Cut Score Analysis and Validity for Writing Prompts

Alan Socha and Raymond Barclay

The College of New Jersey

Paper presented at the 32nd Annual Conference of North East Association for Institutional Research (Saratoga Springs, NY, November 6-8, 2005).

Abstract

A cut score validity analysis for a pretest writing prompt and the SAT Verbal and Math assessments was performed for academic writing course placement. This study was carried out as part of a broader program of assessment aimed at understanding the impact of the writing course and eventually set the baseline for assessing the impact of ‘writing across the curriculum’ initiatives on student learning. The discussion and report will review pertinent findings for the cut score validity analyses and discuss issues related to the scope of methods and presentation of findings. In particular, there are many guides in the literature on undertaking a cut-score analysis and the current study included a myriad of methods to inform its research.

Cut Score Analysis and Validity for Writing Prompts

Curriculum transformation and the implementation of new Liberal Learning courses at The College of New Jersey have led to much discussion of curriculum efficacy and impact. In order to provide faculty and staff with a more accurate picture of student writing skills and to identify students in need of a writing intervention course, we need reliable and valid assessment procedures and a thorough process and outcomes evaluation design. To this end, we are undertaking several studies aimed at understanding how a new writing course (WRI-102) impacts student learning. This particular analysis is part of that larger picture and will review the intervention effect of WRI-102 and determine a cut scores for the SAT Verbal and Math assessments and for a pretest writing prompt to better place students who need WRI-102 intervention.

WRI-102 and the writing prompts were piloted during the same academic year. Because of this it was decided that all incoming students should take both the writing prompts and WRI-102, without an option to place out. A reliability and validity analysis was then conducted on the writing prompts to determine that they are internally consistent and stable and that they are appropriate for the purpose of WRI-102 placement. WRI-102 intervention was also analyzed to determine whether or not WRI-102 is effective in its purpose of enhancing the writing skills of students. After it has been established that the prompts are valid, a cut score analysis was conducted to determine the score to be used for future placement of students. The cut score for the writing prompt and SAT Verbal and Math exams was then tested the following year to determine if the two types of errors produced by the cut score can be minimized further and to determine if the cut scores did indeed discriminate the students well enough to place them correctly.

Literature Review

A cut score is a test score that discriminates between a group of students that have adequate skills and those with inadequate skills relative to a testing situation. Cut scores are often used to make policy decisions, thus leading to many social and political issues, as well as many different procedures for carrying out such an analysis. Choosing an appropriate method depends on the policy being affected, the test, and how practically and theoretically sound the method is in the context where it is applied. Item content, educational consequences, psychological and financial costs, performance of others, and errors due to guessing and item sampling should also play a role in choosing an appropriate cut score method (Hambleton, 1978a). Despite some popular misconceptions, there is no such thing as a true cut score. Making a determination about where to place a cut score on some continuum always entail judgment and inevitably leads to some level of misclassification. This misclassification occurs because we are imposing an artificial dichotomy on what is really a "continuous distribution of knowledge and skills" (Dwyer, 1996). Since tests are not usually perfect and there is usually a group of students whose skills are considered 'borderline' adequate, different cut score analysis procedures can lead to different cut scores, thus making it more important to choose an adequate.

There are two errors which need to be taken into consideration when choosing a cut score, the number of students who achieve a passing score and should not have, and the number of students who do not achieve a passing score and should have. These errors will always exist because "tests are almost never perfect measures of the knowledge and skills they are intended to measure" (Livingston & Zieky, 1982, p. 12). Reducing one of these errors will increase the other, i.e., you can reduce the chance that test-takers will not meet the passing score by lowering the score, but this will place more test-takers into a course where they may not have adequate skills.

The reverse is also true if the passing score is raised to reduce the chance test-takers do meet the passing score. The only way to reduce both errors is to improve the test itself.

Some methods of choosing a cut score use a relative standard, which depends on comparisons between individuals (Livingston, et al., 1982). A relative standard allows the administrator to choose a score that will pass a desired number or percentage of test-takers. This method can insure that cost limitations are met but does not minimize the two errors created by utilizing a cut score. Most cut score procedures are considered absolute standards where it does not matter how well the other test-takers perform.

Nedelsky developed an absolute standard method for multiple choice tests (Livingston, et al., 1982). This method requires the judges to look at each wrong answer for every question and decide which one would most likely be chosen by a borderline test-taker. Typically each judge will go through the test individually and make their judgments. Then the judges will meet and discuss their reasoning for their choices and make a final decision. This method is based on the idea that a borderline test-taker will eliminate the answers they think is wrong and then guess from one of the final answers. To find the expected score of a borderline test-taker one would divide the number of answers the test-taker has to guess from for a particular question from 1 and then add these up for each question.

Angoff suggested a method similar to Nedelsky's that is not restricted to multiple choice tests (Livingston, et al., 1982). The method for computing the expected score of the test is the same as Nedelsky's method. The only difference between Angoff's method and Nedelsky's is that here the judges are making a judgment of the probability a borderline test-taker will answer a question correctly instead of looking at each possible wrong answer (Livingston, et al., 1982; Hurtz & Hertz, 1999). One study suggested that judges can adequately discriminate low-performing students from others but cannot adequately estimate how those students would perform on the test (Impara & Plake, 1998). The judges' ability to make a judgment about probability is the biggest drawback of Angoff's method. Some people have used a modified form of Angoff's method where the judges are presented with a selection of probabilities to choose from instead of making their own. This should only be used when the judges cannot make valid probability judgments, since this method may bias the judges' choices and limits the scale from which they choose from (Livingston, et al., 1982). One study gives evidence that Angoff's method and a modified Angoff's method using deciles instead of the full probability range produce different results (Plake & Giraud, 1998), so the decision of which Angoff's method to utilize should be made wisely. Other extensions of the Angoff method are to compute the expected score a barely adequate student will get, presenting passing rate data to the judges, asking the judges to provide the expected distribution of barely adequate students across the score scale, and presenting adequate and non-adequate score profiles after each stage to ensure that the implications of the standards are understood by the judges (Hambleton & Plake, 1995).

A paper selection method was used to obtain cut scores on open-ended assessments. This was used in conjunction with a multiple choice assessment for a reading assessment in a metropolitan school district in the Midwest (Buckendahl, Plake, & Impara, 1999). First the assessment was broken into subparts, each having unique scoring rubrics and each testing certain skills and knowledge. Overlap between subparts of the multiple choice assessment as well as overlap between subparts of the open-ended assessment and the multiple choice assessment can be used to ascertain if different judges are providing estimates on the same scale. Anchor papers are then selected that demonstrate typical responses and characteristics of each score point.

These papers are coded so the judges will not know the actual score of the paper. The judges will use these papers for training and to discuss what knowledge and skills are required for a barely adequate student. These papers will be used to decide upon an initial cut score estimate by selecting the paper that most accurately portrays the borderline student. Then the judges will be presented with item performance data and the percentages of students who will pass and fail based on their initial cut score estimate. Judges will then discuss and create a second round estimate of the cut score. One study suggested using a blind review where the judge does not know the score of the essay for the first round and an informed review where the judge does know the score for the second round (Cross, Frary, Kelly, Small, & Impara, 1985). This was suggested based on the evidence found that a blind review will produce different results from an informed review.

