Testing Proficiency in Writing at San Francisco State University

Sara Warshauer Freedman and William S. Robinson

At San Francisco State University, we have been involved in testing our students' proficiency in written composition since 1960, long before the current proficiency movement gained momentum. Through our years of experience, we have learned how to design a reasonably reliable, reasonably economical testing program capable of handling large populations of students semester in and semester out. Here we will present an overview of our testing program and discuss in some detail three important features: the selection of essay topics, special issues in scoring, and the development of a counseling program for students who fail.

The examination at San Francisco State is called the Junior English Proficiency Essay Test (JEPET). It consists of a single assigned essay topic, on which every student writes an expository response in one hour. The test is required early in the junior year following a semester of freshman and one of sophomore composition, and students who fail it must take a junior expository writing course. The test is meant to be seen as an upper-division minimal proficiency test of a student's ability to write exposition. (Readers should remember that the term "minimal proficiency" is relative and means whatever local standards cause it to mean.) The essays are graded on a six-point holistic scale following, in general, the procedure developed by Educational Testing Service. Students who fail are invited to see a counselor, whose principal job is to see that no one fails who should have passed.

Before beginning our discussion, we want to caution that in an absolute sense, the notion of testing writing proficiency is nonsense, for proficiency, as noted, is a relative term, and writing comes in as many guises as there are human interests and occupations. A poet may be inept in the world of busi-

Sarah Warshauer Freedman is now Assistant Professor of Writing Research and Instruction in the School of Education at the University of California, Berkeley. She works with teacher-researchers on processes of classroom inquiry. She has published essays in (among other journals) *CCC*, *English Journal*, *Research in the Teaching of English*, and *Educational Psychology*. William Robinson is Professor English and Coordinator of Composition at San Francisco State University. He has served on the English Equivalency Examination and English Placement Test committees in the California State University system, and has acted as a consultant on test development to other campuses in the system.

ness reports, and a published psychologist a disaster at the personal essay. Because writing comes in so many guises, it is important to specify narrowly the kind of writing one will examine and the standards one will employ, and then to place statements about proficiency in a specific context.

Presumably, if one wants to see whether students can write in a particular form, one has them write in that form, but it is a truism of testing that single-item essay tests are "unreliable," and thus that it is dangerous to generalize about a student's overall writing ability (supposing such a thing exists) from his or her performance on only one test essay. Carefully devised multiple-item, multiple-choice tests, on the other hand, are theoretically more "reliable" than essay tests, but whether they measure the ability to perform a particular writing task has not been satisfactorily determined. Obviously, such tests do not directly measure writing proficiency of any kind.[1] Even if it were possible to develop tests of writing proficiency in a multiple-choice format, the development and validation costs of such tests—costs which are ongoing, since one cannot use the same test over and over—are far beyond the resources of most schools. If one is going to attempt to devise one's own test of student writing, the essay test seems to be the most practical for now. With careful topic development, some changes in the usual holistic scoring procedure, and the development of a counseling system, we found it easier to compensate for the potential unreliability of the essay than the questionable validity and high test-development costs of responsibly constructed multiple-choice tests.

Topic Design

We chose to develop only expository, or, in Britton's term, "transactional" topics, since most college work puts little premium on one's ability to write descriptions or narrations. Most expository topics for timed essay tests call for the writing of either exposition based on personal experience or for the analysis of a written passage. To test the latter, one must provide something for the student to analyze, and here it can be difficult to come up with passages both suitable for the purpose and equally accessible to students from all majors and fields of study. Further, although we see reading and writing as related skills, we wanted a relatively easy test of writing, one that would not make students' success depend on reading comprehension abilities. It seemed to us that for our purposes we could test a sufficiently wide range of college writing skills by testing the students' ability to handle exposition based on personal experience. Others might want to select a more complex task involving reading.

In determining how much time to allot to the writing task, we decided that we wanted as long a writing sample as possible given the restraints upon length imposed by holistic reading and by the numbers of test booklets our

readers could evaluate during each administration. A one-hour test seemed about right given these constraints.

