-based terminology for discussion of writing instruction. The domain of NAEP writing tasks is divided into three broad purposes for writing—informative, persuasive, and narrative. This framework encourages writing within each of these purposes involving a "variety of tasks" and "many different" audiences, triggered by a "variety" of stimulus materials (NAGB, 2002). There is no consensus in theory or practice, however, about the proper way to partition the domain of writing tasks, and there has always been a perception of overlap among the categories: Doesn't an author of an "informative" text implicitly intend to persuade a reader of the truth or accuracy of what is being said? Isn't narrative an important technique for both informing and persuading? ("Narrative" has itself evolved out of concerns in earlier versions of the assessment with "personal," "imaginative," or "expressive" writing, in an attempt to capture the genres of literature as well as of personal reflection.)

The problems in terminology extend to state writing assessments, which have often turned to NAEP as a starting point in designing their own assessments. Texas, for example, requires writing for "various audiences and purposes," in a variety of forms, including "business, personal, literary, and persuasive texts." California instead treats these generalized purposes as part of "writing strategies," and specifies a variety of specific genres to be assessed (e.g., at Grade 11, fictional, autobiographical, or biographical narrative; responses to literature; reflective compositions; historical investigation reports; and job applications and resumes.)

There are other alternatives. College entrance exams from the College Board and ACT both assume that good writing is a generic skill, at least in academic contexts; the College Board, for example, advises that high scores will go to "essays that insightfully

develop a point of view with appropriate reasons and examples and use language skillfully" (College Board, 2008).

From the Australian genre-theory perspective, Martin and Rothery point in another direction, with a list of schooled nonfiction genres: recount, report, procedure, explanation, persuasion, and discussion. Their listing, like others from the Australian group, introduces terminology unfamiliar to American readers, and also collapses their original insights about the situated nature of genre knowledge into a generic set of "school" genres that are not all that distant from Britton et al.'s (1975) and Moffett's (1968) subcategories of informational or expository writing.

The Outcome

Lacking a widely accepted way to resolve these problems in definition and categorization, the committees developing the 2011 NAEP framework (NAGB, 2007) proposed organizing the assessment around three broad purposes for writing that are closely related to the distinctions made in earlier assessments:

1. to persuade, in order to change the reader's point of view or affect the reader's action;
2. to explain, in order to expand the reader's understanding; and
3. to convey experience, real or imagined.

These represent an attempt to clarify and elaborate the categories of persuade, inform, and narrate in the previous assessment. The framework also attempts to separate purposes from the ways they are carried out, noting that there are a wide variety of strategies for thinking and writing that writers may use in addressing these purposes, including the traditional modes of narration and description, as well as processes such as analyzing and interpreting, and organizational strategies such as compare and contrast. Taking this notion of choices available to writers even further, the 2011 framework recommends that students in Grades 8 and 12 be allowed to choose the particular genre or form in which they will respond (e.g., letter, essay, brochure), rather than having the form of the response dictated by the writing prompt.

The purposes embodied in this proposal, despite the changes in terminology, will provide an easy transition for other assessments that look to NAEP for guidance. The proposal also acknowledges that there is at present no widely accepted alternative in either theory or practice.

The Issue: What Writing Tasks/Types Should Be Assessed At Each Grade Level?

Tangled with the problem of specifying the domain of writing tasks is the distribution of tasks across grade levels. The framework in place through 2006-2007 assumes that each of the broad purposes for writing is appropriate even for primary grade writers, with development taking the form of the ability to complete ever-more sophisticated or specialized tasks within those purposes. Although

informative writing tasks have been relatively uncontroversial across the grades, arguments have been raised against assessment of persuasive writing at the fourth grade level, and narrative (particularly story) writing at grade 12. At the fourth-grade level, the arguments have been that persuasive writing is

1. too difficult,
2. developmentally inappropriate, or
3. out of step with the curriculum.

At grade 12, the arguments have been that story writing is

1. too easy,
2. no longer relevant to the curriculum of most students, or
3. not consistent with the types of writing expected in college and the workplace.

The current framework addresses this issue by placing more emphasis on persuasive writing in Grade 12, and more on narrative writing in Grade 4.

NAEP itself offers some evidence on these arguments, in that achievement has been somewhat higher on narrative tasks and somewhat lower on persuasive ones. There has been a narrowing of the range of task difficulty over time, however, early assessments showed much greater between-task variation than is presently evident. This is the result of pilot-testing and task-selection procedures that have eliminated tasks that were very easy or very hard at a given grade level. In fact, the current framework cautions against items that are either too hard or too difficult (NAGB, 2002). One result of this has been that it is no longer possible to comment on tasks that lower-achieving students can complete successfully, because these tasks are no longer included in the assessment.

