MALLA Teagle Study: Report of the Reading of Senior Papers, June 2009

Tim Schermer, Director of Institutional Research, Augustana College
September 2009

Introduction

This is a report of the results from reading and scoring papers on writing and critical-thinking rubrics that were developed by the six MALLA institutions. The papers were argumentative papers written by seniors who entered as first-time students in the fall of 2005. Each institution collected approximately 60 senior papers that were brought to Augustana College in June 2009 for a three day session in which faculty from the six institutions read and scored each paper. The same rubrics were used in earlier sessions to score papers from first-year, junior and senior students, making value-added comparisons of growth possible.

The Reading/Scoring Process

The reading of papers was done by 32 faculty from the six institutions. The faculty represented a variety of disciplines including English (11), natural sciences/math (6), social sciences/religion (7), education (3), speech communications (2), and foreign languages (2). At the paper-reading sessions, the faculty were split into two groups with one group scoring the papers using a writing rubric and the other scoring the papers with a critical-thinking rubric. To improve reliability, each group was led by an experienced leader/trainer who conducted a training session in which anchor papers were read and discussed. Each paper was read by two readers for each of writing and critical thinking, and if the two readers disagreed by more than one unit on the Overall/Holistic rating, a third reading was done and the outlier reading rejected. For each rubric scale, the average of the scores from the two closest readings was used in the subsequent analysis.

Covariates for Analysis

To aid in the analysis of the paper scores, each institution provided data on the student authors that included the ACT score, high school rank, end-of academic-year-2008/09 GPA, primary major, and gender. In addition, the length in pages of the body of each paper and the length of the bibliography were computed. We also attempted to gather for each paper the discipline, date the paper was due, the weight of the paper in the final grade, whether the paper was revised after an initial instructor reading, and whether the paper was peer reviewed, but this data was not available from all schools, so was not used in the regression analysis procedures described below.

Sample Representativeness

The samples of papers gathered by institution were convenience samples – basically what suitable papers the IR director or other administrator involved at each institution could cajole from faculty. We found one impediment to this type of study is that the assessment administrators do not have the clout needed with faculty to be able to gather materials meeting design objectives for sample type and representativeness. Indeed, our samples were generally not representative of the senior cohorts as a whole on the basis of ACT score, where the average ACT scores of the student authors was generally at least a point higher, or by gender or major. For example, for school White 41% of the papers were from English majors, a much higher percentage than in the general cohort. Consequently, when comparing

institutional means for seniors, a step in the analysis below was to first use regression to compute adjusted scores after controlling for ACT scores, primary major, etc.

Results – Descriptive Statistics for Papers and Student Authors

In keeping with our confidentiality agreement, the results are discussed below with the names of the schools masked by using colors.

Our target of at least 50 papers per school was substantially but not entirely achieved.

Table 1: Papers Scored by School

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Blue	54	15.8	15.8	15.8
	Gold	64	18.7	18.7	34.5
	Rust	56	16.4	16.4	50.9
	Salmon	49	14.3	14.3	65.2
	Silver	59	17.3	17.3	82.5
	White	60	17.5	17.5	100.0
	Total	342	100.0	100.0	

Females were generally overrepresented in the samples:

