Fighting Back: Assessing the Assessments

George Hillocks Jr.


Stable URL:
http://links.jstor.org/sici?sici=0013-8274%28200303%2992%3A4%3C63%3AFTA%3E2.0.CO%3B2-C

*The English Journal* is currently published by National Council of Teachers of English.
Fighting Back: Assessing the Assessments

Any teachers and administrators have complained vociferously about the growing mania for testing. Recently I heard a speaker complain for over twenty minutes that there is far too much testing, that testing begins too early in children’s educational careers, that teachers should not be held responsible for students’ learning because the tests do not account for what students know when they enter teachers’ classes, that too much valuable school time is spent on testing, that the results of tests serve only to categorize children rather than educate them, and on and on. Surely all of these complaints are valid.

But will simple complaints be adequate to make any change in the testing system? I doubt it. With the new federal education act, testing has become the official driving force to reform education. By and large, Congress and the public have bought George W’s lines, which he used as a mantra throughout his campaign for the presidency. For example, at the NAACP meeting on July 10, 2000, he announced that “a great movement of education [sic] reform has begun in this country,” one that is built on “clear principles.” His principles, however, sound more like dictates: “Raise the bar of standards. Give schools the flexibility to meet them. Measure progress. Insist on results. Blow the whistle on failure. Provide parents with options to increase their influence. And don’t leave any child behind.” This simple-minded view of education, with testing at its center, has been enacted into law. States that do not comply with the demand for testing will fail to receive federal money. Testing is here to stay for a long time, and my guess is that it is likely to increase.

With this explosion in testing and with federal money at stake, teacher complaints about too much testing begin to sound like simple whining, moaning, and groaning, as we used to say. I am certain that, for complaints to have any impact on school administrators and state legislators, they will have to rise above the level of whining. They will have to be based on thoughtful analyses of the learning standards, the test items, the scoring procedures, and, in the case of writing, the scoring criteria and benchmark or anchor papers that illustrate the criteria. When possible, the analyses need to examine the impact that testing has on teaching and the curriculum—that is, on how teachers and administrators decide to prepare students for testing. If enough people listen and are convinced by such analyses, perhaps it will be possible to improve testing practices, if not eliminate them. I am not sure I would bet on change, even in the face of a clear demonstration of the harmfulness of the tests. But I believe that the attempt is worth the effort.

With funding from the Spencer Foundation, I have had the opportunity over several years to study the writing assessments of five states (Illinois, Kentucky, New York, Oregon, and Texas) in considerable detail. My assistants and I interviewed about eighty teachers and administrators in six school districts in each state; examined state, local, and commercial materials related to the assessments; and examined the writing assessments produced by the other forty-three states that have them. The chief
I am certain that, for complaints to have any impact on school administrators and state legislators, they will have to rise above the level of whining.

The finding of this study is that writing assessment drives instruction (Hillocks). It stipulates the kinds of writing that should be taught; it sets the standards for what counts as good writing; and it sets the conditions under which students must demonstrate their proficiency, and, as a result, sets what students learn. The problem is that no matter how foolish the testing, it drives the writing curriculum and instruction in the state. In each state, we interviewed teachers in two large urban districts, as well as two suburban, one small town, and one rural district. The vast majority of these teachers, although complaining about the testing programs, tended to accept the tests uncritically.

Assessing the Assessments

I believe that it takes no special expertise to undertake a critical examination of an assessment’s validity, to determine the extent to which it tests what it purports to assess. Here are some types of basic questions to help you begin the analysis of your own local and state assessments.

Questions about Learning Standards

What are the learning standards upon which particular tests and test items are based? Do the standards appear to be consistent? Do they include all that you think is necessary for high literacy? Are they explicit enough to be useful? Are their priorities reasonable? For example, do learning standards include statements about critical thinking or supporting ideas or providing evidence? If they do not, there is a good chance that the testing programs deal only with the simplest academic skills. Even if they do, you will still need to discover what the standard writers really mean by their phrases about supporting ideas and critical thinking.

Questions about the Theory of Writing

Usually, a state announces some conception of what writing involves, what the types of writing are, and what learning to write involves, but not always. In our study, Oregon had announced no theory at the time of the study. The questions to ask are important because the conception that the state adopts will result in decisions about what will be tested and how it will be tested. Is the theory comprehensive? Does it include attention to all kinds of writing, even those we might not attend to in schools? Does the theory attend to what writing entails, e.g., the writing process? Does it attend to learning processes for writing? Does it have a basis in research? Finally, are all of the tests sound? To what extent do they hold up to careful scrutiny? If they do not hold up well to scrutiny, it is predictable that the writing assessment itself will be a disaster.

