America's Brightest ORANGE

Oklahoma State University

Committee for the Assessment of General Education

And

The Office of University Assessment and Testing

Annual Report, 2012

Committee for the Assessment of General Education
Jon Comer, Chair
Greg Wilber, Vice Chair
Melanie Bayles
Carol Beier
John Gelder
Bridget Miller

Office of University Assessment and Testing
Jeremy Penn, Ph.D., Director
Mark Nicholas, Ph.D., Assistant Director
John D. Hathcoat, Ph.D., Statistical Analyst
Sarah Banks, M.S., Graduate Research Associate
http://uat.okstate.edu
uat@okstate.edu
(405) 744-6687

Table of Contents

Table of Tables

Executive Summary

In the summer of 2012, seven teams of faculty raters scored 481 samples of student work from courses across campus using the critical thinking rubric (see page 41). The purpose of general education assessment is to provide information about students' level of achievement of the general education learning outcomes using this institutional portfolio process.

Key findings:

- Juniors and seniors scored significantly higher than freshmen (effect sizes of 0.41 and 0.51). Stated another way, the average senior scored higher than 69% of freshmen.
- Inter-rater reliability is low for many of the teams of raters on the initial scoring. However, after discussion with the team leader the raters reached an acceptable level of agreement.
- Although significantly higher than critical thinking scores in 2007, scores in 2012 were similar to scores in the other years in which critical thinking was assessed.
- 19.8% of seniors who were assessed in 2012 scored below raters' expectations for minimally competent critical thinking skills in a graduating student. The percentage increases to 28.3% when scores from all years are considered.
- The correlations between critical thinking scores and students' ACT composite and OSU GPAs were small.
- Assignment characteristics affected students' critical thinking scores. Specifically, instructors' level of emphasis of the "own-perspective" dimension of the critical thinking rubric accounted for nearly 47% of the variation in critical thinking scores between course sections.
- Generalizability studies provided evidence that critical thinking scores were distinct from written communication scores; evidence suggests that about 72% of the variance in critical thinking scores may be unique to critical thinking whereas about 25% of the variation in critical thinking consensus scores may be attributable to written communication.

Recommendations:

- Results from 2012 highlighted the importance of high quality assignments. As a result, efforts should be made to improve the quality of assignments given across campus and particularly in general education courses. As Elliot Eisner wrote more than twenty years ago, "Our nets define what we shall catch" (1992).
- Inter-rater reliability continues to be a concern for the raters. Increased emphasis should be given to training raters. Alternative scoring processes (such as synchronous scoring) should be considered if they will improve inter-rater reliability.
- Although critical thinking scores appeared to be distinct from written communication ability, consideration should be given to alternative ways students can demonstrate critical thinking, such as oral presentations, portfolios, or other performances.
- Cross-sectional assessment models (such as the one used in this study) have limitations. Consideration should be given to a cohort longitudinal model to investigate students' development of critical thinking over time.
- Gains on critical thinking scores from freshmen to seniors were not as large as they could or should be. There are many strategies worth considering to address this concern, including creating a General Education Coordinator position (or working with the newly developed position of Assistant Provost for Innovative Education), creating a center for critical thinking, or initiating campus-wide development opportunities. Of course, in the end, improving students' critical thinking requires commitment from faculty members, staff members, and

students – perhaps developing such a campus-wide commitment would be one way to begin improvement efforts.

Assessment of general education is a critical aspect of our work to continuously improve our institution. We are fortunate that Oklahoma State University provides substantial resources to assess students' learning and to consider ways in which that learning might be improved. Our challenge moving forward is clear: to make the most of this investment by using the results to make meaningful changes to our programs.

Thank you for your time and support of general education assessment. Please let me know if you have any additional questions or comments.

Sincerely,

Jeremy Penn, Ph.D.
Director, University Assessment and Testing
Oklahoma State University
February, 2013

Overview

Introduction

Objectives of the General Education program at OSU include[1]:

A. Construct a broad foundation for the student's specialized course of study,
B. Develop the student's ability to read, observe, and listen with comprehension,
C. Enhance the student's skills in communicating effectively,
D. Expand the student's capacity for critical analysis and problem solving,
E. Assist the student in understanding and respecting diversity in people, beliefs, and societies, and
F. Develop the student's ability to appreciate and function in the human and natural environment.

OSU has been involved in assessing general education using institutional portfolios since 2000. Assessment of these objectives occurs via three approaches: institutional portfolios, review of general education course database, and college-, department-, and program-level approaches. This report focuses on OSU's use of institutional portfolios to assess the general education program.

Institutional portfolios provide direct evidence of student performance aligned with the overall goals of general education. Institutional portfolios have been developed in five areas that represent the overall goals of the general education program (letters in parentheses map portfolios to the goals above):

1. Written communication (B and C)
2. Critical thinking (D)
3. Math problem solving (D)
4. Science problem solving (D and F)
5. Diversity (E and F)

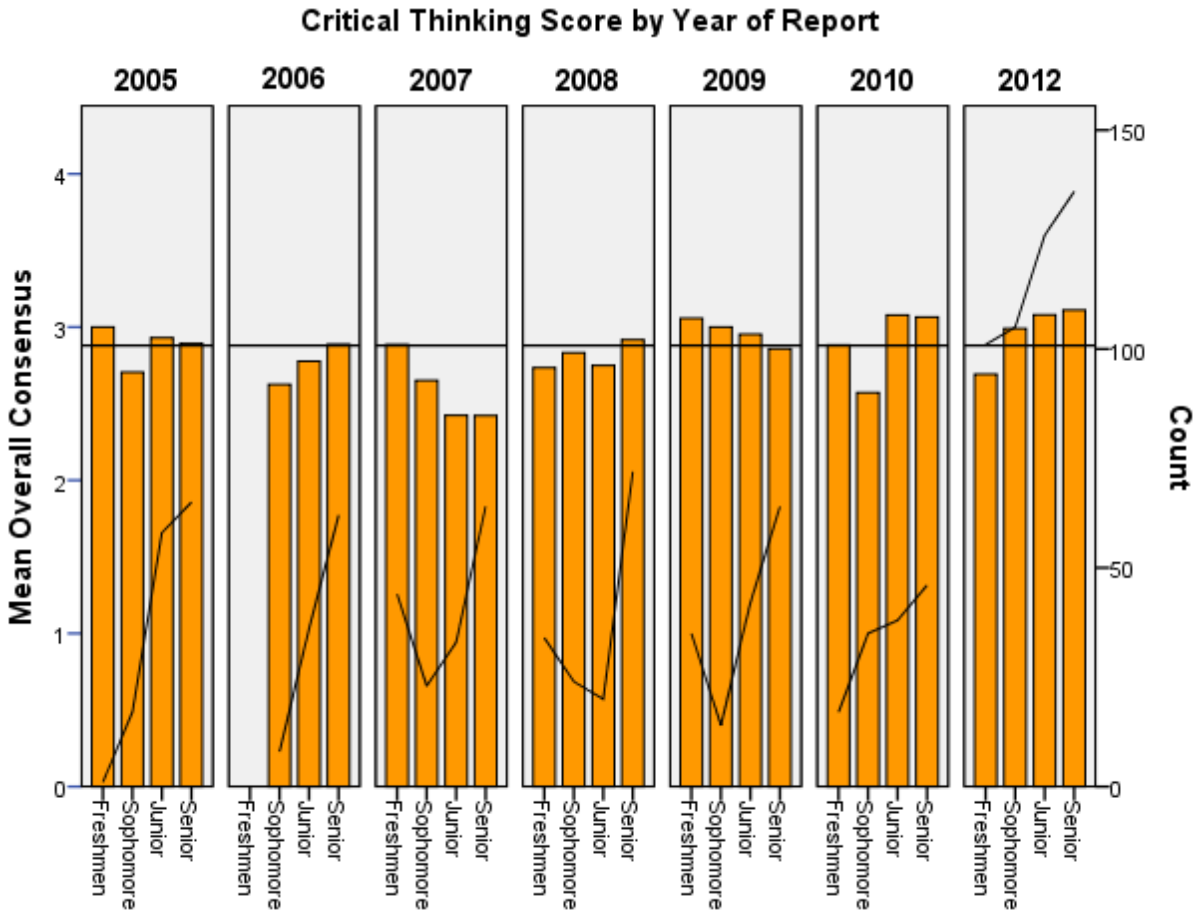
The Office of University Assessment and Testing samples assignments from students (called "artifacts") embedded in existing courses across campus. A panel of faculty members acts as paid raters who provide scores for each artifact using a common rubric developed at OSU. Each rubric has a different number of categories used in the scoring process. However, all rubrics use a 1 to 5 scale where 1 is a low score and 5 is a high score. In 2012, UAT developed one institutional portfolio in the area of critical thinking.

Critical Thinking Results

In 2012, 7 teams of faculty members rated 481 artifacts (samples of student work) for critical thinking. On average, critical thinking scores in 2012 were 2.96 (*SD* = .91). Of the 481 artifacts, 22 (4.6%) were assigned a score of 1, 123 (25.6%) were assigned a score of 2, 207 (43.0%) were assigned a score of 3, 109 (22.7%) were assigned a score of 4, and 20 (4.2%) were assigned a score of 5.

[1] http://academicaffairs.okstate.edu/current-students/64-general-education-overview

Figure 1. Critical thinking scores by year and number of artifacts



The orange bars show the average score by year and classification status (the left y-axis). The black line shows the number of artifacts collected by year and classification status (the right y-axis). The black horizontal line shows the overall average score across all years and classification statuses.

Mean Differences by Grade Classification

Across all years combined, mean differences between undergraduate grade classifications failed to be statistically significant: $F(3, 1316) = 1.49$, $p = .23$, $R^2 = .003$. In 2012, there were 101 freshmen who had an average critical thinking score of 2.69 ($SD = .92$), 105 sophomores with an average critical thinking score of 2.99 ($SD = .85$), 126 juniors with an average critical thinking score of 3.08 ($SD = 1.0$), and 136 seniors with an average critical thinking score of 3.11 ($SD = .82$). These differences were statistically significant: $F(3, 446) = 4.89$, $p = .03$, $R^2 = .036$. Follow-up tests indicated that freshmen, on average, had lower scores than juniors ($SE = .12$, $p = .008$, $d = 0.41$) and seniors ($SE = .12$, $p = .003$, $d = 0.51$).