The Outcome

Most large-scale assessments have too few separate writing items to have a wide range of task difficulty. The NAEP framework for 2011 similarly recommends a focus on tasks that will encourage all students to write at some length, rather than including some unusually easy or unusually difficult tasks.

The NAEP framework for 2011 emphasizes the importance of writing for a wide range of purposes at all grade levels, including some attention to each of three broad purposes included in the assessment framework. In recognition of the shifting demands of the curriculum, however, the framework places somewhat more emphasis on writing to explain and to persuade in the upper grades, and correspondingly less emphasis on writing to convey experience (primarily storytelling and personal experience essays). For all three purposes, the framework recommends increasingly abstract content and more distant audiences in the upper grades.

The specific types of writing to be emphasized at different grades warrants careful consideration in any large-scale assessment. Curriculum has a tendency to narrow around the types that are assessed, often coupled with unintended effects on

what counts as writing well (Hillocks, 2002). Assessments that have to rely on a limited number of tasks at a given grade level might do well to consider designs that sample from a larger range of possible tasks at each grade level assessed, rather than focusing on one or two types.

The Issue: How Can the 12th-Grade Assessment be Structured to Measure Preparedness for Postsecondary Endeavors, Including College, Workplace Training, and Entrance into the Military?

In 2003 the NAGB established the National Commission on NAEP 12th Grade Assessment and Reporting to review the 12th grade NAEP assessment and to recommend improvements to NAGB. The Commission's report (2004) noted that the high school diploma is no longer a culminating degree for most students; 88% of eighth graders report wanting to continue into higher education, and 70% of high school graduates actually do so within 2 years of graduation. At the same time, 45%-55% of entering freshmen are unprepared for college work, as reflected in placements in remedial coursework during their first year in college.

Lacking any other national standard for measuring preparedness, the Commission recommended that new NAEP frameworks for the 12th grade be oriented toward assessing preparedness for the challenges of college, workplace training, and the military. At the same time, the Commission noted that there is little consensus on what "preparedness" means, and that validating measures of preparedness is likely to require extensive follow-up studies exploring how students at various achievement levels do in various post-high school contexts. The NAGB's Assessment Development Committee has endorsed this emphasis on 12th-grade preparedness, while noting that the issue is complex and the message that NAEP will send in this regard is very important.

The history of attempts to shape curriculum and assessment around preparedness for future life or work is not a happy one (Applebee, 1974). Past attempts to inventory necessary skills have tended to converge on simple skills that are easy to itemize (spelling, punctuation) rather than higher-level skills (e.g., thoughtful argument and use of evidence) that virtually everyone cites as essential goals of education. The result was usually a system of curriculum and assessment that focused on basic skills or on generic workplace tasks (e.g., business letter format) that easily degenerated into formulas with little real-world relevance.

The most extensive recent effort to relate high school achievement to preparedness both for college study and for the workplace is the American Diploma Project (2004). Drawing on studies of the skills needed in high-performance, high-growth jobs, as well as the requirements for college-level tasks, the American Diploma Project report emphasizes higher-level skills such as expressing ideas clearly and persuasively, and producing high quality writing resulting from careful planning, drafting, and meaningful revision. The report also includes extensive benchmarks meant to indicate the level of achievement appropriate for high school graduation. The 10 benchmarks for writing cover a wide range, from planning, drafting, and revising; to selecting language appropriate for purpose, audience, and context; to

writing well-structured academic essays and work-related texts; to using appropriate software programs. Benchmarks under other headings also refer to writing tasks, however, including benchmarks labeled as research, logic, informational text, media, and literature. Although the overall emphasis remains on higher-level accomplishments, the benchmarks show some of the problems of earlier attempts, with appropriate citation of print or electronic sources emerging as a benchmark at the same level of importance as writing an academic essay.

The Outcome

The issue of how current performance relates to the demands of future contexts is an important consideration in the development of any assessment. The new NAEP framework addresses this issue by stressing the continuity of skills that will be needed in postsecondary contexts, rather than by emphasizing particular postsecondary types of writing: Good writing at all levels entails appropriate development of ideas, logical organization, language facility, and use of conventions—all shaped by purpose and audience. Postsecondary contexts also emphasize effective analysis, interpretation, and problem-solving, which is reflected in the 2011 framework in a gradual increase in writing to explain and to persuade at Grades 8 and 12.

The Issue: Should the Writing Assessment be Computerized?

Computer use is becoming widespread in American schools, and by the 2011 assessment it should be even more so. In 2003, for example, virtually all schools reported having computers with Internet access, with no differences among schools serving demographically different populations. Student access to such computers for instructional use has also been increasing rapidly; there was one computer with internet access for every 4.4 students in 2003, compared with one computer for every 12.3 students in 1998 (U.S. Department of Education, 2005).

