An Assessment of Writing Outcomes in the First Semester of College at the University of Wisconsin-Madison: A Pilot Study

James A. Wollack
Research Director
Director, Testing & Evaluation Services
Associate Professor, Educational Psychology

Morris Young
Professor, English

Elaine Klein
Chair, University General Education Committee
Assistant Dean, College of Letters & Sciences

Nancy Westphal-Johnson
Associate Dean, College of Letters & Sciences

December, 2011
Introduction

As of May 1996, students entering the UW-Madison are required to satisfy two communications literacy requirements as part of their General Education requirements (referred to as Communication-A and Communication-B). The introductory Communication-A requirement, which students are expected to satisfy by the end of their first year, is offered in a handful of departments across campus. Students may satisfy the Communication-A requirement by passing a Communication-A course or equivalent, or by receiving sufficiently high scores on the University of Wisconsin English Placement Test, Advanced Placement (AP) exam, or International Baccalaureate (IB) exam. Though the nature of the Communication-A courses differs somewhat depending on the department in which they are housed, all are taught around the same three educational objectives: written communication, oral communication, and information literacy.

In 2007, a previous study of the Communication-A requirement and its effects was submitted to the University of Wisconsin General Education Committee (Halaby, Westphal-Johnson, Wollack, & Klein, 2007). This study was based on survey data collected from students about their Communication-A experiences. That study concluded that students satisfying Communication-A through course work completed at UW-Madison exhibited a benefit compared to students satisfying the requirement though the University of Wisconsin English Placement Test, Advanced Placement, or International Baccalaureate or students who had yet to complete Communication-A. In particular, the 2007 study found that first-year students who
took a Communication-A certified course during fall semester 2006-07 reported greater improvement in their communication skills as a result of their educational experiences than did comparable freshmen who did not satisfy the requirement. This was true across the five Communication-A courses, including English 118, the Communication-A course taken by non-native speakers of English. The study concluded that decentralized mode of implementing the Communication-A requirement at UW-Madison is a highly effective means of achieving general education proficiency objectives with respect to writing, speaking, and information-seeking skills. Although these results are in keeping with expectations and are very encouraging, surveys are indirect measures of student learning, and it remains to be demonstrated that the benefits of Communication-A courses manifest themselves in demonstrated improvements in student writing.

In 2009-2010, the Communication A Course Directors Subcommittee identified two projects to pursue as part of its efforts to continue to review and assess the efficacy of the Communication-A requirement. First, the subcommittee asked that assessment funds be requested in order hire a project assistant to collect data from peer institutions about their general education communication/written communication requirements. These data were used to identify similarities and differences between UW-Madison and its peers in how students are placed, assessed, and fulfill the general education requirement (See Benbow, 2010). The subcommittee also requested 2010-2011 assessment funds to develop a pilot study that would be a direct assessment of student writing. This is a report of that pilot study.

The current pilot study has two goals: The first goal is to begin the process of evaluating the writing literacy component of the Communication-A requirement at UW-Madison. The second goal of the study is to allow the campus to evaluate whether the proposed research design
is an appropriate way to assess the Communication-A course. If so, this study will be followed by a larger study of similar design. An additional but related question is also raised about the efficacy of the University of Wisconsin English Placement Test as an appropriate placement or exemption tool. The approach used to address these goals involves studying student writing, not in Communications-A courses, but in a different freshman-level course that enrolls a representative sample of all incoming freshmen, thereby allowing the inclusion of a control group of students needing, but not currently taking Communications-A, as well as an authentic comparison with those students who exempted from the requirement.

Methods

Participants
Counseling Psychology 125 (CP125), *A Wisconsin Experience Seminar*, is a one-credit, freshman-level course taught every fall. The course aims to help students transition to the university by providing them with an overview of the campus, orienting them towards campus resources, and familiarizing them with the expectations of being a student at the University of Wisconsin. In addition, CP125 helps to better prepare students for the rigors of college by asking students to complete two projects, a faculty interview conducted towards the beginning of the semester, and a Significant Learning Project (SLP) at the end of the semester in which students reflect on the knowledge they gained throughout the course. Both projects include a writing component. To accompany the interview, students submit a 2-page paper reflecting on the experience. Students are told that the rationale behind the assignment is to help them see professors as people and to introduce students to the importance of meeting with faculty outside the classroom. For the SLP, students are asked to be creative in developing a project that they
think demonstrates what they have learned about the Wisconsin Experience. Examples of projects include board games, poems, scrapbooks, puzzles, collages, and presentations that represent the knowledge-making process students have experienced in investigating the University of Wisconsin-Madison. The students are asked to write three pages to accompany the project, in which they were to explain the rationale behind the project and how it relates to the Wisconsin Experience.

CP125 is appropriate for all incoming students, and the population of students taking it is generally quite representative of the entire freshman class, hence CP125 offers a reasonable site to undertake a pilot study. As a result, CP125 includes a sufficient number of students who are concurrently enrolled in a Communications-A course (i.e., the target population for this study), but also contains a sample of students still needing to satisfy the requirement but who are not currently taking it, as well as a sample of students who have been exempted from the Communications-A requirement. To illustrate this point, in Fall 2009, of 170 enrolled students, approximately 40 were exempt from Communication-A based upon placement scores, 60 were enrolled in a Communication-A course, and the remainder had either satisfied the requirement through AP or had yet to enroll in Communication-A. Although these sample sizes are not large, they are sufficient to begin to study these issues, especially if this is seen as a pilot that helps us understand if this model might be effective if used on a larger scale.

Participants for this study were recruited during the first two weeks of the Fall, 2010 semester from all ten sections of CP125. These ten sections were all capped at 20 students and enrolled only freshmen, so involved the recruitment of a maximum of 200 students. To participate, students were asked to release to the research team their faculty interview and SLP papers, along with some demographic information and course history data. The consent
documents informed students that their papers would be graded by their instructors for classroom purposes, but, after the semester was over, would be re-evaluated by a team of graders using criteria developed by campus for evaluating freshman-level writing. They were not specifically informed that the study was an evaluation of the Communications-A requirement, nor that their writing would be evaluated with respect to Communications-A writing criteria.

We received consent from 98 students; however, electronic copies of both papers were available for only 81 students. Therefore, the results and analyses in this study were based on the 81 students for whom we had complete data. Table 1 shows the demographic breakdown for the students participating in this study. As one can see, nearly two-thirds of the sample was female. Also, the vast majority of participants were 18 years old, and described themselves as White. The sample is also nearly evenly split between those having satisfied Communications-A on the basis of previous test scores (placement test, AP, or IB) and those who did not exempt from the Communications-A requirement.