Questions about Test Items and Prompts

Examine the test items and writing prompts to see if they reflect the language of the standards. You are likely to find that while the standards reflect some concern for critical thinking, the test items and prompts do not. It is easy to talk superficially about critical thinking but difficult to write prompts and items that actually reflect such thinking. How many items in the test reflect the various learning standards? Are they in proportion to the importance of the standards? Tests like the Chicago CASE exams tend to focus on recall and low level reading such as finding explicit information in texts or on low level inference and identification of terms. Are multiple choice items unambiguous? Are the distractors (the intended wrong answers) defensible as correct answers? Chicago CASE exams last year included several items for which one of the distractors was defensible as a correct answer.

Questions about Scoring Criteria

More and more tests are including what are called “constructed responses.” Most state writing assessments require samples of actual writing, but exams in many academic areas now include constructed responses, as do Kentucky’s in mathematics and science. For all of these, examine the scoring criteria. Do the criteria reflect the language of the stan-
ards, or do they fudge on issues such as critical thinking? Do the criteria really reflect what you would regard as an effective persuasive or narrative piece of writing?

If the standards, tests, scoring criteria, and benchmark examples do not reflect good writing, then you will have discovered a serious complaint against the testing program.

Questions about Benchmark Responses or Compositions

Next, examine the benchmark papers that are supposed to illustrate the criteria. This is where you find what the testing agency really means by the language in the learning standards and the scoring criteria. Do those pieces of student writing at the highest score levels exhibit critical thinking, the use of evidence, or other indications of thought? Or do the papers receiving the highest scores amount to no more than language to fill up the required space? Go back to the testing format. Does the test provide material for the students to write about, or does it leave students to somehow generate concrete material out of nothing? Could you or another educated adult write a thoughtful response to the prompt in the time allowed? If not, it is unreasonable to ask students to do so.

If the standards, tests, scoring criteria, and benchmark examples do not reflect good writing, then you will have discovered a serious complaint against the testing program. Remember that countless hours are spent on preparing for these tests, often over several school years. High school English classes typically meet for forty to fifty minutes a day. In one Chicago high school, English meets for forty-five minutes, 135 hours in a school year of 180 days, precious little time to deal with the complex learning of high literacy that our culture now requires. In many schools, half or more of that time is ripped away to prepare for tests that do far more harm than good to a student’s education.

Applying the Questions

In my study, the writing assessments in Texas and Illinois were the most grievous examples of poorly considered choices for assessment at every turn. Most assessments in the country are comparable to those in Texas and Illinois. Let me use Illinois as an example to illustrate how to examine an assessment.

Standards

Learning standards are usually implied in the legislation that brings assessments into place. Sometimes, legislation presents only a very general outline of the standards upon which the criteria are to be based. The Illinois legislation, for example, indicates that “as a result of their schooling, students will be able to write standard English in a grammatical, well-organized and coherent manner for a variety of purposes” (Illinois State Board of Education [ISBE], Write On, Illinois/178). However, a piece of writing may be grammatical, well-organized, and coherent without being well-reasoned, thoughtfully developed, or effective in any way.

The ISBE Academic Standards are somewhat more explicit. The statement at least mentions the importance of problem solving in a prefatory statement to the standards. The ISBE provides a statement called “Applications of Learning,” which is technically not part of the standards but which presents comments on how they apply to “Solving Problems,” “Communicating,” “Using Technology,” “Working on Teams,” and “Making Academic Connections.” The short, bold-faced version of the statement on problem solving tells us that students should “recognize and investigate problems; formulate and propose solutions supported by reason and evidence” (1–2). The actual learning goals or standards make no allusion to reason and evidence, not even in the brief discussions of the late high school “benchmarks of learning” (4–23). The closest these statements come to explaining what is meant by “reasoning” is a benchmark statement for late high school writing listed as 5.B.5: “Evaluate the usefulness of information; synthesize information to support a thesis; and present information in a logical manner in oral and written forms as individuals and members of a group” (21; online version omits the final 7 words).
The problem with such statements, however well intended, is that they do not explain what will count for “evaluation,” “support,” or “logical manner”; yet, knowing what counts as support, evidence, logic, and reason results in very important differences in what students learn. Clearly, then, there is an initial problem with the Illinois learning standards: they are not explicit about some of the most important goals. We can be suspicious of what we will find in the actual testing program.

