Jon Jonz

USING POOLED JUDGMENTS TO DEVELOP TESTS OF BASIC WRITING

In this paper I describe the technique that we at East Texas State University use to create, administer, and monitor valid and reliable measures of the writing skills of students enrolled in basic writing courses.[1] Our technique meets two major needs: it provides one means by which the instructors of basic writing students[2]may benefit from the judgments of colleagues in assessing the skill levels of their students, and it allows students to have an audience beyond the class in which they are enrolled.

BACKGROUND

Each semester the instructors of our basic writing course prepare alternate versions of a reading/writing test for students seeking to exit the course. The purpose for the test is to provide instructors and students with information, not to certify proficiency in reading and writing nor to evaluate achievement in the course. Instructors add this information to their record of student accomplishment in the course and consider it as but one component of final-grade and course-exit decisions. Were the test to be used to certify proficiency, in fairness to students and in deference to what is well known about the variability in an individual's language production from one occasion and one context to the next, the test itself would need to be longer, to be given under unspeeded conditions, to offer a variety of topics, and to elicit a variety of writing samples, at least.[3]

Jon Jonz, associate professor of Literature and Language, East Texas State University, Commerce, TX, was for five years director of the Communication Skills Center and coordinator of basic writing courses. His work in language testing and the sociology of language has appeared in Anthropological Linguistics, TESOL Quarterly, Language Learning, *and elsewhere.*

© *Journal of Basic Writing*, Vol. 6, No. 2, 1987

DOI: 10.37514/JBW-J.1987.6.2.03

Our test, though clearly a proficiency test, is not used as the single certificate of proficiency. None should be. Our test might most clearly be viewed as simply one element of a larger proficiency measure: the course itself. The portfolio of writing that each student generates during the course is the raw data of this larger proficiency measure, and the judgment that a student's instructor makes is the certification of proficiency upon which we rely.[4]

The fact of the test figures into the conduct of the course and is an element that is viewed positively by student and instructor alike. Students know that the test results are not binding on instructors and that grades do not depend solely on those results. Students also know that their instructor will not be the only audience for their work. They know that they will have at least this one opportunity to be evaluated by outside raters, basic writing instructors who do not know whose papers they are reading. Likewise, instructors know that their students' work will be subject to the scrutiny of departmental colleagues at the conclusion of the term.

The test, sample of which I have provided in Appendix A to this essay, requires students to read a stimulus passage and to prepare a written response. This format is appealing for a number of reasons. First, integrative reading/writing skills are precisely what the course is designed to teach; separating reading from writing for teaching and testing purposes reflects a view of language skills to which we do not subscribe. Second, the test consistently prompts the kind of writing that the course emphasizes: expository prose written for a general academic audience. Third, studies that we have conducted demonstrate that the format is reliable and valid. Finally, and perhaps most importantly, the test-construction process itself capitalizes on the judgmental processes of experienced teachers of basic writing. Each of these points deserves some elaboration.

First, our test format reflects quite completely the published objectives of the course. Therefore, it has validity on its face that discrete-point, objective tests would most certainly lack. One might use a standardized reading test or a multiple-choice grammar and usage test to provide course-exit data. One might even demonstrate that the results of these discretely focused tests match other measures of student skill: essay grades or course grades, for example. But this kind of formal demonstration of criterion-related validity often lacks the intuitively appealing and convincing qualities associated with face validity.

Second, our test appeals to us because it reflects the pedagogical philosophy that underlies our course: we do not assume language proficiency to be equivalent to the sum of discretely identifiable subproficiencies. The committee who designed the course did not create a series of discretely isolable and measurable objectives. In fact, the committee rejected the traditional word-sentence-paragraph-essay approach to teaching reading and writing, devising instead a process-driven course during all stages of which students read and write in discourse units that possess a good deal of contextual integrity. The test we devised as an exit measure in such a course could hardly lack contextual integrity itself, so nothing less than an integrative reading/writing test would do.[5]

Another reason for our satisfaction with our testing procedure is that formal studies we have conducted demonstrate that its format is stable: it tends to produce reliable results each time it is employed. These studies also support the claim that our tests measure exactly what we want them to measure.

Yet another reason underlying our choice of testing procedure is that we have not found better measuring devices to employ in questions of language proficiency than the considered judgments of skilled, experienced language teachers.[6] It is that judgment that is central to holistic scoring, and it is that judgment, in fact, that we rely upon as a final measure of proficiency in the form of a course grade.

