Writing Assessment Test Design: Response to Two *WPA* Essays

Karen L. Greenberg

As a WPA who is primarily concerned with the teaching and the testing of basic writers, I was pleased to see two excellent essays on writing assessment in a recent issue of *WPA:* "Do You Agree or Disagree" by Judith Fishman and "The Phenomenon of Impact" by Lynn Quitman Troyka (*WPA*, Fall-Winter 1984, 17-26 and 27-36). However, as the current Chair of CUNY's Task Force on Writing and as a Director of the National Testing Network in Writing, I was disturbed by several of Fishman's assumptions about writing assessment and about research methodology, assumptions that overlook the findings of recent research in the field.

In her essay, Fishman does not acknowledge the literature and the relevant research on writing assessment on the design of essay test topics. Her analysis of writing assessment instruments is based on personal experience and hearsay, factors that are important but that cannot substitute for knowledge based on theory and research. For example, her primary concern about CUNY's writing assessment test is that "the test does not enable students to write at their best (even considering the constraints of testing) (18)." This concern reflects an assumption that it is possible, and necessary, to develop a test of minimum writing competencies that enables all or most of the test-takers to perform optimally. This assumption is questionable.

First, it is extremely difficult to design a test that will elicit thoughtful and sincere responses from a large and diverse population. The test must present subject matter with which all students—of both sexes, of all ages and races, and of dozens of different native language backgrounds—can become engaged. In addition, the test must be "valid" (i.e., enable a rater to rank students according to the test's criteria and to discriminate among students' differing levels of writing competence), and it must be "reliable" (i.e., yield the same relative magnitude of scores for the same group of students under differing conditions). Research on the CUNY Writing Skills Assessment Test has demonstrated that it satisfies the criteria described above.[1] Can we—CUNY faculty members—refine the test? Of course we can; and research is underway that attempts to do so. Can we, or any other testing committee, refine or redesign the test so that it "enables students to write at their best"? Not if the test's sole purpose is to determine which students are at least minimally competent college-level writers and which are not.

Researchers in writing assessment agree that a test that might enable students to do their best work as writers would have to have the following features:[2]

1. it would allow students to write under circumstances that approximate the conditions under which their best writing is done;
2. it would require students to do several different types of writing tasks (in order to obtain an adequate sample of their best writing);
3. it would provide ample opportunity for students to revise and to edit each of the tasks.

Writing assessment specialists agree that if one were attempting to measure students' growth as writers over a specific time or to diagnose students' writing problems, or to determine the effectiveness of a writing program or methodology, it would be necessary to incorporate the three features above into the instrument that would be used to assess students' writing skills.[3] However, if one's purpose for testing is to sort students into those who need remediation in writing and those who do not, it is not necessary to address these three features.[4] This is particularly true in the case of the CUNY test, because any errors in the test results are found and corrected by faculty on the first day of class: students write a diagnostic essay and faculty can use this essay to change students' course placement.

At many of the CUNY colleges, the test is used to place students into composition courses and to exit them from developmental writing courses. It is important to note, however, that CUNY's Office of Academic Affairs mandates only that the test be taken for placement and, if failed, that it be taken again before the completion of sixty credits. Troyka underlines this point in her essay and remarks that "no test, essay or multiple-choice, can be statistically sensitive enough to measure growth over the short haul." Thus, she adds, "colleges that re-test students after only ten to fifteen weeks of life in college are bound to be disappointed with the results, and the teachers whose final grades are determined by one test are bound to be frustrated" (31).

There are many large-scale postsecondary writing assessment programs that use a test similar to CUNY's for determining whether students have minimally competent writing skills.[5] For example, all students entering a state college in California or in New Jersey have to take a holistically-scored writing test that consists of a single expository question. Research on New Jersey's test indicates that it is extremely accurate in placing students into remedial and non-remedial courses: faculty agree with more than 90% of the placement decisions.[6] The New Jersey test allows students only twenty minutes (unlike CUNY's test which allows fifty minutes). Again, research confirms that brief amounts of time are adequate for the test's purpose: "for the purpose of placement into remedial courses, a twenty minute essay produces as much information as a forty-five minute essay."[7]

Furthermore, several colleges have adopted the CUNY test or a version of it, and they have done so for reasons that contradict another of Fishman's assumptions—that the CUNY test has become a model simply because of its "efficiency" and its "packaging." According to a survey of faculty at Drew University—one of the colleges that uses the CUNY model—the CUNY test is a "demanding" test: "they [Drew faculty] believe the CUNY test to be a much more richly discriminant instrument [than the previous test], particularly at the upper end of the scale" (Salmore, 3). Faculty at other schools using versions of the CUNY test (including Oakton Community College and Malcolm X College) have also been pleased with its validity and reliability.