For writing instruction, the most important computer-based tool has been the word processor. Like the calculator in mathematics, word processing transforms the writing task, simplifying editing and revision and providing embedded tools for spelling and grammar checking. Although most assessments are still paper-and-pencil, computer-based assessment that allows the use of word processing is becoming more widespread. When the Test of English as a Foreign Language (TOEFL) exam was recently revised, for example, is was moved to an Internet-based format that assesses reading, writing, and spoken language skills; and the Canadian province of Alberta has, for a number of years, made provision for optional use of word processors for Diploma exams in English and other subjects (Alberta Ministry of Education, 2008; Russell & Plati, 2002).

Computer-based writing assessment nonetheless raises some difficult issues of equity and access. Writing produced on a computer tends to be longer than writing produced by hand, and longer writing tends to be more highly evaluated than shorter selections, perhaps because of the inclusion of more evidence or elaboration

(Bereiter & Scardamalia, 1987). The bias arguments run in both directions: Not having access to a computer may penalize those who are used to writing on a computer in school and at home; on the other hand, those who are not used to writing on a computer will either be handicapped by poor keyboarding skills, or if they compose by hand by the greater length of essays produced by their computer-using peers.

The research base on the effects of word processors on assessment results is slim and not particularly convincing; arguments that paper-and-pencil tests underestimate achievement of students who are used to writing on word processors treat writing as though it were being evaluated against an external, fixed standard (e.g., Russell & Plati, 2002), when in fact writing rubrics ordinarily reflect the circumstances of production. Rather than an overall increase in performance, a switch to a computerized assessment including word processing software is more likely to lead to changes in the benchmarks at each level in the scoring rubric to reflect the advantages accrued from the new format.

The most extensive study of the effects of computerizing a writing assessment is NAEP's 2002 study of writing online (Sandene et al., 2005). This special study compared performance on two NAEP writing tasks (one informative and one persuasive) at the eighth-grade level, when given as part of the regular paper-and-pencil assessment or given in a special Web- or laptop-based format that also included simple word processing tools. The detailed results show a number of topic-specific differences in performance across formats, but are generally encouraging. There were no equity-related differences in essay quality, although there was a 1% higher response rate for the paper-and-pencil version of one task. Males also wrote significantly longer responses on computer than on the paper-and-pencil version of one task, but their essays were not rated significantly higher.

Students with more hands-on computer skill (as measured by typing speed, error rate, and ability to use word processing tools) did better on both of the computer-based writing tasks; the correlation between their overall writing score and the measure of computer skill was .42; even after adjusting for paper-and-pencil writing achievement, computer skill still accounted for about 11% of the variation in computer-based measures of writing achievement. The "hands-on" computer familiarity measure, however, had a significant literacy component that may account for much of this relationship. Other measures of computer experience, including frequency of completing various kinds of writing assignments on a computer, were unrelated to computer-based writing achievement.

Overall, the authors of the NAEP writing online study conclude that aggregated scores from online assessment do not differ significantly from paper-and-pencil results, although results for individual students may do so.

Although school-level data have recently suggested that equity issues in computer access have been reduced, at the student level issues of access have not been completely resolved. In 2003, for example, there were fewer computers with Internet access available in schools serving high proportions of minority students than in schools with the lowest proportion of minority students (5.1 students per computer vs. 4.1 per computer). Data from the 2002 writing assessment suggest an even larger divide: Some 29% of White students reported using a computer for

writing "a lot," compared with only 19% of Black and 18% of Hispanic students (NAEP Data Explorer, 2002 Writing Assessment).

The Outcome

In designing the 2011 NAEP framework, the committees decided that writing in the 21st century will be computer-based. This is already how most students write, and it is certainly an expectation for writing in the workplace and in postsecondary education. Thus, the 2011 writing framework calls for assessing the ability to write using word processing software at Grades 8 and 12. The framework calls for students to write using "commonly available tools," including the various writing and editing tools widely available in commercial word processing programs. The framework also calls for a computer-based assessment to be phased in at Grade 4 over the life of the framework, as access to and experience with word processing becomes more widespread in the elementary grades.

Computer-based assessment seems almost an inevitable response to the frequency and scale of mandated assessments in all areas of the curriculum. For writing assessment, developers will need to consider how advances in computer use and availability are impacting writing instruction, and what this means for definitions of what it means to write well. If equity issues can be resolved, a computer-based assessment has a number of advantages in measuring writing achievement and in providing accommodations to students who need them (see section on accommodations). Equity issues, however, are much more acute for assessments that report individual scores than they are for NAEP, whose results can serve policy development by highlighting issues of access without penalizing individuals.

The Issue: What Aspects of Writing Achievement Should Be Measured?

