DOCUMENT RESUME

ED 397 424 CS 215 365

AUTHOR	Hayes, John R.; And Others
TITLE	Experimental Approaches to Evaluating Writing. Study 1: In Search of Writing Ability: Exploring Consistency of Student Performance on Holistically Scored Writing Tasks. Final Report.
INSTITUTION	National Center for the Study of Writing and Literacy, Berkeley, CA.; National Center for the Study of Writing and Literacy, Pittsburgh, PA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	May 96
CONTRACT	R117G10036
NOTE	21p.; For study 2, see CS 215 366.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	College Freshmen; *Freshman Composition; Grading; Higher Education; *Reliability; *Student Evaluation; *Writing Achievement; *Writing Evaluation; Writing Research
IDENTIFIERS	Alternative Assessment; *Writing Tasks

ABSTRACT

A study examined college students' responses to writing tasks that were created by their instructors--writing tasks that constituted an important part of the instructors' course designs and that were presented to students as an integral part of the curriculum. In all, approximately 4800 independent evaluations of 796 essays were analyzed. The essays were written by 241 first-year writing students at 2 colleges. The writing samples were scored by raters who were encouraged to use criteria they normally employ in their grading, rather than a special set of criteria imposed by the researchers. Results indicated that the correlation in quality among successive essays written by the same student was quite low; and raters did not tend to grade essays written later in the semester better than essays written earlier in the semester. Findings suggest that the consistency of writing performance in more naturalistic settings may be even lower than previous estimates of writing consistency have indicated. (Contains 10 notes, 14 references, and 2 tables of data. Appendixes present the ranking worksheet and a survey instrument.) (RS)

NATIONAL CENTER FOR THE STUDY OF WRITING AND LITERACY
School of Education • University of California • Berkeley CA 94720-1670
telephone: (510) 643-7022 • e-mail: writ@violet.berkeley.edu
http://www-gse.berkeley.edu/research/NCSWL/csw.homepage.html

FINAL REPORT

EXPERIMENTAL APPROACHES TO EVALUATING WRITING

Study 1

In Search of Writing Ability: Exploring Consistency of Student Performance on Holistically Scored Writing Tasks

John R. Hayes
Jill A. Hatch
Christine M. Silk

Carnegie Mellon University

May, 1996

Project Directors: John R. Hayes and Karen A. Schriver

NOTE: The research reported herein was supported under the Educational Research and Development Center Program (R117G10036 for the National Center for the Study of Writing) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

In Search of Writing Ability: Exploring Consistency of Student Performance on Holistically Scored Writing Tasks

John R. Hayes, Jill A. Hatch, and Christine M. Silk

Carnegie Mellon University

In a large auditorium at a state university, 700 incoming undergraduate students prepare to spend the next 60 minutes writing an impromptu essay. The results of this placement examination will be used to assess the students' present level of writing ability so that they can be assigned to the writing course most appropriate for them.

In a composition classroom at a junior college, a writing instructor collects the second assignment of the semester from her students. Like the first assignment, she expects that this set of papers will reflect a range of ability levels. And while she does hope for at least some improvement in most of her students by the end of the semester, she expects that, for the most part, the students who do well on her final assignment will be the same students who did well on the first.

In a faculty conference room at a private university, instructors from a variety of disciplines gather to read writing portfolios submitted by the university's graduating seniors. Whether students receive their degrees will depend, in part, on whether the writing samples they have provided for this exit test demonstrate that they have attained a sufficient level of writing ability.

At first glance, there may appear to be nothing very remarkable about any of the commonplace examples of writing assessment described above. One reason for this may be that they have, in fact, become so commonplace that many of their underlying assumptions are largely taken for granted. Certainly, one of the most fundamental of these assumptions is that there is consistency of performance from one occasion to the next. It would make little sense to undertake placement or exit testing, for example, if a relationship were not assumed between students' performance on the test and their future writing performance in college courses or in the workplace. Similarly, the practice of responding to students' classroom writing would make little sense if it were not assumed that, without intervention, students would be likely to continue to experience the same difficulties. As such, we believe that the assumption of consistency of performance bears much more

than simply remarking; it in fact demands careful examination. To this end, we undertook a study designed to estimate the consistency of student writing performance on a series of holistically scored writing tasks. While we recognize that the assumption of consistency underlies all forms of writing assessment–for instance, multiple choice testing, primary trait scoring of whole texts, and holistic scoring of whole texts–we chose to focus on the holistic scoring of whole texts because this method is so widely used for three very common assessment purposes: placement testing, classroom evaluation, and exit testing.

