Research Note

When Scores Do Not Increase: Notes on Quantitative Approaches to Writing Assessment

Lisa Rourke, *Brandeis University*

Xuchen Zhou, *The College of William & Mary*

Structured Abstract

- **Aim:** This research note focuses on administering a constructed response test before and immediately after students complete the required First-Year University Writing Seminar (UWS) to assess their ability to incorporate elements of the academic essay into their writing. The institution featured in this note is Brandeis University, a private research university outside of Boston with approximately 3,600 undergraduates. This study analyzes tests of 405 of the 852 incoming students before and after they took the UWS in the 2017-2018 academic year. By analyzing the data using means, paired *t*-tests, Pearson's tests, Analysis of Variance (ANOVA), and Tukey’s Honestly Significant Difference (HSD) test, researchers sought to identify areas of strength and areas for improvement to better guide writing instructors and curricular change. As we will show, the efficacy of such pretest-posttest designs is complicated by known issues in the design of writing assessment episodes such as ours.
- **Problem Formation:** In our study, we controlled for instructor and course variables and sought to determine if we could overcome other challenges relating to validity and reliability with a test administered pre and post UWS. We hypothesized that, by both limiting the scope of the assessment and testing each student, analytics would show that the test was valid with respect to our goals and sample size. Moreover, we assumed that assigning the same group of students to each grader for the pre and post tests would mitigate issues of

DOI: 10.37514/JWA-J.2019.3.1.12

unreliability: although scorers evaluated a different number of students, they evaluated the same students before and after the UWS. In other words, students were exactly matched to the same raters for both tests to minimize variations among scorers. Finally, we attempted to administer the test under similar conditions for each student to ensure fairness.

- **Information Collection:** In May of 2017, two months before starting Brandeis, incoming students took a 45-minute test in which they were asked to write a six to seven paragraph essay that responded to a constructed response task. They took another 45-minute test at the end of the semester in which they responded to a different task. A group of eight experienced writing instructors used a rubric to assess how well students incorporated five elements of the academic essay into their writing: this is the same rubric used by first-year writing instructors to evaluate students' work. Raters used a four-point Likert scale ranging from "strongly disagree" to "strongly agree" to evaluate these five elements, which included: thesis and focus, organization and structure, evidence and analysis, communication/style, and use of source materials. In addition, before evaluating individual elements, raters assigned a holistic score based on an overall impression of the essay. We then matched the post-assessment data with the pre-assessment data and conducted paired *t*-tests to determine if there was a statistically significant difference between the two (positive or negative). The paired *t*-tests revealed no improvement or a decline in scores, a result that we had not anticipated. In additional analysis, we used correlation testing to identify model strength and ANOVA and HSD to gather further evidence regarding inter-rater reliability.
- **Conclusions:** Contrary to our hypothesis, after students completed the UWS, their scores did not provide evidence of overall statistically significant scope gain. Methodological challenges included the following: inter-rater reliability, uneven student incentives on pretests and posttests, unknown differences in pretest and posttest tasks, challenges in rubric design, absence of attention to the cognitive processes by which students learn writing, and absence of robust construct representation. It is our hope that our study will both act as a cautionary tale and discourage institutions that may be contemplating a similar assessment design to avoid misguided inferences about student abilities and curriculum potential.
- **Directions for Further Research:** Although we no longer believe that pretest-posttest writing assessment designs such as ours will yield useful information based on the difficulty of designing and administering reliable and valid tests, we posit that there are opportunities in the field to use analytics to assess specific writing genres such as reflection letters. In the

2019-2020 academic year, we are doing a soft rollout of our new transfer-based curriculum and will use reflection letters from classes taught with the current curriculum as a control group against reflection letters from classes taught with the new curriculum. In particular, we will use lexical analysis to explore the ways in which reflection letters articulate how students might apply the skills learned in their first-year writing classes to a hypothetical research essay in an economics class. In this way, we hope to learn if the ideas of transferability are taking hold.

Keywords: assessment, constructed responses, timed tests, pretest-posttest design, writing analytics

1.0 Aim

Writing program assessment dates to over fifty years ago (White, Elliott, & Peckham, 2015), and re-accreditation requirements for institutions of higher education have increased the pressure to demonstrate effectiveness in achieving learning goals. While external stakeholders may prefer a quantitative over qualitative approach that can favor single pieces of writing, the Conference on College Composition and Communication position statement on writing assessment asserts that best assessment practices use multiple measures, as one piece of writing cannot serve as an indicator of overall writing ability (2009). Nevertheless, time and financial constraints can favor the evaluation of single writing samples, otherwise known as constructed response assessment.

In the present study, our goal was to identify areas of strength as well as areas for instructional and curricular improvement using two constructed responses from First-Year students. The first was a timed response administered before the start of our required first-year writing class; the second was a timed response administered after students completed this class. In particular, we compared the pretest and posttest to evaluate differences in how students incorporated into their writing the five elements of the academic essay on our grading rubric: thesis and focus, organization and structure, evidence and analysis, communication, and use of sources.