Just as there is no widely agreed on definition of the domain of writing tasks, there are many competing approaches to measuring the various interrelated components of writing achievement. Over time, the primary rubric used to measure writing achievement in NAEP has evolved from a holistic rating to a prompt-specific primary trait rating (Lloyd-Jones, 1977) to the current set of purpose-related rubrics (one for each of the three purposes for writing) that can be seen as either generalized primary trait or focused holistic. Although NAEP reports have been organized around separate sections discussing informative, persuasive, and narrative writing, reporting has either remained at the level of individual writing prompts, or has been aggregated to a total writing score. There have been no separate subscales for types of writing in published reports or in the data available online (NAEP Data Explorer).

Other scoring systems have attempted to provide separate ratings for different features of a writing sample. The most widely used today is probably the 6-trait (or 6+1 trait) system disseminated by Northwest Regional Laboratory. This provides separate scores for ideas, organization, voice, word choice, sentence fluency, conventions, and (optionally) presentation. This system emerged out of the work of

Paul Diederich and his colleagues at Education Testing Service (Diederich, French, & Carleton, 1961), and can be a useful tool in reminding teachers and students of the many dimensions of effective writing. As a measurement tool, however, it is not clear that the profiles that result yield psychometrically useful information (Hill, 2001). Diederich's (1974) suggestion was to use the traits for socializing raters to a common standard, and then to drop the traits and focus on total scores.

But there have been many attempts to measure other aspects of writing achievement, including syntactic complexity, ability to edit and revise, mastery of writing conventions (punctuation, capitalization, usage, spelling), organizational ability, and vocabulary level. Such features are arguably of interest in understanding writing achievement, but they have usually required time-consuming scoring procedures and been complicated by the fact that the results are task- and content-specific. Syntactic complexity is usually greater for an analytic or persuasive task than for a narrative task, for example, reflecting the typically embedded nature of clauses in argumentative discourse. Error rates in writing conventions similarly vary with task—with errors tending to increase as tasks become more difficult, presumably as the result of the deflection of cognitive and linguistic resources from one aspect of the task to another.

Many measures of interest that are tedious to derive by hand are very easy to derive by computer. There are now a range of text-analytic software programs available that will report features such as number of words, variety in word choice, syntactic complexity, vocabulary level, and error rates. Many also calculate an overall quality score. If the 2011 writing assessment is computer based, it would allow the assessment of aspects of writing development that can currently only be examined in special studies on limited subsamples of papers.

There is of course another psychometrically efficient option to obtain measures of some of these features. Knowledge of written language conventions and vocabulary level, for example, can be tested quite efficiently in multiple-choice formats. Such measures are highly reliable and have good predictive validity (Breland, Camp, Jones, Morris, & Rock, 1987; Godshalk, Swineford, & Coffman, 1966); however, they have long been resisted by the community of writing educators because of their impact on curriculum and instruction. Such short-answer formats divert the focus of instruction away from student experiences with writing extended text.

Thus, another benefit of computer-based analyses of features of writing is the ability to derive these measures from samples of extended writing rather than from short-answer or multiple-choice formats. This could provide a richer portrait of writing achievement without sacrificing the emphasis on the creation of complete texts.

The Outcome

For the 2011 NAEP, the framework development committees have recommended a focused holistic scoring system with components tailored to the three purposes to be assessed (to explain, to persuade, and to convey experience). Raters will be trained to attend to the development of ideas, to organization, and to language facility and use of conventions, all as appropriate and relevant to the purpose and audience of each task.

The new framework also envisions a "Profile of Student Writing" that would examine in more detail each of these three components. The profile will rely to the extent possible on measures that can be computed automatically from the word-processed writing samples, but will also include analytic scoring of a subsample of student writing for features that cannot be derived from computerized text analyses.

The Issue: What Should Students Write About?

The current framework for the NAEP writing assessment emphasizes writing prompts that are accessible to all students. In practice, this results in an emphasis on common life experience, generic academic content (e.g., favorite books or music, fictional or historical figures, the value of space travel), and on writing that reflects public discourse in a democratic society (e.g., persuasive tasks about community or school issues). If content is provided, it is typically more illustrative than substantive—a brief "story starter," a picture stimulus, or a brief framing of sides of a "controversial" issue. (Real controversies that have political volatility do not make it through the item-review process.) When reading and language difficulties of English-language learners and low-achieving students are taken into account, the push in item development is toward simple and "clean" writing prompts with a low vocabulary load.

At the same time, writing plays a role in virtually all of the other subject area assessments in NAEP. Both short and extended constructed responses comprise major sections of the current assessments in science, history, geography, civics, and reading, as well as the frameworks for new assessments in economics and foreign languages. Rubrics in these assessments bear little similarity to the rubrics in the writing assessment, however, often emphasizing listing of specific content rather than the construction of an argument or explanation.

This creates an artificial separation of writing from content knowledge. As Hillocks (2002) pointed out in his critique of state writing assessments, one of the biggest problems in many assessments is the lack of a substantive content base on which to base the writing. Without a content base, much of the writing that results is formulaic and shallow.