Mean Differences by Year of Report

In 2012 the average critical thinking score was 2.96 (*SD* = .91,). A 95% confidence interval around this value for 2012 is 2.88 to 3.04. Year of data collection accounts for approximately 2.2% of the variance in critical thinking scores: $F(6, 1309) = 5.14$, $p < .001$. Follow up tests indicated that on average, critical thinking scores in 2012 were significantly higher than scores in 2007: (*M* = 2.58, *SD* = .78) *SE* = .07, $p < .001$. The average score in 2007 was on average lower than scores in most years of data collection (i.e. 2005, 2009, 2010, and 2012). As a result, critical thinking scores in 2012 were similar to most other years in which assessment of critical thinking occurred.

Critical Thinking by Transfer Status

Across all years combined transfer students had an average critical thinking score of 2.80 (*SD* = .82), whereas non-transfer students had an average critical thinking score of 2.90 (*SD* = .80). Transfer status accounts for approximately 0.31% of the variance in critical thinking scores: $F(1, 1314) = 3.814$, $p = .043$. A 95% confidence interval for the average critical thinking score for transfer students is 2.71 to 2.89, and the confidence interval for non-transfer students is 2.85 to 2.95. The overlap across these intervals suggests there is insufficient evidence to conclude that critical thinking scores are different across transfer and non-transfer students when analyzing all years combined.

In 2012, the average critical thinking score for transfer students was 2.99 (*SD* = .93), and a 95% confidence interval is 2.84 to 3.14. The average critical thinking score for non-transfer students was 2.95 (*SD* = .93) and the 95% confidence interval was 2.85 to 3.05. The overlap among these intervals, coupled with a significance test $t(479) = .44$, $p = .66$, $d = .04$, indicates that there is insufficient evidence to conclude that critical thinking scores differed across transfer and non-transfer students in 2012.

Relationship between Critical Thinking other Study Variables

Across all years combined, critical thinking scores had small correlations with OSU GPA ($r = .18$, $p < .001$) and composite ACT scores ($r = .21$, $p < .001$). Across all years combined, the observed correlation between critical thinking scores and cumulative credit hours failed to be statistically significant ($r = .04$, $p = .20$).

In 2012, the relation between critical thinking and these variables was similar in magnitude, with the exception of cumulative credit hours ($r = .16$, $p < .001$). In 2012 the observed correlation between critical thinking and OSU GPA was .14 ($p = .002$), and the observed correlation between critical thinking and composite ACT scores was .21 ($p < .001$). GPA in courses with general education designations at OSU was also collected. The correlation between critical thinking and GPA within these courses was .15 ($p = .001$).

Assignment Characteristics

Faculty submitting student artifacts were asked to participate in an online survey wherein they were asked questions about assignment characteristics. Questions included percent of final grade, whether the student received feedback, and the extent to which the assignment reflects each dimension on the critical thinking rubric. Scores from 18 classes containing 343 artifacts

were used in the specification of several models using hierarchical linear modeling. Of primary interest was the extent to which average critical thinking scores varied across classrooms and the extent to which assignment characteristics accounted for this variation.

Results indicated that approximately 18% of the total variation derives from differences in average critical thinking scores between classrooms or variation between assignments. Several models were examined in order to investigate which classroom characteristics account for this variation. The critical thinking dimension, “own-perspective,” accounted for nearly 47% of the variation in average critical thinking scores between course sections. Assignments judged to have higher levels of “own-perspective” have higher critical thinking scores. No other predictors were statistically significant.

Ideally, critical thinking scores should be unaffected by assignment characteristics. This evidence suggests that average critical thinking scores indeed change substantially across assignments and that this may be due to the extent to which an assignment allows a student to provide his or her own perspective within the paper. Therefore, when screening artifacts for a fit with the critical thinking rubric, raters should give extra attention to this dimension.

Reliability of Critical Thinking and Written Communication Scores

Artifacts used for assessing critical thinking consist of written papers. Judges then score these artifacts using a common rubric. A consequence of this strategy is that judges derive critical thinking scores only from writing samples; thus, inferences about critical thinking are strictly located within the realm of written communication. Such a strategy leads to questions about the empirical distinction between critical thinking and written communication.

In order to investigate this issue, two groups of judges scored a set of artifacts for critical thinking and a third group of judges scored the same artifacts for written communication (N = 71). Two analytic procedures are utilized in order to investigate this issue. First, two generalizability studies investigated our ability to reliably estimate critical thinking and written communication average scores, as well as mean differences between critical thinking and written communication. Each generalizability study contains the same design, though they differ in that they control for specific general education designations (i.e. in one analysis only artifacts with an S general designation are investigated and in a second study only artifacts with an H general education designation are investigated).

Results from both generalizability theory studies suggest that the precision of our ability to estimate critical thinking and written communication mean scores and mean differences is below acceptable limits. In one study, approximately 74% of error variance is attributable to the severity of raters assigned to the domain of writing or critical thinking, and in the second study this value was estimated at nearly 85%. In other words, measurement imprecision appeared to primarily result from differences across judges who were assigned to score the same domain. A series of 95% confidence intervals were constructed around means and mean differences between these domains. Across both studies the magnitude of these intervals showed that our estimates of average critical thinking and written communication scores may vary by nearly a whole point when using a 1-5 scale across replications of the measurement procedure.

Results from this study imply that inter-rater reliability, at least prior to consensus, remains a concern. However, in many analyses the scores that are utilized are post-consensus. The procedure for developing consensus scores consists of having a third rater score artifacts for which discrepancies by more than one point were apparent. Although this approach solves the problem pragmatically, the utility of this approach for resolving concerns about score reliability remain unclear. In the absence of clarity on this question, continued steps are needed to ensure that raters are effectively trained. Moreover, this analysis reinforced our concerns that empirical distinctions between critical thinking and written communication may be ambiguous.

Written Communication as a Source of Construct-Irrelevant Variance

Construct-irrelevant variation consists of systematic error that unduly influences observed scores. This investigation sought to understand the extent to which written communication may act as a source of systematic error in critical thinking assessment. This procedure entails an examination of the factor structure accounting for the observed correlations between dimensions of the critical thinking and written communication rubrics (see Appendices A and B for the rubric). Factor scores derived from this analysis are used to make subsequent predictions about critical thinking consensus scores. Squared partial and semipartial correlation coefficients provide a framework for isolating the amount of variation in critical thinking consensus scores attributable to written communication.

The pattern of correlations between these rubric dimensions is accounted for by two orthogonal factors that generally correspond to our theoretical expectations. In other words, dimensions of the critical thinking rubric tend to load on a single factor and dimensions of the written communication rubric tend to load on a distinct factor. Factor scores were generated from this analysis and used to predict to critical thinking consensus scores. Our estimates suggest that approximately 72% of the variance in critical thinking consensus scores may be construct-relevant, in that it is unique to variation in critical thinking. Conversely, approximately 25% of the variance in critical thinking consensus scores appears to be construct-irrelevant, in that is associated with systematic error variance attributable to written communication.

Use of Results and Future Plans

On March 8, 2013, the three committees or councils that have primary responsibility for the general education program (Assessment and Academic Improvement Council, General Education Advisory Committee, and the Committee for the Assessment of General Education) met jointly to hear a summary of this report and to discuss uses of the results and future plans. First and foremost, regardless of the concerns over inter-rater reliability, there was general agreement that students neither write nor critically think at levels most faculty deem acceptable. Thus, we need to continue to engage the university community to develop better ways to help reach out to faculty and provide professional development training in this area. Strategies for reaching out included:

- Faculty development – both institutional level and targeted at the college level, and particularly for instructors teaching general education-designated courses. The *Provost's Initiative* series should be continued, with possible expansion at the college level.
- Wider use of curriculum mapping, both at the course and program level.
- Enhancing the partnership with ITLE and ITLE's newly developed teaching fellows program. Perhaps an emphasis year on critical thinking could bring additional attention to improving students' abilities in this area.
- Approaching the teaching of critical thinking as critical thinking across the curriculum. Students cannot learning critical thinking in only 1 course or only 1 semester.
- Identifying strategies to better support adjuncts', graduate assistants', and teaching assistants' teaching of critical thinking and considering the preparation of future faculty members to support their future teaching of critical thinking.
- Providing feedback at the course or student level to better inform the teaching process.

Second, the attendees at the joint committee discussed inter-rater reliability. At this point the methodology for scoring papers and the reporting of inter-rater reliability and its impact on scores seems sufficient. The joint committee discussed alternative methods for assessing critical thinking, such as performance exams, interviews, standardized testing, or following a cohort over time. The joint group encouraged the General Education Task Force to consider structures to support one or more of these alternative assessment methods that may provide additional information about students' achievement of the learning outcomes and may provide more clear avenues for responding when results suggest students' achievement is insufficient.

Third, attendees at the joint meeting noted critical thinking and writing are shared, joint, and student career-spanning (i.e. freshman to graduation) tasks that simply do not devolve to a few select departments on campus. It cannot be implicit that we create better writers and critical thinkers; it must be an explicit activity all across campus intentionally designed into curricula.

Finally, the joint group recommended the General Education Task Force carefully consider the assessment results when developing recommendations on modifications and improvements to the general education program. The best way to assist students in developing the skills and abilities we desire is a carefully and thoughtfully designed program intentionally designed

around the learning outcomes desired, taught by highly qualified and prepared instructors, assessed by multiple measures and multiple formats, and using assessment results for reflection and regular program updates to address concerns or issues as they arise.

Methods

Artifacts (course assignments embedded in existing courses) were collected by direct request from a random sample of general education designated courses, from faculty members who voluntarily submitted samples of student work, and from faculty members who participated in the *Provost's Faculty Development Initiative: Focus on General Education*. The courses from which artifacts were sampled are shown in Table 1. Artifacts selected for the Institutional Portfolio were coded and all identifying information was removed from the samples. Demographic data were collected for each artifact using the OSU student database; these data were collected for analysis purposes only and the information cannot be used to identify any individual. Students' demographic information associated with the artifacts were not shared with reviewers prior to the reviews.