2.0 Problem Formation

Two assessment issues are important in studies such as ours that seek to provide evidence of score gain: the concept of a constructed response and constructed response measures.

2.1 Meanings of Constructed Response

Constructed response assessments are frequently criticized for their shortcomings. For example, inter-rater unreliability can be high (Baldwin, 2012; Breland, Camp, Jones, Morris, & Rock, 1987), and there is scant evidence that these responses produce notably different results than multiple choice tests (Hogan, 1981; Traub & MacRury, 1990). However, constructed responses have a variety of meanings and, therefore, should be evaluated individually. Baldwin (2012) defines constructed response as "any questions that require the test-taker to provide a response, as opposed to selected-response (most often multiple choice) assessments; these are often also called 'free-response' assessments" (p. 325). This definition spans different possibilities. For example, constructed responses among science test takers can vary from generating a missing word to generating a scientific investigation (Bennett, 1993).

Moreover, a number of subgroups, such as performance assessment, derive from the general genre of constructed responses. Baron (1991) defines performance assessment as a task where test-takers engage with a real-life issue compared with recalling memorized knowledge. So-called authentic assessment, a sub-group of performance assessment, asks test-takers to construct responses related to their professional discipline (Wiggins, 1989). And portfolio assessment is a selection over time of a particular class of constructed responses (Camp, 1993; Wolf, 1993).

The overall concept of a constructed response is important to our study. The concept of a constructed response is implicitly related to construct validity. The span of forms reveals a continuum of test types, from multiple-choice tests to portfolio assessment, that will yield various forms of construct representation, from the narrow to the robust. As such, task type is related to representation of the writing construct (White, Elliott, & Peckham, 2015).

2.2 Measures of Constructed Response Technical Quality

As Baldwin (2012) notes, the key to the technical quality of constructed response is the degree to which the assessment is reliable and valid. A test is reliable if it consistently measures what it was intended to measure (Cherry & Meyer, 1993). Similarly, a test is valid to "the extent to which the intended meanings and uses of test scores are supported" (Baldwin, 2012, p. 326). In other words, a test is valid to the extent that it provides the information that was intended from the assessment design. Ensuring that tests are reliable is just as crucial as ensuring that they are valid. As Cherry and Meyer explain, "A test cannot be valid unless it is reliable, but the opposite is not true; a test can be reliable but still not valid" (1993, p. 110).

Among the many forms of reliability, evidence of inter-rater reliability assures consistency between scorers. At a basic level—a level that is often used in studies such as ours—consistency can be demonstrated by using null hypothesis testing, where the null hypothesis is that variables have not been scored reliably by raters. To reject this hypothesis in favor of an alternative hypothesis (i.e., that variables have been scored reliably), the probability that scores vary among scorers cannot exceed 5%. In other words, scorers must attain a 95% confidence level that agreement has not occurred by change in the consistency of their scores (White, Peckham, & Elliott, 2015). After that confidence level has been established, correlation coefficients are used to estimate the level of reliability and to subsequently draw inferences about rater consistency.

In addition to reliability, Baldwin poses a number of other challenges to constructed response testing (p. 337), including the following questions:

- How do researchers maintain a maximum overall quality?

- Are standardized (e.g., comparable) tasks accompanied by consistent policies on the use of partial second scores, the choice of media to write one's response, the use of grammar- and/or spell-check programs, and other affordances that appear fair to all test-takers?
- Are the tasks scored reliably (including consistent policies and procedures around training and concerns with possible plagiarism or other problematic areas of test preparation)?
- How do researchers meet the technical requirements of fairness (i.e., no differential impact on identified subgroups, including both domestic and international English language learners)?

Baldwin's concern about ensuring fairness in test design is shared by many scholars. David Slomp (2016) argues that ethics should be considered more important than validity in the design of assessment programs. For example, he reminds us that the constructs upon which we base our assessments are socially constructed and potentially unstable across different contexts and highlights the importance of ensuring that all test-takers are provided with equal opportunity to demonstrate their learning. Social justice theory also explores the possibilities of writing assessment to identify opportunities and actionable outcomes within an educational context (Poe, Inoue, & Elliot, 2018).

The technical aspects of constructed responses are equally important to our study. Given the challenges of constructed response assessments, we sought to determine the feasibility of designing a test of this type that would provide varied forms of evidence related to validity, reliability, and fairness (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). Most importantly, we desired actionable results that would help improve our First Year Writing classes with respect to instruction and curriculum. Our design most closely aligns with Baron's characterization of performance assessment because we asked students to use critical thinking skills to evaluate the kinds of academic arguments that they would likely encounter in their college careers. We hypothesized that administering this type of timed constructed response before and immediately after taking First Year Writing would spotlight areas of strength and areas for improvement and speculated that this dual approach would skirt some of the challenges posed by administering a single timed response. As we demonstrate below, we were mistaken.

