Examining the Generalizability of Direct Writing Assessment Tasks

CSE Technical Report 718

Eva Chen, David Niemi, Jia Wang, Haiwen Wang, and Jim Mirocha
CRESST/University of California, Los Angeles

June 2007

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation
Graduate School of Education and Information Studies
University of California, Los Angeles
GSEIS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

EXAMINING THE GENERALIZABILITY OF DIRECT WRITING ASSESSMENT TASKS

Eva Chen, David Niemi, Jia Wang, Haiwen Wang, and Jim Mirocha

CRESST/University of California, Los Angeles

Abstract

This study investigated the level of generalizability across a few high quality assessment tasks and the validity of measuring student writing ability using a limited number of essay tasks. More specifically, the research team explored how well writing prompts could measure student general writing ability and if student performance from one writing task could be generalized to other similar writing tasks. A total of four writing prompts were used in the study, with three tasks being literature-based and one task based on a short story. A total of 397 students participated in the study and each student was randomly assigned to complete two of the four tasks. The research team found that three to five essays were required to evaluate and make a reliable judgment of student writing performance.

Examining the Generalizability of Direct Writing Assessment Tasks

Performance assessment can serve to measure important and complex learning outcomes (Resnick & Resnick, 1989), provide a more direct measurement of student ability (Frederiksen, 1984; Glaser, 1991; Guthrie, 1984), and help guide improvement in instructional practices (Baron, 1991; Bennett, 1993). Of the various types of performance assessment, direct tests of writing ability have experienced the most acceptance in state and national assessment programs (Afflebach, 1985; Applebee, Langer, Jenkins, Mullins & Foertsch, 1990; Applebee, Langer, & Mullis, 1995). Advocates of direct writing assessment point out that students need more exposure to writing in the form of instruction and more frequent examinations (Breland, 1983).

However, there are problems associated with using essays to measure students' writing abilities, like objectivity of ratings, generalizability of scores across raters and tasks (Crehan, 1997). Previous generalizability studies of direct writing assessment

have also found that the variance component for the sampling of tasks tends to be greater than that for the sampling of raters (Breland, Camp, Jones, Morris, & Rock, 1987; Hieronymus & Hoover, 1986). With the time requirement to administer and rate tasks especially when they are multiple tasks and multiple raters, it becomes more desirable and crucial to have writing tasks that have a high level of generalizability. A writing task may have limited validity if it does not provide a basis for making generalizations about the test-taker's writing ability.

The goal of this study was to investigate the level of generalizability across a few high quality tasks and to establish the validity of measuring student writing ability using a limited number of essay tasks. We will examine the magnitude of variability due to the sampling of writing tasks and how to make generalization from the specific writing tasks to the broader domain of student writing ability. The specific research questions are:

- How do students perform across different writing prompts designed to measure general writing ability?
- How many essays should be used to make reliable decisions about students' writing ability?

Literature Review

A review of the literature indicates that there are very few studies that investigate the task generalizability of direct writing assessments. Results from the few existing studies found that performance assessments in writing have a low level of generalizability across different tasks (Baker, Abedi, Linn & Niemi, 1996; Boodoo & Garlinghouse, 1983; Gabrielson, Gordon & Engelhard, 1995; Lamb, 1987). As part of the International Evaluation of Education Achievement tests, Lamb (1987) examined the essay scores from students in New Zealand who were tested in writing skills near the end of their secondary education. Students wrote essays in the following categories: functional letter writing and narrative, persuasive and reflective essay writing. Student performance across tasks and writing dimensions were examined by looking at correlations between scores. Correlations for each dimension of writing were high within tasks but lower between tasks. The author claimed that the students' writing performance varied substantially across each dimension of task depending on how familiar the students were with the task and how comfortable they felt in responding to it. Accuracy in the mechanics of spelling, grammar and usage were also task specific and linked to the student's confidence in responding to each task. Lamb (1987)

reasoned that the different categories of writing tasks demanded students operate at different cognitive levels in order to complete the essays.

