office of Academic Planning & Assessment
University of Massachusetts Amherst

Report on UMass Amherst's Participation in the Second Pilot Year

Valid Assessment of Learning in Undergraduate Education (VALUE) Institute, 2016-2017

In this report, we briefly highlight findings and lessons learned from UMass Amherst's second year (2016-2017) participation in the multi-state, multi-institutional Valid Assessment of Learning in Undergraduate Education (VALUE) Institute to assess student learning outcomes. To learn more about the background on the national project, our decision to participate in it, and the results for the first year, please visit our website.

Background

The VALUE Institute is a multi-state, multi-institutional project coordinated by the Association of American Colleges and Universities (AAC&U). It aims to provide the means for a valid assessment of student learning on key learning objectives (Critical Thinking, Written Communication, and Quantitative Literacy) in ways that allow for comparison of performance by student demographics.

What distinguishes the VALUE Institute from standardized approaches (e.g., the Collegiate Learning Assessment instrument) is that it:

1. uses actual student work, collected from actual course contexts, as the source of evidence of student performance;

2. uses rubrics developed in 2008-2009 by national teams of faculty for key learning objectives (AAC&U VALUE Rubric Development Project); and

3. emphasizes the formative side of assessment (i.e., campus-based involvement in student learning assessment, conversations about how to better teach the assessed skills, and cross disciplinary conversations about teaching).

It is these qualities that led UMass Amherst to participate for a second year in the VALUE Institute. Focusing on student performance is a natural progression from the focus on indirect measures of student learning (student surveys/self-reports, satisfaction measures) that have been a feature of the UMass Amherst Strategic Planning/Unit Planning efforts so far. The VALUE assessment process also echoes the direct assessment process the campus has used to assess student reflective and integrative thinking in the Integrative Experience (IE).

UMass Amherst Participation Specifics

Per the VALUE Institute requirements:

- We focused on upper division (300 or higher) courses and asked for volunteer instructors to submit student work from their courses that they felt met the VALUE Rubric Critical Thinking criteria. (See

Appendix A for VALUE Critical Thinking Rubric.)

- Our target sample size for the VALUE project was 100 pieces of student work from students who had earned at least 75% of the credits for graduation.

- To select the sample, we used a stratified random sample by course and then checked to ensure the sample of students was generally representative of the race/ethnicity, gender, and Pell grant status of the UMass Amherst population as a whole. (See Appendix B for sample/population comparisons by selected demographics.)

- As in our first year of participation, we focused only on Critical Thinking, since that is a learning objective that all Gen Ed courses share, and that is arguably a priority objective for most courses on campus.

Assessment Procedure for Year Two:

- As a result of our internal assessment of the project for the first year, we made two key changes. First, on the basis of scorers' and participating instructors' feedback from the first year, we revised the VALUE Rubric to clarify some descriptors and better align with aspects of critical thinking stressed in UMass courses. (See Appendix C for the UMass Amherst Revised Critical Thinking Rubric.) Second, we revised our guidelines for submitted work so that the assignment demands were more homogeneous (that is, less variation in length and nature of the assignments), and better aligned with the rubric than last year's assignments.

- 19 faculty across 15 departments submitted student work and the corresponding assignment. The highest concentration of participating courses came from HFA (6) with the second highest from both CNS and SBS (3). (See Appendix D.)

- We collected 461 pieces of student work and randomly selected 115 pieces that were submitted to the national project. (114 were coded internally.)

- For internal scoring, we recruited 12 UMass Amherst faculty and 1 representative each from Student Affairs and the Institute for Teaching Excellence and Faculty Development. They participated in a full-day session to familiarize themselves with the rubric and discuss their assessments of sample papers. The scorers then worked independently to score up to 13 student papers apiece, with two readers for each paper. In the case of discrepancies (scores differing by more than one point on two or more criteria), a third reader was used[1]. (See Appendix E for more detail on scoring agreements and discrepancies.) The UMass Amherst scoring was done during Summer 2017 using our UMass Amherst Revised Critical Thinking Rubric. (Details on the procedures for the VALUE Institute national scoring—referred to in this report as "external scoring"—are provided in AAC&U's publication "On Solid Ground: VALUE Report 2017." For that scoring, the Critical Thinking VALUE Rubric was used.)