Table 1. 2012 Collection of Critical Thinking Artifacts

Course No.	Course Name	General Education Designation (if any)	Number of artifacts randomly collected from one assignment	Number of artifacts used in data analysis
AMST 3503	Television and American Society	H, D	15	15
ANTH 3353	Cultural Anthropology	S, I	39	29
BOT 3253	Environment and Society	N	18	10
CIVE 3813	Environmental Engineering Sc.		33	22
ENGL 2413	Introduction to Literature	H, D	24	14
GEOG 1113	Intro to Cultural Geography	S, I	53	39
GEOG 2253	World Regional Geography	S, I	46	33
GEOG 3723	Geography of Europe	S, I	40	26
GWST 2123	Gender and Women's Studies	H, D	22	5
GWST 3450	Topics in Women's Studies		16	8
HIST 1713	Survey of Eastern Civilization	H	57	28
HONR 1043	Western Humanities	H	13	12
MICR 3103	Microbes: Friends or Foes	N	5	5
NSCI 3543	Food and the Human Environ.	S, I	19	19
PHIL 1013	Philosophical Classics	H	64	14
PHIL 1213	Philosophies of Life	H	100	35
PHIL 3920	God, Philosophy, and the Movies	H	26	12
PHIL 4013	Perspectives on Death and Dying	S	10	5
PHIL 4312	Philosophy of the Mind	H	21	16
PSYC 1113	Introduction to Psychology	S	74	24
PSYC 3073	Neurobiological Psychology	N	45	23
SOC 3993	Sociology of Aging	S, D	23	21
SPCH 3733	Elements of Persuasion	S	27	16
ZOOL 3104	Invertebrate Zoology		36	36
ZOOL 4273	Environmental Physiology		14	14
	Total Number of Critical Thinking Artifacts		**840**	**481**

Scoring Process and Reliability Estimation

All portfolio reviewers met for two training sessions where they received an overview of the general education program and the portfolio review process. After reviewing the critical thinking rubric, faculty members reviewed critical thinking artifacts from previous years. Faculty reviewers then rated new student artifacts during the training session so that reviewers could discuss any questions or concerns regarding the use of the rubric and to align raters' scores with each other.

Seven teams of two reviewers scored artifacts independently. Raters were nested within three discipline groups (humanities, natural sciences, and social sciences). Raters in each of these three groups identified themselves as fitting with that group and scored papers that reflected each discipline group. That is, raters from the natural sciences rated papers from natural sciences discipline, and so on. This was done to see if it substantively improved inter-rater reliability and to allow raters to have some familiarity with the content of the artifact being rated.

Each artifact received an overall score ranging from 1-5 wherein higher scores reflect a greater level of critical thinking. Reviewers also assigned a sub-score to each artifact across four components: identification of problem, own perspective, use of supporting evidence, and conclusion. Three additional scoring components on the rubric are optional, and include other, assumptions, and context. When discrepant scores between raters existed, a third member of the team, the team leader, either facilitated a discussion between the two original raters or, if they were unable to reach consensus, broke the disagreement by assigning his or her own score to the artifact. Each team was initially assigned approximately 80 artifacts, ten of which were the same across all teams. Reliability estimates for the first four teams is provided in Table 2 and reliability estimates for the next 3 teams are provided in Table 3.

Table 2. Reliability of Rater Groups 1-4[2]

	Group 1			Group 2			Group 3			Group 4		
Method	Value	SE	C.I.	Value	SE	C.I	Value	SE	C.I.	Value	SE	C.I.
AC1	.48	.06	.36-.61	.93	.03	.87-.99	.38	.06	.26-.51	.33	.07	.20-.47
Kappa	.42	.07	.29-.55	.92	.04	.85-.99	.28	.08	.12-.44	.09	.09	.01-.28
PI	.41	.07	.28-.54	.92	.04	.85-.99	.26	.08	.11-.42	.07	.08	.01-.25
BP	.47	.06	.36-.60	.93	.03	.86-.99	.37	.07	.24-.50	.30	.07	.16-.44

[2] Descriptions of reliability coefficient may be found at http://agreestat.com/research_papers.html. AC1 = variation of Kappa statistic and BP statistic that incorporates the conditional probability that two random rater will agree given no chance agreement; Kappa = omnibus measure of percent agreement among raters when corrected for chance agreement wherein chance is defined as the expected value if ratings were completely independent; PI = probability that a randomly selected rater will classify a randomly selected artifact into specific category. BP = Brennan-Prediger modification of Kappa statistic that incorporates a modification of marginal estimates so that chance is redefined to adjust for the number of possible categories.

Table 3. Reliability Estimates of Groups 5-7[3]

	Group 5			Group 6			Group 7		
Method	Value	SE	C.I.	Value	SE	C.I	Value	SE	C.I.
AC1	**.14**	.06	.02-.27	**.28**	.06	.15-.40	**1.00**	.00	1-1
Kappa	**.08**	.07	.00-.23	**.15**	.08	.00-.30	**1.00**	.00	1-1
PI	**.07**	.07	.00-.21	**.14**	.08	.00-.29	**1.00**	.00	1-1
BP	**.13**	.06	.00-.26	**.26**	.07	.13-.39	**1.00**	.00	1-1

There are numerous ways to evaluate the adequacy of reliability estimates. Though many proposed benchmarks may be too liberal (Gwet, 2010), guidelines given by Altman (1991) are provided below:

- < .20 = Slight Agreement
- .21 to .40 = Fair Agreement
- .41 to .60 = Moderate Agreement
- .61 to .80 = Good Agreement
- .81 to 1.00 = Very Good.

These guidelines indicates that one team had “very good” levels of agreement, one team had “moderate” levels of agreement, and that two teams had a “fair” level of agreement. Two teams had “slight” agreement and one group had perfect agreement.

Use of scores when there is poor inter-rater reliability estimates, before corrected by a third rater, is problematic. However, most analyses use the scores after correction by a third rater. The extent to which this process “corrects” for score inconsistency across raters remains empirically unexamined (since artifacts’ “true” scores are unknown).

[3] Descriptions of reliability coefficient may be found at http://agreestat.com/research_papers.html. AC1 = variation of Kappa statistic and BP statistic that incorporates the conditional probability that two random rater will agree given no chance agreement; Kappa = omnibus measure of percent agreement among raters when corrected for chance agreement wherein chance is defined as the expected value if ratings were completely independent; PI = probability that a randomly selected rater will classify a randomly selected artifact into specific category. BP = Brennan-Prediger modification of Kappa statistic that incorporates a modification of marginal estimates so that chance is redefined to adjust for the number of possible categories.

Table 4. Regression of Overall Consensus Scores on Critical Thinking Dimensions

	Identification		Own Perspective		Supporting Evid.		Conclusion	
Reviewer	mean	β weight	mean	β weight	mean	β weight	mean	β weight
Team 1								
1	3.44	.30***	2.99	.04	2.99	.28**	2.76	.43***
2	3.43	.33***	3.11	.02	3.27	.40***	2.94	.31***
Team 2								
3	2.99	.20**	3.09	.32***	2.89	.26**	2.60	.30***
4	2.97	.33***	3.01	.33***	2.85	.20*	2.80	.22**
Team 3								
5	3.44	.33***	3.36	.11	3.18	.21*	3.20	.45***
6	3.24	.19*	3.07	.08	3.00	.47***	2.87	.30**
Team 4								
7	3.27	.31***	3.19	.25**	3.09	.30***	2.98	.29***
8	3.21	.22***	3.21	.22***	3.16	.45***	3.11	.36***
Team 5								
9	3.51	.28***	3.76	-.02	3.05	.52***	3.30	.26***
10	3.77	.17**	3.72	.08	3.54	.48***	3.49	.32***
Team 6								
13	2.82	.17**	2.61	.27***	2.75	.55***	2.76	.14*
14	3.25	.23**	3.25	.26***	2.68	.48***	2.36	.24**
Team 7								
15	3.39	.23**	3.10	.07	3.04	.50***	2.99	.21*
16	3.69	.23*	3.31	.16	3.28	.20	3.19	.32*

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ on individual-level regression with overall score as the dependent variable.

Results

The tables on the following pages present descriptive and inferential statistics for the critical thinking scores.

Table 5. Critical Thinking Scores from each Review Group

Review Group	Artifact Score	Number of Artifacts	Percent of Artifacts
#1 (69 artifacts scored)	1	2	2.9%
	2	22	31.9%
	3	26	37.7%
	4	15	21.7%
	5	4	5.8%
#2 (79 artifacts scored)	1	7	8.9%
	2	23	29.1%
	3	27	34.2%
	4	17	21.5%
	5	5	6.3%
#3 (62 artifacts scored)	1	3	4.8%
	2	23	37.1%
	3	26	41.9%
	4	8	12.9%
	5	2	3.2%
#4 (70 artifacts scored)	1	0	0.0%
	2	7	10.0%
	3	42	60.0%
	4	20	28.6%
	5	1	1.4%
#5 (70 artifacts scored)	1	2	2.9%
	2	11	15.7%
	3	30	42.9%
	4	19	27.1%
	5	8	11.4%
#6 (66 artifacts scored)	1	8	12.1%
	2	30	45.5%
	3	19	28.8%
	4	9	13.6%
	5	0	0.0%
#7 (65 artifacts scored)	1	0	0.0%
	2	7	10.8%
	3	37	56.9%
	4	21	32.3%
	5	0	0.0%

Table 6. Students' Demographics Associated with Critical Thinking Skills Artifacts, 2005-2012