The usefulness of predictions about student writing performance that are made on the basis of placement or exit testing scores or instructors' evaluations depends upon how consistent writing performance is. Of course, it would be unreasonable to expect that writing performance should be perfectly consistent. Variations in genre, topic, and available writing time, for example, as well as day-to-day changes in motivation and fatigue, will lead to fluctuations in the quality of a writer's performance from one writing task to another. Yet these fluctuations are usually viewed as deviations from some more or less typical level of performance. If these deviations are relatively minor–and thus consistency from one writing task to another quite high–then we can feel comfortable using a single writing sample to place students into writing classes or to determine whether students have acquired certain writing skills. If fluctuations in performance are somewhat greater, and therefore writing performance is perhaps only moderately consistent, then we can still be fairly confident that three or four writing samples–the number of samples often used as the basis for determining course grades and writing portfolio scores–will tell us something useful about students' writing abilities. If, however, writing performance varies widely from one occasion to the next, indicating that the consistency of writing performance is quite low, then we may have to question the utility and appropriateness of many of our writing assessment practices.

Considerable research attention has been devoted to the issue of *interrater reliability*, that is, the degree to which independent judges can agree in assessing the quality of a single writing sample. In contrast, although it is clearly an important topic, the issue of writing consistency, also referred to as *test-retest reliability*, has received surprisingly scant attention. Although achieving high interrater reliability is important because consistency of scoring across judges helps to insure that the quality of a particular writing sample has been accurately assessed, even complete agreement across judges cannot tell us what level of quality to expect from a subsequent text produced by the same writer.

Although test-retest reliability has not been widely researched, there do exist a few studies that provide us with estimates of how consistent students' writing performance may be. These estimates are reported in one of two ways:

as a correlation coefficient or as a generalizability coefficient. The first of these represents the average correlation among the scores that writers obtain on two or more writing samples. These average correlations reflect the extent to which writers who perform well (or poorly) on one writing sample tend to perform well (or poorly) on another. The second type of estimate is expressed as a generalizability coefficient (Brennan, 1977, Crocker & Algina, 1986). Conceptually, these two estimates are closely related. As the average correlation among a set of writing scores increases or decreases, so does the generalizability coefficient for that set of scores. The estimates differ in that the average correlations simply reflect observed relations in the obtained data. In contrast, generalizability coefficients, although based on data, reflect some strong theoretical assumptions. In particular, in calculating generalizability coefficients, it is assumed that the observed writing scores are a random sample of all possible writing scores and that a person's "true writing ability" may be thought of as the average score that the person would receive on all possible writing tasks. The generalizability coefficient is intended to estimate the correlation of the obtained writing scores to theoretical estimates of the writer's "true writing ability."

These studies differ not only in the way in which they calculate and report estimates of writing consistency but also in whether these estimates are for a single writing sample or for a specified number of samples (greater than one). In the review of the literature that follows, we have attempted to facilitate comparisons of results across studies by, when necessary, using the published data to calculate estimates not provided by the authors.

Herman, Gearhart, & Baker (1993) collected narratives and summaries from 34 first-, third-, and fourth-grade students. The writing samples were holistically scored by instructors who were experienced in evaluating the writing of students in these grades and who had received training designed to improve interrater agreement. The test-retest reliability for a single writing sample[1] was 0.33[2]. Using the Spearman-Brown Prophesy, we estimate that it would require a portfolio of nine writing samples per student to achieve a test-retest reliability of 0.80, the level usually regarded as the minimum for satisfactory reliability.

Barrett (1994) collected three writing samples from each of 164 fourth- and fifth-grade students. The samples were holistically scored by the students' instructors. Barrett calculated a generalizability coefficient of 0.50 for these data and, using the Spearman-Brown formula, estimated that five writing samples per student would be required to achieve a test-retest reliability of 0.80.