Table 2: Student Authors by Gender

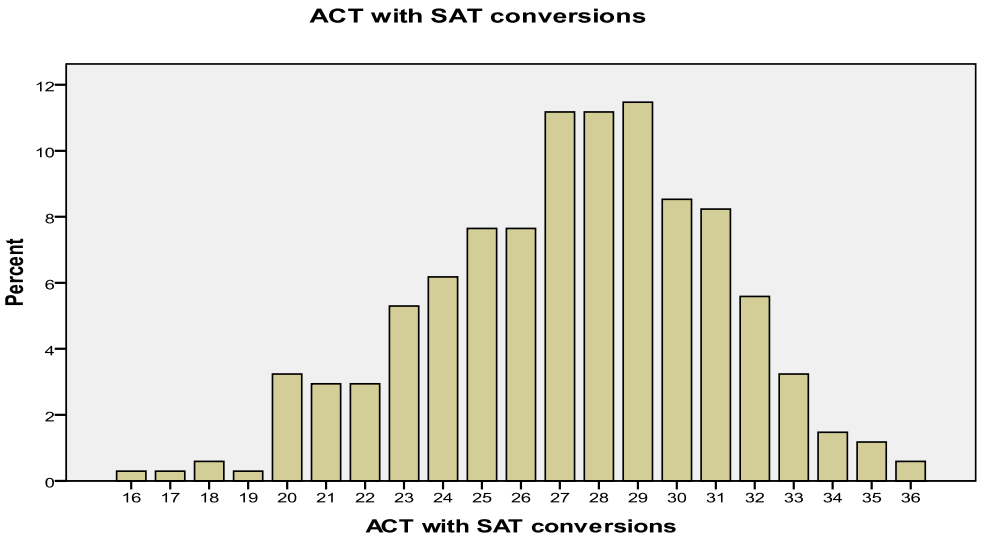
			Gender		
			Male	Female	Total
school	Blue	Count	20	34	54
		% within school	37.0%	63.0%	100.0%
	Gold	Count	18	46	64
		% within school	28.1%	71.9%	100.0%
	Rust	Count	24	32	56
		% within school	42.9%	57.1%	100.0%
	Salmon	Count	17	32	49
		% within school	34.7%	65.3%	100.0%
	Silver	Count	23	36	59
		% within school	39.0%	61.0%	100.0%
	White	Count	24	36	60
		% within school	40.0%	60.0%	100.0%