		2005-10		2012		Years Combined	
		No. of Artifacts	Pct	No. of Artifacts	Pct	No. of Artifacts	Pct
Number of Artifacts	# collected	1336	-	840	-	1648	-
	# scored	866	-	504	-	1370	-
	# used in analysis	858	-	481	-	1339	-
Class	Freshman	131	15.3%	101	21.6%	232	17.6%
	Sophomore	121	14.2%	105	22.4%	226	17.1%
	Junior	227	26.6%	126	26.9%	353	26.7%
	Senior	373	43.8%	136	29.1%	509	38.6%
College	CAS	214	25.1%	257	53.8%	471	35.4%
	CASNR	116	13.6%	17	3.6%	133	10.0%
	SSB	71	8.3%	50	10.4%	121	9.1%
	COE	51	6.0%	26	5.4%	77	5.8%
	CEAT	200	23.3%	70	14.6%	270	20.3%
	HS	195	22.7%	18	3.8%	213	16.0%
	UAS	7	0.8%	40	8.4%	47	3.5%
Gender	Female	447	55.3%	236	49.1%	683	53.0%
	Male	361	44.7%	245	50.9%	606	47.0%
Admit Type	Regular (A, AR, L)	552	68.4%	307	63.8%	859	66.7%
	Alternative Admit (F)	23	2.9%	11	2.3%	34	2.6%
	Adult Admit (G)	2	0.2%	0	0.0%	2	0.1%
	International (J)	12	1.5	2	0.4%	14	1.1%
	Transfer (M, MR)	215	26.6%	131	27.2%	346	26.9%
	Other or Blank	3	0.4%	30	7.3%	33	2.6%
ACT	<22	156	23.4%	84	22.0%	240	22.9%
	22 to 24	162	24.3%	104	27.2%	266	25.4%
	25 to 27	172	25.8%	90	23.6%	262	25.0%
	28 to 30	112	16.8%	73	19.1%	185	17.6%
	>30	65	9.7%	31	8.1%	96	9.2%
OSU GPA	<2.0	37	5.2%	36	7.6%	73	6.2%
	2.0 to 2.49	79	11.2%	51	10.6%	130	11.0%
	2.50 to 2.99	173	24.5%	101	21.0%	274	23.2%
	3.00 to 3.49	205	29.0%	150	31.2%	355	30.0%
	3.50 to 4.00	213	30.1%	138	28.7%	351	29.7%

Table 7. Critical Thinking Scores: 2012

			Score						
			1	**2**	**3**	**4**	**5**	***M***	**N**
Overall Scores	Overall	n	22	123	207	109	20	2.96	481
		%	4.6%	25.6%	43.0%	22.7%	4.2%		
By Class	Freshmen	n	9	32	44	13	3	2.69	101
		%	8.9%	31.7%	43.6%	12.9%	3.0%		
	Sophomores	n	1	29	51	18	6	2.99	105
		%	1.0%	27.6%	48.6%	17.1%	5.7%		
	Juniors	n	7	30	42	40	7	3.08	126
		%	5.6%	23.8%	33.3%	31.7%	5.6%		
	Seniors	n	4	23	67	38	4	3.11	136
		%	2.9%	16.9 %	49.3%	27.9%	2.9%		
By Class (regular admit only)	Freshmen	n	7	27	40	11	3	2.50	88
		%	8.0%	30.7%	45.5%	12.5%	3.4%		
	Sophomores	n	1	24	39	13	6	2.73	83
		%	1.2%	28.9%	47.0%	15.7%	7.2%		
	Juniors	n	1	18	25	26	6	2.99	76
		%	1.3%	23.7%	32.9%	34.2%	7.9%		
	Seniors	n	4	7	36	24	2	3.24	73
		%	5.5%	9.6%	49.3%	32.9%	2.7%		
By Transfer Status	Non-transfer Students	n	17	91	151	74	17	2.95	350
		%	4.9%	26.0%	43.1%	21.1%	4.9%		
	Transfer Students	n	5	32	56	35	3	2.99	131
		%	3.8%	24.4%	42.7%	26.7%	2.3%		

Table 8. Average Component Scores for Critical Thinking: 2012

Component	Problem	Perspective	Support	Conclusion	Others	Assumptions	Context
Average Score	3.28 (N=1114)	3.18 (N=1045)	3.05 (N=1114)	2.92 (N=1102)	2.66 (N = 383)	2.41 (N=346)	2.57 (N=379)

Table 9. Critical Thinking Scores: 2005-2012 (years combined)

			Score						
			1	**2**	**3**	**4**	**5**	***M***	**N**
Overall Scores	Overall	n	45	369	656	245	24	2.88	1339
		%	3.4%	27.6%	49.0%	18.3%	1.8%		
By Class	Freshmen	n	11	69	109	40	3	2.81	232
		%	4.7%	29.7%	47.0%	17.2%	1.3%		
	Sophomores	n	4	68	121	26	7	2.84	226
		%	1.8%	30.1%	53.5%	11.5%	3.1%		
	Juniors	n	15	92	156	83	7	2.93	353
		%	4.2%	26.1%	44.2%	23.5%	2.0%		
	Seniors	n	14	130	262	96	7	2.91	509
		%	2.8%	25.5%	51.5%	18.9%	1.4%		
By Class (regular admit only)	Freshmen	n	8	61	102	37	3	2.84	211
		%	3.8%	28.9%	48.3%	17.5%	1.4%		
	Sophomores	n	1	49	91	20	7	2.90	168
		%	0.6%	29.2%	54.2%	11.9%	4.2%		
	Juniors	n	8	55	96	50	6	2.96	215
		%	3.7%	25.6%	44.7%	23.3%	2.8%		
	Seniors	n	7	55	150	63	3	3.00	278
		%	2.5%	19.8%	54.0%	22.7%	1.1%		
By Transfer Status	Non-transfer Students	n	30	263	493	188	19	2.90	993
		%	3.0%	26.5%	49.6%	18.9%	1.9%		
	Transfer Students	n	15	106	163	57	5	2.80	340
		%	4.3%	30.6%	47.1%	16.5%	1.4%		

Table 10. Comparison of Overall Average Critical Thinking Scores by Year

			Score						
			1	**2**	**3**	**4**	**5**	***M***	**N**
Overall Scores	Overall	n	45	369	656	245	24	2.88	1339
		%	3.4%	27.6%	49.0%	18.3%	1.8%		
By Year	2005	n	2	40	72	26	1	2.89	141
		%	1.4%	28.4%	51.1%	18.4%	0.7%		
	2006	n	4	29	54	19	0	2.83	106
		%	3.8%	27.4%	50.9%	17.9%	0.0%		
	2007	n	13	59	76	16	0	2.58	164
		%	7.9%	36.0%	46.3%	9.8%	0.0%		
	2008	n	1	46	81	24	0	2.84	152
		%	0.7%	30.3%	53.3%	15.8%	0.0%		
	2009	n	1	35	93	24	2	2.94	155
		%	0.6%	22.6%	60.0%	15.5%	1.3%		
	2010	n	2	37	73	27	1	2.91	140
		%	1.4%	26.4%	52.1%	19.3%	0.7%		
	2012	n	22	123	207	109	20	2.96	481
		%	4.6%	25.6%	43.0%	22.7%	4.2%		

Table 11. Comparison of Overall Average Critical Thinking Scores by Classification and Year

		Year							
		2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2012**	**N**
Freshmen	n	1	0	44	34	35	17	101	232
	M	3.00	0.00	2.89	2.74	3.06	2.88	2.69	
Sophomores	n	17	8	23	24	14	35	105	226
	M	2.71	2.62	2.65	2.83	3.00	2.57	2.99	
Juniors	n	58	36	33	20	42	38	126	353
	M	2.93	2.78	2.42	2.75	2.95	3.08	3.08	
Seniors	n	65	62	64	72	64	46	136	509
	M	2.89	2.89	2.42	2.92	2.86	3.07	3.11	

Key Findings

- In 2012, there were 101 freshmen who had an average critical thinking score of 2.69 (*SD* = .92), 105 sophomores with an average critical thinking score of 2.99 (*SD* = .85), 126 juniors with an average critical thinking score of 3.08 (*SD* = 1.0), and 136 seniors with an average critical thinking score of 3.11 (*SD* = .82). These differences were statistically significant: $F(3, 446) = 4.89$, p = .03, $R^2 = .036$. Follow-up tests indicated that freshmen, on average, had lower scores than juniors (*SE* = .12, *p* = .008, *d* = .41) and seniors (*SE* = .12, *p* = .003, *d* = .50).

- Year of data collection accounts for approximately 2.2% of the variance in critical thinking scores $F(6, 1309) = 5.14$, $p < .001$. Follow up tests indicated that, on average, critical thinking scores in 2012 were higher than scores in 2007 (*M* = 2.58, *SD* = .78) *SE* = .07, $p < .001$. The average score in 2007 was on average lower than scores in most years of data collection (i.e. 2005, 2009, and 2010). This evidence suggests that critical thinking scores in 2012 were similar to most other years in which assessment of critical thinking occurred.

- Across all years combined (i.e. 2005, 2006, 2007, 2008, 2009, 2010, and 2012) transfer students non-transfer students had an average critical thinking score of 2.90 (SD = .80). Transfer students had an average critical thinking score of 2.80 (SD = .82). These differences were statistically significant: $F(1, 1314) = 4.09$, $p = .043$, $R^2 = 0.03$. Such differences were not apparent when only examining 2012 consensus scores $t(479) = .44$, $p = .66$, $d = .04$.

- Across all years combined (i.e. 2005, 2006, 2007, 2008, 2009, 2010, and 2012), critical thinking scores had small correlations with OSU GPA ($r = .18$, $p < .001$) and composite ACT scores ($r = .21$, $p < .001$). In 2012, the observed correlation between critical thinking and OSU GPA was .14 ($p = .002$), and the observed correlation between critical thinking and composite ACT scores was .21 ($p < .001$).

- Across all years combined (i.e. 2005, 2006, 2007, 2008, 2009, 2010, and 2012), the observed correlation between critical thinking scores and cumulative credit hours failed to be statistically significant ($r = .04$, $p = .20$). In 2012, the relation between critical thinking and cumulative credit hours was small, but statistically significant ($r = .16$, $p < .001$).

- In 2012, the correlation between critical thinking and GPA in OSU general education courses was .15 ($p = .001$). This information was not collected in previous years in which critical thinking was assessed.

Critical Thinking and Assignment Characteristics

In 2012, UAT asked faculty members who submitted artifacts to complete a short online survey pertaining to assignment characteristics. This survey asked faculty submitting artifacts to provide the percent of the final grade for submitted artifacts, whether each artifact received feedback prior to submission (Yes = 1; No = 0), and the extent to which the artifact reflected each aspect of the critical thinking rubric scored on a 1-5 scale (1 = not much to 5 = a great deal). These variables, given that they are constant for all artifacts nested within a particular classroom, are classroom level variables.