Developing a Scoring Rubric

After final course grades were submitted by the instructors, a subcommittee of the Communications-A working group gathered to develop a scoring rubric that was reflective of the essential learning outcomes (ELOs) for Communications-A writing instruction. The goal of this group was to develop a rubric that could be applied very broadly by instructors across the university to evaluate student writing in their courses along criteria that are consistent with the Communications-A ELOs. However, in the interest of making the rubric maximally useful for
the present purpose, it was understood that the rubric may need to be modified somewhat to more directly apply to the two types of writing which were collected for this study.

The subcommittee began by collecting samples of writing rubrics which had been used for similar purposes at other universities, and ultimately decided to design a rubric based largely on one from the University of Wisconsin-Oshkosh. The agreed upon rubric called for student papers to be evaluated on a 4-point scale along five separate dimensions: (a) ideas, content, and critical thinking (Content), (b) organization (at both the global and local levels), (c) voice, audience, and word choice (Voice), (d) sentence fluency, language and genre conventions (Conventions), and (e) use of evidence (Evidence). Scores of 4 were intended to reflect writing that, with respect to the dimension being evaluated, required little revision, exhibited consistent strengths and was successful in conveying a clear point. Scores of 1 reflected writing that required much revision, was consistently weak and did not convey a clear point sufficiently well. Experience with rubrics suggests that scores of 4 and 1 should be both easy to recognize and classify accurately, and relatively uncommon when compared with the numbers of students earning scores of 2 or 3. Therefore, the subcommittee focused much of its attention on distinguishing those two scores. The subcommittee decided that, on a general level, the assignment of a 2 or a 3 should be made based on a weighing of the relative strengths and weaknesses of the paper (for that particular dimension), such that scores of 3 reflected writing in clear need of some revision where the strengths outweighed the negatives and scores of 2 were reserved for those situations in which the negatives outweighed the strengths. By its very nature, writing assessment is a subjective process; however, in the interest of promoting rater uniformity, for each dimension, the subcommittee developed a detailed description of the
characteristics they would expect to see in a paper receiving each of the four score points. The entire scoring rubric, along with these descriptions, is provided in Table 2.

In addition to developing detailed scoring criteria for each of the five dimensions, the subcommittee also decided that raters should provide an overall holistic score (holistic) on the same 4-point scale (the scale anchors for that overall score are the descriptions provided in the top row of the rubric, immediately below the scores). This holistic grade is intended to capture the rater’s evaluation of the paper when it is considered in its entirety. During training, raters were specifically directed that their holistic score should consider the five dimensions (possibly as well as others), but should not merely average across them, nor should it involve any other formulaic method for weighing the various dimensions.

Prior to the training, the subcommittee graded the papers from all consenting students for whom only one of the two papers was available electronically. The purpose of this exercise was both to gain some experience using the rubric that could be used to make revisions and refinements, as necessary, and to identify a subset of papers that could be used to help train raters on the rubric.

From this exercise, five essays were identified as anchors: two as anchor 2’s, two as anchor 3’s, and one as an anchor 4. The model 2 and model 3 essays both included one faculty interview and one SLP paper. The model 4 was a SLP paper. Although these papers were identified as models, it is important to note that the subcommittee members did not uniformly assign these essays with their anchored score. In the case of the model 2 and model 3 papers, they were selected because there was widespread support for one score, and all disagreements
were no more than one point from the anchored score. In the case of the model 4, it was selected because its average score across subcommittee members was highest among the candidate essays, and half the members deemed it appropriate for a 4. However, there was not uniform agreement that it represented a 4 (let alone a model 4); in fact, one member even rated the essay as a 2. Still, given the limited number of essays from which to choose the anchors, there was general consensus that the chosen paper was of better quality than the others.

**Raters**

Raters were recruited from among the past and present Communications-A instructors in the departments of English (including English as a Second Language), Communication Arts, Life Science Communication, and Engineering Professional Development, the four departments offering courses that satisfy the requirement. Raters were paid at a rate of $17 per hour for two full days of rating.

A total of 17 raters were included. The raters were selected so as to include instructors from several different Communication-A courses. The raters included seven instructors of English 100, seven instructors of Communication Arts 100, two instructors of Life Science Communications 100, and one instructor from English 118 (which is an English as a Second Language course). No instructors from Engineering Professional Development agreed to serve as raters.

**Rating Sessions**

The rating sessions were held from 8:00 – 5:00 on May 23-24, 2011. Raters had been told in advance that it was expected that they attend for the whole day on both days.

The first half of the first day was set aside to orient raters to the study and to train them on using the rubric. Raters were told that the purpose of the study was to evaluate freshman-
level writing, so as to have a better understanding of our students’ skills upon entry and more realistic expectations about how to best help them develop their skills while at the University of Wisconsin. Raters were not specifically told at the outset that this study was aimed at evaluating the Communications-A learning outcomes for writing.

Raters were further informed about the types of writing they would be asked to grade. Because we did not wish to bias raters into scoring end-of-semester assignments higher than beginning-of-semester assignments, we did not tell them that the writing samples were all collected from different sections of a common course, nor did we inform them that they were collected at different times during the semester. Instead, raters were told that we were able to get instructors from a number of different freshman-level classes to work with us to develop some writing samples that allowed them to measure their own course objectives, but also could be used to assess the quality of freshman writing. We informed raters that, as a consequence of working with different classes, they will be looking at a couple of somewhat different genres of writing, but that the goal is to apply the same criteria to all papers, to the extent possible. The research team proceeded to explain the different genres of writing they would encounter from the classes that were selected.

Raters were then divided into three groups, with each group assigned a table leader. The table leaders, two of whom were from the English Department and one from Engineering Professional Development, were all members of the Communications-A subcommittee that developed the scoring rubric. Table leaders distributed the rubric to their raters and led a discussion about the rubric itself, the distinctions between the scores, the definitions of the five dimensions, and the process by which raters should produce a holistic overall grade for each paper. Table leaders emphasized that papers need not be perfect to merit a score of 4, nor do
they have to be horrendous to earn a score of 1. Raters were reminded that the papers they were reading were all written by new freshmen, and that grading criteria should be applied with respect to what can reasonably be expected of that population.

Following the discussion of the rubric, raters were given the five anchor papers in random order. They were given time to read all papers and were asked to use the rubric to assign ratings to each. Because this was their initial experience reading the types of assignments they would be asked to evaluate, and because it was their first opportunity to apply the rubric, they were encouraged to return to earlier papers to modify their grades, as necessary. After all raters in a group had finished grading the papers, the table leader led a group discussion about each of the papers individually, first informing the group of the anchor score for each paper, and then facilitating a discussion about why each rater evaluated the paper in the way they did. Following that discussion, the table leader informed the group about the rationale behind the subcommittee’s anchor rating, and offered advice to group members on how to modify their evaluation criteria to better conform to the rubric.