---

[1]In all of the studies reported here, test-retest reliabilities for single writing samples were measured by average pairwise correlations.

[2]The average pairwise correlation was estimated from Table 9 in Herman, Gearhart, and Baker (1993).

Godshalk, Swineford, & Coffman (1966) collected five writing samples from 11th- and 12th-grade students in 24 schools. For three of the five samples, students were allowed 20 minutes and for the remaining two samples, they were allowed 40 minutes. Each sample was scored by five readers. Godshalk et al. reported a test-retest reliability of 0.84 for all five samples. This figure is consistent with a test-retest reliability of 0.51 for a single sample.

Lehmann (1987) collected four writing samples from each of 1487 eleventh-grade students. The samples were holistically scored by two raters who received training to improve interrater reliability. Lehmann found a test-retest reliability of 0.307 for a single sample. Using the Spearman-Brown formula, we would estimate that ten essays per student would be required to achieve a test-retest reliability of 0.80.

Breland, Camp, Jones, Morris, & Rock (1987) collected six writing samples from each of 267 college students including two narrative, two expository, and two persuasive samples. The narrative and expository samples were composed in class and the persuasive samples were composed both in class and out of class. Each sample was holistically scored by three raters who received training to improve interrater reliability. Breland et al. report a test-retest reliability of 0.42 for a single writing sample. They estimate that six essays would be required to achieve a test-retest reliability of 0.80.

Taken together, these studies suggest that test-retest reliabilities for individual writing performances range from 0.31 to 0.51 and that to achieve a test-retest reliability of 0.80 one would need between five and ten independent, holistically assessed writing samples–a less than encouraging state of affairs. Even less encouraging still, these studies may actually present us with some of the higher estimates of consistency to be had since the writing samples collected i ı the studies may not be representative of the writing done in typical classroom settings. For instance, for those studies that reported assignment length, the majority of these were short, in-class tasks, some as short as 20 minutes. In many cases, students were given the same amount of time to work on each writing task. In none of the studies were the writing tasks designed by the classroom instructor nor was there evidence that the assessment tasks were integrated into the curriculum as instructors' writing assignments usually are. Finally, in most of the studies, the writing samples were scored by raters who were trained to rate samples in ways that promoted interrater reliability–not the typical context for grading writing assignments in most classrooms. Many of these features–for example, giving students relatively brief periods of time for writing and holding them to similar time constraints for different tasks–may serve to minimize the amount of potential variation that samples of students' writing might otherwise display.

With these concerns in mind, the present study was designed to provide evidence about the test-retest reliability of writing performance that we believe more closely reflects the conditions that prevail in typical college classrooms. In particular, we examined college students' responses to writing tasks that were created by their instructors--writing tasks that constituted an important part of the instructors' course designs and that were presented to students as an integral part of the curriculum. In addition, the writing samples were scored by raters who were encouraged to use the criteria they normally employ in their grading, rather than a special set of criteria imposed by the researchers. In all, we analyzed approximately 4800 independent evaluations of 796 essays written by 241 students in 13 first-year writing classes at two colleges. The results of these analyses suggest that consistency of writing performance in more naturalistic settings may be even lower than previous estimates of writing consistency have indicated.

Method

Materials. The materials for the study were collected during the Spring and Fall semesters of 1993, from 13 freshman writing classes--seven at College 1[3] and six at College 2[4]. Writing instructors at the two schools who volunteered to participate were asked to submit duplicates of final drafts of the three or four major essay assignments which they planned to include as a part of the course and which counted toward the final grade. The essays were written under normal course conditions and produced by the students outside of class. Students' consent was obtained prior to data collection. Students were debriefed about the nature of the study by their instructors at the end of the semester.

Of the seven classes taught at College 1, four were sections of Course 100 and the other three were section of Course 101. Within each of these courses, all sections had identical syllabi, readings and essay assignments, with only minor changes made by individual instructors. The six writing courses taught at College 2 did not share a common approach or a common set of assignments. The syllabus, readings and essay assignments in each course were chosen by the instructor who taught it.