Instead of examining essays from different modes of writing, other research focused on student performance on one type of writing task. Gabrielson et al. (1995) examined the effect of writing tasks on student persuasive writing. Fifteen writings tasks, all designed to elicit persuasive essays, were administered to 34,200 11th grade students in Georgia. The tasks were taken from the Georgia High-School Writing Test, a criterion-referenced test designed to provide a direct assessment of student writing competence. A multivariate analysis of variance was conducted using four domain scores as the dependent variables assessing writing quality, and writing tasks as the independent variable. Each composition was judged independently by two raters using the following four domains: a) content and organization, which measures student ability to develop a controlling idea; b) style, which measures ability to control language to establish individuality; c) conventions, which measure ability to use appropriate written language; and d) sentence formation, which measures ability to formulate correct sentences. They found that the writing task variable had a significant effect on all four dependent scales with the largest effect on the content and organization scale and the smallest on the sentence formation scale. The authors claimed that even though all the tasks were constructed to elicit examples of persuasive writing and to be of comparable difficulty, the results indicate that various groups of students reacted to as least some of the writing tasks differently (Gabrielson et al.1995).

One weakness of the Gabrielson et al. (1995) study is that every student wrote only one essay. A better designed study was conducted by Boodoo and Garlinghouse (1983) to assess student competency in persuasive writing. Three writing tasks were administered to a random sample of 34 junior education major college students during a regular class period. The tasks were supposedly and sufficiently structured to reduce bias and the scoring criteria were clearly specified to reduce measurement error in rating, according to the author. Three teachers who were familiar with the course content, questions and examinees anonymously scored the responses using a holistic scoring method. The analysis of variance showed that the factor that contributed to the largest amount of variation was the student by task by rater interaction, followed by student by task interaction. The authors claimed that contents played a large role in the response and the student by task interaction was due to some students doing better on certain topics while doing poorly on other topics. However, the small sample size

could also contribute to the large variance found within the same student across different topics.

Using domain specifications to control topic and rater variability, Baker et al. (1996) conducted a study that involved 69 students in two 11th-grade history classes. The students took three on-demand, multi-step performance tasks one per week for three weeks. For each topic, all students completed a Prior Knowledge Test, read primary source materials, and wrote an essay of explanation. The sequence and types of measures were identical for each of these three tasks. Using a theory-based scoring rubric, four trained raters rated all essays on six dimensions: General Content Quality, Prior Knowledge, Principles, Proportion of Text Detail, Misconceptions and Argumentation. The findings showed that variance components for subjects-by-topics were relatively large for all six dimensions, especially the Prior Knowledge and Misconceptions dimensions. In comparison, the General Content Quality and Argumentation dimensions showed less variability over raters and topics than other dimensions. As there was a low level of generalizability across topics, the authors recommended using multiple topics in any high-stake assessment context and refining topic design to improve the quality of measurement (Baker et al., 1996)

Analytical Scoring Strategies and Task Generalizability

The low task generalizability found in Boodoo and Garlinghouse's (1983) study could be due to the holistic rubric used for scoring. Some researcher recommended analytic scoring of writing products since writing is a multifaceted performance and as such involves attainment of a number of mental traits (Huot, 1990; March & Ireland, 1987; Novak, Herman & Gerahart, 1996; Roid, 1994). Analytical scoring is a trait-by-trait analysis of features important to any piece of discourse. An example of a typical analytical scoring rubric was the one designed by Diederich (1974). Based on the rubric, compositions are scored on a 5-point scale in the following categories: ideas, organization, wording, flavor, and mechanics, which is further divided into usage, punctuation, spelling, and handwriting. The first two subskills, ideas and organization, receive double weighting because of their importance to the success of the essay (Diederich, 1974).

Research conducted by Baker et al. (1996) and Gabrielson et al. (1995) indicated that the writing task variable had a different effect on different dimensions of student writing. In the Gabrielson (1995) study, sentence formation compared to content and organization, was less sensitive to topic change. Since writing performance involves a

number of traits on which individuals differ, multi-trait analytic scoring strategies for writing performance assessment may increase task generalizability over a single holistic score (Crehan, 1997).