Another of Fishman's questionable assumptions is that an essay test question can be sufficiently fine-tuned to enable all of the test-takers to write at their best. Some of the characteristics of a test question that she thinks would allow all students to write better include specification of an audience, specification of a purpose for addressing that audience, and attribution to an actual speaker or writer (23). She states these characteristics as if there is conclusive evidence to support their importance, but there isn't. The research findings are contradictory, and Fishman has not connected her assertions to the relevant theory of research. At first glance, Fishman's assumptions seem to be correct because many professionals have recommended full specification of the rhetorical context in designing writing assignments. However, as James Hoetker wrote in his comprehensive review of the literature on topic design, "so far there are no convincing data to show that extensive fictional contexts have any facilitating effect on students' writing" (386). In fact, Hoetker went on to state that:

> First, such a scenario [a fictional rhetorical context] introduces into the testing situation all of the problems of varying individual interpretations and responses that are associated with the reading of any work of fiction. Second, the sheer amount of language that students must process is increased. Opportunities for confusion, misinterpretation, and creative misreadings are proportionately increased. Third, the more language and information students are given, the more difficult it seems to be for them to get beyond the

language of the topic to discover what they themselves have to say, so that examiners find themselves receiving, not "original response," but their "own prose back in copy speech" (387).

My experimental study of the effects of variations in essay test topics confirmed all three of Hoetker's points, as did Gordon Brossell and Barbara Ash's studies of the rhetorical contexts of essay questions.[8]

On the issue of audience awareness, there is actually some evidence that writing assignments specifying types of audiences do not elicit writing that shows more signs of audience awareness and adaptation than do assignments without any audience specified.[9] This finding has been interpreted as evidence for a difference between a "real" audience and an "ostensible" one. If students know that the only person who will read their writing is the teacher (and in a test situation, the teacher may not even be known to them), many will ignore the ostensible audience specified by the directions and will write for the real evaluators. Moreover, sometimes specifying the audience for a writing assignment or for an essay test may influence students' writing performance in negative ways. There is some research evidence showing that students who are asked to write to a relative or to a friend may write more fragmentary prose or more informal diction and nonstandard grammar than they would write for a teacher, and their writing may be downgraded as a result.[10] This research addresses Fishman's suggestion that students' writing competence should be measured by a question that asks them to write to their "Uncle Harold". Finally, in his review of the research on the effects of audience and role specification, Leo Ruth concluded that "the awareness of specific critical readers, whether they be students or others, may inhibit and complicate rather than simplify the problems of dealing with audience" (84) and that "an exceedingly complex task becomes even more convoluted as the writer realizes that in the real-world context of the assessment situation, his only real reader(s) will be the evaluator(s)" (85). Students who attempt to address both the real and the ostensible audiences at once—whether they are compatible or not—may get very confused.

Fishman also assumes that an increase in the rhetorical specificity of the CUNY test question will elicit better writing, but she believes that the "cue words" in the directions should be less specific. She and some of her colleagues feel that asking students to "agree or disagree" with a statement invites an "essay of opinion" on an "issue that they know little about." Instead, she proposes that students should answer the question, "What do you think?" and that they should "discuss your response [to the statement]." However, I would guess that "What do you think?" would cue students to begin their essays with "I think that..." and would lead to essays that do not ever go beyond personal opinion. Furthermore, as the literature on essay test topic design reveals, decreasing the specificity of the cue words in the question creates unintended traps for students. In her analysis of the "describe trap" and the "discuss trap" of vague directions on essay tests of writing, Catharine Keech explained why these cue words are cases of "underprompting" in which the test-makers do not clarify their expectations about how the students should respond to the prompt (166). This underprompting discriminates against students who cannot psyche out the test-makers' implicit expectations. Moreover, as Keech has shown, "the apparent freedom of this instruction may become a trap if the readers who score these papers have a 'hidden agenda' that rewards conventional school-essay language" (173). In the Fishman example, if students actually wrote "what they think" and if their responses were not analytic or persuasive essays, readers might fail them.