The seven College 1 classes included 127 students--a mean of 18.1 per class. The six College 2 classes included 114 students--a mean of 19 per class. From these classes, we obtained essays written in response to either three,

---

[3]College 1 is a small research-oriented university in Pittsburgh, PA. Average SAT-verbal scores for students at College 1 were 564 at the time of the study.
[4]College 2 is a small business college in Pittsburgh, PA. Average SAT-verbal scores for students at College 2 were 396 at the time of the study.

four, or five assignments (the average was 3.5). We received and evaluated a total of 796 essays from the 241 student participants in the two colleges. If every student had completed every assigned essay, we would have received 856 essays. Thus, the students in our sample completed 93% of the essays assigned.

Judges. The 16 judges who evaluated the essays were experienced writing instructors from College 1 who had taught writing for at least two semesters. The mean number of semesters they had taught writing (not including the semester in which they participated in the study) was 8.8. On average, judges had begun teaching 4.3 years prior to participating in this study. In no case did the judges in the study evaluate papers written in their own class. None of the judges were told the purpose of the study until they had completely finished evaluating the essays.

Procedure. The essays were prepared for evaluation by removing any marks that would reveal students' and instructors' names, dates on which the essays were written, course titles or course numbers. Handwritten essays were typed so that an writer's handwriting could not be identified. However, spelling and grammar errors were preserved. Code numbers were assigned to each essay, rather than to each student, so that judges could not use the code numbers to determine if the same student had written more than one essay.

All of the essays written by the students in the same class in response to a particular assignment were bundled together as one *essay set*. All of the essay sets from a given class were evaluated by the same judges (usually six of them), and each judge evaluated essays from at least three classes. Judges were instructed to work on only one essay set at a time. Once judges finished evaluating a set and moved to another, they were instructed not to re-adjust the scores assigned to the finished set. Sets of essays submitted to the judges were shuffled so that the order of the sets would reveal nothing about the class in which they were written or the order in which they were assigned in that class. We hoped that these procedures would prevent judges from identifying cases in which two or more essays were written by the same student. Our objective was to insure that all of the essays were evaluated independently of each other.

Accompanying each essay set was the original assignment sheet for that essay set with instructors' name, course title, and dates removed. In assessing essays, judges were explicitly instructed to use the same criteria they would use if the essays were written by their own students. Although it is common practice in writing studies to conduct training sessions for the judges so that they will be consistent with each other in evaluating papers, we intentionally did not do so. We made this decision for three reasons:

(1) Instructors are ordinarily not trained to increase interrater reliability when they evaluate writing in their classes. We believed that by training the judges, we would be pressuring them to abandon their own values as to what constitutes good writing in the composition class and to adopt criteria that did not fully reflect either their own or perhaps any of the judges values. Since we wanted our study to approximate normal classroom conditions as much as possible, we decided that such training would be counterproductive.

(2) Our judges, as experienced instructors, have had years of experience in applying their own criteria for evaluating writing. If we had provided them with a few hours of training, we might well have led them to modify their criteria at least at the outset. However, we were concerned that over time, judges might gradually drift back to using their own more familiar criteria. Such drift would increase the variability within each judges' evaluations. For this reason, we again decided that such training would be counterproductive.

(3) Since none of our conclusions depended on comparisons among judges, the issue of interrater reliability is irrelevant to our study.

Two assessment procedures were used: grading and ranking. A judge either graded or ranked all of the essay sets from one class. Judges who graded essays were instructed to use standard letter grades from A+ to E, with A+ indicating excellent and E, failing. Judges who ranked the essays were instructed to divide the set of essays into four groups. The first quartile contained the highest quality essays while the fourth quartile contained the lowest quality essays. Within each quartile, the judges ranked the essays from strongest to weakest in descending order. (A coding sheet is shown in Appendix A). When the total number of essays in a given set was not evenly divisible by four, the coding sheet indicated how the essays were to be divided.

In most cases, each essay set was ranked by three judges and graded by another three judges, but a few sets had only two judges of each kind. In addition, we collected the essay grades from the instructors when they were available[5] and compared the instructors' evaluations of their own students to the evaluations of the judges.

We administered a questionnaire to the judge after they were completely finished evaluating the essays (see Appendix B). The questionnaire asked the judges, among other things, to describe the general criteria they used in evaluating the essays, whether their evaluation would be different if these essays were from a class they were teaching, and whether while reading an

[5]Some instructors did not assign grades to individual essays.

essay they felt they had previously read an essay by the same writer in an earlier set.