After the initial training, table leaders were dismissed and raters were ready to begin rating live essays. The remaining day and a half was divided into 6 rating sessions, each separated by a break of at least 20 minutes. The length of each rating session was either an hour and 45 minutes (long) or 45 minutes (short), depending on where it fit into the schedule. In addition, recalibration sessions were scheduled for the first hour of the second rating day, as well as at the start of the Day 2 afternoon session, if it was deemed necessary. At the end of both days, a short wrap-up was planned. The agenda for the two-day rating is shown in Table 3.
All essays were rated by five different raters, and the papers were randomly divided across raters so that raters were to evaluate between 46 and 49 papers each\(^1\). Rating sessions were developed so that 162 papers to be evaluated were divided into six groups, one corresponding to each of the six scheduled rating sessions. During long sessions, either 32 or 33 papers were rated, whereas during short sessions, 16 papers were rated. For each session, each rater was assigned a packet of materials to read and grade individually (usually 9-11 essays for long sessions, 4-6 for short sessions), using the rubric and the rating sheets provided. Although each rater’s packet was different, all raters assigned to one of the selected papers would receive that paper to grade during the same rating session. In this way, each rating session provided complete data for the papers evaluated, thereby making it easier to monitor rater conformity.

Raters were encouraged to take breaks as necessary to keep themselves fresh, but to be considerate of others and refrain from talking except during scheduled breaks. Once raters completed their packets, they were asked to submit their rating sheets to one of the members of the research team, who was monitoring the ratings throughout the two days and evaluating whether raters used the rubric as intended.

**Quality Control**

During the break between each rating session, a member of the research team would enter the all the ratings from the previous session into a spreadsheet that dynamically tracked ratings and monitored them for conformity. Three different conformity measures were used to identify examinees who were consistently rating differently from their peers who evaluated the same essays. All conformity measures were computed separately for each of the five dimensions, the

\(^1\) Although all raters had agreed to participate on both days, one rater got sick and was unable to participate in the second day of rating. Consequently, the papers that had been assigned to her were re-distributed to other raters so the actual number of papers evaluated by the other 16 raters was between 47 and 51.
holistic score, and a composite summed score across the five dimensions. These three methods are described below:

\textit{Z-Test}

This measure was intended to flag raters who consistently rated essays as either higher or lower than their colleagues who were rating the same essays. The Z-test index first standardizes the five ratings for each essay, as follows:

\[ z_{ij} = \frac{r_{ij} - \bar{r}_i}{s_i} \]

where, \( z_{ij} \) is the standardized rating for person \( j \) on essay \( i \), \( r_{ij} \) is the rating given by person \( j \) to essay \( i \), \( \bar{r}_i \) is the average rating for essay \( i \) across all \( j \) raters to evaluate it, and \( s_i \) is the standard deviation of ratings for essay \( i \) across the raters who evaluated it.

The index \( Z_j \) is given as

\[ Z_j = \frac{\bar{z}_j - \bar{z}_.}{s_{pooled}/\sqrt{n_j}} \]

and expresses the magnitude of the average \( z_{ij} \) value for person \( j \) across all items rated by that individual, \( \bar{z}_j \), in standard deviation units, where \( \bar{z}_j = \frac{1}{n_j} \sum_{i} z_{ij} \), \( \bar{z}_. = \frac{1}{\sum_{j} n_j} \sum_{j} \sum_{i} z_{ij} \), \( n_j \equiv 0 \), \( s_{pooled} = \sqrt{\frac{\sum_{i} n_i s_i^2}{n}} \) is the pooled variance estimator\(^2\) across all \( n \) essays (note that in this study, \( n = 162 \)), and \( n_j \) is the total number of essays rated by rater \( j \).

High values of \( Z_j \) indicate that the rater was consistently too easy (i.e., his/her ratings were higher than those of the other raters evaluating those same items), whereas low values of \( Z_j \) indicate that the rater was consistently too hard. In theory, indexes with this construction should be normally distributed. We adopted a two-tailed \( \alpha \)-level of .05, meaning that raters who are not extreme

\(^2\) Because all papers were evaluated by 5 raters, the general pooled variance formula which weighs each variance by its degrees of freedom reduces to the simple average across all variances.
should only expect to be flagged 5% of the time. Consequently, critical values on $Z_j$ were taken to be ±1.96.

**Sign Test**

This measure parallels the Z-Test and was also intended to flag raters who were consistently more extreme in their ratings than their colleagues. The sign test relaxes the distributional assumptions of the Z-test to allow testing the same hypothesis under a wider array of conditions. The sign test computes a test statistic $S_{+j}$, which equals the sum of two counts, the number of $z_{ij}$ which are greater than 0 for person $j$ and half the number of $z_{ij}$ which are exactly equal to 0 for person $j$. This number can be evaluated for statistical significance by referencing a binomial distribution with $n_j$ trials and a probability of success equal to 0.5. Again, we adopted a two-tailed $\alpha$-level of .05.

**Variance Test**

The final measure was intended to evaluate whether raters were using the extremes of the scale in the same way and whether the amount of variability in their ratings was in keeping with expectation. One of the issues with any rating scale is that individuals vary with respect to their tendency to assign extreme ratings. Another issue is that raters may fixate on a particular category and assign a disproportionate number of papers with a single score. The variance test examined the within-person variance relative to the pooled variance across all raters, hence was sensitive to whether use of the rating scale was consistent across raters. The test is as follows:

$$ V_j = \frac{(n_j - 1)s_j^2}{s_{pooled}^2} $$
where $s_j^2$ is the variance of actual ratings across all essays for person $j$, and $s_{pooled}^2 = \sum_{j=1}^{N_j} (n_j - 1)s_j^2 / \sum_{j=1}^{N_j} (n_j - 1)$, where $N_j$ is the number of raters. $V_j$ is evaluated for statistical significance by referencing a $\chi^2$ distribution with $(n_j - 1)$ degrees of freedom, and $\alpha = .05$.

**Monitoring Ratings**

Prior to the start of each new session on Day 2, the three quality control indexes were computed using data not only from the most recent session, but from all previous sessions. Individuals whose holistic score was flagged by one of the three indexes were approached by a member of the research team and informed about the nature of the flag. At this time, the research team also communicated with the rater about any tests on the individual dimensions or on the Composite measure that were also identified as statistically significant, so the rater gained a better appreciation for which characteristics might have been contributing most to their aberrance. Raters who were not flagged on the holistic index were not approached about potential flags on the Composite or on one of the dimensions.

**Rater Re-Calibration**

Following the two rating sessions on Day 1, the research team had identified a small set of papers that the selected raters had consistently graded at each of the four score points. In addition, the research team identified some papers for which the raters were most in disagreement. At the beginning of Day 2, these two sets of papers were presented to the raters as part of an exercise aimed at re-familiarizing themselves with the rubric. Raters evaluated each of the papers and engaged in a large group discussion in which the previous day’s rating of the item was shared with the group and individual raters were given the opportunity to explain their ratings.