Fishman is also concerned that the "agree-disagree" format "promotes a model of mind that does not reflect what we should be reaching for in institutions of higher learning" (81). I am not sure what she means by this, but I do know that surveys conducted by the National Testing Network in Writing indicate that this test format is preferred by most American postsecondary institutions.[11] It is also used for one of the essays on the College Board's English Composition Test because, according to ETS, and agree-disagree format "measures the student's ability to do the kind of writing required in most college courses, writing in which the student explains a point of view, defends it, or persuades another to accept it.... This kind of expository writing emphasizes precision in diction, logic in the presentation of ideas, and clarity of expression" (3).

Even if all of the experts in topic design could agree on the best format, cue words, and content of a writing test topic, this "ideal" topic might not enable students to write any better than they would on other topics because of a problem discussed in the literature as the "mismatch problem." Any writing assignment offers multiple opportunities for mismatches in the way that test-makers, test-takers, and test-raters interpret it.[12] I discovered this problem in my experimental research on topic design: students' interpretations of the directions and the content of my experimental topics often differed dramatically from those of the teacher-raters.[13] The researchers at the Berkeley Writing Assessment Project have been systematically exploring this mismatch problem, and their findings are of great importance to all of us who are endeavoring to assess writing fairly and accurately. One of this team's recent findings is that students frequently focus on phrases in a writing topic or question that teachers do not feel are important for shaping an adequate response. For example, the students and teachers in one of their studies were interviewed about a writing topic that consisted of one declarative sentence stating a generalization about the topic and one interrogatory sentence asking a question about the topic. Fifty-three percent of the students said that it was important to elaborate on the idea stated in the declarative sentence, but not one single teacher felt that the declarative sentence had to be mentioned in the response.[14]

Experimental research conducted by Gordon Brossell and Barbara Ash confirms the idea that variations in the structure of topics may not have any effect on students' writing performance. After doing extensive field-testing of topics for the College-Level Academic Skills Test (a test that every college student in Florida must take), they found no significant differences in test scores on the different topics: The essays revealed little to suggest that writers had been helped or hindered by particular versions of topics. Subject matter similarly seemed of slight consequence. We came away feeling that as long as topics do not require special knowledge and are suited to the characteristics of the test-takers, neither small syntactical variations nor subject matter has much of an effect on essay examination scores (424).

In fact, the research conducted by Leo Ruth and Sandra Murphy, members of the Berkeley team, suggests that only clarity of statement is critical: test designers must "be sure that the language of the prompt forestalls needless difficulties that arise from ambiguous wording and confusing signals," and that we should be "wary of introducing writing tasks that simulate real life" because these tasks, with their full rhetorical context, may confuse students or may elicit writing samples that will not satisfy the purposes of assessment (418-419). In order to avoid some of these problems, these researchers suggest that we do extensive pilot-and field-testing of any topics that we create and that we interview students to elicit their interpretations of the topics. And Ruth and Murphy call for further research on topic design.

True, Fishman calls for more research too, but I am concerned about the types of investigations that she characterizes as "research." She acknowledges that the work that she and some of her colleagues did was "informal research" consisting of "informal interviews." However, anecdotes and interviews conducted by interviewers with a clear bias do not constitute "research" even at the most informal level. Descriptive research is as productive as controlled experimental research, as long as the research design is adequate in scope, depth, and precision. A biased, informal investigation is not.

There are dozens of textbooks on the methods and the tools of conducting descriptive or experimental research.[15] Most of these texts recommend that research must include, at the very least, the following characteristics:

1. a clear identification of the exact nature and dimensions of the problem;

2. a clear statement of the specific research objectives related to the problem;

3. a discussion of the relevant literature and of existing theory and research evidence;

4. an explanation of the validity and the reliability of the research procedures or instruments;

5. an analysis of competing interpretations for the findings.