Ebel suggested a two-stage procedure that involves classifying the questions into a group for difficulty and a group for relevance or importance (Livingston, et al., 1982). Ebel suggested that there be three difficulty levels labeled “easy,” “medium,” and “hard,” and four relevance categories labeled “essential,” “important,” “acceptable,” and “questionable.” A classification table will be made for the judges’ decisions by crossing difficulty with relevance and then the judges will make one judgment for each of the twelve blocks of the table with the percentage of questions a borderline test-taker will answer correctly. It is recommended that the judges classify each question separately and then meet and discussion their reasoning. This is the same with the percentages for each classification, each judge should derive their percentages for each block separately and then meet and discuss their reasoning for these decisions. A passing score can then be found by multiplying the judged percentage correct for a category by the number of questions in that category and summing these up for every category.

The dominant profile judgment method is designed for polytomous tests. First, judges are asked to state their standard-setting policies by identifying some score profiles for barely adequate test-takers and then they met and discussed their policies. If the judges could not agree then a questionnaire would be mailed to each of them containing the favored policies. The questionnaire would ask the judges to indicate their support for each of the favored policies, indicate their level of confidence in each policy, rank the policies in order of preference, and comment on the process itself (Plake, Hambleton, & Jaeger, 1997).

The Borderline-Group method is similar to Nedelsky's and Angoff's methods except judges will identify actual test-takers as borderline instead of making educated guesses about how a borderline test-taker will perform. A passing score can be set by using the median score of the borderline group of test-takers. The median is used instead of the mean in this case because it is less affected by a few extremely high or extremely low scores (Livingston, et al., 1982; Kane, 1994). This method is easy and simple, but has disadvantages in that borderline test-takers are usually a small percentage of test-takers and that judges may have trouble identifying borderline test-takers. This method is also problematic in that judges may base their decisions on something other than what the test measures and may differ considerably in their standards for judging test-takers.

The Contrasting-Groups method uses judgments of test-takers' knowledge and skills to divide them into two contrasting groups, a "qualified" group and an "unqualified" group. First the administrator needs to define adequate and inadequate levels of the knowledge and skills tested, select a sample of test-takers, and obtain test scores and judgments from the judges about these test-takers. The judges should not know the test-takers' scores. Then the test-takers will be divided at each score level into the "qualified" and "unqualified" groups based on these judgments and the percentage of test-takers at each score level in the “qualified” group will be computed. These percentages will then be smoothed and a passing score chosen. The passing score is typically the test score for which the “smoothed” percent-qualified is exactly 50 percent. A higher score will be more likely to judge a test-taker as qualified rather than unqualified and the reverse is true for a lower score. One study produced evidence that the Contrasting-Groups method will fail less test-takers than Nedelsky’s method mentioned earlier (Mills, 1983). The main advantage with this method is one’s ability to estimate the frequencies of the two types of decision errors (Livingston, et al., 1982).

A modification on the Contrasting-Groups method is the Up-and-Down method which involves scoring the test-takers and then choosing them one at a time based on a score that is assumed to be near what a proper passing score would be and judging whether or not that particular test-taker is qualified or unqualified. If this test-taker is qualified then another test-taker is chosen who has a lower score; and the reverse if the test-taker is unqualified. A passing score is then chosen by taking the average score of the test-takers selected for judging beginning just before the scores start to zigzag and ending with the score of the next person who would have been judged if the procedure had continued. It is important that the judges do not know this method is being used because their judgments may be based on the previous one (Livingston, et al., 1982).

In the judgmental policy capturing method, judges are asked to review scored tests and assign an overall rating (Jaeger, 1995). Then multiple regression is used to calculate the extent to which an individual judge’s overall ratings are predictable. Weighted averages of the individual judges’ regression weights are then computed to form a single equation used for determining the cut score.

The analytical judgment method is similar to the judgmental policy capturing method. This method asks judges to classify student papers into performance categories. The judges will do this for each component part of the test-taker's paper. Typically a judge will assess one component of everyone's paper and then go through the papers again, in a different order, and assess the next component (Plake & Hambleton, 2000). A cut score is then calculated using a boundary paper method which takes the average of papers that are borderline, or a cubic regression method which uses all data to determine the relation between the categorization values and the scores (Plake, et al., 2000).

Judges should meet certain requirements regardless of method. Judges should be qualified, be able to make judgments of the knowledge and skills the test is intended to measure, must be able to make judgments that reflect the test-takers' skills at the time of the testing, and must be able to make judgments that reflect the judges' true opinions. Judges must be able to determine the test-taker's knowledge and skills and must know what level of knowledge and skills a person passing the test should have (Livingston, et al., 1982). Good judges typically excel in their domain. Good judges should not let their group performance expectations affect their cut score. There is some evidence to suggest that judges estimates of a later round of item performance will change in the direction of their initial estimates if this were the case (Buckendahl, Impara, Giraud, & Irwin, 2000). One difficult decision is to decide how many judges are needed. This depends on the cost and benefit of more judges and the method chosen. One study recommends fifteen judges for Angoff's method (Hurtz, et al., 1999). It is important no matter how many judges are selected that they each are in agreement of the behavioral characteristics to use when examining and judging individual test items. There is evidence to suggest that judges using different definitions of the knowledge and skills necessary to pass and who have different perceptions of the target test-taker will set different cut scores (Impara, Giraud, & Plake, 2000). This agreement will result in a more defensible cut score (Impara, et al., 2000).

It is crucial that the test be reliable and valid for its intended purpose. Reliability is necessary to provide evidence of the validity of the cut score process (Kane, 1994). One method of checking this is by having the judges look at the test. The test should be free of bias and groups of test-takers should not differ systematically in their scores unless their knowledge and skills truly differ. The effects of the derived cut score should be observed to ensure that it is appropriate. Reliability measures the internal consistency and temporal stability of a test. This refers to consistency between judges and consistency between judgments. The inter-rater reliability should be average at minimum. Lower reliabilities will lead to lower precision when determining the cut score.

With the increasing number of assessments used for setting cut scores, validity is becoming more important (Buckendahl, et al., 2000). Content validity refers to whether or not the test is appropriate for its intended purpose. Messick (1995) stated that validity is a property of the interpretation assigned to test scores. Validation demonstrates that the cut score represents the minimally adequate level of performance for some purpose (Kane, 1994) and will “influence the kind of test score interpretations that are possible” (Hambleton, Swaminathan, Algina, & Coulson, 1978b). An assessment can be invalid in two ways: it can be underrepresentative and fail to include facets of the construct or it can be irrelevant in that it contains excess reliable variance associated with other distinct constructs (Messick, 1995). One way of measuring content validity of a test is by observing the relationships between test scores and course grade. Another way is to measure how accurately the cut score separates test-takers into mastery states

(Hambleton, 1978a). Observing this relationship also provides a method of modifying the cut score to minimize the two types of errors. Content validity should not vary across different groups of test-takers and should not vary much over time (Hambleton, et al., 1978b). Validity can be measured with correlational and experimental methods. One such example is assigning individuals to two groups, giving one group instruction and the other group none. If the group with instruction obtains higher test scores then this would support the validity hypothesis (Hambleton, et al., 1978b).

A question often asked about methods where judges make decisions based on test-takers rather than the questions is “how many test-takers should be tested?” Usually a tradeoff between costs and benefits is associated with the answer to this question. The costs are getting the judgments and the benefits are better representation of the test-taker population and greater precision in determining the passing score.

There is also the controversial question of whether or not granting exceptions to the decision rule is appropriate. This comes from the two types of errors that result from using a cut score, as well as unforeseen circumstances in the life of a test-taker. Controversy can be reduced by building exceptions into the decision rule or by allowing test-takers to retake the test. Retaking the test may be appropriate because the test-taker may have had a “bad day” on the day of the test. It is also possible for the test-taker’s skills to improve between tests. If the test-taker is allowed to take a test again it should be a different form of the test to ensure that the test-taker is not improving their score because of the memorization of specific test questions. Allowing a test-taker to take a test again is a way to reduce the uncertainty of the skills and knowledge the test-taker is judged with.