In addition to using a holistic scoring rubric, lower task generalizability could also be due to the narrow scale employed (1-5) and range of ratings assigned by the raters (2-5) in the study by Boodoo and Garlinghouse (1983). A longer scale coupled with instructions to raters regulating the use of the entire scale could yield more valid results.

Topic Design and Task Generalizability

One problem with the study conducted by Lamb (1987) is that the topic familiarity was not controlled; the student writing performance varied depending on how familiar they were with the topic. As Hoetker (1982) points out, a topic must not demand information, awareness or special skill that is not likely to be shared by everyone in the population being examined. Brand (1991) recommends two ways to overcome the problem of unequal familiarity with a topic; a) generate a pool of topics of similar difficulty and b) supply the information by providing reading passages in the exam. Generating a pool of prompts of similar difficulty and familiarity can help to ensure the comparability of the measure of examinees' writing performances. The advantage of providing reading passages is that the knowledge base is controlled to some extend. However, because reading means interaction with texts, students need to comprehend and interpret them, which introduces complicated socio-cognitive and affective variables. Their level of reading comprehension could affect their writing performance (Brand, 1991).

Task Specification and Task Generalizability

In the study conducted by Gabrielson et al. (1995) and Boodoo and Garlinghouse (1983), writing tasks were administered to students to examine their skills in persuasive writing. It was not clear in either of these studies how the prompts were designed and if they followed the same task specification. As Liu (1997) emphasized, prompts should be constructed in such a way that they completely follow the same criterion-reference-based test specification, eliciting the same skill or the same composite of multiple skills.

Rationale for This Study

Past research has indicated that student writing performance could vary depending on how familiar the students were with the topic (Hoetker, 1982). One possible explanation for the inconsistency in student writing performance in the studies reviewed above is that the topic familiarity was not controlled. As Hoetker (1982) points out, a topic must not demand information, awareness or special skill that is not likely to be shared by everyone in the population being examined. Built upon the results from past research, this study involved several well-designed writing prompts to investigate 9th-grade students' writing. The topic familiarity was controlled by providing prompts of similar familiarity to the students. Two of the prompts in this study provided reading passages. The advantage of providing reading passages was that the knowledge base could be controlled to some extent.

Research Procedure and Methodology

A total of four different writing tasks were implemented in the study (see Table 1).

Table 1

Writing Tasks Administered in the study

Task	Reading Required	Reading Material	Writing Topic	Summary
1	Students choose a story	Undetermined story	Describe conflict	Undetermined story + conflict
2	Students read a different story provided by CRESST	"The Third Wish"	Describe conflict	Fixed story + conflict
3	Students choose a story	Undetermined story	Describe theme	Undetermined story + theme
4	Students read a non-fiction story provided by CRESST	"On Being Seventeen, Bright, and Unable to Read"	Describe and compare personal difficulty	Fixed non-fiction story + Personal Difficulty

Three of the above four writing prompts were literature-based tasks. In Task 1, students chose a literary work they read in class during the year and wrote about conflict in the story. Similar to Task 1, Task 2 prompted students to write about conflict as well. However, their writing was based upon a short story ("The Third Wish") they read in class. In Task 3, students again chose a literary work they read in class during the year and wrote about the theme of the story. Similarly to Task 2, Task 4 required reading a short story ("On Being Seventeen, Bright and Unable to Read") about a person growing up with dyslexia and prompted the students to write about author's personal experience in coping with dyslexia.

Research Questions

This research collected 9th-grade student writing samples to answer the following research questions:

- How do students perform across different writing prompts designed to measure general writing ability?
- How many essays should be used to make reliable decisions about students' writing ability?

Data Collection Procedure

The study was conducted in 2003 and involved 397 9th-grade students from 19 classes taught by 6 teachers in one urban high school. Those students who had signed the assent and consent forms (by their parents) were included in the study. Since most teachers were not able to spend more than 4 class periods for this research, each student in the study was only required to write two essays out of the four essay prompts. Every class spent two periods for each essay.