Results

Correlations among essays written by the same student. Our major result, shown in Table 1, is that on the average, the correlation in quality among successive essays written by the same student in the same class is quite low--0.106 for grading and 0.209 for ranking for an average of 0.16. The Spearman correlations among ranks were significantly greater than the Pearson Product Moment correlations among grades ($F(1,22)=5.501$, $p=.028$). Thus, ranking may provide somewhat more information about student performance than grading. The correlations between successive essays for students in College 1 were not significantly different from those for students in College 2.

One might expect that successive essays written by a student, e.g., essays 1 and 2 or essays 3 and 4, would tend to resemble each other in quality more than non-successive essays, e.g., essays 1 and 3 or essays 2 and 4. However, the data reveal no tendency for the quality of successive essays to be more strongly correlated than the quality of non-successive ones. In fact, non-successive essays were slightly but not significantly more strongly correlated ($r=.154$) than successive ones ($r=.133$).

Grading. The independent judges knew what the writing assignments were but they did not know the order in which they were assigned. Table 2 shows that the independent judges did not tend to grade essays written later in the semester better than essays written earlier in the semester[6]. This does not necessarily mean that the students were not improving their skills over the semester. It might have been that the instructors were giving more challenging assignments as the semester progressed.

The independent judges assigned significantly better grades ($t= -3.098$, $df= 11$, $p=0.010$) to essays written by students at College 1 (mean grade = 2.71) than those written by students at College 2 (mean grade = 2.20).

Finally, classroom instructors assigned slightly but significantly better grades ($t= 2.638$, $df=5$, $p=0.046$) to their students' writing (mean grade = 2.99) than did the independent judges (mean grade = 2.59).

Discussion

The average test-retest reliabilities that we observed in this study (0.11 for grades, 0.21 for ranks) are substantially lower than those observed in earlier

---

[6] The grades that the teacher assigned were given numerical values as follows: A=4, B=3, C=2, D=1.

studies. We believe that the lower reliabilities that we observed may be the result of the more naturalistic method we used in selecting writing assignments. All of the assignments we evaluated were chosen by the classroom instructors as the three or four most significant assignments of the semester. These assignments were all take-home assignments. In contrast, in the earlier studies, the majority of assignments were written in class with strict time limits. In comparison to in-class assignments, take-home assignments afford the writer greater opportunity for library research, ror reflection, and for revision. Further, they allow for considerable variation in time-on-task both across writers and across writing occasions.

Whatever the cause of the lower reliabilities, these results have strong implications for the use of holistically scored essays for assessment in college classes. Correlations as low as those we found indicate that knowing how well a student did on one essay allows us to predict very little about the quality of another essay that the student writes for the same class. In particular, it will allow us to predict only between 1.5% and 4.5% of the variance. Stated negatively, that means that at least 95% of the variance is unaccounted for.

Implications for placement and exit testing. In most placement testing programs, the entire freshman class is asked to provide a writing sample in a timed and supervised setting. These samples are then holistically graded and the results are used to assign students to classes. In the best programs, considerable attention is devoted to establishing the rater reliability of the holistic grading (see Smith, 1992). However, relatively little attention has been devoted to assessing how well such writing samples predict classroom writing performance.

If we are to draw inferences from our results for placement and exit testing, we need to be concerned about whether the classes we studied represent the range of skills in the group that undergoes placement and exit testing. In both colleges, there is reason to believe that the range of skills in the classes we studied is restricted compared to the total student population. In College 1, students with advanced placement credit in English, about 40% of the student body, were excused from taking freshman writing courses. In College 2, students were assigned to writing classes on the basis of test scores and high school grades. Thus, in both colleges, we would expect that test-retest reliabilities would be somewhat higher if measured over the whole population of freshman students in the college than just those enrolled in freshman writing courses. However, even if we make generous assumptions (for example, that the standard deviation of skill levels in the restricted group is only half that in the unrestricted group), the estimated test-retest reliability of grades for the whole population would be about .12 for College 1 and .22 for College 2. Thus, even for the unrestricted population, performance on one essay predicts less than 5% of the variance in performance on another.