Test scores reveal differences of knowledge and skills. Two test-takers with similar scores are not very different until treated differently, especially since tests cannot be perfectly reliable and valid (Dwyer, 1996). Cut scores create this difference and treat the test-takers differently, resulting in cut score controversy. Some of the resulting decisions from a cut score may not be correct. Granting exceptions and allowing test-takers to take the test again may be appropriate in this sense.

Another issue that often arises is whether or not the standard will change over time. Depending on the type of testing it may be important for the standard to change, e.g. the test is required for certification. If the test changes from one year to the next its level of difficulty may change as well. In this case the cut score should be adjusted to account for these differences in the test and the standard.

Due to the nature of WRI-102, we chose writing prompts for our assessment. None of the above cut score methods were utilized to determine the cut score on this assessment due to the introduction of the course and the writing prompts to the college at the same time. It was necessary that we validate the writing prompts before implementing a cut score. After validating the writing prompts, intervention was used to determine which students were impacted by the course, and based on this information cut scores were selected. Our grading of the writing prompts and training followed that of the paper selection method. Essays would be scored using a writing rubric, with each score representing a certain level of skills and knowledge. Anchor papers were used for training.

Method

Participants and Administration

First year students entering The College of New Jersey in fall 2003 and fall 2004 were evaluated in terms of their efficacy at writing. During Welcome Week in August 2003, all entering first year students responded to a prompt which asked them for an argumentative essay (the “pre-test”). Fall 2003 WRI-102 students responded to another argumentative essay prompt (the “post-test”) during the final exam period in December 2003. Another pre-test prompt was administered in January 2004 to all students taking WRI-102 in the spring semester and a post-test prompt was administered in May 2004 to these students. Intervention and cut score analyses were performed only on those who had less than a 580 on either the SAT Verbal or SAT Math for the following cohort (fall 2004) to assess how the cut score is working.

The same assessment data was collected in 2004-2005 in essentially the same way; the major differences were the lack of a December 2003 capture point and the capture of January and May data only form a third of the first-year students (the third with the lowest SAT scores), since the institution had determined that WRI-102 would be offered in the spring semester only, and that only a third of the first-year students would be required to complete the course.

Scoring

A holistic scoring approach has long been common practice in the assessment of writing performance, whether for programmatic assessment, placement, or other purposes. Holistic scoring generally relies on a rubric. All essays generated by the pre- and post-tests were scored by trained assessors (TCNJ faculty) on the same six point rubric (see Appendix A) developed for the scoring of portfolios from pilot sections of Rhetoric (the pre-transformed writing course). This rubric was in turn based on the Writing Program Administrators (WPA) outcomes for the end of the first year of college (Harrington, Malencyzk, Peckham, Rhodes, & Yancy, 2001), and compared with other rubrics such as those used by ETS for the GMAT and AP-English tests. Before the writing prompts and rubric the only assessment common to all students entering TCNJ was the SAT.

To maximize inter-rater reliability, assessment by a relatively small and well-trained group is suggested as optimal. The assessment team included fulltime, part-time, and term faculty from various disciplines, ethnicities, and gender, all of whom have experience assigning and evaluating writing in their classes. Most of the participants were teachers of Academic Writing, and several of the faculty members have prior experience assessing writing for Educational Testing Services and as part of campus-based writing assessment programs.

Training

For the essay scoring sessions, readers were asked to seat themselves three or four to a table, with certain readers designated by the project's reading leaders as table leaders. For the exit essay scoring, we had two tables of five readers each, each with a designated table leader; this seating arrangement allows for a mix of experience and ability at each table. Each time, readers were (re)acquainted with procedures at the tables and then were sent to a desktop computer to score essays.

Readers were trained at the beginning of each scoring project, and re-trained throughout the scoring at regular intervals. With the table leaders and the reading leaders facilitating the discussion, readers compared each benchmark essay to the description under the appropriate scoring point, and discuss the criteria. Several benchmark (or calibrating) essays were selected to use for the initial training session. Each benchmark or sample essay was identified by a letter and photocopied for each reader and the rubric was distributed to each reader. These sample essays scored by the leaders were given to test the readers' ability to use the rubric, but more importantly, so that through discussion the readers may approach consensus on the forms the criteria may commonly take in a particular set of essays. Moreover, every reading has idiosyncratic essays which resist easy categorization according to the rubric; some of these essays were discussed as samples later in the reading, when the readers became comfortable with one another and the rubric.

Each essay was scored independently by two readers, with a third independent reader in the case of noncontiguous scores. (Noncontiguous scores are commonly found with idiosyncratic essays). All scoring was double-blind: that is, essays were identified only by student identification number; and the readers cannot see one another's scores. Each individual score was entered into the computer by student ID, and tagged with a code to identify the reader. Scoring become faster during the assessment period and training periods shorter, as is usual during holistic scoring. The rate eventually increased and the number of third reads required subsequently diminished over the given scoring period as well as between periods, since even experienced readers become faster and more reliable over several days of scoring.

Results

What follows is a reliability analysis, construct validity of the assessment prompts, intervention analysis, SAT cut score analysis, and proposed cut score analysis. The analysis was conducted separately on data from the 2003-2004 and 2004-2005 academic years. Based on these results an informed decision was made to choose the cut score that most adequately ensures that we capture low performers and not waste the time of those who will not benefit from the intervention as well as keeping costs down for the program.

Descriptive Statistics

First the course grade distributions for WRI-102 were looked at and statistical tests were conducted to determine if there were any differences between students based on gender and EOF status.

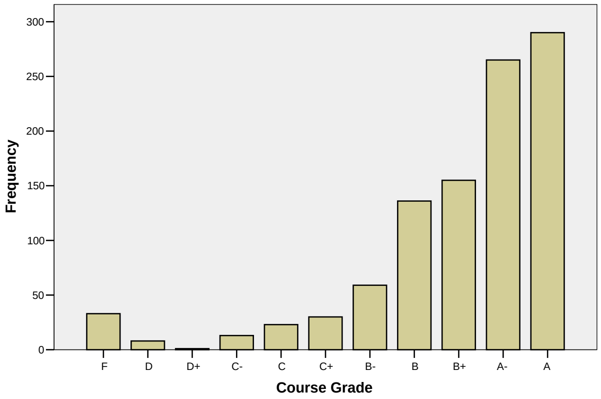
Fall 2003 Cohort

Descriptive Statistics

	N	Mean	Std. Deviation
Gender	1242	.58	.494
Ethnicity SAT Verbal	1242	.24	.425
SAT Math	1238	602.43	117.889
High School Rank	1238	626.87	121.012
Pre-Test	1216	74.08	35.157

Course Grade	Frequency	Percent
F	33	3.3
D-	0	0.0
D	8	0.8
D+	1	.01
C-	13	1.3
C	23	2.3
C+	30	3.0
B-	59	5.8
B	136	13.4
B+	155	15.3
A-	265	26.2
A	290	28.6
Total	1013	100.0

Course Grade Distribution



It was found that females outperformed males on August 2003, January 2004, and May 2004 writing prompts. EOF students performed worse than non-EOF students on August 2003, December 2003, and January 2004 writing prompts.