---

3 Indexes were not computed prior to Session B on Day 1 because the small number of papers read by each rater during Session A would have made these indexes too unreliable.
Research Questions

Because this study took place within a course taken by a broad cross-section of incoming freshmen and that the data collection involved the gathering of two student papers, one from early in the semester and the other late in the semester, it was possible to investigate a number of interesting research questions. The five research questions for this study are as follows:

1. Does the writing quality of students who are currently taking a Communications-A course improve across the semester by more than would be expected of other freshmen not yet having fulfilled the requirement, but having experienced the University of Wisconsin for a semester?

2. After having taken a Communication-A course, is student writing of equal quality to that of students who exempted from the requirement?

3. Prior to completing a Communication-A course, is the writing quality of students who exempted from the Communication-A requirement better than the writing quality of those who did not exempt from the requirement?

4. Is there a difference in writing quality between students who exempted from the Communications-A requirement on the basis of the placement test versus exempting based on the AP or IB tests?

5. Is there a difference in writing quality between students who elect to enroll in a Communications-A course in the fall of their freshman year and those who enroll at a later date?

Statistical Methods

The research questions were answered using Analysis of Variance (ANOVA) techniques (including t-tests, the special case of ANOVA for which there are only two groups) appropriate
to the research question, as described below. For ease of describing the variables involved in the statistical tests, Table 4 presents a coding scheme for discussing the between-subjects variable (exemption/Communications-A status), the within-subjects variable (writing assessments), and the individual assessments for a specific subgroup. In all cases, primary interest is on the holistic score, but results will also be presented for any of the other dependent measures that achieved statistical significance. Results not presented can be assumed to have been non-significant.

| Insert Table 4 About Here |

Question 1 involves the comparison of the improvement from A1FI to A1SLP relative to any improvement from A2FI to A2SLP. This was analyzed by testing the interaction effect in a writing assessments (2) × Communication-A status (2) repeated measures ANOVA with $\alpha = .05$. Because this statistical test involves only two groups, the test statistic was converted to a t-statistic using the relationship $t_{df} = \sqrt{F_{1, df}}$ so that a 1-tailed test could be conducted to maximize the power of finding that writing is improved in the group having taken the Communications-A course (Group A1). Because the between-groups effect (i.e., a test of the mean differences between those taking and not taking Communication-A, averaged across the two writing assessments) was not of interest, results of that analysis will not be presented.

Question 2 involves the comparison of A1SLP to BFI (i.e., B1FI, B2FI, and B3FI all together). This was assessed using an independent samples t-test. Because the logic of standard statistical hypothesis testing is not set up to test for equality of means, it was decided to apply a 1-tailed test to see whether the BFI scores were statistically higher than those for A1SLP. A 1-tailed test of this nature makes it easier to see whether the Group B scores are higher than those for A1, thereby providing a more rigorous test of the hypothesis.
Question 3 involves the comparison of $A_{F1}$ to $B_{F1}$ using an independent samples t-test. Because it is anticipated that students who exempted from the requirement should be better writers, this hypothesis was explored using a 1-tailed test.

Question 4 involves the comparison of $B_{1F1}$ with $B_{2F1}$ using an independent samples t-test. Because no hypothesis exists about the potential direction of any such effect, this test was conducted using a 2-tailed test. In addition, to further study the relationship between the different exemption mechanisms and students’ writing skill upon entry, we examined the correlations between test scores (including ACT scores) and FI essay ratings across all students.

Question 5 involves the comparison of $A_{1F1}$ and $A_{2F1}$ using an independent samples t-test. Because no hypothesis exists about the potential direction of any such effect, this test was conducted using a 2-tailed test.

**Results**

**Rater Conformity and Consistency**

Rater conformity using the three indexes described above was evaluated prior to Rating Sessions C, D, E, and F. Generally speaking, conformity was good, with the number of raters being approached each round about not conforming ranging from 1 to 3. Raters who were informed that their scores were unusually different from those of their colleagues consistently produced index values that were less extreme in subsequent rounds.

The nature of the indexes is such that the more data that are collected, the more power they have to identify if individuals are scoring differently from others. Hence, after the final round of rating, a greater number of raters were identified by the indexes (four as being too easy, one as being too hard, one as having too much variance and one as having too little variance)
than had been identified after any previous round. However, it is important to mention that the 
\( \alpha \)-level being used to flag a rater (\( \alpha = .05 \)) was fairly liberal in light of the number of statistical 
tests being conducted, suggesting that some of these raters may have been flagged unnecessarily. 
In addition, the magnitude of atypicality required to trigger a flag becomes much smaller as more 
data are collected. As an example, the most extreme rater was flagged for producing an average 
standardized rating that was just 0.23 points lower than that of the other raters.

Although rater conformity seemed reasonable, rater consistency—the extent to which 
raters agreed with one another—was more problematic. The design of the study called for each 
rater to be paired up with any other rater about 10-15 times, on average, so it was possible to 
look at the correlations among raters. On the basis of the holistic overall score, the median 
correlation between raters was just \( r = 0.265 \). That number was only marginally improved by 
systematically excluding correlations between raters who were paired up fewer than 8 times (i.e., 
\( r = 0.287 \)). These numbers suggest that raters had a very difficult time assigning the same scores 
to the essays.

One way to mitigate the effects of rater instability and improve the accuracy of the final 
ratings is to involve multiple raters per item. Although using two or three raters is conventional 
for this type of a design, in this study we utilized five raters per item. In this way, the evaluation 
of essays was not too heavily influenced by any one rater who might have been grading 
differently from his/her peers.

There is evidence to suggest that the strategy to adopt multiple raters was effective at 
producing final scores that were meaningful. Although raters struggled to agree with one 
another on the score to assign to an essay, the data suggest that they rarely disagreed by very 
much. Of the 162 essays that were graded, 72.2% of the time, the five raters were all within one
point of one another for the holistic rating. And in no circumstances did we observe an item for which two raters differed by three points (i.e., one rated it as a 1 and another as a 4). In addition, for 33% of the papers, four of the five raters were in agreement and in 8.6% of the papers, all five raters were in agreement. Also, only for 16.7% of the papers did there exist a rater whose score for an essay was more than one point different from the average across all raters. Results for the individual dimensions were very similar to those for the holistic score.

Research Questions

Question 1: The holistic mean ratings for both writing assignments are shown in Table 5 for Groups A1 and A2. As one can see, the gain score for Group A1 was larger than that for Group A2 by 0.247 points. However, the t-test indicated that this result was not statistically significant ($t_{40} = 1.674, p > .05$). Among the remaining DVs, a statistically significant difference was observed for Voice ($t_{40} = 2.422, p = .01$). The mean ratings for Groups A1 and A2 corresponding to the Voice dimension are provided in Table 6.