3.0 Information Collection

3.1 Writing Placement at Brandeis

All incoming students take a 45-minute test the May before starting Brandeis; this test is used to determine their placement in the required first-year writing class. Approximately 85% of the class is placed into UWS, and the other 15% is placed into Composition, which is for students who need more support. While in the past we have used these tests solely for placement purposes, in June of 2017 we evaluated student essays for both placement and ability to incorporate elements of the academic essay. Since our curriculum emphasizes critical thinking with respect to academic writing, the tasks asked questions about excerpts from scholarly articles. The pre-UWS task excerpted a passage from Paulo Freire's *Pedagogy of the Oppressed* about student/instructor relationships, and the post-UWS task offered an excerpt from Claude Levi Strauss's *Tristes Tropique*, arguing that written communication functions to facilitate slavery (see Appendix A for both tasks).

The tests were administered on Moodle, the University-wide platform used by students and faculty. Students were given a two-week window in May to take the test at a time and place of their choosing. Once the test was opened, students had 45 minutes to complete it. The clock ran continuously so that they were not able to pause the test. Students took the post-UWS test in class on the last day of the semester.

3.2 Methodology for Evaluating Constructed Responses

Of the 952 students invited to take the May test, 852 completed it. Of these, 26 received extra time (90 minutes). The study obtained IRB approval, and we analyzed results from the cohort who gave consent. After adjusting for students who did not give consent or attend class on the last day, we had a total of 405 pretests and posttests.

A group of eight experienced UWS instructors assessed the essays. Raters participated in a two-hour training session after which they were provided with randomly generated lists of students to evaluate. Raters first assigned a holistic score from 1–10 following the research of Nancy Robb Singer and Paul LeMahieu which "affirms the common wisdom that readers will more validly score holistically when they can first assess a piece 'as a whole,' that is, before it is scored for its analytic, component parts" (2011). Raters then rated each element of the academic essay on a Likert scale from 1–4, which ranged from "strongly disagree" to "strongly agree." Each rater graded a different number of students because faster raters would grade essays from the lists of slower raters. In particular, one rater graded significantly more essays than other raters, and this grader was disproportionately harsher. However, we attempted to mitigate the imbalance by having raters evaluate the same group of students in the pretest and posttest.

Raters used the following rubric to rate each essay:

1. Holistic score: assesses overall essay before evaluating component parts;
2. Thesis and focus: examines how well essays articulate an original, arguable, persuasive and non-trivial thesis and analyzes whether the thesis is focused and fully developed throughout the essay;
3. Organization and structure: analyzes how well the introduction frames the thesis to the problem or question being addressed. Assesses whether ideas are logically and seamlessly structured around the thesis;
4. Evidence and analysis: assesses the relevance of evidence to ensure that it is not superfluous to the argument. Assesses to what extent the essay effectively expresses how evidence relates to the thesis and clearly states why it's important. Looks for ideas and thoughts that are clear, interesting and original;

5. Communication/style: assesses how well grammar flows from one sentence to the next to determine the command of vocabulary and syntax. Assesses clarity and articulation of ideas; and
6. Use of source materials: assesses how well the essay incorporates evidence from the task.

Table 1 offers a descriptive layout of our results. For example, the pre and post means for thesis and focus (3.08 and 3.01) and organization and structure (3.47 and 3.43) were virtually unchanged.

Table 1

Descriptive Analysis

Variable	Pretest	Posttest

Holistic score	Mean=6.71
	N=405
	Std. Dev.=1.62
		Mean=6.82
		N=405
		Std. Dev.=2.17

Thesis and Focus	Mean=3.08
	N=405
	Std. Dev.=0.76
		Mean=3.01
		N=405
		Std. Dev.=0.83

Org. and Structure	Mean=3.47
	N=405
	Std. Dev.=0.91
		Mean=3.43
		N=405
		Std. Dev.=0.95

Analysis	Mean=2.90
	N=405
	Std. Dev.=0.79
		Mean=2.81
		N=405
		Std. Dev.=0.88

Use of Source	Mean=3.10
	N=405
	Std. Dev.=0.73
		Mean=2.98
		N=405
		Std. Dev.=0.82

Variable	Pretest	Posttest

Communication	Mean=3.11
	N=405
	Std. Dev.=0.72
		Mean=2.91
		N=405
		Std. Dev.=0.85

3.3 Results of Paired *T*-Tests and Pearson's Pre and Post Tests

After calculating the means, we analyzed the data using paired *t*-tests, shown in Table 2, to assess the difference in means between the pre- and post-scores.

Table 2

Paired T-test: Pre vs Post

Post-Pre	Mean of Diff.	Std. Dev.	Std. Error Mean	Lower	Upper	*t*	Sig. (2-tailed)
Holistic Score	0.11	1.82	0.09	-0.07	0.30	1.29	0.22
Thesis and Focus	-0.07	0.85	0.04	-0.16	0.01	-1.64	0.10
Org. and Structure	-0.04	0.68	0.04	-0.11	0.03	-1.13	0.26
Analysis	-0.08	0.90	0.05	-0.18	0.01	-1.82	0.07
Use of Source	-0.12	0.90	0.05	-0.21	-0.03	-2.62	0.01
Communication	-0.20	0.95	0.05	-0.30	-0.10	-4.11	0.00

The results of the t-test for the holistic score and elements of the academic essay scores are explained below.