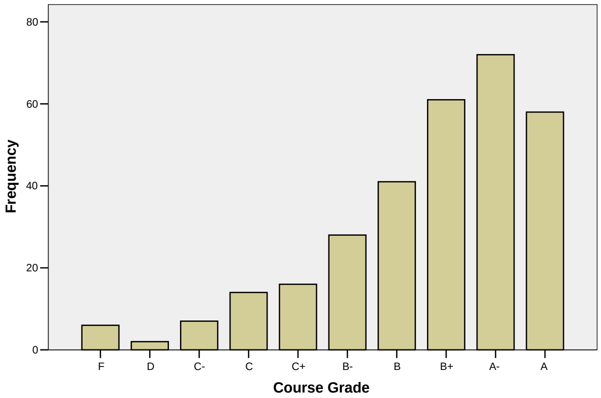
Fall 2004 Cohort

Descriptive Statistics

	N	Mean	Std. Deviation
Gender	1265	.45	.497
Ethnicity SAT Verbal	1265	.77	.424
SAT Math	1246	617.62	83.552
High School Rank	1246	635.53	83.510
Pre-Test	981	89.33	10.001

Course Grade	Frequency	Percent
F	6	2.0
D-	0	0.0
D	2	0.7
D+	0	0.0
C-	7	2.3
C	14	4.6
C+	16	5.2
B-	28	9.2
B	41	13.4
B+	61	20.0
A-	72	23.6
A	58	19.0
Total	305	100.0

Spring 2005 Course Grade Distribution



If was found that females outperformed males on each writing prompt. EOF students performed worse than non-EOF students on September 2004 and May 2005 writing prompts.

Reliability

First we looked at descriptive statistics for each reader to see if any reader stood apart from the others. The combined number of tests graded for September 2004, January 2005, and May 2005, the range of the scores given (maximum score minus minimum score), the minimum score given, the maximum score given, the mean score for that reader, and the standard deviation of the scores given by each reader are given in the following table.

Reader	N	Range	Minimum	Maximum	Mean	Std. Deviation
Reader101	211	4	1	5	3.08	0.875
Reader102	235	6	0	6	3.64	0.948
Reader104	184	5	1	6	3.65	0.869
Reader106	256	5	1	6	3.64	0.976
Reader107	191	5	1	6	3.51	0.994
Reader204	178	4	1	5	3.56	0.876
Reader207	164	4	1	5	2.93	0.869
Reader208	176	5	1	6	3.61	0.997
Reader209	185	5	1	6	3.43	0.942
Reader303	24	2	3	5	4.08	0.776

Next the intra-class correlation coefficients were calculated for each pair of readers for the September 2004, January 2005, and May 2005 exams (all exam scores were pooled together since they were scored at the same time). Values of +0.60 and above are average to above average reliabilities. Low coefficients and negative coefficients are bad reliabilities, meaning there was little to none agreement between the scores of this pair of readers, as seen in the following table containing the reader pairs and their intra-class correlation coefficient.

Readers	Intra-class Correlation Coefficient
101 & 104	0.35
101 & 106	0.65
101 & 107	-1.41
101 & 207	0.79
101 & 208	0.91
102 & 104	0.88
102 & 107	0.75
102 & 204	0.80
102 & 207	0.37
104 & 107	0.67
104 & 209	0.80
106 & 207	0.84
106 & 208	0.74
106 & 209	0.82
107 & 207	0.83
107 & 208	-0.11
107 & 209	0.66
107 & 303	0.71
204 & 207	0.77
209 & 303	0.57

Finally the average intra-class correlation coefficients were calculated for each test for the fall 2003 cohort and fall 2004 cohort. Every intra-class coefficient was above average, as seen in the following table of correlation coefficients.

Test Date	Average Intra-class Correlation Coefficient
Fall 2003 Cohort	
August 2003 (pre-test)	0.643
December 2003 (post-test)	0.705
January 2004 (pre-test)	0.654
May 2004 (post-test)	0.770
Fall 2004 Cohort	
Summer Placement Exam	0.529
September 2004	0.666
January 2005 (pre-test)	0.765
May 2005 (post-test)	0.702

Construct Validity

Construct validity refers to the consistency between a given test or questionnaire and the accepted theoretical constructs related to the subject being studied. Bivariate correlations were utilized to understand the degree that the pre-test, SAT Verbal, SAT Math, post-test, and WRI-102 interval level grade related to one another for first-time students. All of the variables were significantly correlated with each other for the fall 2003 cohort except Post-Test with SAT Verbal, SAT Math, and WRI-102 Interval Level Grade.

Fall 2003 Cohort Correlations

		SAT Verbal	SAT Math	Pre-Test	WRI 102 Interval Level Grade	Post-Test
SAT Verbal	Pearson Correlation	1	.816**	.236**	.193**	.025
	Sig. (2-tailed)		.000	.000	.000	.491
	N	1165	1165	676	879	746
SAT Math	Pearson Correlation		1	.140**	.194**	.017
	Sig. (2-tailed)			.000	.000	.646
	N		1165	676	879	746
Pre-Test	Pearson Correlation			1	.269**	.092*
	Sig. (2-tailed)				.000	.028
	N			676	671	571
WRI 102 Interval Level Grade	Pearson Correlation				1	.043
	Sig. (2-tailed)					.247
	N				879	744
Post-Test	Pearson Correlation					1
	Sig. (2-tailed)					
	N					746

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

*. Correlation is significant at the 0.05 level (2-tailed).

The next step of the analysis utilized stepwise regression techniques to better understand how specific variables functioned as predictors of the dependent variables: Pre-Test; WRI-102 Interval Level Grade; and Post-Test. Independent variables included Student Entering Status (Continuing or First-Time Freshman), Sex (Female or Male), Ethnicity (White or Non-White/Minority), SAT Verbal, SAT Math, and the Pre-Test. Descriptive statistics for the independent variables follow:

In this stage of the analysis, we entered all significant student-level bivariate correlations into a stepwise regression model for each of the dependent variables. The bivariate correlations function as a screening tool whereby only the student-level measures that significantly relate to the dependent variable are entered into the stepwise regression analysis. This stage of the analysis can only be conducted for the fall 2003 cohort, where all students took WRI-102.

Each table provides key information for the entered variables: (1) order of entry, (2) cumulative percent of total variance explained for the dependent variable, and (3) respective proportion of the explained variance accounted for by the variable.

DV: Pre-Test

Significant bivariate correlations for the dependent measure Pre-Test included Major (<.05), Gender (<.05), SAT Verbal (>.001), and SAT Math (>.001). We entered these into a stepwise model and SAT Verbal and SAT Math significantly contributed to the explained variance for the dependent variable.

Variable	*Order of Entry*	*Percent of Additional Variation Explained by the Variable*	*Proportion of Explained Variance Accounted for By Variable*
SAT Verbal	1	5.6%	86.2%
SAT Math	2	6.5%	15.4%

DV: Post-Test

Significant bivariate correlations for the dependent measure Post-Test only included Pre-Test (>.05). Pre-Test was entered these into a stepwise model and was found that it significantly contributed to the explained variance for the dependent variable.

Variable	*Order of Entry*	*Percent of Additional Variation Explained by the Variable*	*Proportion of Explained Variance Accounted for By Variable*
Pre-Test	1	0.8%	100%

DV: WRI-102 Interval Level Grade

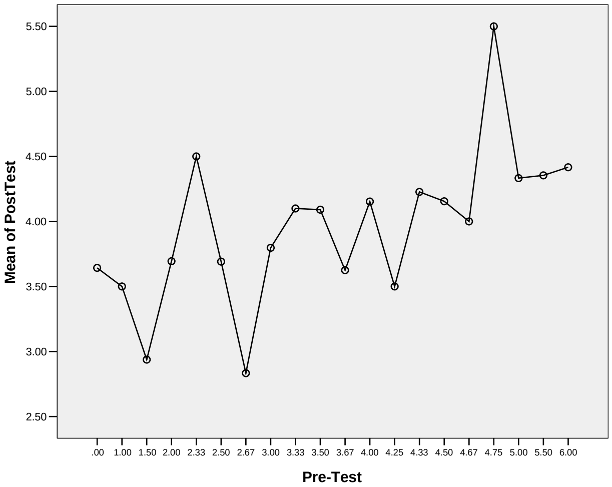
Significant bivariate correlations for the dependent measure Pre-Test included Gender (<.001), Ethnicity (<.001), SAT Verbal (>.001), and SAT Math (>.001). We entered these into a stepwise model and SAT Math, Gender, and Ethnicity significantly contributed to the explained variance for the dependent variable.