[1] An important methodological note: as discussed earlier, for our internal process each paper was read at least twice. The double scoring of work is a recommended practice for qualitative assessment, a practice the national project fully supports. However, due to resource constraints, the national project was only able to score each paper once for reporting purposes.

- As we did in the first year of the project, to follow up on the scoring, we also interviewed a number of the instructors who provided student work. As part of these interviews, we provided each instructor with two of the student papers from her/his class (one scored high and one low), asked each to rank-order the papers, and discussed that instructor's evaluation of each in relation to both rubrics (the VALUE Rubric and the UMass Amherst Revised Critical Thinking Rubric). This encouraged the instructors to look closely at the rubrics in relationship to their aims for critical thinking in their courses and their expectations for students and also enabled us to compare their rank-ordering with the assessment scores for each paper.

- During the two years that UMass Amherst has participated in the Value Institute, 29 faculty provided student work and 19 participated as scorers. We have student work from 30 courses, 20 departments, and 6 colleges (Colleges/Schools that did not participate: Nursing, Education, College of Information and Computer Sciences).

Key Findings from Participating

In this second pilot year, findings about specific aspects of the assessment and its overall relative potential are as important as scoring results. Thus, we report both process and scoring results.

Findings: Assessment Process

- For both years, the common rubric and the process of reading student work from UMass Amherst courses in a cross-disciplinary group opened up opportunities to talk together about definitions of critical thinking, expectations of students, and pedagogy. Certainly, it engaged the faculty scorers in talking about teaching and learning in ways that a standardized test cannot.

- Feedback from scorers and analysis of assignments for submitted work validated the revisions that had been made in the guidelines for student work; there was more comparability amongst the assignments and, in turn, the student work submitted across the sample, thus making the results more meaningful.

- Asked about the value of an assessment process such as this, scorers saw merit in it at the university level and, even more so, at the department level.

- During the first pilot year, 2015-2016, a statistical analysis of both the internal and external scores showed a statistically significant correlation of student work scores to assignment length and number of secondary sources, two factors that are not rubric criteria. This finding led to changes for 2016-17 to obtain more commonality in types of assignments, particularly in terms of length and references. In line with these changes, the correlations shrunk for 2016-17, particularly for internal scoring. Indeed, the correlation with internal scores was so small as to not be statistically significant; for external scores, the correlation was weaker than in 2015-16, but still statistically significant. The difference in external and internal correlations may also be attributed to having two scorers for the internal scoring instead of one; thus, the score is not dependent on just one person's judgment. (See "Table 2: Correlation values for number of pages and references by year" in Appendix F.)

- "Critical thinking" is a multi-faceted construct, the components of which can vary by discipline and assignment. Neither the VALUE nor the UMass Amherst rubric encompass all aspects or definitions of critical thinking. Still, similar to last year, the scorers agreed that the UMass Amherst rubric does

assess generally valued traits of critical thinking and does work for evaluating student work across disciplines. At the same time, scorers did mention the difficulty of assessing work in disciplines where they had no expertise (e.g., lack of knowledge of content, scholarship, or specialized analytic methods). This awareness prompted discussion of whether scoring should be based in disciplines or pair a disciplinary expert with an "outside" reader. On balance, scorers still felt it feasible in most instances to assess work according to the rubric criteria, using two readers, and that doing so has the added benefit of prompting cross-disciplinary conversations.

- Interviews with instructors who submitted student work reinforced our sense of the validity of scoring work from outside their own discipline, except in very specialized circumstances. In these interviews, instructors were asked to rank-order two pieces of student work on the basis of the Critical Thinking Rubric criteria. In a majority of instances, their rank-ordering was in alignment with our internal scoring of those papers; that is, in 8 out of 11 instances, the rank-ordering mirrored the scores given by the UMass Amherst scorers. In the three instances where they did not align, the assignments were very specialized either in terms of disciplinary or specific course expectations.

- Scorers that participated in both years indicated that the UMass Amherst Revised Critical Thinking Rubric is a better fit for scoring the student work, although there are still revisions that can be made to clarify some descriptors and so that it better matches the nature of critical thinking valued across disciplines at UMass Amherst.