3.3.1 Holistic score. Results of the paired *t*-test for overall test scores showed no statistically significant difference in the pretest and posttest mean test scores ($p = 0.22$). The overall test scores were equivalent in the pretest and posttest populations.

3.3.2 Thesis and focus score. Results showed no statistically significant difference in the pretest and posttest mean test scores ($p = 0.10$). The overall test scores were equivalent in the pretest and posttest populations.

3.3.3 Organization and structure score. Results showed no statistically significant difference in the pretest and posttest mean test scores ($p = 0.26$). The overall test scores were equivalent in the pretest and posttest populations.

3.3.4 Evidence and analysis score. Results showed no statistically significant difference in the pretest and posttest mean test scores ($p = 0.07$). The overall test scores were equivalent in the pretest and posttest populations.

3.3.5 Use of source material score. Results showed a statistically significant difference in the pretest and posttest mean test scores ($p < 0.01$). The use of source materials had declined from the pretest to posttest populations.

3.3.6 Communication score. Results showed a statistically significant difference in the pretest and posttest mean test scores ($p < 0.01$). The communication score had declined from the pretest to posttest populations.

3.4 Correlation Analysis

Although the paired *t*-tests unexpectedly revealed no improvement or a statistically significant decline in scores, we analyzed our data using Pearson’s product moment correlations, as shown in Table 3 and Table 4. Such analysis is commonly used to examine model strength by investigating statistically significant relationships between variables (Elliot, Rudniy, Deess, Klobucar, Collins, & Sava, 2016)

Table 3

Pearson's Pretest

Pearson's r	Holistic Score	Thesis and Focus	Org. and Structure	Analysis	Use of Source	Communication
Holistic Score	1	.72**	.55**	.71**	.67**	.65**
Thesis and Focus	.72**	1	.39**	.70**	.63**	.68**
Org. and Structure	.55**	.39**	1	.35**	.39**	.45**
Analysis	.71**	.70**	.35**	1	.72**	.66**
Use of Source	.67**	.63**	.39**	.72**	1	.59**
Communication	.65**	.68**	.45**	.66**	.59**	1

Note: ** Correlation is significant at the 0.01 level (2-tailed)

Table 4

Pearson's Posttest

Pearson's r	Holistic Score	Thesis and Focus	Org. and Structure	Analysis	Use of Source	Communication
Holistic Score	1	.80**	.56**	.80**	.66**	.73**
Thesis and Focus	.80**	1	.47**	.69**	.59**	.68**
Org. and Structure	.56**	.47**	1	.39**	.37**	.50**
Analysis	.80**	.69**	.39**	1	.71**	.66**
Use of Source	.66**	.59**	.37**	.71**	1	.48**
Communication	.73**	.68**	.50**	.66**	.48**	1

Note: ** Correlation is significant at the 0.01 level (2-tailed)

In both the pretest and posttest, each variable is correlated at a statistically significant level. Correlations range from low to high, with correlations somewhat higher in the posttest. It therefore appears that the construct model, as evidenced in the scores, was consistently strong in both the pretest and the posttest, with some indication that the model strengthened in the posttest. While scores did not increase, model strength remained strong, with some growth. As a constructed response assessment, the variables therefore were well integrated.

3. 5 ANOVA and Tukey's Honestly Significance Difference

We administered two additional tests to gather evidence on inter-rater reliability: Analysis of Variance (ANOVA) measures overall levels of consistency among scorers, while Tukey's Honestly Significance Difference (HSD) measures the consistency between individual raters. Table 5 shows the results of the ANOVA test on the post-UWS overall test scores. The *p*-value (denoted in column "Sig") of the tested factor, "Rater," is less than 0.05, which is rounded to 0.00. As such, there was a statistically significant difference between the raters or, to put it another way, a high degree of inter-rater unreliability. The limitation of ANOVA is that it only reveals that the raters graded unevenly without specifying which raters differed. Thus, we used the HSD, shown in Table 6, on the post-UWS overall scores to identify where the difference occurred between scorers.

Table 5

ANOVA for Inter-rater Reliability Post Overall Score

Source	Type III Sum of Sq[a].	df	Mean Square	F	Sig (*p*-value)
Corrected Model	299.59[b]	7	42.80	10.61	0.00
Intercept	13867.92	1	13867.92	3437.42	0.00
Rater's Name	299.59	7	42.80	10.61	0.00
Error	1601.65	397	4.03		
Total	19385	405			
Corrected Total	1901.24	404			

a. Type III Sum of Sq. is the default selection by SPSS. It assumes the presence of interactions. When the data is well-balanced, there should be no difference among Type I, II, and III Sum of Sq.

b. R Squared=.16 (Adjusted R Squared =.14)

Table 6

Interrater Reliability: Post-Hoc Test Using Tukey's HSD

Grader	N	Subset 1	Subset 2	Subset 3
Grader 1	126	5.40		
Grader 2	33	6.42	6.42	
Grader 3	52	6.62	6.62	
Grader 4	53	6.77	6.77	6.77
Grader 5	16		7.13	7.13
Grader 6	60		7.33	7.33
Grader 7	41		7.73	7.73
Grader 8	24			8.08
Sig. (*p*-value)		0.07	0.10	0.10