Theory of Writing

Usually, but not always, states base their selection of writing prompts on some theory that provides an explanation of what kinds of writing might be taught and tested. Write On, Illinois! lists forty-five teachers from around the state who developed a theory of writing, claiming that the categories of persuasive, expository, and narrative represent all writing domains. The rationale claims the test to be comprehensive, to cover the whole of writing. However, the typology excludes advertising, poetry, drama, jokes, and many other kinds of writing. The initial decision about what counts as writing has powerful implications for what will be tested. As it turns out, fiction is also excluded, since the narrative prompts focus on personal writing. The theory does not attend to the writing process or on learning to write. Nevertheless, our interviews revealed that teachers largely adopted this outline from the state as their “theory” of writing and found it suitable in supporting the kind of writing program they wanted for their schools.

By contrast, the Kentucky assessment is based on a far more complex theory of writing espoused by James Britton, et al. It allows for many kinds of writing, providing a far richer basis for a curriculum that includes all kinds of imaginative writing, several kinds of personal writing, and many kinds of transactional writing. In addition, the Kentucky assessment emphasizes the writing process, the idea of writing for a real audience, and circumstances necessary for learning to write.

Writing Prompts

Illinois makes use of three kinds of writing prompts: narrative, expository, and persuasive. I confine my comments here to expository and persuasive. There are many problems with these prompts, chief of which is the poor differentiation among them. Write On, Illinois! presents the following description of expository prompts: “Students are asked to explain, interpret, or describe something based on background experiences or information provided in the prompt.” Two of the sample prompts in Write On, Illinois! provide a list of possible topics under a superordinate category: workers and historical periods. Topics under workers include mechanic, doctor, waitress, and so on. Students are to choose one of those or any other and write a composition explaining how the work is “important to your school, your community, or the country as a whole.” The prompt under historical periods suggests “the time George Washington lived,” “the time of the Wild West,” “more recent times when your parents were young.” Students are to choose a time from the list or one of their own and write a report in which they “name the time,” “give reasons why it is important,” and “explain what people did and/or the things that happened during that time and why those things are important.” Students are told to “be sure to report about things that actually happened” (70). Finally, a third sample prompt, for sixth grade and up, sets up a set of environmental problems considered in science class and asks students to “name and describe one problem in our environment” and then to “explain why it is a problem and how it hurts the environment” (102). All of the supposed “expository” topics involve “explaining” why something is important or a problem.

Let us turn to the persuasive topics. In one prompt, students are asked to pick a person who should win the “best-relative-of-the-year award.” Others involve persuading someone about the best place to live and whether to support a proposal to support Saturday school. Two of these are quite similar, making a case that something is the best. The third, however, is different in kind. It goes beyond making a judgment to deciding on a course of action and supporting that decision.

These persuasive prompts are representative of two of the three types of argument discussed by Aristotle in the Art of Rhetoric: argument concerned with praise or blame or judgment (epideictic) and argument concerned with policy, what to do in a particular situation, which Aristotle calls deliberative argument. (His third type has to do with establishing the facts in a case and is referred to as forensic argument.)

Establishing praise or blame (the best relative or the best place to live—or the worst for that matter) necessarily requires establishing or assuming some definition of the terms of the praise or
blame (e.g., what are the essential characteristics of a good relative and why are those characteristics essential?). On the other hand, deciding what to do in a particular situation has to do with the moral, educational, economic, or other principles that come into play and resolving the conflicts that are likely to exist among them as a result of the situation. Almost inevitably arguments of policy require arguments of praise or blame. For example, if we wish to argue for Saturday school, it may be necessary to argue that the additional time for education is a good thing and that the existing calendar for schooling is bad or inadequate. If we can establish that the additional time is a good thing, then the next argument is whether to put a new schedule into effect and whether the additional time should be added on Saturday or in some other way. This argument will depend on a host of exigencies inherent in the particular situation: the community's ability and willingness to pay, teacher union rules, parental wishes; and so forth.

Of the expository prompts presented, one calls upon students to explain why some type of work is important, another to explain why a historical period is important, and a third to explain why some sort of pollution is a problem. Each calls for an argument of judgment. The use of the term explain suggests that all is self-evident, that no persuasion is necessary. One problem revealed by this comparison is that there is no real difference between expository and persuasive writing as conceived by the Illinois system of assessment. All of the Illinois supposedly expository prompts are really persuasive and require the same kinds of arguments as do the persuasive prompts. The phoney difference lies in the use of the term explain that assumes no argument is necessary because in expository writing we write only about what is self-evident, an epistemological assumption that is not defensible.