The Outcome

For the 2011 NAEP, the framework committees have recommended a continued focus on generic, easily accessible content, including short reading passages, visual stimuli, or graphics. The one major change in the content of the writing tasks is the recommendation that students at Grades 8 and 12 be allowed to choose the genre or form they consider most appropriate to the audience and purpose specified in the prompt. The framework recommends pilot-testing items in a variety of formats (with form specified, without form specified, and with a choice of forms specified) in order to better understand the interaction between purpose, choice of genre or form, and student performance in an assessment context.

Other large-scale assessments vary in the degree to which they rely on generic, easily accessible content. Although many use items very similar to those in NAEP, others, such as New York State, base writing on extended reading passages, or

include at least some classroom-based writing as part of the assessment (Kentucky, Vermont). A more general issue for assessment developers is whether it would be useful to increase the content load of student writing prompts, and if so, how this could be done within current assessment frameworks or through extensions of them. One possibility, particularly if writing and other assessments become computerized, would be through the adoption of some common metrics for assessing quality of writing across assessments in different content areas.

The Issue: How Should the Framework Address the Question of Time?

Time to write has been an issue for successive NAEP writing framework committees, and has led both to changes in time allotments and to special studies. From 1970 to 1979, NAEP writing assessments had items of variable length, from a few minutes for completing forms to nearly 30 minutes on some essay tasks. The move to BIB spiraling in the 1984 assessment reduced the maximum time to 15 minutes. Beginning with the 1992 assessment, this was increased to 25 minutes (with a subset of 50-minute writing tasks that was eliminated in the 2002 assessment).

Two issues usually dominate discussions of writing time: Do the results misrepresent overall writing achievement because students have too little time to write? And does the limited time allowed penalize some groups of students, particularly those whose classrooms have emphasized an extended process of writing and revision? (Conversely, will extended time frustrate lower achieving students and exacerbate achievement gaps?)

The issue of time has been driven by a tension between the constraints of assessment and the conventional wisdom on instruction. One of the accomplishments of the writing process movement in instruction was to remind teachers and students that writing takes place over time—that there are identifiable strategies for generating ideas, drafting, revising, editing, and sharing that shape and reshape a final written text. During the past 30 years of writing assessment, the proportion of teachers claiming to emphasize process-oriented approaches to writing instruction has risen sharply; by 1998 it was central to the instruction of 70% of fourth- grade teachers surveyed, and used to supplement instruction by another 28% (Applebee & Langer, 2006). Comparable figures were reported by Grade 8 and 12 students in the 2002 assessment. (Background questions and grade levels at which they are asked vary from assessment to assessment so there is no single set of data on which to draw.)

Given the constraints of large-scale assessment, NAEP has always emphasized that the writing assessment focuses on first-draft writing (as do the College Board and ACT in their college entrance examinations). Given the overall design of the assessment, when NAEP has included 50-minute tasks the trade-off has been these tasks have not been scalable. (With a 50-minute prompt, each student completes only one task, so interrelationships among tasks cannot be determined.) In 1998, the results of these longer tasks do not seem to have even been reported.

Previous NAEP studies of the impact of additional time have yielded mixed results. One special study compared 11th-graders' performance on a persuasive

writing task given in 16- or 50-minute time blocks but mixed together for scoring with identical rubrics. As common sense might suggest, the students who had more time for writing scored higher—although the gain was less than might have been expected: 45.4% produced adequate or better responses in 50 minutes, compared with 33.8% in the 16-minute format. The benefits of extra time were not equally distributed among students, however; the extra time made little difference to the weaker writers, increasing the performance gap between the two groups (Applebee, Langer, & Mullis, 1989). The 1992 assessment reported results for 50-minute as well as 25-minute prompts, with achievement noticeably higher on a 50-minute informative writing task than on the other, 25-minute informative tasks. But comparable increases in achievement did not occur on 50-minute narrative and persuasive tasks included in the assessment; the report concluded that the differences were likely to be topic-related rather than a function of the increased response time (Applebee, Langer, Mullis, Latham, & Gentile, 1994).

These results do not mean that time is not an important factor in quality of writing; simply that the effects of time within the constraints of NAEP writing prompts as they are currently designed are not as large as might be thought, and may be topic-specific. It may be that for meaningful effects of time to emerge, the nature of the tasks would need to be radically reconstrued to incorporate, for example, significant content to be examined or reviewed, or significant feedback to be provided after an initial draft.

Related to the issue of time is whether to provide any special supports for students as they write, particularly supports related to how students use the time available to them. The current NAEP assessment format, for example, includes a blank space that students are encouraged to use to plan their writing. Students also receive a booklet, "Ideas for Planning and Reviewing Your Writing," that suggests planning and revision strategies.