The average of overall post scores ranges from 5.40 to 8.08 for the eight scorers. The HSD test groups the scorers into three subsets. The scorers who belong in the same subset are considered to be grading using the same standard. Grader 1 to Grader 4 are subset 1, Grader 2 to Grader 7 are subset 2, and Grader 4 to Grader 8 are subset 3. The *p*-values of each subset are not statistically significant. Grader 1 and Grader 8 never appear in the same subset; hence, Grader 1 is statistically harder than Grader 8 because there is a 2.68-point difference between them. By contrast, the difference between Grader 2 and Grader 7, who are both in subset 2, is 1.31, which is not statistically significant.

In addition, each grader assessed a different number of tests. For example, Grader 1 assessed 125 essays whereas Grader 8 only assessed 24 essays. The combination of inter-rater unreliability and the disproportionate number of assigned essays contributed to the results of the paired *t*-test.

4.0 Conclusions

Perhaps unsurprisingly, most published pretest and posttest studies report measurable statistical gains as a result of an intervention (Pagano, Bernhardt, Reynolds, Williams, & McCurrie, 2008). Yet even when statistically significant score gains are rerecorded, qualifications are always in order. As Pagano and colleagues concluded after recording score gains on an inter-institutional assessment model, "Writing programs are complex, and we should not load too much weight on a score differential" (p. 300). In addition, results showing gains may be overstated or misinterpreted by researchers failing to account for alternative explanations besides the intervention, including design elements such as absence of a control group, maturation, and inattention to the phenomenon of regression to the mean (Marsden & Torgerson, 2012). Designs without control groups may not account for possible "contemporaneous effects of 'normal' educational experience or innovations in practice that may account for some or all of the observed changes" (Marsden & Torgerson, 2012, p. 584); thus, designs without control groups may overstate improvements that are attributable to other factors. In addition, increased maturity, particularly over longer intervals of time, often leads to improved learning outcomes, which may contribute to statistical gains. Finally, scores in the higher and lower ends of a study tend to have a higher error rate. When subjects are re-tested, scores in the lower and higher quartiles tend to rise less than on the initial test. This phenomenon is known as regression to the mean and can affect the data because scores in the lowest ranges may show gains that are not reflected upon re-testing.

In addition to methodological issues that can skew results, Carol Weiss (1998) points to ways that design choices for pretest and posttest studies can be used for political purposes. For example, she analyzes D. Breslau's 1962 study of the effects of training programs on the earnings of chronically unemployed workers to demonstrate how the choice of study design granted the federal staff administering the study influence over the program and policy. Such cases are not far removed from educational measurement research.

4.1 Test Design Issues

While Marsden and Torgerson (2012) and Weiss (1998) point to common issues with pretest and posttest designs, our results were affected by six design considerations, some of which were flagged by the statistical tests: scorers, incentives, different pretest and posttest tasks, rubric design, cognitive processes by which students learn to write, and construct representation.

4.1.1 Scorers. The Writing Program hired eight experienced scorers who had taught UWS to assess the 852 pretests in a one-week period. The time span was limited because tests were used for placement purposes as well as for the UWS pretest, and we needed to finish in time for the student lottery that placed students into their top choice seminars. This pressure to finish quickly, along with financial constraints, led to a number of other grader-related issues. First, scorers were not adequately trained and normed, which resulted in a high degree of inter-rater unreliability. In hindsight, scorers should have collectively reviewed essays of many different types, a process that likely would have taken a full day. Second, only one grader assessed each test when two would have given more reliable results. A third issue was that some scorers finished more quickly than others: when one grader completed the list, he or she would help others to finish. Finally, not all scorers were able to assess essays for the entire week. As a result, some scorers assessed many more essays than others. For example, one grader assessed 126 of the 405 essays compared to another who only assessed 24 (see Table 6). The grader who assessed the most tests was also the harshest grader, which further skewed our results.

4.1.2 Incentives. Students had a strong incentive to give their best effort on the pretest because this test determined their placement into either UWS or Composition. For the posttest, students who wrote the two highest scored essays in each class received a $10 Amazon gift card. However, this sum meant relatively little to students and failed to incentivize them in the same way as, for example, tying a small part of their grades to the posttest. Moreover, the posttest was administered in class on the last day of the semester, a time when motivation is extremely low.

4.1.3 Tasks. The pretest asked students to consider a passage from *Pedagogy of the Oppressed* by Paolo Freire, while the posttest offered a passage from Claude Levi Strauss's *Tristes Tropiques* (see Appendix A). Both passages required students to grapple with excerpts from scholarly texts in a similar manner to what they were asked to do in their writing seminars. However, scores may have declined if students struggled more with the posttest than the pretest. We could have circumvented this potential issue by dividing the cohort into two and having half take the pretest with the Freire and the other half take the pretest with the Levi-Strauss and then switching the tasks for the posttest. Structuring the assessment in this manner would have offered the opportunity to use statistical analysis to determine whether the task itself influenced outcomes.