A final part of our rationale is that the test format appears to stimulate the sort of writing that we expect it to stimulate. By far the most frequent mode or type of written response that the test evokes is exposition. The test itself, however, specifies no particular mode, and the suggestions that each version of the test makes regarding possible responses are not designed specifically to evoke expository prose. The test format does not specify a rhetorical situation, yet students rarely fail to write the expository prose of uncertain and unsophisticated learners. Students rarely appear to be writing for anyone other than a panel of English teachers as their audience with the purpose of demonstrating reading and writing proficiency.[7] The reason for these facts is clear. Students know that the course is designed to prepare them for the standard sequence of writing courses in the department. They also know that their instructors have encouraged them to produce writing that would stand up to scrutiny in those courses: exposition, argumentation, persuasion. Students also know that the exit examination is a feature of the course, and they will have encountered exit examinations from previous semesters as classroom activities in most sections of the course.

CONSTRUCTION OF THE TEST

The test-construction process begins about midway through each academic term when a call for topics and passages is circulated among course instructors. Each instructor of the course is asked to submit three passages suitable for use in test prompts. In our call, we make no suggestions for topics, give no counsel regarding appropriateness, and specify no parameters within which instructors are to limit their choices. Instructors employ only their own inherent judgmental processes. Previous versions of the test are on file, however, and are readily available to instructors to use as a guide to acceptability.

Once passages have been collected, they are photoduplicated and circulated among the instructors with the request that they assign a rating to each passage on a scale of 1 to 4. A rating of "4" means that the passage in question is an excellent selection, that it will work well as a test passage, and that it will require little, if any, editing to make it appropriate. A rating of

"3" means that the selection has potential as a test passage, but that it would require a good deal of editing to be truly suitable. A rating of "2" means that the selection should be ignored, that it has none but the slightest hint of potential, and that it would be more trouble than it merits to turn it into a suitable passage. A rating of "1" means that a selection has no potential whatsoever.

We feel that this stage of the selection process approximates pretesting of the passages. Under other circumstances, established test-construction procedure would require a field trial of individual passages. If we used the test results for purposes other than those I discuss here, we would consider pretesting each prompt. However, given our circumstances and our track record (see below), it is uneconomical and unnecessary to do anything more elaborate than to pool the judgment of the course instructors.

After the passages have been rated and returned, the results are tallied, and we generally find that three to five passages have attracted top ratings. At this point any of these top-rated selections could be turned into a fine prompt. However, at least two or three instructors will have expressed reservations or will have made editorial suggestions concerning even the most highly rated selections. Therefore, we further capitalize on instructors' judgment by again circulating the top-rated selections. This time each instructor is asked to rank order the passages and to make editorial suggestions. Instructors are also asked to offer appropriate headnotes and suggestions to students to be included in the test. Based on these final rankings and editorial suggestions, we select the top two passages and draft the tests. Every suggestion is incorporated into the drafts, and in the rare event that suggestions from two instructors conflict with one another, we confer and compromise.

After the two tests are drafted, they are circulated among the instructors for final suggestions, revisions, and additions. This final step is advisable because passages interact with their lead-in material, with the suggestions to students, and with the general directions. These interactions usually prompt a final barrage of suggestions from the instructors, and once again (and for the final time) we consider, collate, and confer until all reservations are removed. The tests are then ready to be printed.

We have three reasons for creating more than one test for each examination period. First, the exams are not administered to all students at the same time, so we have something of a security problem. Second, with more than one topic the instructors who read and rate the writing do not have to read dozens of samples on the same topic. Although some test-development experts have suggested that the need for uniformity requires that all students write on the same topic, our experience has been that raters enjoy a little variety. As I will presently show, we have not been able to detect inexplicable differences in the ratings assigned samples prompted by one topic rather than another.

And that fact, in a way, is the product of our third reason for having at least two test passages: the results produced by one passage can be studied in comparison to the results produced by the other. One test serves as a criterion against which to judge the other. The test passages are randomly assigned to classes, and judges are randomly assigned to rate sets of writing samples; thus, we could reasonably attribute any great differences in results produced by the test passages to variation in the passages themselves.

MONITORING TEST RESULTS

With minor and interpretable exceptions, the tests that we have developed have worked quite well. In our scrutiny of test results we check four characteristics of test scores and course grades: (1) the percentage of students failing the tests, (2) the percentage of students failing the course, (3) the degree to which readers of the students' writing samples concur in their ratings, and (4) the degree to which students' ratings on the test match their course grades. Tables 1 through 3 in Appendix B summarize such data for a recent two and one-half year period.

In three of the five semesters for which data are presented (that is, semesters I, II, IV), the data provide strong evidence for the stability and trustworthiness of the testing procedure. For example, in each of these three semesters each test produced a similar failure rate (Table 1). We interpret this to mean that students found the multiple forms of the test to be equally challenging. An alternate interpretation, of course, is that raters simply display a thoroughgoing bias to judge half of the writing samples as passing and half as failing. However, we have been able to discount this interpretation by carefully studying raters in training sessions where we have manipulated the pass/fail ratio of the writing samples. Raters tend to approximate closely in their own ratings the pass/fail ratio that has been purposely built into the samples.