The Outcome

For the moment at least, the NAEP writing assessment remains constrained by a 25- or 30-minute format. Other large-scale assessments have the option to explore formats that go well beyond this, however, and to investigate the effects of variations in time and administrative procedures on student performance. New York State, for example, provides substantive material for students to read and write about, using extended, 3-hour time blocks. Kentucky pairs classroom-based writing with an assigned task, and also insists that some of the classroom-based writing come from subject areas other than English. Hillocks (2002) commented favorably on both of these assessments in his critical look at the quality of writing elicited by various approaches to writing assessment at the state level.

The Issue: What Accommodations Should Be Made for English-Language Learners, Students With Disabilities, And Low-Achieving Students?

NAEP policy is to include as many students as possible in all of its assessments, without altering the construct being measured. In practice, this is accomplished by careful item-development procedures and, where necessary, by providing accommodations to students with disabilities and English-language learners. Typical NAEP accommodations include more testing time, small-group testing, and other appropriate accommodations depending on the NAEP subject being tested.

As noted earlier, recent writing framework committees have been concerned with making all writing prompts as accessible as possible to all students. This has usually meant a lightening of the vocabulary load and of content provided through the prompt, so that these students would not be put off by problems in understanding before even beginning with their own writing.

The inevitable consequence of this accommodation in the current booklet-based testing format has been that there has been little room to experiment with alternative formats that have the possibility of providing a more substantive context for at least some of the writing tasks.

A computer-based assessment in 2011 would open up a variety of new possibilities, particularly if paired with writing analysis software that could make rapid initial judgments about writing proficiency of individual students. A simple "range finder" task, for example, might be used to place students in alternative formats adjusted to their general literacy levels. (New Zealand, for example, uses a very simple and quick initial task in its reading assessment; see NEMP, 2000). Or the response level on the first task administered to each student could be used (with computerized scoring) to select a second task of appropriate difficulty. This could serve to provide accommodations for students who need them, and also to provide greater challenges for higher-ability students.

The Outcome

The 2011 NAEP writing framework recommends typical accommodations such as large-print booklets, extended time, or one-on-one testing when needed. It also emphasizes item-development procedures that will ensure that every item is presented in a simple and clear format accessible to all students.

A move to a computer-based administration for NAEP and other large-scale assessments opens up the possibility of more tailored accommodations in the future, however. By taking advantage of the computer platform, future assessments might be able to individualize such factors as reading load and vocabulary level in ways that are not possible with paper-and-pencil assessment booklets. Assessment developers need to continue to give serious consideration to the effects of accommodations for poor readers and English-language learners on the overall content of the assessment, and look for alternatives that might provide a richer array of assessment options for all students.

Conclusion

The framework for the NAEP writing assessment has evolved significantly over the years, in the nature of the writing prompts, in the time available for each task, and in its emphasis on rhetorical features such as audience and purpose. The issues considered in developing a new framework for the 2011 writing assessment have no easy answers, but the changes recommended for 2011 represent an updating that reflects recent changes in scholarship and practice, and that will also return NAEP to its position as a leader in assessment practice and assessment technology. The most significant change involves the movement from paper-and-pencil booklets to a computer-based assessment, which carries with it potential changes in many different aspects of the assessment: in the underlying construct that is being assessed, in possibilities for analyzing the writing samples and reporting on student performance, and in adaptive testing. The challenges will be large, but the opportunity for improving our understanding of student performance is equally large.

States and other groups developing writing assessments will have to confront similar issues in the design of their own assessments, though the particular answers they reach will vary in response to their varying purposes and constraints.

Appendix: The Evolution of the NAEP Writing Framework

Cross-Sectional Writing Assessments

1969-1970 Assessment

1. Write to communicate adequately in a social situation.
2. Write to communicate adequately in a business or vocational situation.
3. Write to communicate adequately in a scholastic situation.
4. Appreciate the value of writing.

1973-1974 and 1978-1979 Assessments

1. Demonstrates the ability in writing to reveal personal feelings and ideas (through free expression and through the use of conventional modes of discourse. [For 1978-1979, reinterpreted as "ability to engage in writing for expressive purposes."])
2. Demonstrates the ability to write a response to a wide range of societal demands and obligations. Ability is defined to include correctness in usage, punctuation, spelling, and form or convention as appropriate to particular writing tasks (social, business/ vocational, scholastic). [For 1978-1979, interpreted as explanatory or persuasive writing done for a particular audience.]
3. Indicates the importance attached to writing skills (recognizes the necessity of writing for a variety of needs, writes to fulfill those needs, and gets satisfaction, even enjoyment, from having written something well).