Insert Tables 5 and 6 About Here

Question 2: The sample mean SLP score for students currently enrolled in a Communications-A course (i.e., $\overline{A_{SLP}}$) was 2.81 (based on 24 students) compared with a mean FI score of 2.95 (based on 39 students) among those exempt from Communications-A (i.e., $\overline{F_{I}}$). The t-statistic for this comparison was $t_{61} = 1.168 (p > .05)$, indicating no significant difference in overall writing quality between those who completed a Communications-A course and those who had exempted.
With respect to the other dependent measures, a statistically significant result was found for Evidence, where $\overline{A_{1\text{SLP}}} = 2.84$ and $\overline{B_{FI}} = 3.12$ ($t_{61} = 2.417$, $p = .009$), suggesting that at the end of a Communications-A course, students in this study were not as strong with respect to the expectations regarding evidence as were those who had previously exempted from the Communications-A requirement.

Although in Research Question 1 we found very few statistical differences between SLP scores for students not exempt from Communications-A (after controlling for FI scores), as a follow-up analysis to those performed for Research Questions 1 and 2, we examined whether SLP scores for students needing, but not taking a Communications-A course (i.e., $A_{2\text{SLP}}$) differed from the $B_{FI}$ scores. For parallelism with the previous analyses, these analyses were done using 1-tailed independent samples $t$-tests. For the holistic score, it was found that $\overline{A_{2\text{SLP}}} = 2.57$ (based on 18 students) and $\overline{B_{FI}} = 2.95$, resulting in a $t_{55} = 2.614$, $p = .006$. Statistically significant differences were also observed for Content ($\overline{A_{2\text{SLP}}} = 2.61$, $\overline{B_{FI}} = 2.88$, $t_{55} = 1.783$, $p = .04$), Voice ($\overline{A_{2\text{SLP}}} = 2.67$, $\overline{B_{FI}} = 2.90$, $t_{55} = 1.767$, $p = .041$), Conventions ($\overline{A_{2\text{SLP}}} = 2.58$, $\overline{B_{FI}} = 2.88$, $t_{55} = 2.072$, $p = .021$), Evidence ($\overline{A_{2\text{SLP}}} = 2.67$, $\overline{B_{FI}} = 3.12$, $t_{55} = 3.178$, $p = .001$), and Composite ($\overline{A_{2\text{SLP}}} = 2.61$, $\overline{B_{FI}} = 2.90$, $t_{55} = 2.224$, $p = .015$).

Question 3: For the holistic score, $\overline{A_{FI}} = 2.452$ (based on 42 students), compared to $\overline{B_{FI}} = 2.954$ (based on 39 students). The $t$-statistic for this comparison was $t_{79} = 4.627$ ($p < .0001$), indicating that students who exempted from the Communications requirement wrote essays that were significantly better than those that were written by students still needing to complete the requirement. In addition, similar results were observed for all of the other dependent measures: for Content ($\overline{A_{FI}} = 2.51$, $\overline{B_{FI}} = 2.877$, $t_{79} = 3.311$, $p = .0007$), Organization ($\overline{A_{FI}} = 2.30$, $\overline{B_{FI}} = 2.708$, $t_{79} = 3.311$, $p = .0007$), Voice ($\overline{A_{FI}} = 2.543$, $\overline{B_{FI}} = 2.903$, $t_{79} = 3.726$, $p = .0002$),
Conventions ($A_{FI} = 2.433$, $B_{FI} = 2.877$, $t_{79} = 4.506$, $p < .0001$), Evidence ($A_{FI} = 2.652$, $B_{FI} = 3.123$, $t_{79} = 4.315$, $p < .0001$), and Composite ($A_{FI} = 2.488$, $B_{FI} = 2.900$, $t_{79} = 4.415$, $p < .0001$).

There appears to be strong evidence that the Communications-A exemption criteria are set up so as to produce exempt students who are better writers than those who do not exempt.

Question 4: For the holistic score, $B_{FI} = 2.982$ (based on 11 students), compared to $B_{FI} = 2.686$ (based on 14 students). The $t$-statistic for this comparison was $t_{23} = 1.725$ ($p > .05$), indicating no statistical difference between the holistic overall scores for the two groups of students. However, statistically significant differences were observed with respect to Voice ($B_{FI} = 3.073$, $B_{FI} = 2.657$, $t_{23} = 2.676$, $p = .013$), Conventions ($B_{FI} = 3.091$, $B_{FI} = 2.614$, $t_{23} = 2.709$, $p = .013$), Evidence ($B_{FI} = 3.255$, $B_{FI} = 2.914$, $t_{23} = 2.242$, $p = .035$), and Composite ($B_{FI} = 3.026$, $B_{FI} = 2.683$, $t_{23} = 2.307$, $p = .030$). For each test in which statistically significant results were found, students exempting on the basis of the University of Wisconsin English Placement Test produced higher writing scores than students exempting through either the AP or IB test.

To further explore the relationship between test scores and writing, Pearson correlations were computed between FI ratings and standardized test scores, provided test scores were available for at least 10 students. These results are shown in Table 7. Note that these correlations are across all participants, not merely those who exempted from the Communication-A requirement. Further note that the sample sizes vary considerably from test to test, thereby impacting the stability of the correlations and the critical values needed to achieve 1-tailed statistical significance at the $\alpha = .05$ (*) or $\alpha = .01$ (**) levels.

The data from Table 7 largely corroborate the findings from the above analysis, namely that UW Placement Test Scores (ENGL) are more highly associated with incoming writing skill
Communication-A Writing Assessment 23

(as measured by the ratings on the FI paper) than are AP scores. What is interesting, however, is how similar the correlations are between the ENGL and the ACT-ENGL test (especially since the sample sizes are nearly identical). Also of interest are the very high correlations between the SAT-WMC and SAT-Writing tests and the FI writing ratings (although these correlations are based on only 11 individuals), and the very low correlations with the ACT writing test (in spite of the large sample size).

Question 5: For the holistic score, $\overline{A_{FI1}} = 2.533$ (based on 24 students), compared to $\overline{A_{FI2}} = 2.478$ (based on 18 students). The $t$-statistic for this comparison was $t_{40} = 0.369$ ($p > .05$), indicating no statistical difference between the holistic scores for the two groups of students. In addition, none of the tests for other dependent variables identified any statistically significant differences; therefore, there is no statistical evidence for differences in writing ability between the students who elect to take Communications-A courses in the fall and those who wait to take it later.