A writing sample collected in a timed and supervised setting is likely to be a poorer predictor of performance on writing assignments within a course than is performance on another writing assignment in that same course. Therefore, our data should provide an upper bound on the ability of placement tests to predict performance in composition classes. That is, we would expect, a typical placement test to predict less, and perhaps, much less, than 5% of the variance in the quality of single essays written in freshman composition classes.

Of course, one would not want to change university policies on the basis of a single, potentially fallible study. However, our results suggest that one should not simply assume that placement and exit testing tell us very much about students' writing skills.

Implications for portfolio assessment. Portfolio assessment is a very popular topic in the composition literature, but, as Calfee and Perfumo (1992) point out, it means different things and has different purposes for different groups. In this section, we will discuss the implications of our results for two approaches to the assessment of writing, both of which make use (but different use) of portfolios. In one view, the primary goal of writing assessment is the measurement of writing ability–a trait that facilitates writing performance in a wide variety of writing tasks. An alternative view is that the primary goal of writing assessment is to identify profiles of skills that writers possess. This difference in approach bears a resemblance to a long-standing controversy in the field of intelligence testing. Intelligence tests assess human performance on a variety of tasks including memory span, vocabulary, and spatial relations. The ability to do well on any one of these tasks could, conceivably, be independent of the ability to do well on any of the others. In fact, though, there are correlations among the various tasks used to measure intelligence. One view, proposed by Spearman (1904), is that these correlations can be accounted for by a single factor which he called general intelligence. The concept of writing ability is similar to the concept of general intelligence in the sense that both are assumed to account for correlations in peoples' performance on a variety of tasks. Thus, people with high general intelligence will be expected to perform well on memory, vocabulary, and problem solving tasks, and people with high writing ability will be expected to perform well in writing summaries, arguments, reports, etc.

An alternative to Spearman's position, put forward by Thurstone (1938)[7], is that the correlations among tasks in intelligence tests are best accounted for by postulating a number of special abilities, for example, abilities in handling numbers, words, spatial relations, etc. Thus, Thurstone believed that a

[7]Recently, Gardner (1983) and Sternberg (1985) have championed similar notions.

person's intelligence is best describes by a profile of special abilities rather than by a single ability. Clearly, Thurstone's view is similar in spirit to the view that the goal of writing assessment should be identifying profiles of writing skills.

These views, although different, are not incompatible. One can imagine that individuals have a mix of general and special abilities both for writing and for taking intelligence tests. However, in the case of writing, we find very little evidence that the concept of general writing ability is useful in accounting for writing performance in college classes, at least in so far as that performance is measured by holistic assessment.

Our study was not designed to identify special writing abilities. However, we believe that attempting to identify such abilities is probably the most profitable direction for assessment research to take. By identifying special writing abilities, one could provide a solid basis for creating useful profiles of writing skills.

Implications for the classroom. We noted in the results section that the students' instructors gave them slightly higher grades–by less than half a letter grade–than did the judges. We doubt that this difference has any important educational consequence. Of more significance, though, may be a tendency on the part of some instructors to expect, and even to see, more consistency in the quality of students' writing than is there. Instructors may be subject to a "halo" effect. That is, they may classify their students as "good" or "poor" writers on the basis of their first few essays and bias their scoring of the students latter essays on the basis of this categorization. Thus, instructors might perceive greater consistency among their students than is warranted. There is some evidence in our data that halo effects may have influenced the instructors in our sample. We found a marginally significant tendency ($t=2.022$, $df=7$, $p=.083$) for instructors to find greater test-retest reliability in their students' essays than did the judges.

Because a student's performance on one essay is so weakly related to that same student's performance on other essays, neither the instructor nor the student should draw strong conclusions from any individual performance. In particular, our results also suggest that a student should not take a poor initial grade as a reason to give up. Palmquist & Young (1992) have shown that many students believe that writing is a gift, that is, that one can either write well or not and that little can be done about it. This belief implies that the quality of separate writing performances are strongly correlated. A student holding this belief might take a poor initial performance as indicating that they don't have the gift and that greater effort devoted to writing will not result in better grades. Our results suggest that the "writing as a gift" theory does not describe writing performance in the classes we have studied.

Students might be able to make better judgments about themselves as writers--and their ability to improve as writers--if they knew this.