1983-1984 and 1987-1988 Assessments

1. Students use writing as a way of thinking and learning (for subject knowledge and self-knowledge).
2. Students use writing to accomplish a variety of purposes (informative, persuasive, and literary). [Literary was variously interpreted as "imaginative" and as "personal /imaginative narrative" in reports on these assessments.]
3. Students manage the writing process (generate, draft, revise, edit).
4. Students control the forms of written language (organization and elaboration, conventions).
5. Students appreciate the value of writing (for interpersonal communication, for society, and for self).

1991-1992, 1997-1998, 2001-2002, and 2006-2007 Assessments

1. Students should write for a variety of purposes: narrative, informative, and persuasive.
2. Students should write on a variety of tasks and for many different audiences.
3. Students should write from a variety of stimulus materials, and within various time constraints.
4. Students should generate, draft, revise, and edit ideas and forms of expression in their writing.
5. Students should display effective choices in the organization of their writing. They should include detail to illustrate and elaborate their ideas, and use appropriate conventions of written English.
6. Students should value writing as a communicative activity.

2010-2011 Assessment and Beyond

The 2011 NAEP writing assessment will assess the ability

1. to persuade, in order to change the reader's point of view or affect the reader's action;
2. to explain, in order to expand the reader's understanding;
3. to convey experience, real or imagined.

Beginning in 2010-2011, the assessment will be administered using commonly available word processing tools at Grades 8 and 12, with a similar assessment being phased-in at Grade 4 by 2018-2019.

Long-Term Trend Assessments

1969-1979 through 1983-1984

Writing prompts developed using the 1969-1970 framework were re-administered to study long-term trends through 1983-1984, although trend reports have reinter-

preted prompts in light of the writing objectives in place at the time of reporting. Trends were analyzed at the item level rather than using scaled scores.

1983-1984 through 1995-1996

Writing prompts developed using the 1983-1984 framework were re-administered to study long-term trends through 1996, again with reinterpretation of prompts in light of later revisions to the writing framework. Two assessments (1993-1994 and 1995-1996) were limited to long-term trends. The last writing long-term trend assessment administered and reported was for 1995-1996. Although writing long-term trend data were collected in 1999, results were not reported due to instability of the score scale. NCES and NAGB determined that the writing long-term trend assessment should be discontinued because too few prompts were administered to enable reporting of viable trend results.

2010-2011 and Beyond

Writing prompts and procedures developed for the 2011 assessment will be used to establish a new trend line.

Framework References

Norris. E. L. (Ed.). (1969). *Writing objectives*. Ann Arbor, MI: Committee on Assessing the Progress of Education.

National Assessment of Educational Progress. (1972). *Writing objectives: Second assessment*. Denver, CO: Education Commission of the States.

National Assessment of Educational Progress. (no date). *Supplement to the 1973-74 writing objectives*. Unpaginated insert to Writing objectives; Second assessment.

National Assessment of Educational Progress. (1982). *Writing objectives: 1983-84 assessment*. Princeton, NJ: Educational Testing Service.

National Assessment Governing Board. (2002). *Writing framework and specifications for the 1998 National Assessment of Educational Progress*. Washington, DC: U.S. Government Printing Office.

References

Alberta Ministry of Education. (2008). *About provincial testing*. Retrieved November 2005, from http://education.alberta.ca/admin/testing.aspx

American Diploma Project. (2004). *Ready or not: Creating a high school diploma that counts*. Washington, DC: Achieve, Inc.

Applebee, A.N. (1974). *Tradition and reform in the teaching of English: A history*. Urbana, IL: National Council of Teachers of English.

Applebee, A.N. (2005). *NAEP 211 writing assessment: Issues in developing a framework and specifications*. Washington, DC: National Assessment Governing Board.

Applebee, A.N., & Langer, J.A. (2006). *The state of writing instruction in America's schools: What existing data tell us (*A report to the National Writing Project and the College Board). Albany, NY: Center on English Learning and Achievement.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1989). *Understanding direct writing assessments: Reflections on a South Carolina writing study.* Princeton, NJ: Educational Testing Service.

Applebee, A. N., Langer, J.A., Mullis, I.V.S., Latham, A.S., & Gentile, C. A. (1994). *NAEP 1992 writing report card*. Washington, DC: U.S. Government Printing Office for the National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. (1987). *Assessing writing skill*. New York: College Entrance Examination Board.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. (1975). *The development of writing abilities*. London: Macmillan Education.

College Board. (2008). *Strategies for success on the SAT essay*. Retrieved November 2005, from http://www.collegeboard.com/student/testing/sat/pre_one/essay/pracTips.html

Cope, B., & Kalantzis, M. (Eds.). (1993). *The powers of literacy: A genre approach to teaching writing*. Pittsburgh, PA: University of Pittsburgh Press.

Diederich, P. B. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, John W., & Carlton, Sydell T. (1961). *Factors in judgments of writing ability. Research Bulletin RB-61-15*. Princeton, NJ: Educational Testing Service.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Halliday, M.A.K., & Martin, J.R. (1993). *Writing science: Literacy and discursive power*. Pittsburgh: University of Pittsburgh Press.