Variable	*Order of Entry*	*Percent of Additional Variation Explained by the Variable*	*Proportion of Explained Variance Accounted for By Variable*
SAT Math	1	3.8%	45.2%
Gender	2	6.3%	29.8%
Ethnicity	3	8.4%	26.2%

Intervention

When several scores are taken on the same student, the scores tend to be correlated with each other. This correlation can be taken into account through repeated measures analysis of variance. Repeated measures are useful in dealing with interventions, i.e. where you compare values of a dependent variable before and after you try something like a program. Here repeated measures is used on the pre and post-tests to better understand WRI-102 intervention and the groups of students impacted most by this intervention. Repeated measures is also used to better understand FSP impact for the fall 2003 cohort, since we have a post-test for FSP students for this group. The intervention is an extension of the construct validity analysis because of the further analysis of the pre-test writing prompt's relationship with WRI-102.

We first started by conducting a one-way ANOVA and looking at the means plot to better understand how WRI-102 impacted students in groups based on their pre-test scores. Here is the means plot:



As can be seen in the fall 2003 cohort means plot, those with a pre-test score higher than 4 were helped the least by WRI-102. Next the repeated measures analysis of variance was conducted to determine if WRI-102 had a significant intervention. WRI 102 seemed to have a positive intervention effect on the fall 2003 cohort, as seen below. WRI-102 intervention is significant at the 0.1 level for the fall 2003 cohort. The mean post-test score was higher than the mean pre-test score, which also suggests positive intervention from WRI-102.

Fall 2003 Cohort

Parameter Estimates

						95% Confidence Interval	
Dependent Variable	Parameter	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
Pre-Test	Intercept	3.517	.038	93.698	.000	3.443	3.591
Post-Test	Intercept	3.992	.035	113.358	.000	3.923	4.061

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
score	Sphericity Assumed	65.851	1	65.851	112.067	.000
	Greenhouse-Geisser	65.851	1.000	65.851	112.067	.000
	Huynh-Feldt	65.851	1.000	65.851	112.067	.000
	Lower-bound	65.851	1.000	65.851	112.067	.000
Error(score)	Sphericity Assumed	342.573	583	.588		
	Greenhouse-Geisser	342.573	583.000	.588		
	Huynh-Feldt	342.573	583.000	.588		
	Lower-bound	342.573	583.000	.588		

Proposed Cut Score Descriptive Findings

First, we reviewed frequencies by pre- and post-test score breakdowns for the fall 2003 cohort WRI-102 participants. Then T-Tests were used to make determinations about the significance level of differences relative to the mean for the four variables. T-Tests were conducted for each possible cut score on the pre-test. These analyses were conducted for each cohort.

The following table shows the means for SAT Verbal, SAT Math, WRI-102 Interval Level Grade, and Post-Test for each score on the Pre-Test. As shown in the table, those with a pre-test score higher than 3.50 received post-test scores lower than their pre-test score. This suggests that WRI-102 positively intervened with those who scored 3.50 or less on the pre-test.

Pre-Test Scor	SAT Verbal		SAT Math		WRI 102 Interval Level Grade		Post-Test	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean
.00	8	510	8	560	8	2.92	8	3.64
.50	1	500	1	440	1	2.67	1	
1.00	3	560	3	613	3	2.00	3	3.50
1.50	11	492	11	527	11	2.76	11	2.94
2.00	27	488	27	542	27	2.26	27	3.69
2.33	4	568	4	625	4	2.92	4	4.50
2.50	72	567	72	609	72	3.14	72	3.69
2.67	3	607	3	717	3	3.56	3	2.83
3.00	107	576	107	611	107	3.32	107	3.80
3.33	6	582	6	665	6	3.45	6	4.10
3.50	140	597	140	632	140	3.43	140	4.09
3.67	10	654	10	671	10	3.53	10	3.63
4.00	151	598	151	622	151	3.46	151	4.15
4.25	1	580	1	800	1	4.00	1	3.50
4.33	14	596	14	664	14	3.45	14	4.23
4.50	67	623	67	645	67	3.63	67	4.15
4.67	4	660	4	690	4	2.59	4	4.00
4.75	1	630	1	500	1	4.00	1	5.50
5.00	33	622	33	630	33	3.77	33	4.33
5.50	9	599	9	574	9	3.48	9	4.35
6.00	4	685	4	653	4	3.09	4	4.42
Total	**676**	**588**	**676**	**620**	**676**	**3.35**	**676**	**3.99**

The following table shows the means for SAT Verbal, SAT Math, WRI-102 Interval Level Grade, and Pre-Test for each score on the Post-Test.

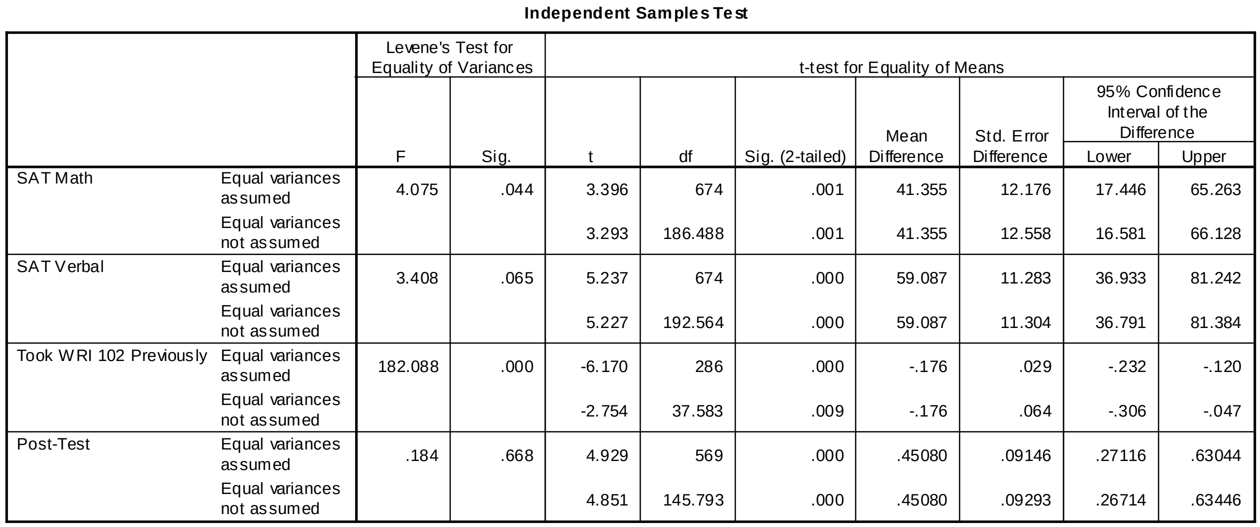
Post-Test Scor	SAT Verbal		SAT Math		WRI 102 Interval Level Grade		Pre-Test	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean
.00	3	640	3	727	3	3.89	3	3.50
1.00	3	567	3	637	3	3.56	3	2.67
1.50	5	526	5	558	5	3.20	5	2.56
2.00	5	622	5	610	5	3.20	5	3.25
2.50	20	537	20	564	20	3.30	20	3.00
3.00	80	565	80	608	80	3.24	80	3.07
3.33	10	588	10	606	10	3.40	10	3.75
3.50	147	576	147	616	147	3.34	147	3.52
3.67	4	570	4	613	4	3.00	4	3.33
4.00	172	575	172	613	172	3.43	172	3.48
4.33	15	597	15	623	15	3.04	15	3.61
4.50	135	593	135	629	135	3.51	135	3.66
4.67	13	610	13	646	13	3.36	13	3.50
5.00	92	598	92	633	92	3.55	92	3.77
5.33	1	690	1	660	1	3.00	1	6.00
5.50	32	628	32	633	32	3.48	32	3.97
6.00	9	668	9	644	9	3.89	9	4.33
Total	**746**	**584**	**746**	**619**	**746**	**3.41**	**746**	**3.53**