Twenty-one instructors provided information through the online survey. However, three of courses were removed from subsequent analysis due to having a small number of observations within each course. This led to 343 artifacts scored for critical thinking that are nested within 18 courses or assignments (see Table 12 for descriptive statistics).

The following research questions are addressed within this section:

1. How much do classrooms vary in their average critical thinking score?
2. What is a range of plausible values for the average critical thinking score across each classroom?
3. How much variation in average critical thinking consensus scores between classrooms can be independently explained by percent of final grade, feedback, and the extent to which the artifact assignment is judged to be aligned with each dimension of the critical thinking rubric?
4. How much within-group variation in critical thinking consensus scores can be accounted for by OSU GPA?
5. What is the average regression equation across classrooms when predicting critical thinking consensus scores from OSU GPA?
6. To what extent does the intercept and slope of this regression equation vary from classroom to classroom?
7. What classroom level variables classroom level predictors predict variation in the intercept and slope between classrooms?

Hierarchical linear modeling (Raudenbush & Bryk, 2002) is utilized in order to address these research questions. Examination of these research questions entails an investigation of several distinct models. Discussion of each model, as well as the parameter estimates characterizing these models, will proceed as each research question is investigated.

Table 12. Descriptive Statistics for Level 1 (artifacts) and Level 2 (classroom) Variables: HLM

Variable Name	Mean	Standard Deviation
Level 1 Variables		
Consensus	2.94	0.91
OSU GPA	3.13	0.62
Level 2 Variables		
Percent of Final Grade	22.68	21.31
Feedback	0.44	-----
Identify Problem	4.39	0.70
Own Perspective	4.11	1.08
Supporting Evidence	4.61	0.70
Conclusion	4.22	0.81

Note: There are 343 Level 1 observations and 18 Level 2 observations. Feedback is dummy coded so 0.44 indicates that 44% of classes provided feedback prior to submission.

Unconditional Means Model (Random Effects ANOVA)

The general purpose of this model is to investigate the extent to which the average critical thinking score varies across each classroom. Additionally, this model will investigate a range of plausible values for the average critical thinking score between each classroom. Formally, this model may be broken down into two levels. Level 1 is defined as the student level, whereas Level 2 is defined as the classroom level. Symbolically each level is represented as follows:

Level 1 Model

$$Consensus_{ij} = \beta_{0j} + r_{ij}$$

Level 2 Model

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

This model thus characterizes consensus scores a function of an intercept β_{0j}, which in this case is an average consensus score, and unspecified error variance r_{ij}. At Level 2, each classroom's average consensus score is specified as a function of the grand mean γ_{00} and error variance μ_{0j}. Variation in the average consensus scores between classrooms is denoted as τ_{00}, whereas variation of consensus scores within each classroom is denoted as σ^2.

Results of this analysis indicate that variation in the average consensus scores between classrooms is estimated at τ_{00} = .15, whereas the within-classroom variation is σ^2 = .69. An intraclass correlation coefficient indicates that approximately 18% of the total variation in critical thinking consensus scores exists between classrooms. The grand mean across each classroom is estimated at γ_{00} = 2.94. Thus, across each classroom the average critical thinking consensus score is 2.94. A 95% confidence interval around this value indicates a plausible estimate of this mean may be anywhere from 2.18 to 3.70 across each classroom. The reliability of the sample mean for each classroom, when estimating their true value .80. This analysis therefore indicates that substantial variation in the average critical thinking consensus score exists between classrooms $\chi^2(17) = 85.05, p < .001$.

Regression with Means as Outcomes Model

Examination of this model is motivated by asking whether variation in β_{0j},or the mean critical thinking consensus score across each classroom, is a function of classroom level characteristics. Classroom level characteristics considered in this analysis include the survey response items provided by faculty who submitted student artifacts. This model was tested separately for each classroom level, or Level 2 predictor. Including Level 2 predictors of β_{0j} therefore allows us to examine whether high percent of total grade, feedback, and the extent to which each dimension of the critical thinking rubric is associated with high levels of average critical thinking scores across each classroom, predicts mean differences in critical thinking scores. Additionally, investigation of this model provides an examination the between-classroom variation in average critical thinking consensus scores that is accounted for by each predictor.

Formally, this model is represented as follows:

Level 1 Model

$$Consensus_{ij} = \beta_{0j} + r_{ij}$$

Level 2 Model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Predictor) + \mu_{0j}$$

Interpretation of the Level 1 model remains the same. However, now at Level 2 the intercept, or the average critical thinking score for each classroom, is a function of distinct predictors γ_{01}. This model was tested independently for each of the classroom level variables under consideration. Table 13 provides a summary of this series of analyses.

Table 13. Variation in Consensus Score Means between Classrooms as Function of Level 2

Predictor	% Variance Explained	γ_{01}	t-value	DF	p-value
Percent of Final Grade	13%	-.01	-1.74	16	.10
Feedback	6%	.29	1.56	16	.14
Identification of Problem	0%	.08	0.54	16	.60
Own Perspective	47%	.25	3.29	16	.01
Supporting Evidence	0%	.08	0.83	16	.42
Conclusion	0%	.09	0.88	16	.39

Note: Each critical thinking dimension reflects the extent to which faculty participating in the online survey judged the artifact assignment to a respective dimension.

When conducting each analysis, only the extent to which each artifact was rated as illustrating the student's own perspective was a statistically significant predictor of between-class variation. In other words, higher scores on own perspective predict increases in the average critical thinking score between classrooms, which accounts for approximately 47% of the between-group variation. Nevertheless, a substantial amount of variation in critical thinking consensus scores between classrooms remains unexplained by this predictor $\chi^2(16) = 52.11, p < .001$. The reliability with which we can discriminate among classrooms with the same own perspective score is .69. No other predictors were statistically significant.

Random Coefficient Model

The random coefficient model may be used in order to estimate an average regression equation across each classroom. For this analysis we chose OSU GPA as a predictor of critical thinking consensus scores. Thus, when predicting critical thinking from GPA, this analysis provides us with an estimate of the average intercept and slope across each classroom. We may also investigate variation in these averages across each classroom. This model may be formally summarized as follows:

Level 1 Model

$$Consensus_{ij} = \beta_{0j} + \beta_{1j}(GPA) + r_{ij}$$

Level 2 Model

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

Within this equation consensus scores are a function of β_{0j} and β_{1j}. The intercept, or β_{0j}, reflects an adjusted group mean for a particular classroom (OSU GPA has been centered around the group mean). The slope, or β_{1j}, indicates the predicted effect of OSU GPA on critical thinking scores for a particular classroom. Two parameters are of particular interest, and include γ_{00}, which reflects that average intercept across all classrooms and γ_{10}, which reflects the average slope across classrooms.

Once again, the average intercept across each classroom was estimated at 2.94 t (17) = 28.39, p < .001. The average slope across each classroom was estimated at .31 t (17) = 5.64, p < .001. On average, OSU GPA is related to critical thinking consensus scores across each classroom. An intraclass correlation coefficient indicates that OSU GPA accounts for approximately 4% of with within-class variation. Stated differently, within-class variation in critical thinking scores is reduced by 4% when accounting for OSU GPA. Aligned with the findings from the unconditional means model, there is substantial variation in the intercept across each classroom $\chi^2(17) = 89.10, p < .001$. However, there is insubstantial variation in the slope across each classroom $\chi^2(17) = 7.18, p = .50$. The extent to which we may reliably estimate the slope as a function of regression equations derived within each class is also extremely low (.03).

Given this information, modeling variation in the slope appears to be unproductive. Since a series of models have already examined variation in the intercept, no additional analyses were conducted.

Summary of Hierarchical Linear Modeling Analysis

The series of models tested in the hierarchical linear modeling analyses indicates that there is substantial variation in the average critical thinking score across each class. Several aspects of assignment characteristics were modeled in order to assess the extent to which they contribute to this variation. However, the only statistically significant predictor was the extent to which faculty indicated that the assignment reflected the students' own perspectives. This judged similarity to the critical thinking rubric accounts for approximately 47% of the variation in average critical thinking scores between the 18 classes included within the analysis. There are various possible explanations for this finding, all of which remain empirically unexamined. For example, it is possible that faculty members who were paid to score student papers are in part influenced by the extent to which a student's perspective is evident within a paper. Sampled papers that fail to have this quality therefore, on average, tend to have lower scores than papers with this quality. If subsequent investigations suggest that this is indeed the case, efforts should be made to control for this assignment quality by sampling student papers that are aligned with this aspect of the critical thinking rubric. A second possibility pertains to variation in the other critical thinking dimensions. Scores for other critical thinking dimensions did not differ as much as the own perspective dimension. This may be because faculty raters have screened artifacts for a fit with the rubric prior to scoring. This restricted variation has two implications (a) measuring these dimensions may be unnecessary in subsequent research and (b) screening efforts may benefit from explicit consideration of the extent to which a student's own perspective is manifest within the paper.

This interpretation however, also remains questionable. For example, when regressing overall consensus scores on each dimension of the critical thinking rubric for individual raters, the own perspective dimension fails to be a statistically significant predictor of consensus scores for 4 of the 7 teams (see Table 4 above). This implies that although the extent to which instructors believe the artifact reflects a student's own perspective predicts differences in the average critical thinking score between assignments, the extent the extent to which most judges actually use this dimension in determining an overall score may be minimal.

It is also of interest to note that all other classroom variables failed to be a statistically significant predictors of variation in critical thinking scores between classrooms. Several additional models, which included various interaction terms and multiple predictors simultaneously, were tested (results not reported here). Details about these analyses are excluded because they failed to substantially change the underlying meaning of the results that have been reported. However, this reinforces an important point. There is substantial variation in critical thinking scores between classrooms, which is synonymous with saying that substantial variation in average critical thinking scores exists between assignments. However, the vast majority of observed variables related to assignment characteristics may be unimportant to understanding this variation.

Ideally, assignment characteristics should not systematically influence critical thinking scores. For example, if, on average, assignment X tends to have higher scores than assignment Y, then critical thinking scores may be influenced by assignment characteristics irrespective of the existing screening procedure. Additional research should investigate this issue, though such efforts may benefit from seeking different observations related to assignment characteristics.