Total		Count	126	216	342
		% within school	36.8%	63.2%	100.0%

The ACT scores and college GPAs were generally higher for the sampled authors than for the senior cohort as a whole. The average weight of the papers in the grade for the courses was 28% and the

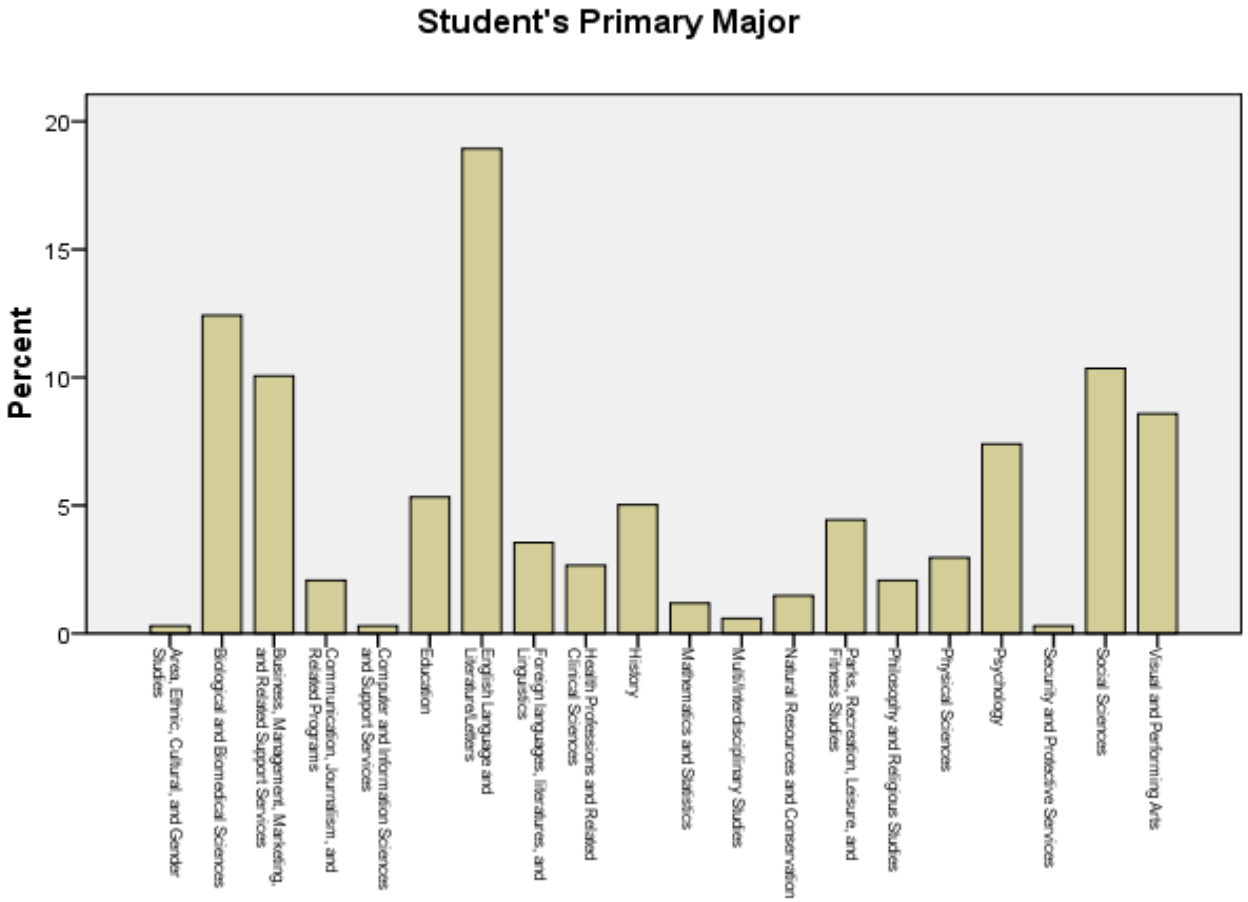
papers had an average length of 9.1 pages. The variations among the schools in paper length and student author characteristics were considerable.