- The assessment process we followed could be useful not only as a cross-disciplinary project, but also as a departmental one. In fact, faculty who submitted work and those who scored student work mentioned the value of doing so in their departments.

- For both years, scorers emphasized the value of the process. The discussion about critical thinking, the Critical Thinking Rubrics, the review of student work from other disciplines, and the consideration of assignments triggered scorers to think about how they define critical thinking, the ways they teach for critical thinking, how effective their assignments are, and what their standards are and should be. The results, both from the internal and external scoring, provide some data about student performance on the critical thinking criteria. Equally important, these two years of participation have broadened the cadre of instructors with experience with this kind of assessment and have enhanced our capacity to conduct meaningful student learning assessment.

Findings: Scoring Results

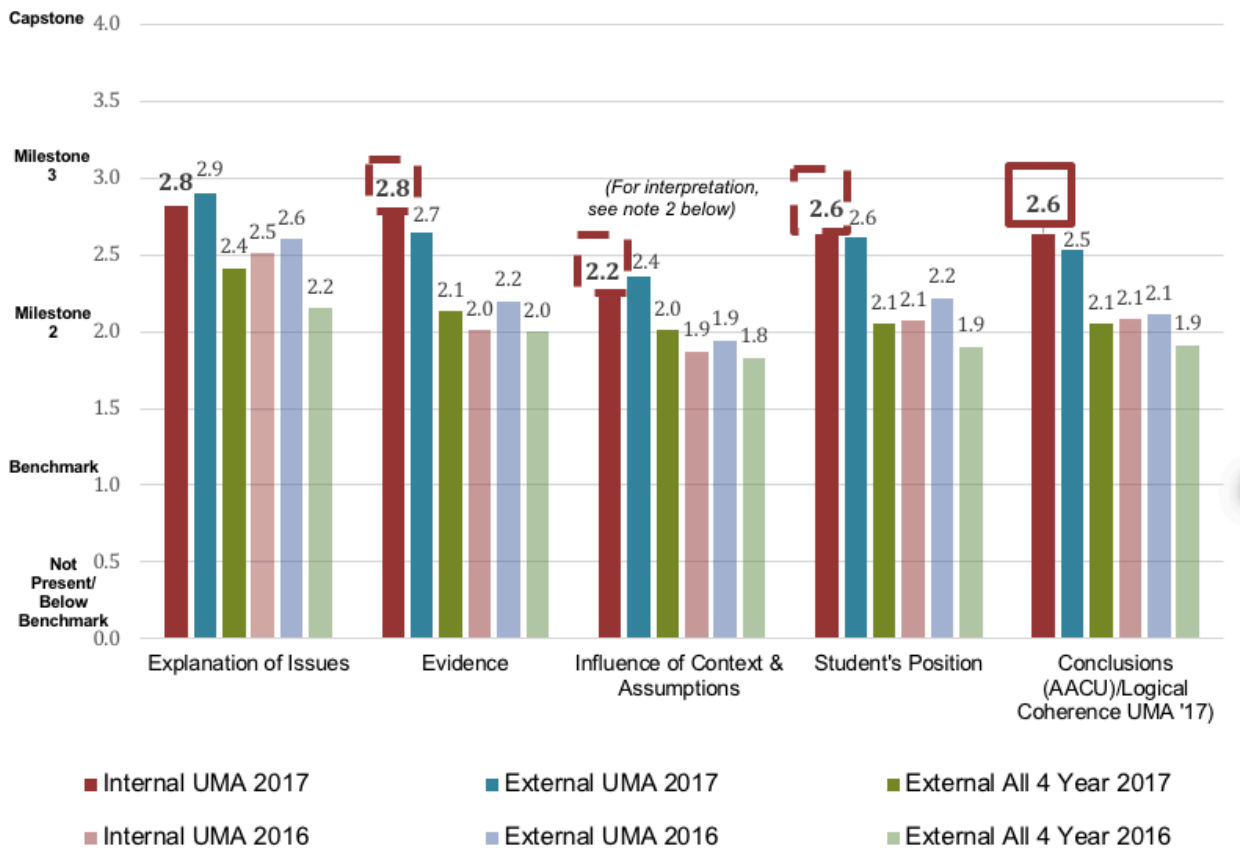
Participation in the national initiative provides us with scoring results not only from our own UMass Amherst internal scoring, but also from the national scoring of the same student work. This enables external comparisons of our students' performance as judged by both internal and external scorers and with those of a sample of upper-division students at other four-year institutions.