Those who receive a pre-test score of 3.50 or less had mean SAT Verbal and Math scores of 572 and 611 respectively, whereas those who received higher than a 3.50 had means of 610 and 632. These frequency tables suggest using a 3.50 as a cut score on the writing assessment and utilizing SAT score scores to around a 580 on the SAT Verbal and 620 on the SAT Math.

T-Tests suggested utilizing a 3.00 cut seemed to be the best option. This cut also suggests that cut scores of 550 and 590 on the SAT Verbal and SAT Math respectively should be utilized.

Group Statistics

	Pre-Test	N	Mean	Std. Deviation	Std. Error Mean
SAT Math	>= 3.00	547	628.10	123.184	5.267
	< 3.00	129	586.74	129.477	11.400
SAT Verbal	>= 3.00	547	599.71	115.214	4.926
	< 3.00	129	540.62	115.561	10.175
Took WRI 102 Previously	>= 3.00	250	.01	.089	.006
	< 3.00	38	.18	.393	.064
Post-Test	>= 3.00	469	4.0750	.83341	.03848
	< 3.00	102	3.6242	.85425	.08458



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
SAT Math	Equal variances assumed	4.075	.044	3.396	674	.001	41.355	12.176	17.446	65.263
	Equal variances not assumed			3.293	186.488	.001	41.355	12.558	16.581	66.128
SAT Verbal	Equal variances assumed	3.408	.065	5.237	674	.000	59.087	11.283	36.933	81.242
	Equal variances not assumed			5.227	192.564	.000	59.087	11.304	36.791	81.384
Took WRI 102 Previously	Equal variances assumed	182.088	.000	-6.170	286	.000	-.176	.029	-.232	-.120
	Equal variances not assumed			-2.754	37.583	.009	-.176	.064	-.306	-.047
Post-Test	Equal variances assumed	.184	.668	4.929	569	.000	.45080	.09146	.27116	.63044
	Equal variances not assumed			4.851	145.793	.000	.45080	.09293	.26714	.63446

Construct Validity Revisited

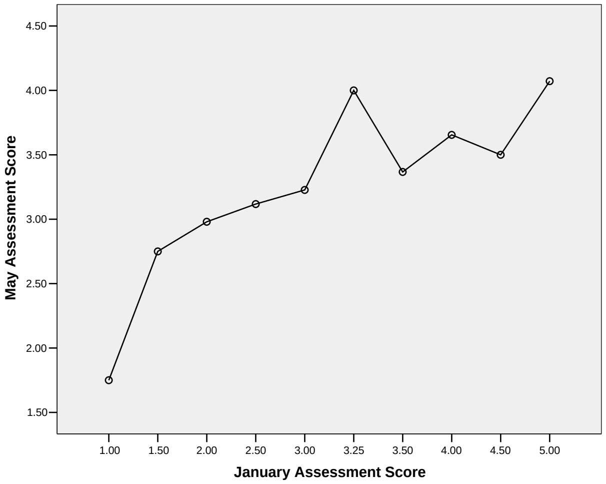
Bivariate correlations for the fall 2004 cohort were only conducted on first time freshman who did not have an SAT Verbal or Math score of 580 or better. This is because a 580 cut score on both the SAT Verbal and SAT Math was in place for this cohort. Here SAT Verbal is significantly correlated with SAT Math and the Pre-Test is significantly correlated with the Post-Test. The lack of correlations that were present with the fall 2004 cohort is due to the reduction in variation from selecting only those students who did not meet the SAT cut score. These correlations indicate that the current cut score is working, but don’t tell us to what extent. The Pre-Test and Post-Test are still good indicators of WRI-102 intervention.

Correlations

		SAT Verbal	SAT Math	Pre-Test	WRI 102 Grade (W=missing)	Post-Test
SAT Verbal	Pearson Correlation	1	.252**	.056	.111	-.045
	Sig. (2-tailed)		.000	.627	.299	.702
	N	387	387	77	90	76
SAT Math	Pearson Correlation		1	-.099	.072	-.099
	Sig. (2-tailed)			.392	.499	.393
	N		387	77	90	76
Pre-Test	Pearson Correlation			1	.030	.340**
	Sig. (2-tailed)				.790	.005
	N			81	79	66
WRI 102 Grade (W=missing)	Pearson Correlation				1	.210
	Sig. (2-tailed)					.069
	N				95	76
Post-Test	Pearson Correlation					1
	Sig. (2-tailed)					
	N					79

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

The fall 2004 cohort means plot shows that those with a pre-test score of 3.50 or higher were impacted least by WRI-102. WRI 102 did not have a positive intervention on the fall 2004 cohort, but rather only with those who scored lower than a 3.50 on the pre-test. The intervention on this group is shown in the ANOVA tables and estimate mean scores tables below. WRI-102 is significant at the 0.001 level for the fall 2004 cohort for WRI-102. The mean post-test score was higher than the mean pre-test score, which also suggests positive intervention from WRI-102.



Estimates

Measure: MEASURE_1

SCORE	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	2.586	.045	2.497	2.674
2	3.092	.084	2.926	3.258

1 = January (pre-test), 2 = May (post-test)

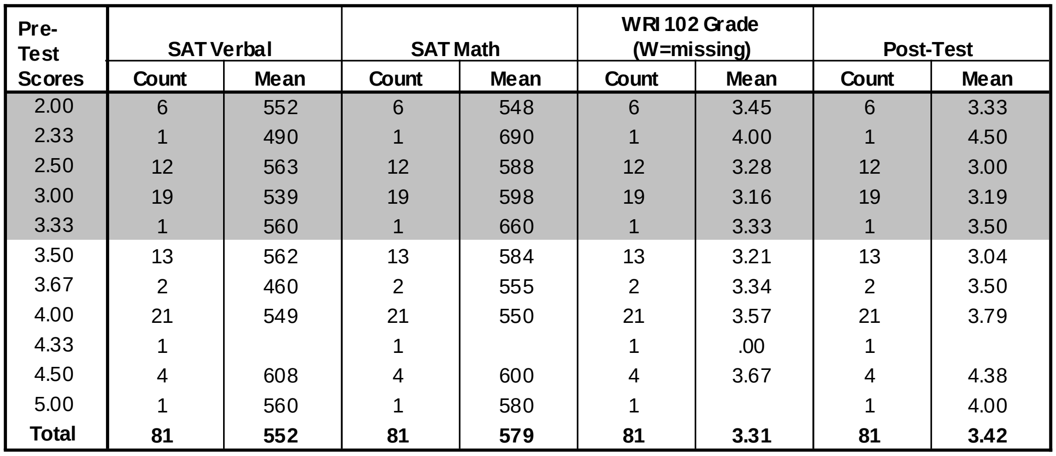
Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
SCORE	Sphericity Assumed	14.627	1	14.627	34.148	.000
	Greenhouse-Geisser	14.627	1.000	14.627	34.148	.000
	Huynh-Feldt	14.627	1.000	14.627	34.148	.000
	Lower-bound	14.627	1.000	14.627	34.148	.000
Error(SCORE)	Sphericity Assumed	48.404	113	.428		
	Greenhouse-Geisser	48.404	113.000	.428		
	Huynh-Feldt	48.404	113.000	.428		
	Lower-bound	48.404	113.000	.428		