Means of Author Variables and Assignment/Paper Variables by School

School		ACT with SAT conversions	HS RANK	EOAY 0809 GPA	Percent of assignment weight in course grade	Number of pages in the paper body	Number of pages in the bibliography
Blue	Mean	27.6	86.8	3.4	24.1	6.7	.7
	StdDev	3.4	16.9	.5	5.6	2.2	.7
Gold	Mean	27.4	89.8	3.6		7.2	.8
	StdDev	3.3	13.3	.3		3.3	.5
Rust	Mean	26.6	79.4	3.3	27.7	7.5	.7
	StdDev	3.0	18.2	.4	9.5	2.1	.5
Salmon	Mean	26.3	85.6	3.5		13.0	1.5
	StdDev	4.0	11.5	.4		8.0	1.8
Silver	Mean	29.3	87.2	3.5	23.7	7.7	.8
	StdDev	2.9	11.1	.4	12.4	3.1	.8
White	Mean	26.7	82.5	3.3	35.3	12.9	1.2
	StdDev	4.6	14.5	.5	22.9	6.4	.9
Total	Mean	27.3	85.5	3.4	28.1	9.1	.9
	StdDev	3.7	14.7	.4	15.5	5.3	1.0



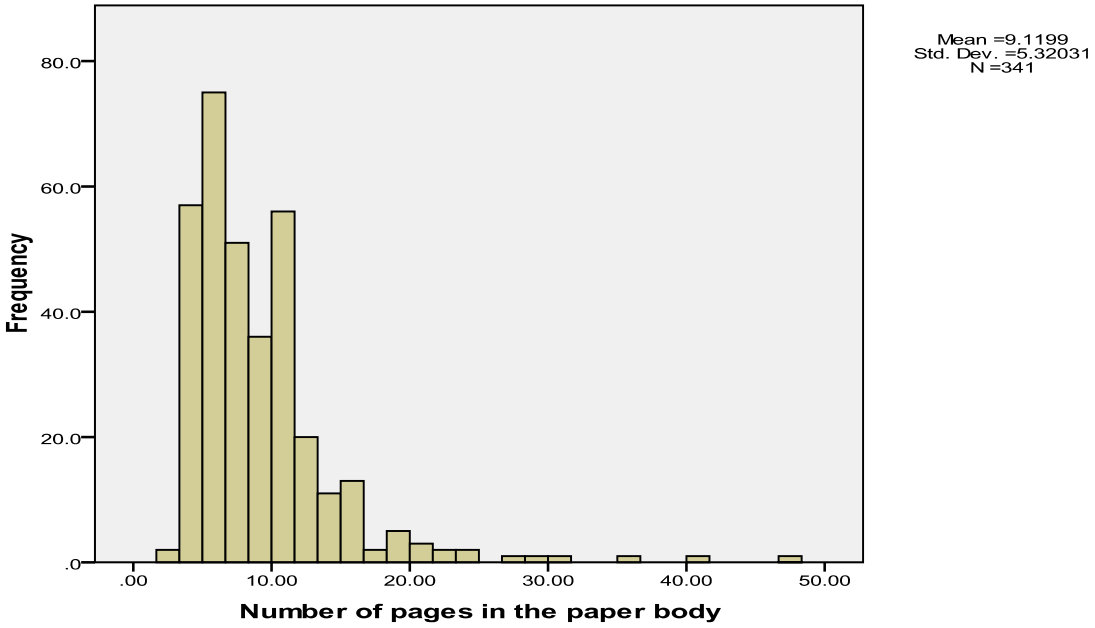
The principle majors of the student authors were varied, but English and Foreign Language majors appear to have been overrepresented.



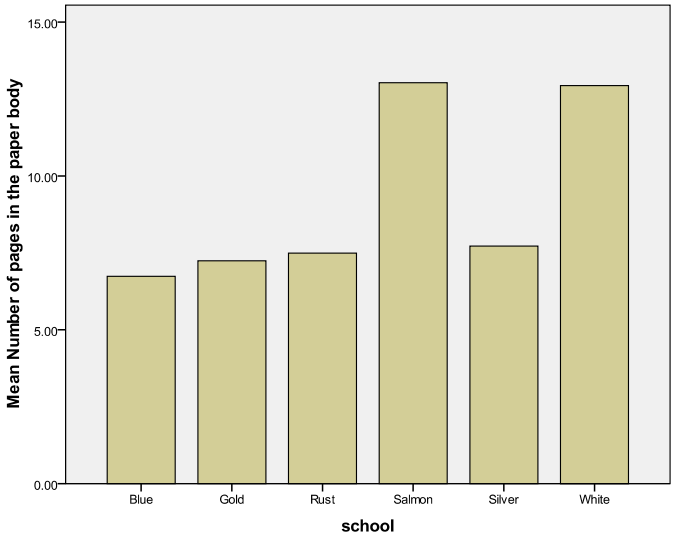
							Total
		Visual and Performing Arts	English/Foreign Language/Philo-sophyReligion	Natural Sciences/Math	Pre-Professional	Social Sciences	
Blue	Count	8	7	20	8	11	54
	% within school	14.8%	13.0%	37.0%	14.8%	20.4%	100.0%
Gold	Count	4	17	12	21	10	64
	% within school	6.3%	26.6%	18.8%	32.8%	15.6%	100.0%
Rust	Count	4	13	15	12	10	54
	% within school	7.4%	24.1%	27.8%	22.2%	18.5%	100.0%
Salmon	Count	2	7	8	16	14	47
	% within school	4.3%	14.9%	17.0%	34.0%	29.8%	100.0%
Silver	Count	9	13	5	11	21	59
	% within school	15.3%	22.0%	8.5%	18.6%	35.6%	100.0%
White	Count	2	26	11	7	12	58
	% within school	3.4%	44.8%	19.0%	12.1%	20.7%	100.0%
Total	Count	29	83	71	75	78	336
	% within school	8.6%	24.7%	21.1%	22.3%	23.2%	100.0%

Twenty-nine percent of the papers in our senior sample were resubmitted after required revision; 11% were peer reviewed.