Still, there are caveats, among them being the following: first, work from 115 students cannot be considered representative of the senior student population at UMass Amherst, although these 115 students do look similar to the larger population on a number of demographic variables. Second, there is still "noise" in the data, particularly for the "Influence of Contexts and Assumptions" criterion. A majority of instructors who submitted student work indicated that this criterion was not relevant to their assignment; however, scorers assessed all student work on all criteria, not aware of instructor designations. Third, few of the participating four-year institutions included in the external comparison group are Research Universities or institutions to

whom we would generally compare ourselves. That is, in 2016, only six other Research One Universities were included in the 41 four-year institutions whose student results are included in the external comparisons. However, the results are still informative.

Graph One below provides a summary of overall results for each criterion from the first two years of our participation. The results from the most recent year (2017) are on the left, starting with the UMass internal scoring means shown in maroon, followed by the external scoring means shown in dark blue, and then the national benchmark means, representing the results for all four-year colleges and universities who participated in 2017, shown in dark green. The final three bars show the results for 2016 (our first year of participation).

Graph One. Internal and External Scoring, 2016 and 2017 Comparison of UMass Amherst Means, Including with Participating 4-Year Institutions



Notes on Interpreting Graph One

1. Results represent work from students who had earned at least 75% of the credits for graduation. Results for 2016 include 115 pieces of student work, from 14 departments representing 6 schools/colleges; results for 2017 include 115 pieces of student work, from 15 departments representing 5 schools/colleges.

2. In 2017 the UMass Amherst team revised the Critical Thinking Rubric we used for internal scoring. The means for the criteria that had minor word changes are indicated with the broken-line red box. Therefore, comparisons between 2017 internal and external scores for those criteria should be made with caution; comparisons on the last criterion are not appropriate since this criterion is

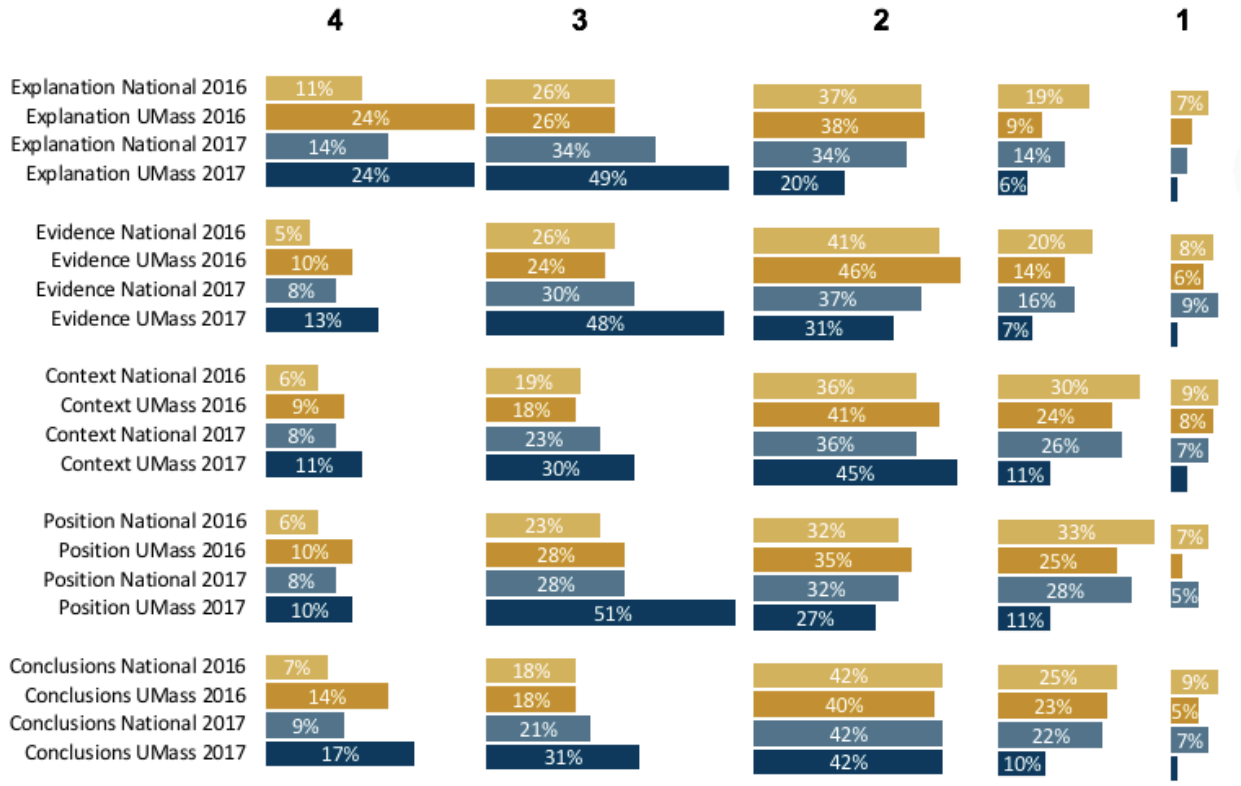
substantially different for internal scoring in 2017. (This mean is indicated with a solid red box.) (See Appendix A & C for the two rubrics used.)