Although Fishman calls the results of her interviews "preliminary" (24), I think that she and her colleagues made several mistakes in interviewing that greatly undermine their assertions. They did not adequately develop a detailed, uniform interview guide; and they failed to establish safeguards against interviewer bias. Neither did they conduct sufficient practice interviews to make sure that interviewers had acquired the needed skills nor did they make any provisions for calculating the reliability of their interview data. Fishman's observational research techniques were also flawed: she did not do any random sampling and she did not use any check on the reliability of her observers. Since Fishman does not describe her "survey" in any detail, one cannot judge accurately the design and administration of her questionnaire. However, it is clear that she did not attempt to obtain a random sample, but instead, selected her survey sample on the basis of convenience. WPAs who are not experienced in conducting research on writing or on writing assessment need to know that problems like the ones described above undermine the validity and the usefulness of any research findings.[16]

Fishman is absolutely right that any teaching or testing program "must be continually reassessed, reevaluated, studied, and probed, questioned, and requestioned" (24). In her essay, Troyka points out that the office that oversees CUNY's entire testing program has offered, and continues to offer, assistance to all faculty interested in conducting research on writing or on writing assessment. Currently, a research subcommittee of the CUNY Task Force on Writing is planning an ethnographic study of the topics on the CUNY test.[17] This study will attempt to discover the various ways in which students interpret the current topic type (and various other types of topics) and how these interpretations affect students' writing processes and products in the context of the classroom and the testing situations. We are trying to design research that is valid and reliable and that can affect both testing and teaching. We believe that the ethnographic model is most appropriate because it takes into account the importance of the writing context and because it is a multimodal enterprise that incorporates surveys, interviews, participant observations, case studies, and protocol analyses.

Both Fishman and Troyka make reference to my essay on "Competency Testing: What Role Should Teachers of Composition Play?" In that essay, I wrote that:

> ...if we are dissatisfied with the content or the planned uses of tests to be given in our schools, we must be fully prepared to document our discontent with evidence that will be convincing to writing program administrators and testing directors. Impassioned speeches

about the corrosive effects of writing tests on students' creativity do not constitute convincing evidence; data on the number of incorrect course placements resulting from the test of data showing very low test reliability do (374).

Fishman seems to have missed my point.

Notes

[1]For research reports on the validity and the reliability of the CUNY Writing Skills Assessment Test, see Mara Zibrin, *The 1979 Audit of the Writing Assessment Test* (New York: CUNY Office of Academic Affairs, 1980); Susan Ryzewic, *The CUNY Writing Assessment Test: A Three-Year Audit Review* (New York: CUNY Office of Academic Affairs, 1982); and Karen Greenberg, *CUNY Writing Faculty: Practices and Perceptions* (New York: CUNY Office of Academic Affairs, 1983). The complete list of research monographs on the testing program can be obtained from the CUNY Instructional Resource Center, 535 East 80th Street, New York, NY 10021.

[2]For a discussion of the characteristics of effective tests of writing, see Charles Cooper and Lee Odell, Eds., *Evaluating Writing: Describing, Measuring, Judging* (Urbana, IL: National Council of Teachers of English, 1977); Charles Cooper, Ed., *The Nature and Measurement of Competency in English* (Urbana, IL: National Council of Teachers of English, 1981); and Davida Charney, "The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview" *Research in the Teaching of English*, 18 (February 1984): 65-81.

[3]For a discussion of the relationships between a writing test's characteristics and its purpose, see the works listed in Note 2 above.

[4]Charles Cooper and Lee Odell, "Introduction," *Evaluating Writing: Describing, Measuring, Judging* (Urbana, IL: National Council of Teachers of English, 1977), 20.

[5]For the results of a survey on the writing assessment practices of NTNW member institutions, write to me at National Testing Network in Writing, CUNY, 535 East 80th Street, New York, NY 10021.

[6]For research on the New Jersey College Basic Skills Testing Program, write to William Lutz, Chair, English Department, Rutgers University, Camden, NJ 08102.

[7]William Lutz, *Statewide Testing in New Jersey* (Camden, NJ: Rutgers University, 1979), 6.