The average length of the body of the papers was 9.1 pages, and the average length of the bibliography was 0.9 pages. Schools were requested to keep the length of the papers down to about 15 pages for convenience in scoring, so the average of 9 pages in our sample is not necessarily representative of the typical length of senior papers at the institutions. The analysis below of our sample indicates a positive correlation between scores and the paper length, so an unintended consequence of our attempt to limit the length of papers in the sample may have been to underestimate the general value added growth of all seniors. Longer pagers, such as seminar papers, were rare in our sample.



The number of pages also varied considerably by school:



Rubric Analysis

The writing and critical-thinking rubrics used for scoring were developed by the MALLA faculty, five from each school. The critical-thinking rubric consisted of scores in eight sub areas plus a holistic rating, and used a six-point scale. The writing rubric consisted of scores in six sub areas plus an overall impression score, and used a five-point scale. Because of the differences in the number of points in the scales, scores on the two rubrics are not directly comparable even when concepts overlap.

A factor analysis of the raw scores for writing extracted only one component, as did a factor analysis of the critical-thinking scores, so from a statistical point of view each rubric appears to represent only one major construct. The sub scores thus may help guide the reader in the scoring by providing more concreteness to the intended qualities of good writing or critical-thinking to be measured, and thereby add to score reliability, but the sub scores did not emerge as independent dimensions for analysis.

The tables below indicate the means for each of the rubric sub items and the average of all the items.

Mean Scores for Critical-thinking (Scale = 1 to 6)

	Mean	N	Std. Deviation
average of 9 critical-thinking scores	3.84	335	1.09
holistic rating	3.87	335	1.19
problem	4.02	335	1.04
central/main idea	4.00	335	1.09
perspective(s)	3.61	335	1.10
supporting data/evidence	3.95	335	1.18
depth of thought	3.73	335	1.23
reasoning	3.79	335	1.16
development	3.85	335	1.18
conclusions/consequences	3.72	335	1.19

Overall, the papers were scored highest for recognizing a problem to be addressed and presenting a central or main idea to address the issue raised. The papers were scored lowest for consideration of alternative salient perspectives.

Mean Scores for Writing (Scale = 1 to 5)

	Mean	N	Std. Deviation
average of 7 writing scores	3.37	331	.85
overall impression	3.32	331	.96
main idea	3.33	331	.97
Argument	3.13	331	.93
Evidence	3.47	331	.98
Organization	3.25	331	.92
Readability	3.50	331	.87
Conventions	3.61	331	.89

On the writing skills rubric, papers scored highest for basic readability and use of conventions, and lowest for presenting an argument.

There has been discussion in the literature about whether critical-thinking and writing skills are positively correlated. To look at this for our sample the average of the seven sub scores for writing and of the nine sub scores for critical-thinking were computed for each student. The correlation of these averages for our sample was 0.619 with significance $p < 0.001$, two tailed, indicating a strong positive correlation. This correlation is almost identical to the 0.61 (significance = 0.195, N = 6) we found in our cross-sectional analysis three years ago.

Descriptive Statistics for the Raw Scores

The average scores varied considerably by the primary major, as shown in the table below where majors have been clustered along typical divisional lines.

Scores by Primary Major Group

Primary Major Group		Average of 9 critical-thinking scores	Average of 7 writing scores
Visual and Performing Arts	Mean	3.87	3.23
	N	27	28
	Std. Deviation	1.08	.88
English/Foreign Languages/Philosophy/Religion	Mean	4.26	3.76
	N	82	82
	Std. Deviation	1.07	.77
Natural Sciences/Math/Computer Science	Mean	3.68	3.30
	N	69	67
	Std. Deviation	.94	.79
Business/Education/Recreation	Mean	3.34	2.87
	N	75	72
	Std. Deviation	1.02	.77
Social Sciences/Psychology/History	Mean	3.96	3.54
	N	76	77
	Std. Deviation	1.09	.78
Total	Mean	3.83	3.37
	N	329	326
	Std. Deviation	1.08	.84

In our sample, students majoring in English language and literature/foreign languages /philosophy/ religion scored the highest on both writing and critical-thinking and those in business/ education/ recreation scored the lowest, on average. The individual major scoring the highest was English, for both writing and critical thinking.