4.1.4 Rubric design. Tables 4 and 5 display highly positive correlations between a number of rubric elements. In other words, the rubric elements were not mutually independent. As such, the relatively high correlation rates created redundancy. However, Tables 4 and 5 also illustrate the capability of the rubric to capture variables of the writing construct under examination. Hence, on the one hand, the difficulty of designing mutually unrelated elements points to inherent challenges of this evaluation method; on the other hand, correlations between variables demonstrate model strength. Contradictory inferences demonstrate the complexities involved in pretest and posttest designs of the kind recorded here.

4.1.5 Cognitive processes. Even if we had optimally administered the test, it is likely that our results would not have been much different because of the cognitive processes engaged in the learning of writing. Specifically, Ronald Kellogg's seminal article, "Training Writing Skills: A Cognitive Developmental Perspective" (2008), argues that demonstrated progress from learning new writing skills happens over a period of years, not months. As such, writing often gets worse before it gets better because of the limited capacity of our central executive function. On some levels, this phenomenon is intuitive: an athlete asked to adjust a technique typically shows a decline in performance until the technique has been mastered. In the same way, students process and incorporate their learning over years and not in a single semester. Thus, before- and after- assessment designs administered over the course of a semester—or even a year—will not measure student achievement because student learning is not linear.

4.1.6 Construct representation. Finally, it is doubtful that any timed test or series of tests will yield valuable insights into writing competencies because the form itself is too constraining to adequately represent writing ability (Condon, 2013). For example, timed tests do not resemble whole construct writing since they are shorter than essays written for class; our shortest essay is 1250–1500 words compared to the average 45-minute timed test essay of 300 words. Assessments that yield the greatest insight into writing ability have the fewest constraints. Condon offers a useful hierarchy of assessment design that reflects construct representation (2013). Assessments at the lowest level do not relate the context of the writing classroom to the test and yield little more than a score or ranking. Assessments in the middle area relate the test to some kind of writing context and start to assess what a writer can accomplish. Our pretest and posttest design reside in this area because the scores relate to the kinds of critical thinking and interpretation of scholarly texts that our courses require. Assessments at the highest level seek a fuller evaluation of student writing through vehicles such as portfolios and have the highest yield of the three areas in the hierarchy. As Condon's work shows, even the most elegant experimental design enacted to examine pretest and posttest comparisons will not compensate for a writing task that underrepresents the construct.

4.2 Final Thoughts

We hypothesize that pretest and posttest studies specifically with writing interventions are largely unreported because of unfavorable results such as ours. Nevertheless, our experience offers three larger lessons. First, it spotlights the importance of publishing such studies in the humanities so that other institutions can benefit. Although some institutions have published pre and post studies that showed statistical improvement in scores on untimed constructed responses over a semester-long composition course (Pagano et al., 2008), few have reported on pretest and posttest studies that documented statistically significant declines over the same timeframe. Second, this research note demonstrates how the application of analytics can pinpoint problem areas or validate results in a study. Finally, this research note articulates the many challenges to assessment designs comprised of constructed responses administered on a pretest and posttest basis. Most importantly, it suggests that designs of this type do not yield information that will be beneficial to administrators seeking to make programmatic change to support students. Worse, conclusions drawn from pretest and posttest designs may lead to incorrect inferences about curriculum potential and students' abilities.

Longitudinal studies, such as those using portfolios, offer assessment opportunities that account for development over time. However, some studies point in a different direction. For example, a two-year study conducted in Australia suggests that assessment be expanded to include domains such as curriculum knowledge and diverse ways of teaching and learning (Wyatt-Smith & Cummings, 2003). Thus, even "successful" assessments may not be comprehensive regarding the knowledge they seek to disclose.

5.0 Directions for Future Research

We identified six design considerations with our timed constructed response assessment. While we could theoretically remedy the first four in future iterations, we cannot change the process by which students learn and the ways that timed assessments underrepresent the writing construct. Based on our experiences and subsequent analysis, we question the ability of timed constructed responses to yield evidence that is valid, reliable and fair.

However, although our research spotlights limitations of timed constructed response assessments, there may still be advantages to untimed constructed response assessments using single genres of writing, such as reflection letters. Can valid, reliable, and fair assessment studies using these letters be designed that will yield productive and actionable results? How can we capitalize on analytics as a means to expand assessment opportunities for constructed responses? Pressures from accreditation agencies and other internal and external sources may make desirable an assessment of this type that can be conducted over a relatively short time period compared to a longer longitudinal study.

Research demonstrates that reflection letters are effective for both pedagogical and assessment purposes. In *Reflection in the Writing Classroom*, Kathleen Yancey argues, “reflection is a critical component of learning and of writing specifically; articulating what we have learned for ourselves is a key process in that learning-in both school learning and out-of-school learning” (1998, p. 7).