Before essay writing started, using student ID numbers, every student in each class was randomly assigned to the following six experimental conditions: a) Essay #1 and #2; b) Essay #1 and #3; c) Essay #1 and #4; d) Essay #2 and #1; e) Essay #3 and #1; f) Essay #4 and #1. Students in the testing conditions a, b, and c wrote Essay # 1 first, followed by one of the other three essays. To counter balance the task sequence effect, students in the testing conditions d, e, and f were given one of the other three essays to write first, followed by Essay #1.

One week before a class was scheduled to write the essays, the teacher received the writing prompts from CRESST. No advance preparation was required and the teachers were informed not to provide any instruction. The students wrote the essays during their regular English class periods. Once they finished writing the first essay, the second essay was to be completed in class within a two-week period. We explained to the teachers that the second essay needed to be completed within a short time to avoid the effect of student maturation on the quality of the writing. The students were asked to write down their school ID numbers on the essays, which were used later to compare the scores from the two essays they had written.

Raters and Scoring

After the writing samples were collected, four former high-school English teachers were recruited to participate in a three-hour training session, in which they were instructed on how to score student essays. A holistic scoring rubric developed by CRESST was used during and after the training to evaluate the compositions. With the rubric, the raters judged the overall quality of the essay by examining the following aspects of writing: content, organization, comprehension, and mechanics. Each essay was assigned a score of 1-4. Rater reliability was checked during and after the training. Each student essay was scored by all four raters and the average essay score was used for data analysis.

Data Analysis

Analysis of variance was conducted using the generalizability framework with a crossed person by rater by task design. The variables of interest were the variance components and the generalizability coefficients observed in the essay scores. The analysis yielded information on the following variance components: person, rater, task, person by task interaction and rater by task interaction. Large variance components of task would indicate the writing prompt effect for all persons due to their inconsistencies in writing from one essay to another as their scores systematically differed across two prompts. Large variance components of task by person interaction would indicate inconsistencies from one prompt to another in some students' writing performance.

Results and Discussion

There are 328 students with 2,525 ratings in the original data set. Among them, 287 students wrote two essays that were each scored by four raters. However we do not have background information on six of these students. Therefore the final data we used contain 281 students, each wrote Essay 1 and another essay, and each essay was scored by four raters.

In the final data set 51% of the students are female, 61% are English language learners, 3% are in special education, and 20% are in honors classes. Fifty-three percent of the students are White, 32% are Hispanic and 15% are of other ethnic background. The distributions of student background variables for each essay are shown in Table 2.

Table 2

Student Distribution by Background Variables by Essays

			N			
Background Variables	Total N (281)	%	Essay 1 (281)	Essay 2 (94)	Essay 3 (95)	Essay 4 (92)
Male	138	49%	138	52	41	45
Female	143	51%	143	42	54	47
White	148	53%	148	60	44	44
Hispanic	91	32%	91	29	34	28
Other	42	15%	42	5	17	20
Non-ELL	110	39%	110	37	41	32
ELL	171	61%	171	57	54	60
Non-Special Ed.	272	97%	272	90	93	89
Special Ed.	9	3%	9	4	2	3
Non-Honors	225	80%	225	80	80	65
Honors	56	20%	56	14	15	27

Table 3 indicates the mean scores of each essay by student background variables. In general, girls scored higher than boys, students in honors classes scored higher than other students, Spanish students scored lower than white and other students, and English language learners and students in special education scored lower than other

students. The magnitude of the differences slightly varies across the four essays. The correlations between students' scores on Essay 1 and scores on other essays are moderate, in the range of 0.61 to 0.68.