Systematic and Unsystematic Error Variance in Critical Thinking Scores

The assessment of critical thinking, as an aspect of general education, involves sampling student papers across the campus. Raters then score these artifacts using a common rubric. A consequence of this strategy is that raters derive critical thinking scores from writing samples; thus, inferences about critical thinking are strictly located within the realm of written communication. Theoretically, critical thinking is expressible in a range of indicators aside from writing artifacts (e.g. portfolios, videos, oral presentations, etc.). Restricting samples to student writing, as opposed to selecting observations from a range of possible indicators, may lead to questions about whether such a strategy adequately represents the breadth of critical thinking. Additionally, this strategy leads to questions pertaining to the *empirical distinction between critical thinking and written communication*.

Construct-irrelevant variance consists of systematic error variance that affects observed score variation (Messick, 1989). For example, if a test were administered via computer and via paper and pencil, test developers would hope that test scores were unaffected by mode of administration. The extent to which mode of administration affects observed scores is an example of construct-irrelevant variance. The introduction of construct-irrelevant variance, given that this denotes systematic error variance in observed scores that is distinct from the intended aim of a measurement procedure, is a fundamental threat to score-based interpretations and entailed uses of test scores (Messick, 1989). Alternatively, it is possible to conceive of construct-relevant variance, which reflects systematic variation in observed scores that are indeed attributable to the intended target of an assessment procedure. For example, if a test aims to assess critical thinking then construct-relevant variance reflects the extent to which differences in these scores are indeed attributable to critical thinking.

To understand this concern within our current assessment program, briefly consider two students who in reality have the same level of critical thinking, though they differ with respect to written communication. The first student has a high written communication score whereas the second student has a low written communication score. It is conceivable that judges assign the first student a higher critical thinking score than the second student because of poor written communication as opposed to true differences in their critical thinking. Put differently, to what extent are critical thinking scores unduly influenced by “good” or “bad” writing? Can we, and more importantly should we, disentangle aspects of written communication from critical thinking scores?

It is important to recognize that construct-irrelevant variance reflects systematic error, whereas most reliability estimation procedures attempt to quantify unsystematic error variance. Both sources of error variance are a concern however, and ultimately detract from measurement precision and validity.

These issues lead to questions about the empirical distinction between critical thinking and written communication given the existing methodology. This section will therefore investigate sources both sources of error variance as it pertains to critical thinking and written communication assessment:

1. To what extent can we reliably estimate the average critical thinking and written communication scores?
2. To what extent can we reliably estimate mean differences between critical thinking and written communication?
3. How much variance in critical thinking consensus scores is attributable to written communication as a source of systematic error variance?
4. How much variance in critical thinking consensus scores is attributable to critical thinking, after removing the effect of written communication?

The first two questions are addressed using generalizability theory, whereas the second two questions reflect an investigation of written communication as a source of construct-irrelevant variation.

Using generalizability theory to investigate reliability

Two generalizability theory studies (G-study) were conducted in order to assess the reliability of average critical thinking and written communication scores, as well as the reliability of mean differences between these two domains. Given that generalizability theory may be unfamiliar to most readers, this section provides a brief conceptual overview of this framework.

Generalizability theory utilizes analysis of variance techniques in order to partition error into distinct sources of variation. Estimating these sources of variation is of central interest in a *g-study* given that these estimates allow us to ascertain the extent to which identified sources of error pose problems in a given assessment procedure. These different 'factors' are considered 'facets' in generalizability theory terminology. Facets therefore define a theoretical universe of possible observations from which a researcher wishes to generalize. In the context of our current assessment methodology, we are interested in estimating an average critical thinking and written communication score, as well as the mean differences between these scores. We wish to estimate these values across both raters and students, thus raters and students are facets defining our universe of generalization. Similar, to classical test theory, generalizability theory allows us to estimate the ratio of universe score variation to observed score variation.

Within the context of generalizability theory, two coefficients are important. First, is the *generalizability coefficient*, which is utilized when making relative decisions about an object of measurement (e.g. differences between means, differences between students, etc.). The second coefficient is the *dependability coefficient*, and this coefficient is important when making absolute decisions about an object of measurement (e.g. locating an object on a scale). The primary difference between these two coefficients resides in their estimation of error. It is more difficult to locate an object of measurement on a scale, thus all sources of error contribute to an estimation of the dependability coefficient. However, there are some sources of error that are irrelevant to making relative decisions about an object of measurement, thus these sources of error are removed prior to estimating the generalizability coefficient. Both generalizability coefficients and dependability coefficients however, range from 0-1, with values above .80 considered an acceptable level of measurement precision for most practical purposes.

Overview of G-Study Design

Each G-study employed the same basic design. Within each study, UAT assigned a set of artifacts to a group of two raters to score for critical thinking. An independent group of two raters scored these same artifacts for written communication. This design may therefore be summarized as follows: a x (r:d)

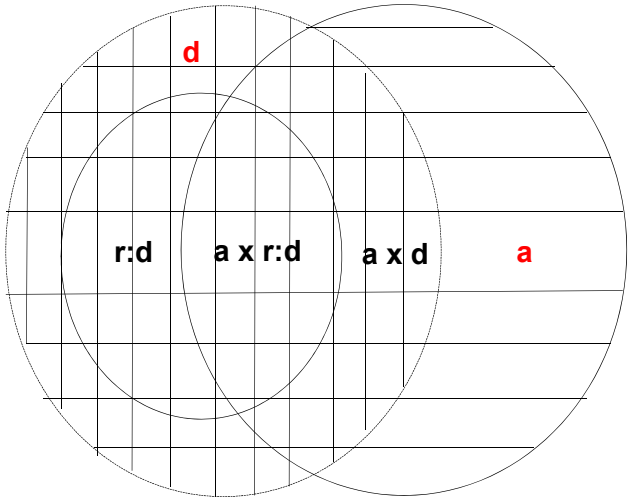
a = artifacts

r = raters

d = domain

In other words, each G-study had a set of artifacts scored by two groups of raters nested within either the domain of critical thinking or written communication. Figure 2 provides the variance attribution diagram for this design.

Figure 2. Variance attribution diagram for a x (r:d) design.



Note: Hatched lines indicate sources of variation that contribute to relative error; vertical lines indicate sources of variation that contribute to absolute error. D = fixed since both writing and critical thinking exhaust our theoretical interests. R and A are both random facets.

Below is a description of each source of variation:

- d = reflects differences in the average critical thinking and writing score across persons and raters.
- r:d = reflects differences in the average rating for each judge within a domain. This source of variation confounds domain differences and a rater by domain interaction.

- a x d = interaction between artifacts and domain; extent to which rank-ordering of artifacts changes across writing and critical thinking.
- a = differences in the average score assigned to each artifact across raters and domains.
- a x r:d = interaction between artifacts and raters within each domain; reflects tendency of judges within each domain to rate each artifact differently. This source of error confounds an artifact by rater by domain interaction, as well as all other unspecified sources of error variation

Study 1 – Social Sciences
Within the first study, two groups of raters scored 32 artifacts for both critical thinking and written communication. Each group consisted of two members. All artifacts were sampled from courses with a Social Sciences general education designation.

Across all artifacts, the average critical thinking score was 2.47 (*SD* = .86) whereas the average written communication score was 3.03 (*SD* = .72). The g study estimated each variance component in order to determine the reliability of estimating means and mean differences from the given assessment procedure. Table 14 provides the results of this analysis. The differentiation variance is estimated at 0.064, which is at least three times larger than all other estimates. This implies that substantial differences exist within the universe of generalization.

The generalizability coefficient was estimated at 0.68 (*SEM* = 0.17), which indicates that the precision with which mean differences may be estimated is a little below acceptable limits. Within the current sample, there is a mean difference of 0.56 points between critical thinking and written communication. When assuming that error is normally distributed, a 95% confidence interval implies that this mean difference may be anywhere between 0.31 to 0.81 points. When estimating mean differences, 74.5% of the error is attributable to variation of raters nested within each domain and 15% is attributable to an artifact by domain interaction. This implies that rater averages within each domain were heterogeneous, or more generally that rater severity tended to vary substantially within each domain.

The dependability coefficient was estimated at 0.61 (*SEM* = 0.20), which indicates that judgments about average critical thinking and written communications scores may fluctuate beyond acceptable limits across replications of the assessment procedure. A 95% confidence interval suggests that the average critical thinking score may be anywhere from 2.07 to 2.87 points. Similarly, a 95% confidence interval indicates that the average written communication score may be anywhere between 2.63 and 3.43 points. The magnitude of these intervals implies that estimates of average critical thinking and written communication scores may vary by nearly a single point when using a 1-5 scale.

Table 14. Generalizability Study for Social Sciences[4]

Source of Variance	df	MS	Differentiation Variance	Error Variance	% Relative	% Absolute
Domain	1	10.125	0.064	-----	-----	-----
Artifacts	31	1.629	-----	0.011	-----	26.8
Raters within Domains	2	1.656	-----	0.023	74.5	54.6
Artifact by Domain	31	0.496	-----	0.004	15.0	11.0
Artifacts by Raters within Domain	62	0.205	-----	0.003	10.5	7.7

Study 2 – Humanities

Within the second study, two groups of raters scored 39 artifacts for critical thinking and written communication. Each artifact was sampled from courses with a Humanities general education designation. As before, raters are nested within the domain of critical thinking or written communication.

Results of the G-study analysis are provided in Table 15. In this case, the differentiation variance, which reflects universe score variance, was estimated at -0.021. Within a generalizability theory framework, it is technically possible to derive negative variance estimates, though such estimates fail to be meaningful (see Cardinet, Johnson, & Pini, 2010). There are various ways to handle these estimates (see Brennan, 2001), one of which is to set this value to zero. Nevertheless, the magnitude of this estimate will be very close to zero, which indicates that there are little differences in the average critical thinking and written communication scores to estimate. A potential implication of this finding is that there is little empirical distinction between these observations within the current assessment procedure.

[4] Whimbey's (1969) correction for mixed model designs is applied when calculating variance estimates.