Likewise, the high agreement between raters (Table 2) is strong evidence for the stability of the testing procedure in semesters I, II, and IV. Not only do raters tend to assign approximately equivalent numbers of failing marks to a given set of writing samples, they also tend to assign the same ratings to the same samples. Similarly, course grades (Table 3) match test results for three out of every four students for semesters I, II, and IV.

The data for two semesters of the study (semesters III and V), however, present a different array of results. In semester III, for example, writing on the second topic was apparently harder for students than writing on the first topic (Table 1). Instructors, however, adjusted for this problem quite nicely, not even knowing at the time that a problem existed: they awarded course grades in concert with test ratings in only 65% of the cases (Table 3), correcting for the artificially depressed exam ratings. If the test were to have been the only factor determining course grades, such a difference between the results produced by the two test forms could possibly have warranted testing again using different forms. Our testing and grading procedures, however, compensated nicely for the problem, and no harm resulted. In fact, two benefits were realized. First, the failure rate on the second form (74%) represents a serendipitously derived confirmation that no "50% pass/50% fail" rating bias existed. Second, there is evidence that the strong relationship between test ratings and course grades (Table 3) is a substantial one, not one resulting from "hyperrespect" for test results. If such were not the case, the relationship between test rating and course grade would have stayed at previous levels and would not have dipped to compensate for the unexpected results produced by the second test topic.

The results from the testing and grading in semester V also represent a deviation from previous patterns. At first glance, it appears that the alternate test topic once again proved to be somewhat more difficult than the first (Table 1). The rest of the data, however, do not tend to support such a conclusion. The elevation in failure rate for the exam (Table 3) to some ten percentage points above the previous high rate is attended by a similarly sharp increase in the failure rate for the course over the previous spring's (semester III) rate. Additionally, raters registered the largest percentage of agreement (Table 2) of all semesters considered in the study, a result that tends to confirm that the quality of student writing was below that of previous terms. In the case of semester V, then, students had produced writing of significantly lesser quality on the test than ever before, the raters agreed that they had, and instructors awarded course grades accordingly.

CONCLUSION

The procedure I have described in this essay is a stable and robust procedure that we have used with confidence to develop alternate forms of exit-test prompts. However, if such tests were to be the sole criterion by which student writing proficiency were judged, the alternate forms would need to be pretested and demonstrated to be of equivalent difficulty. A preferable solution, and the solution to which we subscribe, is to collect evidence of writing proficiency from as many sources as possible on as many occasions as possible before assigning final course grades.

(Continued)

Appendix A

Sample Exit-Test Prompt

> The author of the following passage tells about a self-discovery game that led her to examine the different roles that she plays. The passage leaves the reader with a question about when a person is his or her "real" self and when he or she is simply playing a role. As you read the passage, be thinking about all of the roles you play.

We all play many roles in our lives. At a recent group meeting, the members were asked to list their roles on cards: husband, father, mother, teacher, student, sister, daughter, tutor, friend, volunteer worker, and so forth. We all felt we had too many roles. Then, one-by-one, we were asked to discard our roles by throwing one card at a time on the floor. I happily threw away "student." It's a role that I don't like very much. Next, I threw away one of my part-time jobs; I have too many, anyway. Throwing away roles was fun, and it seemed to make my life much less complicated until I got down to the last roles: Mother and friend.

"I'm not playing anymore," I said. "I have to have these roles to make life worth living."

That statement led the group to discuss what roles were the most important to them. Some picked the role of husband, some picked the role that they play at work, and other picked the role of student. It was an interesting exercise,and we all agreed that it was quite important for us to carefully examine our roles from time to time.

Write a well-organized, detailed response to some narrowed aspect of the subject of the reading passage. Your response should express <u>your</u> ideas and should not simply restate the points made in the passage; your response should not be merely a summary.

In writing your response, you might want to consider, for example, how the various roles you play differ from one another. On the other hand, you might want to tell why you prefer one role to another. Do you prefer your role as a student, for example, to your role as a family memeber, friend, employee, ball player, sorority or fraternity member, or date?

Another possibility is for you to consider the way that other people whom you know play their roles. Perhaps you know somebody who has a job or a talent or a hobby that always keeps him or her before the public. Does the person you know in private show up in that person's public role?

Whatever narrowed aspect of the subject you choose to write about, remember that your response should reflect your own experiences and point of view,and it should clearly display mastery of the skills that you have learned in English 100.

For your final draft, use only the answer sheet that your instructor provides. Do not make any marks on your answer sheet that you do not want the faculty judges to take into consideration as they rate your work.