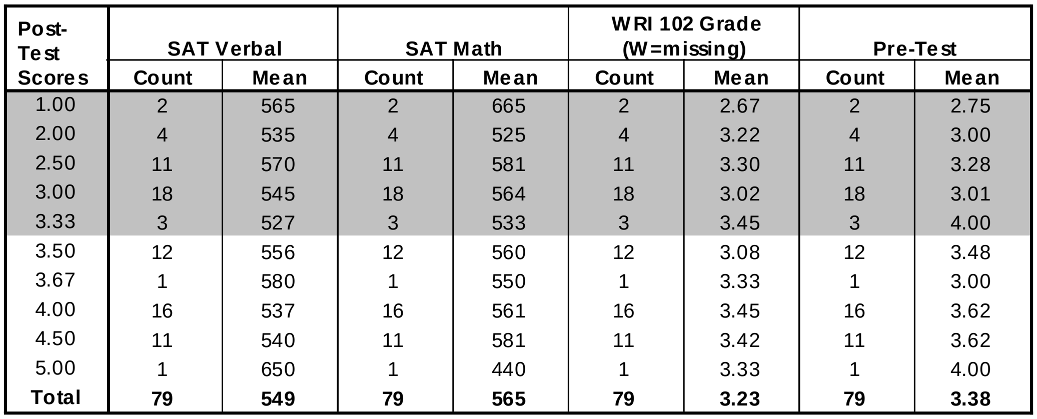
Refined Cut Score Analysis

Unlike 2003, students were allowed to place out of WRI-102 this year if they had a 580 on both the SAT Verbal and SAT Math, or a 4.00 or better on the writing pre-test. The same cut score analysis was performed on this year to see if the cut score would change. The following table shows the means for SAT Verbal, SAT Math, WRI-102 Interval Level Grade, and Post-Test for each score on the Pre-Test. As shown in the table, those with a pre-test score of 3.50 or higher received post-test scores lower than their pre-test score. This suggests that WRI-102 positively intervened with those who scored less than a 3.50 on the pre-test.



Pre-Test Scores	SAT Verbal		SAT Math		WRI 102 Grade (W=missing)		Post-Test	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean
2.00	6	552	6	548	6	3.45	6	3.33
2.33	1	490	1	690	1	4.00	1	4.50
2.50	12	563	12	588	12	3.28	12	3.00
3.00	19	539	19	598	19	3.16	19	3.19
3.33	1	560	1	660	1	3.33	1	3.50
3.50	13	562	13	584	13	3.21	13	3.04
3.67	2	460	2	555	2	3.34	2	3.50
4.00	21	549	21	550	21	3.57	21	3.79
4.33	1		1		1	.00	1	
4.50	4	608	4	600	4	3.67	4	4.38
5.00	1	560	1	580	1		1	4.00
Total	**81**	**552**	**81**	**579**	**81**	**3.31**	**81**	**3.42**

The following table shows the means for SAT Verbal, SAT Math, WRI-102 Interval Level Grade, and Pre-Test for each score on the Post-Test.



Post-Test Scores	SAT Verbal		SAT Math		WRI 102 Grade (W=missing)		Pre-Test	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean
1.00	2	565	2	665	2	2.67	2	2.75
2.00	4	535	4	525	4	3.22	4	3.00
2.50	11	570	11	581	11	3.30	11	3.28
3.00	18	545	18	564	18	3.02	18	3.01
3.33	3	527	3	533	3	3.45	3	4.00
3.50	12	556	12	560	12	3.08	12	3.48
3.67	1	580	1	550	1	3.33	1	3.00
4.00	16	537	16	561	16	3.45	16	3.62
4.50	11	540	11	581	11	3.42	11	3.62
5.00	1	650	1	440	1	3.33	1	4.00
Total	**79**	**549**	**79**	**565**	**79**	**3.23**	**79**	**3.38**

Those who receive a pre-test score of 3.33 or less had mean SAT Verbal and Math scores of 548 and 592 respectively, whereas those who received higher than a 3.33 had means of 555 and 567. These frequency tables suggest using a 3.33 as a cut score on the writing assessment and utilizing SAT score scores to around a 550 on the SAT Verbal and 590 on the SAT Math.

T-Tests suggested utilizing a 4.00 cut seemed to be the best option. This cut also suggests that cut scores of 550 and 580 on the SAT Verbal and SAT Math respectively should be utilized.

Group Statistics

	Pre-Test	N	Mean	Std. Deviation	Std. Error Mean
SAT Math	>= 4.00	25	558.80	64.635	12.927
	< 4.00	52	588.65	81.265	11.269
SAT Verbal	>= 4.00	25	558.80	91.802	18.360
	< 4.00	52	548.08	62.780	8.706
WRI 102 Grade (W=missing)	>= 4.00	25	3.4400	.79798	.15960
	< 4.00	54	3.2535	.56770	.07725
Post-Test	>= 4.00	23	3.8986	.66237	.13811
	< 4.00	43	3.1667	.82375	.12562

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
SAT Math	Equal variances assumed	2.511	.117	-1.607	75	.112	-29.854	18.579	-66.865	7.157
	Equal variances not assumed			-1.741	58.454	.087	-29.854	17.150	-64.177	4.469
SAT Verbal	Equal variances assumed	3.003	.087	.601	75	.550	10.723	17.846	-24.828	46.274
	Equal variances not assumed			.528	35.169	.601	10.723	20.320	-30.522	51.968
WRI 102 Grade (W=missing)	Equal variances assumed	.014	.908	1.189	77	.238	.18648	.15683	-.12581	.49877
	Equal variances not assumed			1.052	35.678	.300	.18648	.17731	-.17323	.54620
Post-Test	Equal variances assumed	.832	.365	3.669	64	.000	.73188	.19945	.33343	1.13034
	Equal variances not assumed			3.920	54.072	.000	.73188	.18670	.35759	1.10618

Discussion

T-tests for the 2003-2004 academic year yield a cut score of 3.00 for the writing prompt and SAT cut scores of 550 and 590 on the verbal and math respectively. The intervention analysis and frequency tables show that there is intervention with those who scored higher than a 3.00 on the pre-test. T-tests for the 2004-2005 academic year suggest leaving the cut score at 4.00 for the writing prompts and SAT cut scores of 550 and 580 on the verbal and math respectively.

Being that a cut score can make two types of errors, we would prefer the error of placing students who do not need WRI-102 as opposed to not placing those who do need it. Cost is another factor in choosing the final cut scores. Leaving the SAT cut score at 580 for both the verbal and the math, we expect to assess approximately 30% (260-265) of all incoming students with a pre-test for placement purposes. Utilizing a cut score of 4.00, we expect approximately 25% (35-40) students in this group will be exempt from the pre-test. This leaves approximately 220 to 225 students who will need a place in a section of WRI-102. Utilizing these cut scores seems to provide a good balance between cost and error.