To explore approaches to this area of assessment, in future assessments we will measure self-efficacy and self-regulation (MacArthur & Graham, 2016) by using reflection letters as a means to evaluate the ability of students to articulate concepts of transfer (National Research Council, 2012). Specifically, in the 2019-2020 academic year, we will ask students to write reflection letters that discuss how they would apply what they have learned in their writing class to an essay from a different discipline (see Appendix B). Put differently, we will ask students how they will transfer knowledge from their first-year writing class to classes in their major and beyond.

In the 2019-2020 academic year we will have a soft rollout of a new curriculum that focuses on making transparent the transferability of skills learned in first-year writing to other disciplines and professional writing. As such, approximately half the classes will teach the current curriculum while the other half will teach the new transfer-based curriculum. The classes teaching the current curriculum will act as a control group for the transfer-based curriculum. As Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson explain in “Estimating Causal Effects Using Experimental and Observational Designs” (2005):

> The effect of a cause must always be evaluated relative to another cause. In a controlled experiment, for example, the outcomes for a given treatment or intervention (one cause) are always defined relative to an alternative treatment or control condition (a second cause). Thus, in evaluating whether an innovative mathematics program is effective in increasing mathematics achievement, the outcomes of the program must be compared with the outcomes from some

> existing program. The question is not simply whether a program is effective but whether it is more effective than some other program. (p. 17)

We hypothesize that these reflection letters will allow students to recall their thought process when taking classes in other disciplines, thereby allowing them to capitalize on skills learned in their writing classes. Importantly, the nature of this assessment design allows us to skirt a number of common issues associated with case and control group designs. For example, because students will not know which classes teach the current versus new curriculum, the study will be perfectly randomized. In addition, we will avoid issues of attrition or non-compliance since first-year writing is a required class that students are not permitted to drop.

To assess the ability of students to articulate ideas related to transferability, we will perform lexical analysis to identify the difference in frequency with which transfer-related keywords and phrases, such as *genre*, *discipline/disciplinary* and *citation style*, appear in student responses in conjunction with statistical methods to determine the significance of any changes. From this data we hope to learn the facility with which students can articulate what they have learned and apply it in order to spotlight areas for curricular improvement and illuminate transfer-related differences between the current and new curriculums. At the same time, we do not want to sacrifice instruction on the elements of the academic essay, so we will similarly measure the frequency with which elements-related keywords and phrases such as *thesis*, *structure*, *evidence* and *analysis* appear in student responses.

In the spring of 2019, we conducted a beta test of this assessment using the methods described above with three experienced instructors: one instructor had 54 students and taught the transfer-based curriculum; the other two instructors had 54 students combined and taught the old curriculum. We thus evaluated a total of 108 reflection letters; letters from the old curriculum served as the control group. The results were encouraging. In the transfer-based curriculum, twice as many students used transfer-related words and phrases in comparison with the old curriculum letters. Moreover, the number of students using words and phrases associated with elements of the academic essay was nearly identical in both groups.

These preliminary findings suggest that students are achieving an awareness of transfer not found in the old curriculum that does not come at the expense of foundational writing concepts. As well, these findings suggest that which may occur when scoring studies are abandoned in favor of richer designs using evaluative processes.

Author Biographies

Lisa Rourke is the Director of First-Year Writing at Brandeis University.

Xuchen Zhou is the Assistant Director of Institutional Research at The College of William & Mary.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baldwin, D. (2012). Fundamental challenges in developing and scoring constructed-response assessments. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 325–339). New York, NY: Hampton Press.

Baron, J. B. (1991). Performance assessment: Blurring the edges of assessment, curriculum and instruction. In G. Kulm & S. M. Malcolm (Eds.), *Science assessment in the service of reform* (pp. 247–266). Washington, D.C.: American Association for the Advancement of Science.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Hillsdale, NJ: Erlbaum.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Erlbaum.

Breland, H. M., Camp, R., & Jones, R. J. (1987). *Assessing writing skill*. New York, NY: College Entrance Examination Board.

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

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 183–212). Hillsdale, NJ: Erlbaum.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M. W. Williamson & B. Huout (Eds.), *Validating holistic scoring in writing assessment: Theoretical and empirical foundations* (pp. 109–141). Cresskill, NJ: Hampton Press.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*, 100–108.

Conference on College Composition and Communication. (2009). Writing assessment: A position statement. Retrieved from https://cccc.ncte.org/cccc/resources/positions/writingassessment

Elliot, N., Rudniy, A., Deess, P., Klobucar, A., Collins, R., & Sava, S. (2016). ePortfolios: Foundational issues in measurement. *Journal of Writing Assessment*, *9*(2).

Hogan, T. P. (1981). *Relationship between free-response and choice-type tests of achievement: A review of the literature.* Washington, DC: National Institute of Education.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1–26.

MacArthur, C. A., & Graham S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham, & J. Fitzgerald, (Eds.), *Handbook of Writing Research*, 2nd ed. (pp. 24–40). New York, NY: Guilford Press.

Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and posttest research designs: Some methodological concerns. *Oxford Review of Education*, *38*(5), 583–616.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.

Pagano, N., Bernhardt, S. A., Reynolds, D., Williams, M. & McCurrie. M. K. (2008). An inter-institutional model for college writing assessment. *College Composition and Communication*, *60*(2), 285–320.