Hill, R. (2001). *Analysis of the scoring of writing essays for the Pennsylvania system of student assessment*. Dover, NH: National Center for the Improvement of Educational Assessment.

Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning*. New York: Teachers College Press.

Kinneavy, J. (1980). *A theory of discourse: The aims of discourse*. New York: W.W. Norton.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33-66). Urbana, IL: National Council of Teachers of English.

Martin, J. R., & Rothery, JJ. (1980). *Writing Project Report No. 1*. Sydney: University of Sydney, Department of Linguistics.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech, 70*, 151-167.

Moffett, J. (1968). Teaching the universe of discourse. Boston: Houghton Mifflin.

National Assessment of Educational Progress (NAEP). (1982). *Writing objectives 1983-84 assessment*. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

National Assessment Governing Board. (2002). *Writing framework and specifications for the 1998 National Assessment of Educational Progress.* Washington, DC: U.S. Department of Education.

National Assessment Governing Board. (2007). *Writing framework for the 2011 National Assessment of Educational Progress* (Prepublication ed.). Washington, DC: National Assessment Governing Board.

National Commission on NAEP 12th Grade Assessment and Reporting. (2004). *12th grade achievement in America: A new vision for NAEP. A report to the National Assessment Governing Board*. Washington, DC: National Assessment Governing Board.

New Zealand National Education Monitoring Project (NEMP). (2000). *Reading and speaking assessment results 2000*. Retrieved November 2005, from http://nemp.otago.ac.nz/read_speak/2000/

Russell, M., & Plati, T. (2002). Does it matter with what I write? Comparing performance on paper, computer, and portable writing devices. *Current Issues in Education* (Online) 5(4).

Sandene, B., Horkay, N., Bennett, R.E., Allen N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project*. Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2005). Internet access in U.S. public schools and classrooms: 1994-2003. *National Center for Education Statistics*. Retrieved November 2005, from http://nces.ed.gov/surveys/frss/publications/2005015/2.asp

Issues in ESL writing assessment. College ESL. 6.1.52â€“72.Google Scholar.Â Issues in evaluating and maintaining an ESL writing assessment program. In Hamp-Lyons, L. (ed.) Assessing second language writing in academic contexts.Â Lumley, Tom 2002. Assessment criteria in a large-scale writing test: what do they really mean to the raters?. Language Testing, Vol. 19, Issue. 3, p. 246. Large-scale Assessments in Education. An IEA-ETS Research Institute Journal. About.Â This issue showcases innovative research and development in the area of log-file and process data with a focus in international large-scale assessments, including the latest methodological and theoretical approaches, and critical ethical considerations. Edited by: Qiwei (Britt) He, ETS, Irwin Kisch, ETS, Caroline McKeown, ERC, Jude Cosgrove, ERC Collection published: 29 May 2020. 2017. Results, methodological aspects and advancements of the Programme for the International Assessment of Adult Competencies (PIAAC). This special issue presents results of the PIAAC study regarding socio-economic d 1. Writing Assessment Scale 1. Assessment criteria. | marking a piece of writing for an exam. For the B1 Preliminary for Schools exam, these are: Content, Communicative Achievement, Organisation and Language. 2. Writing Assessment subscales. 2. Assessment categories. Each piece of writing gets four sets of marks for each of the subscales, from 0 (lowest) to 5 (highest).Â Cambridge English writing examiners are extensively trained to assess learnersâ€™ writing using these assessment scales, bands and descriptors. The quality and consistency of their marks is closely monitored by a team of senior examiners through an annual certification process and during live. | testing sessions. scale in writing assessment, which is one of the conditions to. be considered when developing rating scale. Moreover, when.Â It collected quantitative and qualitative data from twenty native-English-speaking undergraduate students at a large Midwestern university. The quantitative data consisted of scores earned by the participants upon completing two writing tasks: one which included the new fluency intervention and one which served as the control condition. The reliability of writing performance assessment has been improved over the years through a combination of: Â· training (McIntyre, 1993; Weigle, 1994a; 1994b); Â· better speciÂ®cation of scoring criteria (Jacobs et al., 1981; Alder-. son, 1991; Hamp-Lyons, 1991; Weigle, 1994a; North, 1995; North and Schneider, 1998); and perhaps, to some extent, Â· tasks (Hamp-Lyons, 1990; 1996; Kroll and Reid, 1994).Â Concerns focus on issues such as the superÂ®ciality of rating scales in comparison with the complexity of written texts and the readings made of them. Charney (1984), in an inÂ¯uential article, raised a num-ber of questions about holistic rating. She recognized a range of fac-tors as relevant to the rating process. She categorized certain factors.