Table 3

Mean Essay Scores by Students' Background Variables

	Essay Scores			
Background Variables	Essay 1	Essay 2	Essay 3	Essay 4
Male	1.90	2.00	1.79	1.88
Female	2.04	2.05	2.02	2.06
White	2.10	2.10	2.11	2.07
Hispanic	1.72	1.76	1.65	1.78
Other	2.07	2.55	1.95	2.04
Non-ELL	2.11	2.20	2.10	2.04
ELL	1.88	1.91	1.78	1.94
Non-Special Ed.	1.98	2.04	1.92	1.99
Special Ed.	1.64	1.66	1.63	1.63
Non-Honors	1.86	1.94	1.84	1.78
Honors	2.44	2.51	2.35	2.44
All Students	1.97	2.02	1.92	1.97

We conducted three separate generalizability studies. Study One compared scores of Essay 1 and Essay 2 written by the same students. Study Two compared scores from students who wrote both Essay 1 and Essay 3. Study Three examined scores from the same students on Essay 1 and Essay 4. Variance component analyses were performed to examine the source of the variances in essay scores in each study (see Table 4).

Table 4

Variance Components Estimates for Study One, Study Two and Study Three

Source of Variance	Study One (N=94)		Study Two (N=95)		Study Three (N=92)	
	Variance Components	Percentage	Variance Components	Percentage	Variance Components	Percentage
Student	0.185	36.3%	0.189	37.2%	0.209	39.7%
Essay	0.003	0.6%	0.000	0.0%	0.000	0.0%
Rater	0.005	1.0%	0.000	0.0%	0.000	0.0%
Student by essay	0.054	10.7%	0.071	13.9%	0.096	18.2%
Student by rater	0.086	16.9%	0.032	6.4%	0.025	4.7%
Essay by rater	0.001	0.3%	0.012	2.3%	0.009	0.0%
Error (student by essay by rater)	0.175	34.3%	0.204	40.2%	0.188	35.7%
Total	0.51	100.0%	0.51	100.0%	0.53	100.0%

The results from the three studies were similar. In general, the differences in student writing ability accounted for 36–40% of the variance in essay scores. The next significant source of variance originated from student by essay by rater interaction (34-39%). Last, the variance from student by essay interaction accounted for 11-18% of the total variance. We found that the four raters scored reliably. The effect of essay by rater was essentially zero. There was no important systematic difference among raters. For Study Two and Study Three, the interaction of student by rater (5% and 6%) was much smaller than the interaction of student by essay (14% and 18%), but for Study One the interaction of student by rater (17%) was larger than the interaction of student by essay (11%).

Decision studies were conducted to compute how many essays were needed to achieve a generalizability coefficient of 0.80, a minimum value for the purpose of making important decisions about individual student writing competency. For Study Two and Study Three with four raters, the number of essays needed was 3. However,

since the variance from the student by rater interaction was larger in Study One, using four raters, five essays were required to reach a generalizability coefficient of 0.80. The effects of different numbers of essays on the generalizability coefficient on the three studies with four raters are shown in Figure 1.