Poe, M., Inoue, A., & Elliot, N. (Eds.). (2018). *Writing assessment, social justice, and the advancement of opportunity*. Fort Collins, CO: University Press of Colorado.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, J. (2005). *Estimating causal effects using experimental and observational designs: A think tank white paper*. Washington, DC: American Educational Research Association.

Singer, N. R., & LeMahieu, P. (2011). The effect of scoring order on the independence of holistic and analytic scores. *The Journal of Writing Assessment*, *4*(1).

Slomp, D. (2016). An integrated design and appraisal framework for ethical writing assessment. *Journal of Writing Assessment*, *9*(1).

Traub, R. E., & MacRury, K. A. (1990). *Multiple-choice vs. free-response in the testing of scholastic achievement* (K. Ingekamp & R. S. Jager, Eds.). Toronto, Ontario: Ontario Institute for Studies in Education.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation*, *19*(1), 21–33.

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Logan, UT: Utah State University Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *The Phi Delta Kappan*, *70*(9), 703–713.

Wolf, D. P. (1993). Assessment as an episode of learning. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 213–241). Hillsdale, NJ: Erlbaum.

Wyatt-Smith, C. M., & Cummings, J. J. (2003). Curriculum literacies: Expanding domains of assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(1), 47–59.

Yancey, K. (1998). *Reflection in the writing classroom*. Logan, UT: Utah State University Press.

Appendix A

Administered Pretest and Posttest Constructed Response Tasks in the UWS

Pre-UWS Administered in May of 2017 Before the Start of the Semester

In his book *Pedagogy of the Oppressed* (1970), Paulo Freire, a Brazilian educator and philosopher, reflects on different approaches to teaching. In particular, he critiques a particular model of teacher-student relationship:

A careful analysis of the teacher-student relationship at any level inside or outside the school, reveals its fundamentally narrative character. This relationship involves a narrating Subject (the teacher) and patient, listening objects (the students). The contents, whether values or empirical dimensions of reality, tend in the process of being narrated to become lifeless and petrified. Education is suffering from narration sickness…Narration (with the teacher as narrator) leads the students to memorize mechanically the narrated content. Worse yet, it turns them into "containers," into "receptacles" to be "filled" by the teacher. The more completely she fills the receptacles, the better a teacher she is. The more meekly the receptacles permit themselves to be filled, the better students they are.[1]

Drawing on your own experience with or ideas about teacher-student relationships within your own life-setting, write a 6-7 paragraph essay that either agrees with, disagrees with, or nuances Freire's view. In addition to restating in your own words what Freire argues, use his quotes as well as evidence from your own experience to make your case. Remember that a good argument is one that might have a reasonable chance of persuading those whose opinions might differ from your own.

[1] Paulo Freire, *Pedagogy of the Oppressed,* New York: Continuum Books, 1993.

Post-UWS End of Semester Administered on the Last Day of Class in the Fall and Spring Semesters (December 2017 and May 2018)

In his 1955 memoir, *Tristes Tropiques* (*Sad Tropics*), anthropologist Claude Levi-Strauss reflects on his travels primarily through Brazil, although the memoir references many other countries as well. While *Tristes Tropiques* is a travelogue of sorts, it also discusses what Levi-Strauss considers to be the oppressive role of writing in societies. In particular, Levi-Strauss argues that "the primary function of written communication is to facilitate slavery":

The only phenomenon with which writing has always been concomitant is the creation of cities and empires, that is the integration of large numbers of individuals into a political system and their grading into castes or classes. Such, at any rate, is the typical pattern of development to be observed from Egypt to China, at the time when writing first emerged: it seems to have favoured the exploitation of human beings rather than their enlightenment. . . My hypothesis, if correct, would oblige us to recognize the fact that the primary function of written communication is to facilitate slavery. The use of writing for disinterested purposes, and as a source of intellectual and aesthetic pleasure, is a secondary result, and more often than not it may even be turned into a means of strengthening, justifying or concealing the other.[1]

Drawing on your own experience with or ideas about writing and education within your own life-setting, write a 6-7 paragraph essay that either agrees with, disagrees with, or nuances Levi-Strauss's view on writing. Use quotes from Levi-Strauss as well as evidence from your own experience as you make your case. Remember that a good argument is one that might have a reasonable chance of persuading those whose opinions might differ from your own.

[1] Claude Levi-Strauss, *Tristes Tropique*, tran. J. and D. Weightman (New York: Atheneum, 1975), 299.

Appendix B

Reflection Letter

Imagine that you have been given the following and asked to write a 10-15 page research paper for an economics class:

“The current federal minimum wage is $7.25 with some variation among the states. Make an argument about the effects of increasing the minimum wage.”

In **1.5 double spaced pages**, write one full paragraph addressing each question.

1. What are the steps you need to take to write a successful paper?
2. Given that you may never have taken a class in this discipline, what discipline-specific knowledge would you need to learn?
3. In what ways will the skills you learned in your UWS help you to write this paper and answer the first two questions?