We decided against giving students a choice of topics. The more topics one offers, the more costly the testing process becomes: the greater the number of topics one must develop, the more difficult and time-consuming the topic-development task becomes, the greater the number of topic-responses readers must evaluate, and the more difficult and time-consuming their evaluation task is. Indeed, to help assure reliability and facilitate the training of raters, holistic rating is normally performed on only one topic at a time, and so each topic in a multiple-topic format would require a separate rating session.

When students are bound by a single topic, that topic must be constructed carefully, and provisions must be made for students who seem to perform poorly because of having little to say about the topic. To construct a single topic for a test, we gather a committee of composition instructors, each of whom writes three topics. The group reads them, chooses several that it believes have the best chance of working, and makes any revisions in wording, format, and the like that seem advisable. The topics are then used as part of the final examination in junior-level writing courses. The instructors pretesting the topics rate their students' responses and comment on the correlation of these essays with their students' previous work. Finally, the composition and JEPET coordinators read the student responses and the instructor reactions and make the final decisions about which of the trial topics to use.

In developing the topics, we adhere to five criteria, which we will outline with reference to the following two topics used successfully at two different test administrations in the past.

(1) Everyone has a gripe about the community in which he or she lives. Whether that problem be major or minor, a matter of rising neighborhood burglaries or of inadequate parking facilities on campus, most of us feel that some community need is being ignored by local officials. What's your gripe? How does it affect your everday life, and how would you suggest correcting it?

(2) Although we all have definite character strengths, we also have weaknesses, be they silly, such as an uncontrollable craving for ice cream or chocolate candy, or more significant, such as impatience or jealousy. Describe one of your weaknesses, show how it affects your life, being as specific as you can, and discuss how you would discourage it.

First, the topic must be accessible to all—that is, it must address concerns one can reasonably expect to be common to all those taking the test. The topics quoted above address such concerns, and fewer than one percent of the tested students found nothing to say on either. To further ensure accessibility, we provide for choice within the large framework of the topic, asking the students to write about any kind of gripe, major or minor, or any kind of character weakness, silly or significant. This expands the range of responses for students and offers a welcome fringe benefit to the examination readers,

inviting a greater number of amusing and imaginative responses than one receives on topics of greater weight and seriousness.

The second criterion for topics is that they must result in an examination that is not "speeded" for most students—that is, one in which they do not find themselves working under great time pressure. When time is a serious factor for most of a test population, one learns from the results not who is capable of performing the task but who is *most* capable of performing it. Since we want to know only who the capable are, we must minimize the potential ill effects of the timed test situation. Thus we pretest our one-hour topics in forty-five minute sessions; if virtually all the pretest population can write a complete essay in three quarters of an hour, the population being tested should be able to respond in a full hour. To aid the student in writing within these time limits, we build a loose organization into the topic itself (e.g., What's your gripe? How does it affect your everyday life? How would you suggest changing it?) And for those mental pumps that need priming we offer examples of possible subjects (neighborhood burglaries, inadequate parking).

Third, we try to construct topics that will not readily call forth platitudinous or "canned" responses. For instance, a response centering on deeply held religious beliefs may be difficult to evaluate since little of the language may be the student's own. In this connection, notice that the slant of both the sample topics is negative; they ask the student to write about a problem or a weakness. In our experience, this negative slant, though not essential, seems to help students write better papers than most positively-slanted topics, perhaps because it serves better to discourage platitudes and clichés.

At least as important as eliciting the students' best responses, the topic should elicit the kind of response we wish to evaluate—in our case, an expository response. Thus the questions at the end of the topic are supposed to guide the student to write exposition rather than narration or polemic. Although the middle direction in each of the samples (How does the community problem affect your life? Show how your character weakness affects your life.) would seem to permit, if not elicit, narrative, the last question (How would you suggest correcting the community problem? How would you discourage your weakness?) brings students back to an expository evaluation of their experiences. When we pretest topics, we make sure they do not promote too much narrative.

Finally, and most important, topics must discriminate accurately between good and poor writers. Our only measure of how well the test discriminates comes from our pretesting. As we have already mentioned, we pretest all possible topics to see whether they do what they should do—produce responses from all students, produce expository responses, and produce responses commensurate with the abilities of the writers. We hope that the topics will work as well for the population being tested as they did for the population who took the pretest.