Table 15. Generalizability Study for Humanities[5]

Source of Variance	df	MS	Differentiation Variance	Error Variance	% Relative	% Absolute
Domain	1	1.852	0.000[a]	-----	-----	-----
Artifacts	38	1.993	-----	0.009	-----	17.9
Raters within Domains	2	3.391	-----	0.036	84.6	69.5
Artifact by Domain	38	0.524	-----	<0.000	0.4	0.3
Artifacts by Raters within Domain	76	0.509	-----	0.006	15.0	12.3

Nevertheless, we can still see that much of the error variance is attributable to raters nested within domains when making either relative (84.6%) or absolute decisions (69.5%). This finding is therefore aligned with the previous analysis, which indicated that rater severity tended to fluctuate within each domain. Within this sample, the average critical thinking scores was 3.44 (*SD* = 1.36) and the average written communication score was 3.23 (*SD* = 0.83). A 95% confidence interval around these mean differences indicates that these estimates may range from -.09 to .81. Put differently, these differences may be estimated as ranging from nearly a single point (1-5 scale) to nearly zero with repeated sampling. When estimating average critical thinking or written communication scores, they will fluctuate, on average, by 0.23 points across repeated sampling. The 95% confidence interval around the average critical thinking score is estimated at 2.99 to 3.89 and the same interval for the average writing score is estimated at 2.78 to 3.68. Once again, the magnitude of these intervals implies that estimates may vary by nearly a single point on a 1-5 scale across repeated replications of the assessment procedure.

Construct-Irrelevant Variance: Writing as a Source of Systematic Error Variance

The purpose of this investigation is to examine our ability to empirically distinguish critical thinking and written communication scores. Unlike the previous section, which focused upon an investigation of unsystematic error variance, this section entails an examination of writing performance as a systematic source of error when assessing critical thinking. Prior to presenting our rationale for the analytical framework, recall that artifacts are scored for specific dimensions (i.e. Writing = content, organization, style, and mechanics; Critical thinking = identification of problem, own perspective, supporting evidence, and conclusion), as well as an overall score.

[5] Whimbey's (1969) correction for mixed model designs is applied when calculating variance estimates. [a] indicates that the variance estimate was -0.021. Since negative variance estimates are not meaningful, this value was set to 0.00.

Rationale. Investigating the pattern of correlations among dimensional scores should provide insight into our ability to empirically distinguish critical thinking and written communication scores. A two-factor solution should account for the correlations between dimension scores wherein each factor is principally defined by critical thinking and written communication dimensions respectively. In other words, critical thinking dimensions should principally load on a single factor and written communication dimensions should principally load on a separate factor.

Before proceeding to a further investigation of construct-irrelevant variance, the section examines the following questions:

1. What is the relationship between the dimensional scores of critical thinking and written communication?
2. What it the number and nature of factors that account for observed correlations between critical thinking and written communication dimension scores?

Table 16 provides the correlations between critical thinking and written communication dimension scores. Scores for writing content was moderately related to identification of a problem (CT) (r = .53, p < .001), own perspective (CT) (r = .47, p < .001), use of supporting evidence (CT) (r = .63, p < .001), and conclusion (CT) (r = .57, p < .001). Scores for writing organization were also associated with identification of a problem (CT) (r = .37, p = .001), own perspective (CT) (r = .38, p = .001), use of supporting evidence (CT) (r = .49, p < .001), and conclusion (CT) (r = .49, p < .001). Scores for writing style were similarly associated with identification of a problem (CT) (r = .34, p = .003), own perspective (CT) (r = .25, p = .028), use of supporting evidence (CT) (r = .49, p < .001), and conclusion (CT) (r = .53, p < .001). Finally, scores for writing mechanics had statistically significant correlations with identification of a problem (CT) (r = .27, p = .018), own perspective (CT) (r = .26, p = .022), use of supporting evidence (CT) (r = .45, p < .001), and conclusion (CT) (r = .51, p < .001).

An exploratory factor analysis was conducted in order to investigate both the number and nature of factors that may account for the pattern of relationships presented in Table 17. Extraction criteria included the K-1 rule, percent of total variance accounted for, as well as the theoretical meaningfulness of extracted factors.

An exploratory factor analysis was conducted initially using an obliman rotation (delta = 0) to allow each factor to be correlated. Under this rotation, a two-factor solution was judged as optimal and an examination of pattern coefficients implied that the structure of each factor tended to coincide with our theoretical expectations in that dimensions associated with writing tended to load on a writing factor and dimensions associated with critical thinking tended to load on a critical thinking factor. However, this rotation was principally employed to examine the correlation between each factor, which was estimated at .48 (p < .001).

Table 16. Correlations Between Critical Thinking and Written Communication Component Score

Dimension	CONT. (WR)	ORG. (WR)	STYLE (WR)	MECH. (WR)	IDENT. (CT)	PERSP. (CT)	EVID. (CT)	CONCL. (CT)
CONT. (WR)	----	.71***	.68***	.59***	.53***	.47***	.63***	.57***
ORG. (WR)	----	----	.66***	.50***	.37**	.38**	.49***	.49***
STYLE (WR)	----	----	----	.80***	.34**	.25*	.49***	.53***
MECH. (WR)	----	----	----	----	.27*	.26*	.45***	.51***
IDENT. (CT)	----	----	----	----	----	.70***	.78***	.75***
PERSP. (CT)	----	----	----	----	----	----	.76***	.70***
EVID. (CT)	----	----	----	----	----	----	----	.82***
CONCL. (CT)	----	----	----	----	----	----	----	----

*** = $p < .001$; ** = $p < .01$; * = $p < .05$. WR = written communication; CT = critical thinking.

Though this was statistically significant, a varimax rotation would force this relationship to approximate zero. This rotation allows us to extract two independent factors, which provides some statistical advantages when using these factors to make subsequent predictions. Once again, a two-factor solution was judged as optimal (see Table 17). After rotation to a final solution, the extracted sum of squared loadings account for 73.5% of the common variance.. Examination of the structure coefficients corresponds with our theoretical expectations in that each rubric dimension tends to load on their own respective factor. There are two exceptions however, which include conclusion (CT) and content (WR) which have cross-loadings > .40. This implies that though we can force the correlation between these two factors to be close to zero, the substantive meaning of each factor is partly contaminated by either writing or critical thinking.

In summary, the above analysis implies that the correlations between critical thinking and written communication dimension scores can be explained by two orthogonal factors. The substantive meaning of these factors generally supports an empirical distinction between critical thinking and written communication, as indicated by the pattern of structure coefficients wherein dimensions of the critical thinking rubric generally clustered together on a common factor and dimensions of written communication generally clustered together on a distinct factor. This provides favorable evidence that indeed empirical distinctions may broadly be made between critical thinking and written communication.

Factor scores were then generated using a regression method. However, the pattern of structure coefficients implies that critical thinking factor scores may be denoted by written communication, particularly scores for writing content. Conversely, aspects of critical thinking may define written communication factor scores, particularly the content dimension scores. This reasoning indicates that some contamination may still exist with the generated factor scores, despite an orthogonal rotation.

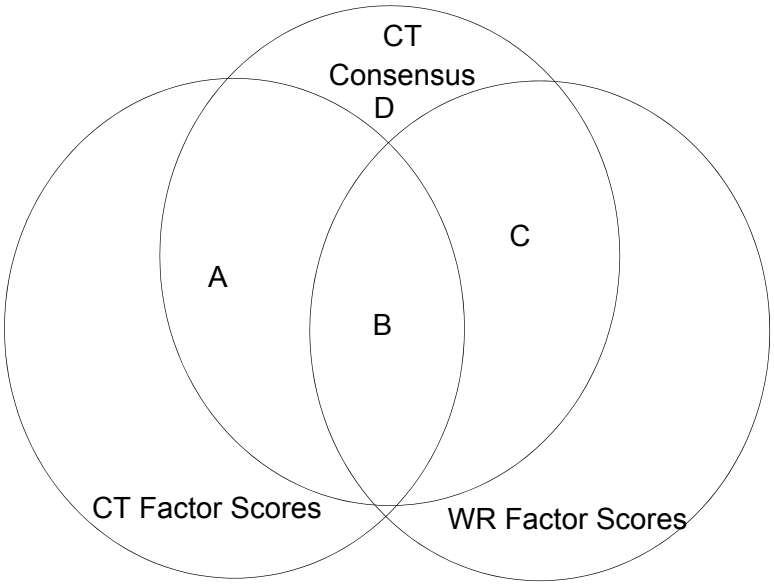
Table 17. Exploratory Factor Analysis on Critical Thinking and Writing Rubric Component Score

Dimension	Critical Thinking	Written Communication	Communalities
Supporting Evidence (CT)	.86	---	.89
Own Perspective (CT)	.83	---	.72
Identification (CT)	.83	---	.73
Conclusion (CT)	.78	.41	.77
Style (WR)	---	.95	.93
Mechanics (WR)	---	.78	.64
Content (WR)	.44	.69	.67
Organization (WR)		.66	.54
Rotated Sum of Squared Loadings	3.08	2.80	
% of Variance	38.5%	35.0%	

Note: Structure coefficients presented after a varimax (i.e. orthogonal) rotation. The correlation between each factor is therefore forced to be 0.0. Percentage of variance for each factor is presented after rotation to a final solution. Structure coefficients < .40 are not reported.

Nevertheless, and though the exploratory factor analysis was promising evidence, we sought to examine the extent to which written communication has a systematic "effect" on critical thinking *consensus scores*. We propose that partitioning sources of construct-irrelevant and construct-relevant variation may be approximated through the statistical framework provided in Figure 3. This framework utilizes the concept of squared partial and squared semipartial correlation coefficients in the context of regression analysis. To understand the rationale behind this framework, a brief overview of these statistical concepts is provided as it relates to Figure 3.