Below are the results by gender. The differences in the means are not statistically significant.

Means by Gender

Gender		Average of 9 critical-thinking scores	Average of 7 writing scores
Male	Mean	3.95	3.40
	N	125	121
	Std. Deviation	1.14	.83
Female	Mean	3.77	3.35
	N	210	210
	Std. Deviation	1.05	.86
Total	Mean	3.84	3.37
	N	335	331
	Std. Deviation	1.09	.85

The means varied by paper length, and though the reasons for that are not clear, longer page counts allow students to develop ideas in greater depth and detail, which could be reflected in the scores. Longer papers may also correspond to assignments that have greater weight in the final grade and that students are able to work on for a greater length of time and on which they exert more effort.

Averages by Paper Length

Body length		Average of 9 critical-thinking scores	Average of 7 writing scores
3-6	Mean	3.11	2.95
	N	87	89
	Std. Deviation	.74	.82
6-9	Mean	3.88	3.38
	N	107	106
	Std. Deviation	1.00	.78
9-12	Mean	4.05	3.53
	N	82	82
	Std. Deviation	1.10	.74
12-15	Mean	4.38	3.77
	N	26	25
	Std. Deviation	.99	.74
15-18	Mean	4.50	3.73
	N	14	14
	Std. Deviation	1.11	.97
18-21	Mean	5.01	3.89
	N	8	7
	Std. Deviation	.97	1.10
21-24	Mean	4.29	3.57
	N	4	3
	Std. Deviation	1.24	1.81
Total	Mean	3.84	3.37
	N	335	330
	Std. Deviation	1.09	.85

As might be expected, the average scores tended to increase with increasing college GPAs, as shown in the following Table. Surprisingly, however, the students in the 2-2.5 range scored higher than those in the 2.5-3.0 range, a result that one might surmise is due to small sample sizes.

Scores by GPA Range

GPA Range		Average of 9 critical-thinking scores	Average of 7 writing scores
2-2.5	Mean	3.39	3.24
	N	12	12
	Std. Deviation	.96	.65
2.5 - 3.0	Mean	3.28	2.73
	N	42	43
	Std. Deviation	.94	.84
3 - 3.5	Mean	3.69	3.26
	N	86	86
	Std. Deviation	1.12	.81
3.5 - 4.0	Mean	4.05	3.57
	N	193	188
	Std. Deviation	1.05	.79
Total	Mean	3.84	3.37
	N	335	331
	Std. Deviation	1.09	.85

A regression analysis of the scores (discussed below) found that a predictive equation for the scores included the student's ACT score, college GPA, and primary major, and the paper length. Thus, before comparing the institutional average scores, it is helpful to look at the scores after controlling for these variables. Nonetheless, the raw scores by school may be of interest and are presented below:

Means by School

School		Average of 9 critical-thinking scores	Average of 7 writing scores
Blue	Mean	3.64	3.29
	N	54	51
	Std. Deviation	1.12	.82
Gold	Mean	3.62	3.27
	N	61	64
	Std. Deviation	.93	.81
Rust	Mean	3.47	3.03
	N	55	55
	Std. Deviation	.89	.83
Salmon	Mean	4.06	3.36
	N	48	45
	Std. Deviation	1.18	.93
Silver	Mean	4.04	3.58
	N	58	59
	Std. Deviation	1.12	.70
White	Mean	4.20	3.68
	N	59	57
	Std. Deviation	1.11	.87
Total	Mean	3.84	3.37
	N	335	331
	Std. Deviation	1.09	.85