Factors that may limit the generality of the findings. The sample that we selected, 13 classes, formed a substantial proportion (about 15%) of the freshman composition courses taught at the two universities in the years studied. However, they did not constitute a representative national sample of classes taught across the country. Similarly, although the judges in the study were all experienced instructors, they did not constitute a representative national sample of university composition instructors[8]. Thus, although we selected the classes, instructors, and assignments without intentional bias, we cannot guarantee that the results will generalize to other universities or other parts of the country.

The results of the study apply only to text quality as measured by holistic assessment. Other methods of measuring quality may be more reliable or less reliable. However, holistic assessment is the measure of quality that is most frequently used in composition classes.

It is possible that some judges perceive more consistency in the quality of student essays than do others. Indeed, we found one judge, HR, whose judgments showed higher correlations across essays than did other judges. HR's correlations over the three classes she assessed as part of our study averaged r=0.381. Having noticed her high correlation, we asked her to assess three additional classes. Her average correlation over these additional classes was r=0.335 yielding an overall average of r=0.358. These results suggest that HR's judgments were more consistent than those of the other judges. However, a questionnaire in which judges were asked to describe their procedures for making judgments failed to reveal any apparent systematic differences between HR and the other judges. In any case, if all of the judges had performed with the same consistency as HR, a student's performance on one essay would still predict only about one eighth of the variance in performance on another essay. Thus, even in the best case, our results indicate that the holistically assessed writing quality of freshman student essays is quite inconsistent from one essay to the next.

Comparison to other academic performances. We wondered if these correlations were really lower than one would expect of other kinds of academic performances. Perhaps the demands of college life on students' time, attention, and motivation are sufficiently variable that we should not expect substantial correlations in any type of academic performance. To provide a comparison with our observations on writing, we examined

---

[8]Generally, it is prohibitively expensive to obtain representative national samples. None of the studies of the stability of writing performance that we reviewed above attempted to obtain such samples.

students' grades on four one-hour exams in a large freshman psychology class. The exams were taken about three weeks apart during the course of the Fall semester. Each of the exams covered a different study unit but all required the students to write short (three to five sentence) responses to five or six questions. Product-moment correlations were calculated for each pair of exams. The average of these pairwise correlations was r=0.45. A parallel analysis of three one-hour exams taken in a 60 person freshman physics course yielded an average pairwise correlation of r=0.55. It appears, then, that students' performance on at least some academic tasks is more consistent than their performance on holistically assessed writing assignments.

Sources of variability in writing. Why might writing performance be especially variable in comparison to other academic tasks? The following are some of the possibilities. First, the various writing tasks that instructors assigned may have called on different writing skills. Typically, the instructors' assignments required students to write in a variety of genres over the course of the semester. Thus, one assignment might have placed heaviest emphasis on descriptive skills, a second, on narrative skills, and a third, on argumentative skills. If students have different profiles of writing skills, such task variability would lead to variability in student performance.

Second, because writing is often a less well defined task than many other academic tasks, it might be that writing quality is more sensitive to variations in the students' level of motivation than are better defined tasks.

Third, a major source of variability may be located in the holistic scoring process itself. The holistic scoring of a writing assignment typically requires the judge to consider a large number of factors and to keep them in mind during the relatively extended period required to read a multipage essay. Scoring an exam question, in contrast, requires the judge to focus rather narrowly (often just on factual accuracy) during a relatively brief reading period.

Summary

In this study, we have found very low test-retest reliabilities for holistically assessed essays written in freshman composition classes. This result suggests that placement and exit testing based on a single sample may not provide much information about students' writing skills. Further, our results suggest that general writing ability, viewed as a common factor contributing to all writing performances, played, at most, a very small role in students' holistically evaluated writing performance. This does not mean that one couldn't find evidence of general writing ability by using other methods, but it does suggest that in classes such as the ones we studied, the concept of general writing ability is not very useful in accounting for the holistic assessments that the students' essays received.

References

Barrett, T. J. (1994, November) *Generalizability of Writing Tasks at Fourth Grade in the Riverside Unified School District.* Paper presented at the Annual Meeting of the California Educational Research Association. San Diego, CA.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987) *Assessing Writing Skill.* (Research Monograph No. 11), New York: College Entrance Examination Board.