Figure 3. Venn diagram depicting construct-irrelevant and construct-relevant variation



Interpretation Key:

- R^2 = (A + B + C) / (A + B+ C+ D)
- D = unaccounted for variation in critical thinking consensus scores.
- A = part of consensus scores uniquely attributed to critical thinking.
- B = joint contribution of critical thinking and written communication to consensus scores.
- C = part of consensus scores uniquely attributed to written communication.
- Proportion of variance that is construct-relevant = A / (A + D)
- Proportion of variance that is construct-irrelevant = C / (C + D)

As previously indicated, squared partial and semipartial correlation coefficients provide a statistical framework for partitioning construct-relevant and construct-irrelevant variation into distinct sources. In the case of a partial correlation coefficient the effect of one independent variable is simultaneously removed from a second independent variable and the dependent variable. For example, in this context the partial correlation coefficient for critical thinking factor scores reflects the relationship between critical thinking factor scores and critical thinking consensus scores after removing written communication factor scores from both. When squaring these coefficients this would amount to removing the effect B and C in Figure 3. A semipartial correlation coefficient, unlike the partial correlation coefficient, examines the relationship between X and Y after removing the effect of a Z variable from X only. So for example, in this context, a semipartial correlation coefficient for critical thinking factor scores reflects the relationship between critical thinking factor scores and critical thinking consensus scores after removing the effect of written communication from critical thinking factor scores only (see B in the diagram).

This framework provides a rationale for disentangling construct-relevant and construct-irrelevant variance within critical thinking consensus scores. First, recall that construct-relevant variance would constitute systematic variation in critical thinking consensus scores that is indeed reflective of critical thinking. Construct-irrelevant variance reflects systematic differences in critical thinking scores attributable to variation in written communication. Consequently, two considerations imply that a squared partial correlation coefficient is applicable to partitioning this variation whereas a squared semi-partial correlation coefficient is inappropriate. First, the structure coefficients from the exploratory factor analysis suggest that aspects of the critical thinking factor scores are partly denoted by writing content, and conversely the conclusion dimension of the critical thinking rubric in part defines that written communication factor scores. Thus, the removal of section B in Figure 3 is paramount, though this should be minimal given the orthogonal rotation in the exploratory factor analysis. Secondly, the critical thinking consensus score is a function of both critical thinking and written communication factor scores. This entails that when examining critical thinking factor scores, section C from the diagram must be removed along with B; and alternatively when examining written communication section A must be removed with along B.

This leads to an important conclusion. The ratio A / (A + D) indicates the proportion of variation in critical thinking consensus scores that is unique to critical thinking (since it has removed the effect of written communication from both). Consequently A / (A + D) should approximate the proportion of variation in critical thinking consensus scores that is construct-relevant. The ratio C / (C + D) indicates the proportion of variation in critical thinking scores that is unique to written communication (since it has removed critical thinking from both). The ratio C / (C + D) should therefore approximate the proportion of variation in critical thinking consensus scores that is construct-irrelevant.

Results from an ordinary least squares regression analysis indicated that critical thinking and written communication factor scores simultaneously accounted for 74.6% of the variance in consensus scores F (2, 72) = 105.50, $p < .001$. When controlling for written communication scores, critical thinking factor predicted an increase in consensus scores of .87 ($p < .001$). When controlling for critical thinking scores, written communication factor scores predicts an increase in consensus scores of .29 points ($p < .001$). Examination of the squared partial correlation for critical thinking factor scores [A / (A + D)] indicates approximately 71.6% of the variance in consensus scores is construct-relevant. When examining the same estimate for written communication scores [C / (C + D)] approximately 25% of the variation in critical thinking consensus scores may be attributed construct-irrelevant variance. Put differently, approximately 25% of the variation in critical thinking consensus scores is the result of systematic error variance reflected by written communication.

Discussion of Results

The results of this section are mixed. In both generalizability studies, an estimation of a critical thinking and written communication average scores are likely to vary by nearly a single point across replications of the assessment procedure. The magnitude of this interval may be of some concern. In one study, an estimation of both the generalizability coefficient and dependability coefficient were below acceptable levels of precision. In the second study, these values were

not estimated given the extremely small estimate of universe score variation. This implies that there is not much differentiation in critical thinking and written communication scores across persons and raters within the second sample. In both cases however, substantive amounts of error variance appear to derive from average rater scores within a particular domain, or differences in rater severity.

It should be noted however, that the scores analyzed in the generalizability studies are those prior to an examination by a third rater. The extent to which a third rater, by having them score a paper independently for which prior discrepancies exist, improves measurement precision is unknown. Nevertheless, such evidence implies that score consistency across raters continues to be a concern.

Finally, examination of the extent to which written communication is a source of construct – irrelevant variance is promising. Correlations between written communication and critical thinking may be accounted for by two orthogonal factors whose substantive meaning generally aligns with theoretical expectations. There may be particular aspects of written communication, such as scores for writing content, which coincide with critical thinking scores. However, evidence derived from an examination of squared partial correlation coefficients indicates that a substantial amount of variation in critical thinking consensus scores appears to be unique to critical thinking. In other words, after removing the overlap of written communication, nearly 72% of the variance in critical thinking consensus scores may be attributable to critical thinking dimensions. Though this estimate assumes that all relevant sources of construct-irrelevant variance have been identified, which may not be likely, it is still promising evidence that the sampling of critical thinking via written communication may not be as problematic as initially suspected. Nevertheless, it is difficult to argue that written communication contributes little systematic error, given that it is estimated that 25% of the variance in critical thinking consensus scores may be construct-irrelevant variance. How to handle this issue remains a task for subsequent research.

Critical Thinking Rubric

Learning Outcome: Graduates will be able to critically analyze and solve problems.

	Characteristics A -D: Essential Characteristics	Level of Achievement				
		1	2*	3	4**	5
A	**Identification** and/or summary of the **problem/question** at issue.	No identification and/or summary of the problem.		The main question is apparent or implied, but not clearly stated.		The main question and subsidiary, embedded, or implicit aspects of a question are identified and clearly stated.
B	Presentation of the **STUDENT'S OWN perspective and position** as it is important to the analysis of the issue.	The student's own interpretation or position relative to the question is not provided.		The student's own interpretation or position on the question is implied or unclearly stated.		The student's own interpretation or position on the issue is clearly stated.
C	Use of **supporting data/evidence.**	No supporting data, logical argument or evidence is used.		Evidence and logic are used, but source(s) of evidence are not evaluated for accuracy, precision, relevance, and completeness. Inferences of cause and effect are stated, but not completely or entirely accurately. Facts and opinions are stated although not clearly distinguished from value judgments.		Evidence is identified and carefully examined. Source(s) of the evidence are questioned for accuracy, precision, relevance, and completeness. Accurately observes cause and effect. Facts, opinions and arguments are stated and clearly distinguished, and value judgments are acknowledged.
D	Discussion of **conclusions**, implications and consequences.	Conclusions are not provided.		Conclusions are provided without discussion of implications or consequences. Some reflective thought is provided with regards to the assertions.		Conclusions are clearly stated and discussed. Implications and consequences of the conclusion are considered in context, relative to assumptions, and supporting evidence. The student provides reflective thought with regards to the assertions.
E – G: Optional Characteristics (evaluated where appropriate)						
E	Consideration of **OTHER salient perspectives and alternate positions** that are important to the analysis of the issue.	Does not acknowledge possible alternate perspectives.		Acknowledges possible alternate perspectives although they are not clearly stated.		Uses alternate perspectives and additional diverse perspectives drawn from outside information.
F	Assessment of the key **assumptions and the validity of the supporting/ background information**.	Does not identify the key assumptions and/or evaluate the given information that underlies the issue.		The key assumption(s) that underlies the issue is clearly stated. Necessary data or other background data is identified but not evaluated for validity, relevance or completeness.		The key assumption that underlies the issue is clearly stated and the validity of the assumption that underlies the issue is assessed. Key data and background information is evaluated for validity and used in a way consistent with this evaluation.
G	Consideration of the influence of the **context** on the issue (including, where appropriate, cultural, social, economic, technological, ethical, political, or personal context).	The problem is not connected to other issues or placed in context.		The context of the question is provided although it is not clearly analyzed. Limited consideration of the audience is provided. Little consideration of other contexts is provided.		The issue is clearly analyzed within the scope and context of the question. An assessment of the audience is provided. Consideration of other pertinent contexts is provided.

* 2 - Exhibits most characteristics of '1' and some characteristics of '3'
** 4 - Exhibits most characteristics of '3' and some characteristics of '5'

OSU Written Communication Rubric

Learning Outcome: Graduates will be able to communicate effectively in writing.

	Skill	Level of Achievement 1	2*	3	4**	5
A	Content	Topic is poorly developed; support is only vague or general; ideas are trite; wording is unclear, simplistic; reflects lack of understanding of topic and audience; minimally accomplishes goals of the assignment.		Topic is evident; some supporting detail; wording is generally clear; reflects understanding of topic and audience; generally accomplishes goals of the assignment.		Topic/thesis is clearly stated and well developed; details/wording is accurate, specific, appropriate for the topic & audience, with no digressions; evidence of effective, clear thinking; completely accomplishes the goals of the assignment.
B	Organization	Most paragraphs are rambling and unfocused; no clear beginning or ending paragraphs; inappropriate or missing sequence markers. No clear over-all organization		Most paragraphs are focused; discernible beginning and ending paragraphs; some appropriate sequence markers. Overall organization can be inferred and is appropriate for the assignment		Paragraphs are clearly focused and organized around a central theme; clear beginnings and ending paragraphs; appropriate, coherent sequences and sequence markers. Overall organization is clearly marked and is appropriate for the assignment
C	Style and mechanics	Inappropriate or inaccurate word choice; repetitive words and sentence types; inappropriate or inconsistent point of view and tone. Frequent non-standard grammar, spelling, punctuation interferes with comprehension and writer's credibility.		Generally appropriate word choice; variety in vocabulary and sentence types; appropriate point of view and tone. Some non-standard grammar, spelling, and punctuation; errors do not generally interfere with comprehension or writer's credibility.		Word choice appropriate for the task; precise, vivid vocabulary; variety of sentence types; consistent and appropriate point of view and tone. Standard grammar, spelling, punctuation; no interference with comprehension or writer's credibility.
D	Documentation	Intext and ending documentation are generally inconsistent and incomplete; cited information is not incorporated into the document.		Intext and ending documentation are generally clear, consistent, and complete; cited information is somewhat incorporated into the document.		Intext and ending documentation are clear, consistent, and complete; cited information is incorporated effectively into the document.

* Exhibits most characteristics of '1' and some of '3'

** Exhibits most characteristics of '3' and some of '5'

revised 5-14-12