3. The substantial improvement in UMass Amherst results in 2017 is likely related to our efforts to collect student work from assignments that more appropriately aligned with the rubric criteria.

4. The score for "Contexts and Assumptions" should not be taken as a meaningful indication of students' skills because of scoring issues with this criterion: a majority of instructors who submitted student work indicated that this criterion was not relevant to their assignment; however, scorers assessed all student work on all criteria, not aware of instructor designations. As a result, some of the lower scores on this criterion really reflect the fact that the criterion was not called for in the assignment.

Another way of considering these results is to look at the distribution of scores by score-point for each criterion, as shown in Graph Two of the national external scoring. Graph Two shows the score distributions on each criterion for all participating four-year institutions and for UMass Amherst.

Graph Two. Comparison of external score distributions by score-point on each criterion for all participating four-year institutions and UMass Amherst for 2016 and 2017



In 2016-2017, UMass Amherst student work earned higher scores than the UMass Amherst scores in 2015-2016 and the national sample. Over 60 percent of UMass Amherst student work scored at the "Capstone" (4) or "Milestone" (3) level on the "Explanation," "Evidence," and "Student Position" criteria. In the previous year, none of the criteria had more than 50 percent of student work scored at "Milestone" or "Capstone" level. For 2016-2017, there are two criteria that have less than 50 percent of student work at the levels expected of rising seniors (Milestone or Capstone levels), namely, "Context" (41%) and "Conclusions" (49%). For 2016-2017, the UMass Amherst results were substantially higher than those for the national four-year institution results on all criteria.

One of the persistent issues surrounding learning assessment tools is the question of how much the tool captures information about student performance beyond that information captured in more standardized and readily available measures of performance (i.e., SAT scores, Cumulative College GPA). We conducted an initial exploration into this question by looking at the relationship between the VALUE Institute results and the individual student's SAT scores and College GPA. Correlational analyses show that the cumulative Critical Thinking VALUE scores (both External Scores and Internal Scores) have small positive correlation (none larger than r=.22) with students' Math SAT, Verbal SAT, Combined SAT, and College GPA, with two of these correlations reaching statistical significance at $p < .05$) (See Appendix F). As a comparison, the correlation between combined SAT scores and the overall score for the standardized Collegiate Learning Assessment (CLA) test is more than twice as large (r=.46 with CLA score overall, r=.54 for the Performance Task component of the test) (Collegiate Learning Assessment, n.d.). The CLA correlations have raised questions about the extent to which the CLA actually measures student learning gain in college versus simply replicating student performance on pre-college standardized tests.

While preliminary, these correlational results indicate that the VALUE Institute scores are related to other measures of student performance, but that they reflect additional aspects of student performance not captured in the currently available standardized measures of student performance. The findings suggest the potential usefulness of these kinds of direct assessments in capturing information about students' performance in college.