Regression to Control for Variances of the Cofactors Among Institutions

A key objective of our study is to determine if there are any institutional differences in the scores that would identify any outlier institutions in terms of being particularly effective (or ineffective) in developing student writing or critical-thinking skills. To be closer to the actual impact of the institution, we would like to do these comparisons with adjusted scores that are net of the student and paper variables discussed above. Consequently, linear regression was used in a two phase process in which the average of the writing scores and the average of the critical-thinking scores were used in separate regression analyses as the dependent variables. In the first phase, the student and assignment variables discussed above were entered stepwise to construct a predictive score based on gender, major, ACT score, etc. In phase 2, which is to identify net institutional differences, the residuals from phase 1, which are the differences between the actual and expected score for each student, were used as the dependent variables in regression where the independent variables were dichotomous variables for the six schools. The primary majors were aggregated in groups along typical divisional lines so that each group would include enough papers to be useful for analysis (see table of primary majors above).

Regression results for writing. The phase 1 regression of the average of the seven writing scores resulted in a model with r square = .370 and the coefficients shown below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	T	Sig.	Tolerance	VIF
5	(Constant)	.142	.384		.371	.711		
	ACT with SAT conversions	.044	.013	.194	3.461	.001	.743	1.346
	Number of pages in the paper body	.054	.008	.338	6.962	.000	.989	1.011
	Major-Langs Lit Phil Relig	.360	.100	.183	3.609	.000	.906	1.104
	Major-PreProf(Busn/Educ/Rec)	-.437	.106	-.214	-4.126	.000	.863	1.158
	EOAY 0809 GPA	.445	.108	.225	4.117	.000	.776	1.289

Variables not entering into the model were gender, primary major- natural sciences/math/computer science, major- social sciences, major-visual/performing arts, and number of pages in the bibliography. Thus, higher writing scores are predicted by higher academic performance indicators (ACT and GPA), longer papers, and relative to the other majors as the baseline, a primary major in English/foreign languages/philosophy/religion. Conversely, relative to the other majors, lower scores are predicted by majoring in business/education/recreation.

The phase 2 regression that looked at institutional effects resulted in a model with r square = 0.017, so accounts for only 1.7% of the variance, and with coefficients:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	T	Sig.	Tolerance	VIF
1	(Constant)	.059	.041		1.459	.146		
	Rust	-.238	.101	-.130	-2.366	.019	1.000	1.000

a. Dependent Variable: Writing Adjusted Score (Unstandardized Residual)

Thus, with all other institutions not in the equation and representing the base line, students from Rust predict -0.238 lower.

Regression results for critical thinking. Similarly, performing regression with the average of the 9 critical-thinking scores as the dependent variable, results in a model with r square = 0.350 and coefficients:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
5	(Constant)	-.091	.495		-.184	.854		
	Number of pages in the paper body	.083	.010	.407	8.362	.000	.989	1.011
	ACT with SAT conversions	.070	.017	.236	4.194	.000	.743	1.346
	Major-Lang Lit Phil Relig	.404	.129	.160	3.135	.002	.906	1.104
	Major-PreProf (Busn/Educ/Recreation)	-.381	.137	-.145	-2.784	.006	.863	1.158
	EOAY 0809 GPA	.364	.139	.144	2.615	.009	.776	1.289

Excluded from the model were gender, primary major- natural sciences/math/computer science, major-social sciences, major-visual/performing arts, and number of pages in the bibliography.

The phase 2 regression resulted in no school being entered into the model--that is, the differences among schools were not sufficiently significant for any school to enter into a predictive equation.

To summarize, the search for school outliers showed that after adjusting for student and assignment variables, no school was an outlier for critical-thinking and Rust lagged the other institutions for writing skills in a model that accounted for only 1.7% of the variance.