You may use a dictionary if you wish.

Appendix B

Table 1

Failure Rate by Test Passage

Semester	Passage 1 (n)	Passage 2 (n)	Passage 3 (n)	χ^2 (df)
I	51% (55)	57% (46)		.13(1), ns
II	46% (46)	52% (66)	50% (40)	.38(2), ns
III	58% (40)	74%) (23)		1.06(1), ns
IV	61% (41)	50% (50)		.69(1), ns
V	65% (26)	78% (41)		.73(1), ns

Table 2

Agreement Between Raters

Semester	Percentage of Agreement
I	79%
II	71%
III	75%
IV	73%
V	83%

Table 3

Failure Rates and Agreement Between Exam Rating and Course Grade

Semester	Final Exam Failure Rate	Course Failure Rate	Agreement Between Final Exam and Course Grade
I	54%	32%	74%
II	49%	29%	76%
III	63%	38%	65%
IV	55%	44%	76%
V	73%	49%	70%

Notes

[1] In the time since this essay was first prepared, a university committee has recommended, over the objection of the basic skills staff, that the *necessary condition* for exiting the basic writing course be a passing mark on a holistically scored essay written under timed-test conditions. Unfortunately, this recommendation has been implemented.

[2] All new undergraduate students (including those who transfer fewer than 21 semester hours of credit) whose ACT composite score falls between 14 and 17 (SAT combined verbal and math score of 700-790) and whose English subtest score is 13 or below (SAT verbal 310 or below) are required to enroll in our basic reading/writing course. Students who do not earn "C" or better in this required course within two semesters of enrollment are suspended from the university for a period of one calendar year.

[3] Lee Odell, for example, advises us to give students the opportunity to make their best showing so that our judgments are not limited and misleading. He gives the following guidelines:

> Have students write under circumstances that approximate the conditions under which important writing is done; ask them to do more than one kind of writing—that is, have them write for more than one audience and purpose; provide them with information about audience and purpose for which a given piece of writing is intended; assess the demands of our writing assignments, especially when we create more than one assignment; base our judgments on an adequate amount of students' writing. (113)

[4] Elbow and Belanoff report using panel-judged portfolios as the basis for course-exit decisions. Students are not allowed to exit the course with the minimum passing grade (C) until their coursework portfolios are judged passing by at least one instructor in addition to the student's own. In contrast, the testing system I report here leaves that final judgment in the hands of the student's instructor.

[5] Readers of this essay might ask how we know whether a student has a "reading problem" or a "writing problem" and how we dare appear "antidiagnostic" in times of great attention to detail in diagnosis and prescription. In our experience it is the rare student who has *either* one sort of problem or the other and who also meets East Texas State University's admission standards. It is also the rare student whose reading or writing skills are so intractably underdeveloped that an integrative approach does not make powerful inroads into the improvement of both.

[6] Charles Cooper warns:

> There is, of course, a serious reliability problem. To overcome it, groups of teachers or researchers have to work together to train themselves as raters. They have to cooperate further to obtain multiple independent ratings of at least two pieces of a student's writing. (21)

We are convinced that we comply in fact and in spirit with this advice. The study I report in this essay emphasizes the interjudge reliability of our testing procedure as well as our preference for using test results only in conjunction with other assessments of student writing proficiency.

[7] Our testing situation approximates what Hoetker refers to as the *real* rhetorical situation (see also Hoetker and Brossell, 329):

> Most students, regardless of what role they are asked to assume or what audience they are asked to imagine, write for what they imagine is their real audience—hypercritical English teachers. Their ideas about the readership of the test and about what will impress that readership are often stereotyped and faulty. I suggest that it would be better to establish accurately and fully the *real* rhetorical situation. What sorts of people will be reading the papers? What will they be looking for? How will they be evaluating? How will the readers probably respond to first-person essays? To elevated diction? To mechanical errors? And so forth. (387)

Works Cited

Cooper, Charles. "Holistic Evaluation of Writing." *Evaluating Writing.* Eds. Charles Cooper and Lee Odell. Urbana, IL: NCTE, 1977. 3-31.

Elbow, Peter, and Pat Belanoff. "Portfolios as Substitute for Proficiency Examinations." *College Composition and Communication* 37 (1986): 336-39.

Hoetker, James. "Essay Examination Topics and Students' Writing." *College Composition and Communication* 33 (1982): 377-92.

Hoetker, James, and Gordon Brossell. "A Procedure for Writing Content-Fair Essay Examination Topics for Large-Scale Writing Assessments." *College Composition and Communication* 37 (1986): 328-35.

Odell, Lee. "Defining and Assessing Competence in Writing." *The Nature and Measurement of Competence in English.* Ed. Charles Cooper. Urbana, IL: NCTE, 1981. 95-138.