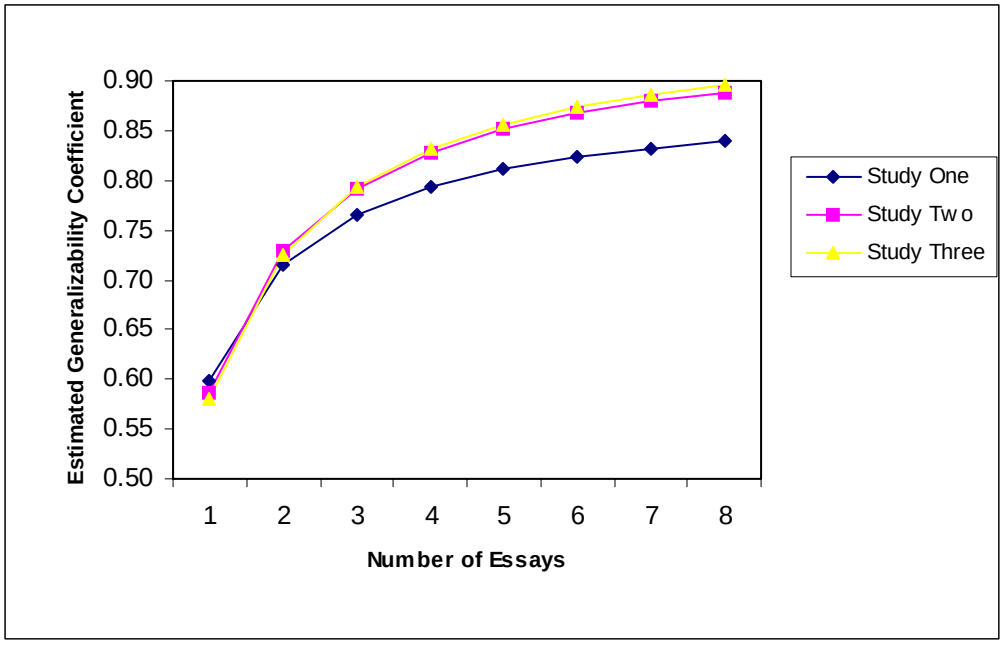

Figure 1

Effects of additional essays on the generalizability coefficient of students with four raters

In this study, due to classroom schedule, each participating student wrote only two essays. Our research revealed that two essays were not sufficient to make judgment of student writing ability. The students performed differently when responding to different writing tasks. Since the errors of measurement attributable to the sample of essays were larger than the errors attributable to raters, the best way to improve the reliability of scores is by increasing the number of essays. Therefore, to rely on performance assessment for high-stake testing, more essays are needed to provide a reliable evaluation of students. Based on our data analysis, to obtain a reliable measurement of student writing ability, students need to write three to five essays.

This study suggests that performance assignment prompts can be used in the classroom in addition to other measurement tools for teachers to gauge a student's strengths and weaknesses in writing, but individual essay exams should not be used alone for making important decisions about student placement. Administering performance assignments quarterly and examining scores from the three to five essays written by a student over the year can provide schools with reliable measurement of a student's writing ability, upon which high-stake decisions can be made.

The lower task generalizability found in this study could be due to the holistic rubric used for scoring. Past research indicated that the writing task variable had a different effect on different dimensions of student writing (Baker, et al., 1996; Gabrielson et al., 1995). Since writing performance involves a number of traits on which individuals differ, multi-trait analytic scoring strategies for writing performance assessment may increase task generalizability over a single holistic score (Crehan, 1997). In addition to using a holistic scoring rubric, lower task generalizability could also be due to the narrow scale employed (1-4) and range of ratings assigned by the raters (2-3) in the study. A longer scale coupled with instructions to raters regulating the use of the entire scale might yield more valid results in future studies.

Conclusion

As essay tests have become more acceptable in high stake student assessments, more research is called upon to examine the validity and reliability of the measurement derived from student essay exams. This research on writing task generalizability has shed light on how well writing prompts could measure students' general writing ability and how well student performance from one writing task could be generated to other similar writing task. Based on our findings, three to five essays are required to evaluate and make reliable judgment of student writing performance. Future studies should focus on how to measure different dimensions in student writing with an analytical rubric for scoring.

References

Baker, E., Abedi, J., Linn, R., & Niemi, D. (1996). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research*, 89(4), 197-205.

Boodoo, G., & Garlinghouse, P. (1983). Use of the essay examination to investigate the writing skills of undergraduate education majors. *Educational and Psychological Measurement*, 43(4), 1005-14.

Crehan, K. D. (1997). An investigation of the validity of locally developed performance measures in a school assessment program. *Educational and Psychological Measurement,* 61(5), 841-848.

Gabrielson, S., Gordon, & Engelhard (1995). The Effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education*, 8(4), 273-90.

Hoetker, J. (1982). Essay examination topics and students' writing. *College Composition and Communication*, 33, 377-392.

Lamb, H. (1987). Student performance across the domain of school writing. (ERIC Document Reproduction Service No. ED286194)

Resnick L. B., & Resnick, D. P. (1989). Assessing the thinking curriculum: New tools for educational reform. In B. B. Gifford & M. C. O'Conner (Eds.), *Future assessment: Changing views of aptitude, achievement and instruction* (pp. 37-75). Boston, MA: Kluwer.

Shavelson R. J., & Webb, N. M. (1991). *Generalizability theory.* London: Sage Publications.