Both kinds of prompts demand similar but quite complex thinking. Consider the argument of judgment as an example. Deciding on the best relative or the best place to live or the best or worst of anything demands considering which qualities permit such a decision. Those who rank the "most liveable" cities in the US spend no little effort in determining what are the most important qualities, weighting and justifying them. We can imagine that they begin by asking what are the most important qualities communities need for the good life of their citizens. They would have to ask what constitutes a good life. They would have to engage to some extent in considering the philosophical dimensions of the qualities they use. Further, each quality will involve a scale of some sort, quantitative or qualitative, that will allow for a comparative analysis of cities in terms of that quality. Finally, they will have to show how each city judged to be among the most livable meets some criterial level of excellence on each quality.

One problem revealed by this comparison is that there is no real difference between expository and persuasive writing as conceived by the Illinois system of assessment.

Given that complexity, would it be possible for you to write a carefully considered response in forty minutes? The Illinois prompts present no material for the student to use in thinking through the problem for writing. Writers must draw on their existing knowledge. They might know enough about relatives to develop evidence for claiming a relative is a good one, but how does one decide on who is the best relative? If we were asked to write about whether schools should require uniforms, where would we find the data to use as evidence in the forty minutes allowed? We find the same kinds of writing prompts in Texas and in many other states. Such prompts indicate a serious problem with the assessment. If writers cannot find evidence to support a case to persuade an audience, what will count as good writing in that case? We will turn to the Illinois criteria and sample compositions to see.

Scoring Criteria

Write On, Illinois! explains that all "three domains" of writing are evaluated using a thirty-two point scale made up of four analytical scores plus a holistic score for each piece of writing. The analytical scales provide scores on focus (1–6), support and elaboration (1–6), organization (1–6), and mechanics (1–2). The holistic score (1–6), called integration, is based to some extent on the success of students on the preceding scales. If students score
at least 2 on mechanics and 3 on the others, integration may be doubled. Interestingly, for a score of 32, expository pieces must meet the same criteria. The criteria shape what will be taught in classrooms.

**The Five Paragraph Theme**

Perhaps the most important criterion in shaping what will be taught appears in the *focus* scale. For a score of 6 on focus, both the persuasive and expository rubrics require that the "subject/issue is clear and maintained," that the "position(s)/opinion(s) are explicitly announced in the opening and maintained throughout the paper," and that "major points of support are explicitly previewed in the opening." If the writer does not wish to write such an explicit opening but prefers to be somewhat more subtle, perhaps to keep the reader in suspense, the score will go down to a 4, for which the rubric declares the "position(s)/opinion(s) ... may be arrived at inductively" (16). But, clearly, doing it this way would not result in a top score.

To illustrate how the preview of reasons should appear and be used in a composition, *Write On, Illinois!* provides teaching materials for demonstrating what focus means. This consists of diagrams that use a solid square, circle, and triangle to represent reasons. In one diagram, all three figures appear in the slot for the first paragraph and then appear in the same order on separate lines farther down the page. This diagram is reiterated for expository and persuasive writing (214 and 229). Clearly it calls for at least three reasons (square, circle, and triangle) for a fully developed paper. The insistence on the preview for a top score and the suggestions for teaching that appear in the 1994 scoring guide lead to the nearly universal teaching of the five paragraph theme in Illinois.

In 1999, Illinois changed part of the scoring rubric. It is no longer mandatory for the highest score to preview the major points of support in the first paragraph, thus eliminating the criterion that helps to prompt teaching the five paragraph theme. However, the criteria for focus state that the opening paragraph "may or may not include specific preview" (*Illinois Standards Achievement Test* 152).

**Benchmark Papers**

To determine what the scoring criteria really mean, it is necessary to examine the benchmark papers, the compositions presented in the scoring rubrics to illustrate the criteria. The only paper presented illustrating a perfect score on a persuasive/expository essay at the high school level in the updated *Illinois Standards Achievement Test* is a five paragraph theme with a preview of the major points of support in the opening paragraph (172–74). The criteria for elaboration/support for the top score require "extensive, in-depth support using multiple strategies (e.g., explanation, evidence and example)" (153). The new rubric calls for "all points [to be] logically presented and interrelated" (154). The sample paper illustrating the highest score is in response to a prompt that tells students that the state legislature is considering making a high school exam a requirement for graduation. It states, "Write a persuasive paper stating whether or not you agree that students should be required to pass an exam to graduate from high school. Give reasons why you think as you do" (157). The author of the highest rated paper for this prompt opens with the following paragraphs:

I strongly believe that students should not have to take an exam and pass it to graduate. Students who don’t care what happens in their life after high school aren’t going to start caring now because of the exam they have to pass. Students who strive to be burger flippers after high school aren’t going to care whether or not they pass. Teachers are already testing the seniors of their knowledge in almost every class. What a high school course teaches an individual is important, but is it important enough to affect whether or not they can go to college?