Value Added Comparisons with Earlier Paper Scores

Value added for Writing. During the past four years, the MALLA institutions have read and scored papers using the same writing rubric three times: at the first-year level from the 2005 cohort, at the junior level for the 2003 cohort and, now, at the senior level for the 2005 cohort. The table below shows that the general pattern has been increasing scores going from the first-year to the senior level. Ignoring the Junior 2003 Cohort measure, the first-year to senior level change for the two independent samples from the 2005 cohort showed a raw score change of plus 0.68.

Writing Averages	First-Year 2005 Cohort	Junior 2003 Cohort	Senior 2005 Cohort	Change FY to Senior 2005 Cohort
Blue	2.76	3.18	3.29	0.53
Gold	2.68	3.04	3.27	0.59
Rust	2.76	3.02	3.03	0.27
Salmon	2.40	2.74	3.36	0.96
Silver	3.13	3.16	3.58	0.45
White	2.51	3.10	3.68	1.17
Total	2.69	3.03	3.37	0.68

To provide perspective on the overall change of 0.68 in the raw scores, we can compute the effect size by dividing by the pooled estimate of the standard deviation. This is 0.85 based on the total of 605 papers read from the first year and senior samples. Dividing gives an effect size of 0.76, which corresponds in a normal distribution to a percentile shift of 28 percentiles going from the first-year to senior level. In comparison, our previous analysis, the estimate of growth from the first-year to junior level for the six schools taken collectively was a 19 percentile shift. A benchmark for comparison is given by the meta-analysis done by Pascarella and Terenzini of research studies done in the 1990s. They report an average effect size for first-year to senior growth of 0.77 for "English (reading and literature, writing)," almost the identical value in this study.

The institutional value added measures should be taken with some caution, since, as noted above, the representativeness of the samples from some institutions is questionable.

Value added for Critical Thinking. In a manner similar to writing, papers have been scored during the past four year for critical-thinking at the first-year, junior/senior level and, now, at the senior level, with average raw scores as indicated below. As was noted in earlier reports, the sample from Blue at the first-year level came from papers from a first-year seminar in which the papers were highly processed through cycles of submission, peer review, and rewriting. Thus the score for Blue comes from a different type of product than the scores at the other levels, and this is likely to be a contributing factor to Blue's otherwise mystifyingly low value added change measure.

Critical-thinking Averages	First-Year 2006 cohort	Junior/Senior 2004/2003 Cohorts	Senior 2005 Cohort	Change FY to Senior 2005 Cohort
Blue	3.70	3.41	3.64	-0.06
Gold	3.08	3.35	3.62	0.54
Rust	3.26	3.80	3.47	0.21
Salmon	2.66	3.15	4.06	1.40
Silver	3.62	3.88	4.04	0.42
White	2.92	3.71	4.20	1.28
Total	3.22	3.55	3.84	0.62

The standard deviations for the scores at each level were close to 1.0, so 1.0 is a good estimate of the pooled standard deviation. Using this estimate, the effect size for the first-year to senior change is 0.62, not as large as the 0.76 observed for writing, but still a sizeable shift, and well above the national estimate of 0.50 provided by Pascarella and Terenzini after a meta-analysis of research studies conducted in the 1990s.

Summary of Findings

- Although the writing and critical-thinking rubrics consisted of scores in several sub areas, these sub areas were so highly correlated that, from a statistical perspective, each rubric only represented one actual dimension.
- The writing and critical-thinking scores given to papers were positively correlated, r square = 0.62, suggesting that good writing and good thinking are related.

- The writing and critical-thinking scores were positively correlated with students' ACT scores and college GPAs, and with the length of the paper.
- Average scores varied significantly by major. English and foreign language had the highest average scores, and business and education majors had the lowest average scores.
- After controlling for student and paper characteristics, the residual differences for the six MALLA institutions were minimal.
- The overall MALLA results from the reading of papers at the first-year and senior levels showed value-added effect sizes equal to the best available comparative data for writing quality and above the best available comparative data for critical thinking.