Scoring Adjustments

We chose holistic scoring over a detailed analytic system or a primary trait system because we are interested not in a possible "diagnostic" function for the test—we are not interested in describing the weaknesses in the writing of individual students—but only in making global decisions about overall proficiency, and because holistic scoring is the most economical and speediest method and involves the least amount of special training for readers.[2]

But we modified holistic scoring in three ways. First, we cannot assume a normally-distributed population. In readings administered by Educational Testing Service, a normally-distributed population is expected and so criteria for upper, middle, and low scores can be developed largely from an inductive examination of the responses themselves. Our different test administrations frequently attract different types of students; for instance, our March test draws heavily from students who were admitted to San Francisco State as freshmen and who were not required to do any remedial work, whereas the August test has a very heavy population of new community-college transfer students. Although we must take care not to be biased against any group, our scoring must take into account the diverse abilities of different test populations. If a greater proportion of test-takers deserved to pass than fail or vice versa, we needed to have the scoring allow such heavily-skewed outcomes. Thus, we select our training essays according to criteria largely external to the test and, when scoring, seek first to distinguish between failing and passing papers and then to assign a score within the pass and fail categories. One must always remain sensitive, however, to possible peculiar effects of particular topics on student writing and on readers and must not penalize students for such effects; both the training essays and those read in the search for samples must be scrutinized carefully for such effects.

Second, as is usual, all essays are read independently by two raters. But we break the usual tradition by having essays that, on the six-point scale, receive a four (passing score) from one reader and a three (failing score) from another read by a third reader of proven expertise who reconciles the scores. Papers that receive two-point split scores across the pass-fail line (5/3's, 4/2's and the like) are also read a third time. All essays to be given third readings are read at the same time and their initial scores are concealed so that the third reader does not know whether he or she is deciding a 4/3 case or a two-point discrepancy. Parenthetically, whereas a group of novice readers might produce a two-point discrepancy rate as high as twenty percent, an experienced homogeneous group can produce a two-point discrepancy rate as low as one percent.

Third, during the reading itself, readers are instructed to look for papers by students who were clearly having trouble with the topic or the testing situation, papers on which the students wrote either nothing or so little as to indicate that they were blocked, and papers containing diatribes against the

test, the testing situation, or "the system." To such students we send a special letter inviting them to talk over their problems with the examination coordinator and to take a retest at our expense—if need be, under special circumstances to be mutually agreed upon.

Counseling System

Finally we offer counseling for students who fail. Even though we pretest topics to make them as fair to the students as possible and even though we use the fairest scoring system we can, we realize that we will still inadvertently fail some students who do not deserve to fail—either because they have not done their representative work or because our readers have erred. We hope to catch these students through counseling. All who fail the test receive a notice informing them that they may speak with a counselor about their test. Counselors may recommend that a failing grade be reversed, they may authorize a retest, or they may uphold the failing grade. We do not require that students who fail see a counselor, and in fact, we word our notice so as not to encourage everyone to do so. The result of this procedure has been that about ten percent of those failing have made counseling appointments. Generally, these are students who truly want to learn why they failed, or who feel that their essays may have been unfairly failed, or who are congenitally disgruntled.

Although we are the first to admit the inherent problems of assuring reliable scores in a single-item essay test, we believe that we have devised some ways to minimize those problems. From our experience with proficiency testing at San Francisco State, we know that it is possible to test large numbers of students and to carry out the procedure with a reasonable cost.

Notes

1. Several studies by CEEB and ETS have examined the correlation between multiple-choice tests and essay tests; however, none claim that the multiple-choice test directly measures writing proficiency. See, for example, F. Godshalk, F. Swineford, and W. Coffman, *The Measurement of Writing Ability* (New York: College Entrance Examination Board, 1966) and more recently H. Breland, *A Study of College English Placement and the Test of Standard Written English* (Princeton: Educational Testing Service, 1977).

2. Although primary trait scoring has been used recently in large-scale evaluations, Lloyd-Jones notes, "For purposes such as classroom placement or equivalency credit, where there are other procedures to assess individual exceptions, the ETS method (holistic scoring) and its variants (analytic scoring) are probably adequate and relatively simple" (Richard Lloyd-Jones, "Primary Trait Scoring," in C. Cooper and L. Odell, ed., *Evaluating Writing: Describing, Measuring, Judging*. [Urbana, IL: National Council of Teachers of English, 1977, p. 37]).