The state would like a high school graduation exam to be added to the many graduation requirements, but are they willing to accept the costs? Unfortunately, every high school will have an exceptionally large amount of students who don’t pass the test, and it is going to cost the state more money to put them through another year of high school. The students who don’t pass the test are going to become majorly discouraged and may only become a burden to our society when we are paying for them while they are on welfare.

The second paragraph begins with a rhetorical question that carries an implicit claim, that the state will not want to accept the costs. This is supported by three sub-claims that large numbers of students will fail the tests, that this failure will require another year of high school, and that those who fail will become so discouraged that they will go on welfare. All of these generalizations are pure speculation with no support. How they support the main claim that the state will be unwilling to accept the costs escapes me.
None has anything to do with the state’s willingness or unwillingness to accept the costs. They are attempts to support a claim that the costs will be large. But the scoring manual states that this paper is exemplary because it “features reasons that are fully and evenly developed through specific detail and multiple strategies” (175). These empty claims are examples of what the state considers “fully developed” reasons. Where are the specific details?

Recall also that one of the criteria for the highest rating calls for “all points [to] be logically presented and interrelated.” The two paragraphs above do not pass the test of logic. If so many students do not care about their lives after high school and aspire only to be “burger flippers,” why should they be greatly discouraged by failing the test? Why would failure of an exam put them on welfare rather than in a burger kitchen? These are simply unsupported, poorly related, claims. The remainder of the piece of writing holds up no better under scrutiny than the first two paragraphs. The third paragraph explains that teachers are testing in classes all the time, while the fourth paragraph develops the third point in the introduction that the graduation exam could keep certain people from attending college. The first point in the introduction, about the low aspirations of students, is never developed.

Someone might contend that students cannot be expected to do better than that, that the circumstances of the testing limit their writing. That is exactly the point. The test limits what students can do. The scoring rubrics cannot be too demanding. The limited time for writing and the inaccessibility of information relevant to the issues in the prompts must allow students to respond in vague generalizations. By the same token, it is also predictable that writing in vague generalities will become the standard to which schools and teachers aspire.

**What Happens in Illinois Classrooms?**

In school after school, from elementary to high school, we found over 70 percent of the Illinois teachers interviewed were hammering away at the five paragraph theme. In this district and in all other Illinois districts in which we interviewed, this formula had become the central staple of the writing program. Under the conditions of the test, Illinois students receive a topic about which the state provides no specific information. They have forty minutes to produce a “developed” response. They must manufacture blather to fill up the space. The five paragraph formula provides a way to organize the blather, but it ignores thoughtful development. It focuses attention on three “reasons” supposedly in support of a main idea, not on whether the three points actually support that idea, or whether the three points themselves require support, or whether
the three points make sense. In Illinois and Texas, teachers intentionally teach their students how to develop specious points to fill in the formula.

Further, when students have been subjected to this instruction for eight to ten years, they come to see the five paragraph theme and the shoddy thinking that goes with it as the solution to any writing problem. Directors of freshman English at three Illinois state universities have complained about the extent of the problem. The English department at Illinois State University publishes a manual advising their incoming freshmen that while the five paragraph essay may have been appropriate in high school, it is not appropriate in college and should be studiously avoided. It shuts down thinking.

This is a crucial time in American democracy. We are faced with problems that demand critical thinking of all citizens. We need to help students examine specious arguments and know them for what they are. Our tests encourage the opposite. They encourage blurry thinking and obfuscation. As a society, we cannot afford to spend valuable classroom time on vacuous thinking and writing. We need to tell citizens and legislators what these problems are and insist that they be addressed.

Works Cited


GEORGE HILLOCKS JR. is a professor in the Department of English Language and Literature at the University of Chicago.

---

**EJ 20 YEARS AGO**

**Segregating the Gifted and Nongifted**

"I'd like to think that we know (which means that I'm not at all sure we do) that the best schooling, the most honest and realistic and democratic schooling, emphasizes both our similarities and our differences, our similarities because we must live with each other, our differences because we must live with ourselves. . . . I worry about some of the consequences of establishing separate classes for the gifted and the less than gifted. I wonder why classes below the gifted emphasize the immediate and the practical, which operationally means grammar and workbooks and newspapers, none particularly immediate or practical. I wonder why those lower classes can't emphasize literature and composition just as much as classes for the gifted. To succeed, teachers in those classes had better emphasize more than the practical."