Discussion

Many of the questions that surround the Communications-A requirement have been difficult to answer. This difficulty stems from two primary sources. First, students who exempt from the requirement do not need to take the course. Similarly, those who are not exempt but are not currently enrolled in a Communication-A course are, quite obviously, also not enrolled in the course. Therefore any assessment done within a Communications-A class will not capture the writing for a significant proportion of the freshman population, thereby failing to provide an
adequate control group against which to compare students who completed a Communication-A course. In addition, writing cannot readily be assessed by asking students to complete a decontextualized, standardized writing assessment because such assessments are too brief and inauthentic, and fail to adequately address the role of revision in the writing process.

The purpose of this study was to establish a proof-of-concept for a direct assessment of the writing component of the Communication-A requirement that overcomes the traditional obstacles described above. This study achieves that goal by assessing the writing of students enrolled in a different class with respect to the ELOs for Communication-A courses. As a result of selecting a class that enrolls students from across the university, this design allows us to collect data on students who are currently taking a Communication-A course with those who are still needing to take one, as well as with those who have exempted from the requirement altogether. In addition, by assessing writing on actual course assignments, we are ensured that the writing is relevant to the students and that it will reflect, if not their best, at least a representative amount of effort. Consequently, we are able to compare growth in student writing for freshmen who have not satisfied the Communication-A requirement, as a function of whether those students are currently enrolled in a Communication-A course (Research Question #1). In addition, we can evaluate whether the writing quality of students who complete a Communication-A course is comparable to that of students who exempted from the requirement (Research Question #2), as well as whether students who exempt, and who have long since been assumed to be better writers, are indeed better writers (Research Question #3) and whether that varies based on the means by which the student exempted (Research Question #4).

The statistical findings from this study were mixed. The good news is that it was found that after students completed a Communication-A course, their holistic score was not statistically
different from that of the students who exempted. In addition, on that same end-of-semester writing assessment, the students who needed but did not take Communications-A were significantly worse than those who exempted. This was true with respect to the holistic score, as well as the Composite score and four of the five dimensions. However, in spite of these encouraging results, and in spite of the fact that the average gain in the holistic writing score for students completing Communication-A was approximately a quarter point higher than the average gain for those who needed but did not take Communication-A, no statistically significant differences were observed between those two groups. A statistically significant difference was observed, however, with respect to the Voice dimension.

The results were fairly compelling with respect to the question of whether students who exempt from the Communication-A requirement enter the university as better writers than students who do not exempt. Statistical tests of the holistic score, Composite, and all five dimensions were statistically significant. Some data exist to suggest that the students exempting on the basis of the University of Wisconsin Placement Tests are better writers than those who exempt on the basis of the AP test, at least with respect to Voice, Conventions, Evidence, and Composite. However, at the level of the holistic score, no difference was observed between the two groups.

Analyses of the correlations between various standardized tests and the early-semester assessment were interesting in that they appeared to confirm our prior belief that stand-alone writing assessments are of limited value in assessing college writing, as indicated by the very low correlation with the ACT writing score. However, the correlations with the SAT writing test were quite large—larger, in fact than correlations with any other measures—but it is important to interpret these correlations with caution because they are based on a sample size of only 11
students. One of the more surprising results is that the AP-Language test correlated so poorly with the writing scores in the study, although the sample sizes here were also very small. AP-Literature correlations were a fair bit higher, though they still appeared, on average, to be somewhat lower than the correlations with the ACT-ENGL or the placement test.

As this was but a proof-of-concept study, our primary goal was to explore whether or not this framework was viable for addressing the types of questions we posed. Finding answers to the research questions themselves was but a secondary goal. As such, the results of this study should be interpreted with caution, and we are hopeful that lessons learned throughout will help us to design a better study the next time around. In particular, we believe a number of features of this study contributed to reduce our power to detect significant differences.

First, and most obviously, the sample sizes in this study were quite low. Because only 81 students were included in this study, all of the subgroups were fairly small. Some of the research questions that involved collapsing across categories were appropriately powered, but questions that involved the comparison of any two individual groups (e.g., comparisons of the two non-exempt groups, comparisons of students who exempt by different means, etc.) were clearly underpowered.

A second limitation is that the early assessment, the paper to accompany the faculty interview, although it was assigned at the very beginning of the semester, was not collected until the eighth week. By that time, students enrolled in a Communication-A course had already completed half their course. Therefore, it is reasonable to hypothesize that the FI papers for those students reflected some of the Communication-A curriculum. To the extent that this is true, and to the extent that the Communication-A curriculum helps improve student writing, our inability to collect a better assessment of students’ incoming writing skill likely decreased the
magnitude of the observed effect. Moving forward, it will be important to work with course instructors so that we can collect a writing sample much earlier in the semester, ideally within the first couple of weeks.

One of the strengths of this design is that it provides an authentic writing assignment to grade, one that related to the studied material and contributed towards the course grade. However, in the case of both assignments, the writing task was somewhat different than what would typically be encountered in most classes, and certainly was different from the types of writing assignments in Communication-A courses. In both cases, the short writing assignment was designed to accompany a larger project, either a faculty interview or a capstone project at the end of the course. As such, it is not clear to what extent the students in CP125 recognized that their papers would be evaluated on the basis of the writing quality, as opposed to the quality of the ideas expressed therein. Certainly, with most stand-alone writing assignments, students can safely assume that the writing quality is being evaluated along with the content; however, the informal nature of these particular assignments makes one question students’ mindsets as they approached these writing tasks.

Similarly, the genres of writing were quite different from what is typically taught in Communication-A courses, and the directions underlying the assignments appeared not to be uniform across sections. As a result, raters expressed some concern that it was very difficult to apply the rubric to certain papers. As an example, for the SLP, in some classes, instructors asked students to select from a list of types of projects, whereas in others, instructors were less prescriptive. Consequently, without knowing the specific task that had been assigned, raters struggled to evaluate papers on the basis of originality. Also, without knowing exactly what materials had been assigned as part of the course, or without knowing the exact instructions to
the students for the paper, it was very difficult to reliably assign a score for the Evidence criterion.

Finally, as anyone involved in test development knows, it is very difficult without using statistical techniques like equating, to develop assessments that are of comparable difficulty. Because this study used two separate assessments, it is entirely possible—likely, even—that those writing prompts were not of equal difficulty. That is, even had the assignments been given at the same time to the same group of students, it is likely that one of them would produce scores that were somewhat higher than the other. This can either hurt the power, if the more difficult writing task were the one at the end of the semester, or could artificially inflate the power, if the more difficult writing task were the one at the beginning of the semester. In an ideal design, the assessments would be counter-balanced so that each was given first to half the students in each group. Obviously, because these are actual assignments for a course, such counter-balancing is impractical. However, it is important to recognize this potential confound, and to work to develop assessments that are as uniform with respect to difficulty as possible.