The scoring rubric describes a “3 paper” as one that “may attempt to argue a position that is uneven in focus or development.” Students who are writing at this level of sophistication have only a “limited” facility with “the core of academic discourse”—“the well-ordered argument, grounded in efficiently presented evidence, clearly articulated in elegant, coherent prose.” The scoring rubric describes a “4 paper” as one that “argues a position, provides supporting detail, and generally demonstrates control of the elements of writing.” Unlike a 3 paper, a paper written at the 4 level must have all of the strengths identified in the score point. Students writing at a 4 level in the first year of college should be prepared to fulfill The College’s mission to “create, preserve, and transmit knowledge, the arts, and wisdom” through writing in progressively more sophisticated levels within their undergraduate curricula. While students at the 4 score point may not write the most well supported arguments, their prose “[d]isplays overall organization” and “[d]emonstrates unified and coherent ideas within paragraphs.” Unlike students who write at the 3 level, students writing at the 4 level exhibit “generally clear and effective control of language” and demonstrate “competence with the examination.” Students writing at the 3 level at the beginning of their college career may have difficulty expressing in writing their intellectual master of coursework.

Given the strong reliability and validity findings relative to developing the prompts, prompt scoring procedures, WRI-102 course implementation, construct validity, and descriptive findings, a cut score of 4.00 functions as an adequate benchmark for making determinations about placements in WRI-102. Given the strong relationship of SAT Verbal and SAT Math with the WRI-102 interval level grade and pre-test and post-test findings we suggest utilizing a 580 cut score on both the SAT Verbal and SAT Math to make a determination about who should or should not sit for a placement assessment. What this means is that a student with a 580 on both the SAT Verbal and SAT Math will not need to sit for the prompt.

APPENDIX A: SCORING GUIDE version 6

6 Outstanding

A 6 paper skillfully argues a clear and specific position supported with relevant evidence and demonstrates excellent control of the elements of writing. A typical paper in this category exhibits all of the following characteristics:

a. Presents a compelling, clear and debatable claim which is focused and specific;

b. Provides ample relevant, concrete evidence and persuasive support for every debatable assertion by synthesizing information and arguments from multiple, reliable sources, summarizing them fairly, and assessing them critically;

c. Displays clear and consistent overall organization that relates all of the ideas together;

d. Develops ideas cogently, organizes them logically within paragraphs, and connects them with highly effective transitions;

e. Demonstrates outstanding control of language, including effective word choice and sentence variety;

f. Demonstrates superior facility with the conventions of standard written English (i.e. grammar, usage, and mechanics) but may have minor errors.

5 Strong

A 5 paper competently argues a position, provides relevant supporting detail, and demonstrates good control of the elements of writing. A typical paper in this category exhibits all of the following characteristics:

a. Presents an interesting, clear, and debatable claim;

b. Provides relevant, concrete evidence and persuasive support for most debatable assertions by using multiple reliable sources, but does not always assess them critically;

c. Displays clear and consistent overall organization that relates most of the ideas together;

d. Develops unified and coherent ideas within paragraphs with clear transitions;

e. Demonstrates strong control of language including appropriate word choice and sentence variety;

f. Demonstrates facility with the conventions of standard written English (i.e. grammar, usage, and mechanics) but may have minor errors.

4 Adequate

A 4 paper argues a position, provides supporting detail, and generally demonstrates control of the elements of writing. A typical paper in this category exhibits all of the following characteristics:

a. Presents a claim that raises some debate, but may lack some specificity;

b. Provides evidence and support for most assertions by using sources, some of which may be unreliable and used uncritically;

c. Displays overall organization, but some ideas may seem illogical and/or unrelated;

d. Demonstrates unified and coherent ideas within paragraphs with generally adequate transitions;

e. Demonstrates generally clear and effective control of language;

f. Demonstrates competence with the conventions of standard written English (i.e. grammar, usage, and mechanics) but may have some errors.

3 Limited

A 3 paper may attempt to argue a position that is uneven in its focus and/or development; or it may demonstrate uneven control of the elements of writing. A typical paper in this category exhibits one or more of the following characteristics:

a. Presents a claim that is vague, limited and/or barely debatable;

b. Provides little analysis or persuasive reasoning, uses limited sources, and/or relies predominantly on sweeping generalizations, narration, description, or summary, or goes off its claim or focus;

c. Demonstrates uneven and/or ineffective overall organization;

d. Generally develops and organizes ideas in paragraphs which are not necessarily connected with transitions;

e. Displays problems in word choice and sentence structure which sometimes interfere with meaning; sentences may be inadequately varied;

f. Contains occasional major or frequent minor errors in grammar, usage, and/or mechanics that can interfere with meaning.

2 Seriously Limited

A 2 paper may assert a position that is unfocused, and/or undeveloped; or it may demonstrate little control of the elements of writing. A typical paper in this category exhibits one or more of the following characteristics:

a. Presents a claim that is not clear, consistent, or debatable;

b. Lacks analysis or persuasive reasoning and/or relies solely on narration, description and/or summary of sources; the essay is likely to go off its claim or focus;

c. Displays no consistent overall organization;

d. Does not develop ideas cogently, organize them logically within paragraphs and/or connect them with clear transitions;

e. Displays problems in word choice and sentence structure that frequently interfere with meaning; sentences are inadequately varied in structure;

f. Contains a combination of errors in grammar, usage, and/or mechanics that frequently interfere with meaning.

1 Fundamentally Deficient

A 1 paper attempts to address the topic, but without success. A typical paper in this category exhibits one or more of the following characteristics:

a. Presents no claim;

b. Presents no relevant support;

c. Presents ideas non-sequentially;

d. Uses language and style that are inappropriate for the given audience, purpose, and/or occasion;

e. Contains few sentences that are free of errors which consistently interfere with meaning.

0 Off topic

- Keystrokes; written in a foreign language; or no reference to topic.

<u>N/A Not Applicable</u>

- Students who should be out of sample (e.g., taken ill, submitted a completely blank document).

Reward the writers for what they do. The score for an exceptionally well-written paper may be raised by one point above the otherwise appropriate score. In no case should a poorly written essay be scored higher than a three

References

Buckendahl, C., Impara, J. C., Giraud, G., & Irwin, P. M. (2000). The consequences of judges making advanced estimates of impact on a cut score. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Buckendahl, C. W., Plake, B. S., & Impara, J. C. (1999). Setting minimum passing scores on high-stakes assessments that combine selected and constructed response formats. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Cross, L. H., Frary, R. B., Kelly, P. P., Small, R. C., & Impara, J. C. (1985). Establishing minimum standards for essays: Blind versus informed reviews. *Journal of Educational Measurement, 22 (2),* 137-146.

Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment, 8 (4),* 360-362.

Hambleton, R. K. (1978a). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement, 15 (4),* 277-290.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8,* 41-55.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978b). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48 (1),* 1-47.

Harrington, R. M., Malencyzk, R., Peckham, I., Rhodes, K., & Yancy, K. B. (2001). WPA outcomes statement for first-year composition. *College English, 63(3),* 321-325.

Hurtz, G. M., & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and psychological Measurement, 59 (6),* 885-897.

Impara, J. C. Giraud, G., & Plake, B. S. (2000). The influence of providing target group descriptors when setting a passing score. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Impara, J., & Plake, B. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35 (1),* 69-81.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8 (1),* 15-40.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64 (3),* 425-461.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50 (9),* 741-749.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. *Journal of Educational Measurement, 20 (3),* 283-292.

Plake, B. & Giraud, G. (1998). Effect of a modified Angoff strategy for obtaining item performance estimates in a standard setting study. Paper presented at the annual meeting of The American Educational Research Association, San Diego, CA.

Plake, B. S., & Hambleton, R. K. (2000). A standard setting method designed for complex performance assessments: Categorical assignments of student work. *Educational Assessment, 6,* 197-215.

Plake, B.S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field test results. *Educational and Psychological Measurement*, *57,* 400-412.