Brennan, L. (1977) Generalizability Analysis: Principles and Procedures. (Technical Bulletin No. 26). Iowa City, Iowa: ACT.

Calfee, R., & Perfumo, P. (1992). A survey of portfolio practices. Berkeley, CA: University of California, Berkeley, Center for the Study of Writing.

Crocker, L. & Algina, J. (1986) *Introduction to Classical and Modern Test Theory.* Fort Worth: Harcourt, Brace, Jovanovich.

Gardner, H. (1983) Frames of mind: The theory of multiple intelligences. New York: Basic Books.

Godshalk, F. F., Swineford, F., & Coffman, W. (1966) The Measurement of Writing Ability. (Research Monograph No. 6). New York: College entrance Examination Board.

Herman, J. L., Gearhart, M, & Baker, E. L. (1993) Assessing Writing Portfolios: Issues in the Validity and Meaning of Scores. *Educational Assessment,* 201-224.

Lehmann, R. H. (1987, April) *Reliability and Generalizability of Ratings of Compositions.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C.

Palmquist, M. & Young, R. (1992) The notion of giftedness and student expectations about writing. *Written Communication,* 9, no. (1), 137-168.

Smith, W. (1992) Teaching experience and placement skill. In Hayes, Young, Matchett, McCaffrey, Cochran, & Hajduk (eds) Reading Empirical Research Studies: The rhetoric of research. Hillsdale, NJ:Lawrence Erlbaum Associates.

Spearman, C. (1904). General Intelligence, objectively determined and measured. *American Journal of Psychology,* 15, 201-293.

Sternberg, R. J. (1985) Beyond IQ: A triarchic theory of human intelligence. New York: Cambridge University Press.

Thurstone, L. L. (1938) Primary Mental Abilities. *Psychometric Monographs* (Whole No. 1).

	Grading	Ranking	Average
College 1	0.062	0.208	0.140
College 2	0.149	0.211	0.185
Both Colleges	0.106	0.209	0.163

Table 1. Average correlations in holistically assessed quality among assignments written by the same students for the same class.

			Essay Number				
	# of classes	# of essays	1	2	3	4	Mean
College 1	3	3	2.78	2.90	2.85		2.84
	4	4	2.68	2.69	2.48	2.62	2.62
College 2	4	3	2.47	1.82	1.89		2.06
	2	4	2.45	2.35	2.09	2.49	2.35
Both Colleges	7	3	2.63	2.36	2.37		2.45
	6	4	2.61	2.57	2.35	2.58	2.53

Table 2. Average grades assigned by independent judges to successive essays written by the same students. Classes requiring 3 essays are presented separately from classes requiring 4 essays. There is no obvious tendency for essay quality to increase as the number of essays that the student has written for the class increases.

Appendix A[9]

Judge_____

Ranking Worksheet

Group 1 -- Highest Quality Quality

1. _____
2. _____
3. _____
4. _____
5. _____

Group 2 -- Medium High

1. _____
2. _____
3. _____
4. _____

Group 3 -- Medium-Low Quality Quality

1. _____
2. _____
3. _____
4. _____

Group 4 -- Lowest

1. _____
2. _____
3. _____
4. _____
5. _____

[9]The worksheet shown here was used for ranking 18 essays. The number of response slots varied with the number of essays to be graded. The four groups were as evenly divided as possible. When the number of essays was not divisible by four, extra essays were placed first in Group 1, then Group 4, and finally in Group 2.

Appendix B[10]

Survey

1. Please list and briefly describe, if necessary, the general criteria you used in ranking the essays, numbering the criteria from most to least important.

2. Did you use the same set of criteria for all the essay sets? If not, how did your criteria vary?

3. Did you evaluate the essays the way you would if these were papers for a class you were teaching? If not, how was your evaluation different?

4. In general, do you think the writers of the essays did what the assignments asked of them? If not, how did this affect your evaluation of the essays (i.e., how did you treat those writers who misinterpreted the assignment)?

5. While you were reading an essay, did you ever have the feeling that you had previously read an essay by the same writer in an earlier set?

[10]These questions were spead over two pages to provide the judges with adequate space to respond.