In addition, it is important to recognize that our decision to assess writing within CP125 does limit somewhat our ability to generalize the results of this study to all Communication-A courses. It is quite likely that this study did not include a representative number of students (if any) taking the Engineering Professional Development or Life Science Communication courses, since students in these units are often enrolled for other freshmen seminar experiences. However, we know from earlier Communication-A studies that differences among the courses offered do not seem to produce significant differences in students’ perceptions of learning outcomes, so we are hopeful that the same is true of college writing.
The results of the study would clearly indicate that this is a worthwhile design to pursue further. A follow-up study to this one should be designed to replicate the findings of this study, and to improve the power to detect actual differences. Ideally, it will be possible to collaborate with one of the many courses on campus that serves a large number of freshmen so that more students could be included in this study. The advantages to working with a single large class as opposed to several small ones are that (a) it is possible to recruit more students, (b) it is easier to work with the instructor to develop writing tasks that are conducive to being evaluated with the Communication-A scoring rubric, and (c) it is easier to standardize the writing tasks for all students, and to communicate those instructions to the research team. Balancing design considerations with available resources, it would be ideal to add a third day to the rating and repeat the study with 125-150 students. Given the difficulty raters experienced in consistently evaluating essays, it will be important to continue having a minimum of 5 individuals rate each paper. Also, by recruiting within a large class, it will hopefully be possible to get consent to collect a pool of papers considerably larger than the 125-150 needed for the study itself so that the research team can create a better, more instructive set of papers on which to train the raters.

It is likely that the failure of the research team to be able to identify good anchor papers limited the ability of the raters to interpret the scale as intended. Also, it will be important that the tasks be spaced appropriately in the syllabus, so that the early semester assessment is collected within the first few weeks and the end-of-semester assessment is collected in the last few weeks. Finally, the Communications-A subcommittee should use the experiences learned from this study to revisit the rubric itself and revise it so that it is more straightforward and easier to apply consistently.
The study described here describes our initial attempt to evaluate the writing literacy component taught in Communication-A courses. It is important to recognize that written communication is but one aspect of the Communication-A curriculum (the others are oral communication and information literacy). The decision to focus this study exclusively on writing literacy is based on former Research Director Halaby’s advice that such an initial study should not be made overly complex and focus on only one of the elements. Although we hoped to learn more about the logistics involved in managing the data for a pre/post design, we also felt it would be too daunting to undertake the challenges of recording student presentations for the oral communication component. Furthermore, the CP125 course contains instruction in information literacy that would likely have had an impact on our ability to separate Communication-A instruction from CP125 instruction in this area. Evaluation of the remaining components of Communication-A is left for future study.

References


Table 1. Demographic Information for Participants

<table>
<thead>
<tr>
<th></th>
<th>Not Exempt from Communications A</th>
<th>Exempt from Communications-A</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Currently Enrolled</td>
<td>Not Currently Enrolled</td>
<td>By Placement Test Scores</td>
</tr>
<tr>
<td>Count</td>
<td>24</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>Female</td>
<td>12</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Male</td>
<td>12</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Asian</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Black/African American</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Native Hawaiian/Other Pacific Islander</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>White</td>
<td>17</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>Int. Student</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18 years</td>
<td>19</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>19 years</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SCORE</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>-------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Trait</strong></td>
<td><strong>Writing exhibits consistent strengths and is successful in conveying a clear point. Little to no revision necessary.</strong></td>
<td><strong>Strengths outweigh weaknesses. Some revision is needed.</strong></td>
<td><strong>Weaknesses outweigh strengths. Considerable need for revision.</strong></td>
</tr>
<tr>
<td><strong>Ideas, Content, Critical Thinking</strong></td>
<td>Essay is grounded by a strong thesis or sense of purpose, is supported by effective evidence when appropriate. Displays fresh, original, and nuanced critical thinking.</td>
<td>Essay has a satisfactory thesis statement or sense of purpose, and utilizes effective evidence when appropriate. Ideas may not be original, but are sound and enrich the overall theme.</td>
<td>A clear and/or purposeful thesis is absent; the essay defines the topic, but does not develop it well. The evidence utilized is unconvincing, and the essay is not yet focused beyond the obvious.</td>
</tr>
<tr>
<td><strong>Organization (at both the global and local levels)</strong></td>
<td>Organization enhances and showcases the argument and purpose of the essay; the structure is compelling, clearly linked to the writer’s central purpose, and moves the reader through the text. Conventions or structure are appropriate to the rhetorical situation, transitions are thoughtful and effective, and support for key ideas is unified and consistent. Paragraphing is sound and reinforces the essay’s organization and overall meaning.</td>
<td>Organization is strong enough to move the reader through the essay; the paragraphs typically employ conventions adequately such as topic sentences, transitions, and support. Overall, the structure includes an introduction and conclusion.</td>
<td>Organization emerges in the essay, but it is insufficient or disconnected. Structure may be too formulaic (such as a “Five Paragraph Theme”). Transitions, topic sentences, and text markers are used ineffectively or not at all.</td>
</tr>
<tr>
<td><strong>Voice, Audience, Word Choice</strong></td>
<td>Word choice, tone, and genre are appropriate for the content, audience, and purpose of the essay. The essay reflects awareness and respect for the audience, offers a compelling and engaging narrative or presents its argument with fluidity and confidence. A sense of personality emerges beyond the formal elements of the piece.</td>
<td>Essay illustrates a general awareness of audience and uses a tone appropriate to its purpose. Typically, the word choice and style effectively illustrate the essay’s message, and the essay reflects a commitment to its topic. Although a point of view is evident, a sense of personality or personal style does not emerge.</td>
<td>Essay strives for a sense of voice and attempts an engagement with the audience, but falls short. Overall, the language is merely functional, and generalities are favored over personal insights or specific examples. Attempts at colorful language may result in a stilted style. Jargon, clichés, and generalities mark this essay’s style.</td>
</tr>
<tr>
<td><strong>Sentence Fluency, Language and Genre Conventions</strong></td>
<td>Essay demonstrates a solid grasp of writing conventions; sentences are well-constructed and utilize effective and varied structures that facilitate the reader’s engagement with the essay’s ideas. Errors are few and the writer may, at times, manipulate convention for stylistic effect.</td>
<td>Overall, the writing flows well and is unencumbered by error. Although some errors exist, the essay ultimately utilizes sentence structures, language, and genre conventions to its benefit.</td>
<td>Sentences lack variety and are often clumsily constructed. A sense of genre conventions exists but mostly is ineffective. Spelling and other conventions are used with some familiarity on the writer’s part, but often display errors or awkwardness.</td>
</tr>
<tr>
<td><strong>Use of Evidence</strong></td>
<td>The essay demonstrates, when appropriate, an especially effective selection, balance, and integration of ideas, descriptions, examples, or information from secondary sources. If used, direct quotations, paraphrase, and summary of relevant external sources support key ideas or initiate inquiry and are appropriately documented.</td>
<td>Generally, the essay uses appropriate supporting materials to enhance the essay’s thesis and supporting points. When used external evidence is surrounded by the writer’s analysis. More variety of direct quotations, paraphrase, and summary may be needed.</td>
<td>Essay mirrors an appropriate genre, but does not achieve an effective balance between the writer’s point of view and the ideas and language of external sources. Examples or sources may be strung together in extended direct quotations or dropped into the essay rather than being synthesized.</td>
</tr>
</tbody>
</table>
### Table 3. Agenda for Rating Sessions