[8]I found that specification of a fuller rhetorical context confused students. See Karen Greenberg, *The Effects of Variations in Essay Questions on the Writing Performance of CUNY Freshmen* (New York: CUNY Office of Academic Affairs, 1981). Brossell and Ash found that changes in the writer's role and purpose had little effect on students' writing or on their test scores. See Gordon Brossell and Barbara Ash, "An Experiment with the Wording of Essay Topics," *College Composition and Communication* 35 (December 1984): 423-426.

[9]For a discussion of the effects of audience specification in a topic, see Patrick Woodworth and Catharine Keech, *The Write Occasion* (Berkeley: Bay Area Relationships: Analysis or Invention," *College Composition and Communication* 31 (May 1980): 221-226; Douglas Park, "The Meanings of Audience," *College English* 44 (March 1982): 247-257; and Leo Ruth, "Sources of Knowledge for Designing Writing Test Prompts," *Properties of Writing Tasks*, Ed. Leo Ruth, ERIC ED 230 576.

[10]For research on audience specification, see Note 9.

[11]To obtain this survey, see Note 5.

[12]For research on "mismatches" in the interpretation of essay topics, see Karen Carroll and Sandra Murphy, "A Study of the Construction of the Meanings of a Writing Prompt by Its Authors, The Student Writers, and the Raters," *Properties of Writing Tasks*, Ed. Leo Tuth, ERIC ED 230 576 and see Leo Ruth and Sandra Murphy, "Designing Topics for Writing Assessment: Problems of Meaning," College Composition and Communication 35 (December 1985): 410-422.

[13]For information on my study, see Note 8.

[14]For information on this study, see Note 12.

[15]Three of the most accessible texts on research methodology and tools are Earl Babbie, *Survey Research Methods* (Belmont, CA: Wadsworth, Inc, 1973), John Best, *Research in Education* (Englewood Cliffs, NJ: Prentice-Hall, Inc., 1970), and Deobold Van Dalen, *Understanding Educational Research* (New York: McGraw Hill, Inc., 1973).

[16]For further information on the various techniques for conducting valid and reliable research, consult the last twenty-five issues of *Research in the Teaching of English*. This journal includes excellent examples of all of the many techniques for studying writing: controlled experiments, quasi-experiments, correlational analysis, longitudinal analysis, case studies, protocol analysis, and ethnographic studies.

[17]For an overview of ethnographic research in English, see Kenneth Kantor, Dan Kirby, and Judith Goetz, "Research in Context: Ethnographic Studies in English Education," *Research in the Teaching of English* 15 (December 1981): 293-310.

Works Cited

Brossell, Gordon. "Current Research and Unanswered Questions." *Writing Assessment: Issues and Strategies*. Eds. Karen Greenberg, Harvey Wiener, and Richard Donovan New York: Longman, Inc., forthcoming.

Brossell, Gordon, and Barbara Ash. "An Experiment with the Wording of Essay Topics." *College Composition and Communication* 35 (December 1984): 423-426.

Educational Testing Service. *Report on The English Composition Test with Essay*. Princeton, NJ: Educational Testing Service, 1979.

Greenberg, Karen L. "Competency Testing: What Role Should Teachers of Composition Play?" *College Composition and Communication* 33 (December 1982): 366-376.

Hoetker, James. "Essay Examination Topics and Students' Writing." *College Composition and Communication* 35 (December 1984): 377-392.

Keech, Catharine. "Practices in Designing Writing Test Prompts: Analysis and Recommendations." *Properties of Writing Tasks.* Ed. Leo Ruth. ERIC ED 230 576.

Ruth, Leo. "Sources of Knowledge for Designing Writing Test Prompts." *Properties of Writing Tasks.* Ed. Leo Ruth. ERIC ED 230 576.

Ruth, Leo, and Sandra Murphy. "Designing Topics for Writing Assessment:- Problems of Meaning." *College Composition and Communication* 35 (December 1985): 410-422.

Salmore, Barbara. *A Comparison of the TSWE and CUNY Writing Tests.* Drew University, May 1979. This Report is available from NTNW, 535 East 80th Street, New York NY, 1002.