**Study PIs:** Jim Wollack (Educational Psychology/Testing & Evaluation Services)  
Morris Young (English)

**Monday, May 23**
- 8:00 - 12:00 Rater Training  
  -- Introduction to Rubric  
  -- Review and Discussion of Model Essays  
  -- Practice Rating  
- 12:00 - 1:00 Lunch  
- 1:00 - 2:45 Rating Session A  
- 2:45 - 3:05 Break  
- 3:05 - 4:50 Rating Session B  
- 4:50 - 5:00 Wrap Up

**Tuesday, May 24**
- 8:00 - 9:00 Rater Re-Calibration  
- 9:00 - 9:45 Rating Session C  
- 9:45 - 10:15 Break  
- 10:15 - 12:00 Rating Session D  
- 12:00 - 1:00 Lunch  
- 1:00 - 2:00 Rater Re-Calibration (if necessary)  
- 2:00 - 2:45 Rating Session E  
- 2:45 - 3:05 Break  
- 3:05 - 4:50 Rating Session F  
- 4:50 - 5:00 Wrap Up
Table 4. Codification of Targeted Groups and Assessments

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Exempt from Communications A</td>
<td>Exempt from Communications-A</td>
</tr>
<tr>
<td>A1</td>
<td>Currently Enrolled</td>
<td>B1</td>
</tr>
<tr>
<td>A2</td>
<td>Not Currently Enrolled</td>
<td>B2</td>
</tr>
<tr>
<td></td>
<td>By Placement Test Scores</td>
<td>B3</td>
</tr>
<tr>
<td></td>
<td>By AP/IB Test Scores</td>
<td></td>
</tr>
<tr>
<td></td>
<td>By Both Placement and AP/IB Scores</td>
<td></td>
</tr>
</tbody>
</table>

Early Assessment: Faculty Interview
- A1<sub>FI</sub>
- A2<sub>FI</sub>
- B1<sub>FI</sub>
- B2<sub>FI</sub>
- B3<sub>FI</sub>

Late Assessment: SLP
- A1<sub>SLP</sub>
- A2<sub>SLP</sub>
- B1<sub>SLP</sub>
- B2<sub>SLP</sub>
- B3<sub>SLP</sub>

Table 5. Mean Ratings for Research Question 1: Holistic Overall

<table>
<thead>
<tr>
<th></th>
<th>FI</th>
<th>SLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Faculty Interview</td>
<td>Significant Learning Project</td>
</tr>
<tr>
<td>A1 Concurrently Enrolled in Comm-A</td>
<td>2.450</td>
<td>2.808</td>
</tr>
<tr>
<td>A2 Not Concurrently Enrolled in Comm-A</td>
<td>2.456</td>
<td>2.567</td>
</tr>
<tr>
<td>Total</td>
<td>2.452</td>
<td>2.705</td>
</tr>
</tbody>
</table>

Table 6. Mean Ratings for Research Question 1: Voice

<table>
<thead>
<tr>
<th></th>
<th>FI</th>
<th>SLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Faculty Interview</td>
<td>Significant Learning Project</td>
</tr>
<tr>
<td>A1 Concurrently Enrolled in Comm-A</td>
<td>2.508</td>
<td>2.942</td>
</tr>
<tr>
<td>A2 Not Concurrently Enrolled in Comm-A</td>
<td>2.589</td>
<td>2.667</td>
</tr>
<tr>
<td>Total</td>
<td>2.543</td>
<td>2.824</td>
</tr>
</tbody>
</table>
Table 7. Correlations Between Test Scores and Faculty Interview Ratings

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Overall</th>
<th>Content</th>
<th>Organization</th>
<th>Voice</th>
<th>Conventions</th>
<th>Evidence</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENGL</td>
<td>79</td>
<td>.527**</td>
<td>.464**</td>
<td>.419**</td>
<td>.526**</td>
<td>.506**</td>
<td>.526**</td>
<td>.547**</td>
</tr>
<tr>
<td>AP-Lang</td>
<td>16</td>
<td>.178</td>
<td>.145</td>
<td>.239</td>
<td>-.102</td>
<td>.182</td>
<td>.277</td>
<td>.167</td>
</tr>
<tr>
<td>AP-Lit</td>
<td>26</td>
<td>.435*</td>
<td>.427*</td>
<td>.499**</td>
<td>.315</td>
<td>.360*</td>
<td>.406*</td>
<td>.441*</td>
</tr>
<tr>
<td>ACT-ENGL</td>
<td>77</td>
<td>.560**</td>
<td>.501**</td>
<td>.454**</td>
<td>.534**</td>
<td>.594**</td>
<td>.441**</td>
<td>.564**</td>
</tr>
<tr>
<td>ACT-READ</td>
<td>77</td>
<td>.313**</td>
<td>.284**</td>
<td>.247*</td>
<td>.310**</td>
<td>.393**</td>
<td>.169</td>
<td>.311**</td>
</tr>
<tr>
<td>ACT-WRSUB</td>
<td>71</td>
<td>.186</td>
<td>.233*</td>
<td>.176</td>
<td>.137</td>
<td>.095</td>
<td>.031</td>
<td>.154</td>
</tr>
<tr>
<td>SAT-Essay</td>
<td>11</td>
<td>.033</td>
<td>.108</td>
<td>.143</td>
<td>.019</td>
<td>.102</td>
<td>-.134</td>
<td>.064</td>
</tr>
<tr>
<td>SAT-Verb</td>
<td>11</td>
<td>.516</td>
<td>.287</td>
<td>.454</td>
<td>.378</td>
<td>.494</td>
<td>.468</td>
<td>.450</td>
</tr>
<tr>
<td>SAT-WMC</td>
<td>11</td>
<td>.753**</td>
<td>.559*</td>
<td>.687**</td>
<td>.393</td>
<td>.698**</td>
<td>.611*</td>
<td>.646*</td>
</tr>
<tr>
<td>SAT-Writing</td>
<td>11</td>
<td>.709**</td>
<td>.561*</td>
<td>.687**</td>
<td>.381</td>
<td>.706**</td>
<td>.505</td>
<td>.628*</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01; all p-values are 1-tailed.