As preliminary and suggestive as these results are, it is reassuring to see that UMass Amherst students' performance is higher than the four-year sample as a whole, even though only a small number of the comparison institutions represent institutions that could be considered UMass Amherst peers. (See AAC&U, 2017, p. 20 for the list of participating four-year institutions.)

Next Steps

UMass Amherst will continue participating in the VALUE Institute national assessment. For this following year, 2017-2018, UMass Amherst will conduct a national and internal assessment of Critical Thinking and Written Communication skills (see all VALUE Rubrics here.)

The UMass Amherst Revised Critical Thinking Rubric will be reviewed to incorporate scorers' and participating instructors' feedback.

References

Association of American Colleges & Universities (AAC&U). (2017). On Solid Ground: VALUE Report 2017. Retrieved from https://www.aacu.org/sites/default/files/files/FINALFORPUBLICATIONRELEASEONSOLIDGROUND.pdf

_____. (2010). VALUE Rubric Development Project. Retrieved from https://www.aacu.org/value/rubrics

Collegiate Learning Assessment. (n.d.) CLA: Frequently Asked Technical Questions. Retrieved from: https://air.sfsu.edu/sites/default/files/CLA%20Technical%20FAQs.pdf (p. 8).

CRITICAL THINKING VALUE RUBRIC

for more information, please contact value@aacu.org

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion.

Framing Language

This rubric is designed to be transdisciplinary, reflecting the recognition that success in all disciplines requires habits of inquiry and analysis that share common attributes. Further, research suggests that successful critical thinkers from all disciplines increasingly need to be able to apply those habits in various and changing situations encountered in all walks of life.

This rubric is designed for use with many different types of assignments and the suggestions here are not an exhaustive list of possibilities. Critical thinking can be demonstrated in assignments that require students to complete analyses of text, data, or issues. Assignments that cut across presentation mode might be especially useful in some fields. If insight into the process components of critical thinking (e.g., how information sources were evaluated regardless of whether they were included in the product) is important, assignments focused on student reflection might be especially illuminating.

Glossary

The definitions that follow were developed to clarify terms and concepts used in this rubric only.

- Ambiguity: Information that may be interpreted in more than one way.
- Assumptions: Ideas, conditions, or beliefs (often implicit or unstated) that are "taken for granted or accepted as true without proof." (quoted from www.dictionary.reference.com/browse/assumptions)
- Context: The historical, ethical, political, cultural, environmental, or circumstantial settings or conditions that influence and complicate the consideration of any issues, ideas, artifacts, and events.
- Literal meaning: Interpretation of information exactly as stated. For example, "she was green with envy" would be interpreted to mean that her skin was green.
- Metaphor: Information that is (intended to be) interpreted in a non-literal way. For example, "she was green with envy" is intended to convey an intensity of emotion, not a skin color.

CRITICAL THINKING VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

Definition

Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion.

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone	Milestones		Benchmark
	4	3	2	1
Explanation of issues	Issue/problem to be considered critically is stated clearly and described comprehensively, delivering all relevant information necessary for full understanding.	Issue/problem to be considered critically is stated, described, and clarified so that understanding is not seriously impeded by omissions.	Issue/problem to be considered critically is stated but description leaves some terms undefined, ambiguities unexplored, boundaries undetermined, and/or backgrounds unknown.	Issue/problem to be considered critically is stated without clarification or description.
Evidence *Selecting and using information to investigate a point of view or conclusion*	Information is taken from source(s) with enough interpretation/evaluation to develop a comprehensive analysis or synthesis. Viewpoints of experts are questioned thoroughly.	Information is taken from source(s) with enough interpretation/evaluation to develop a coherent analysis or synthesis. Viewpoints of experts are subject to questioning.	Information is taken from source(s) with some interpretation/evaluation, but not enough to develop a coherent analysis or synthesis. Viewpoints of experts are taken as mostly fact, with little questioning.	Information is taken from source(s) without any interpretation/evaluation. Viewpoints of experts are taken as fact, without question.
Influence of context and assumptions	Thoroughly (systematically and methodically) analyzes own and others' assumptions and carefully evaluates the relevance of contexts when presenting a position.	Identifies own and others' assumptions and several relevant contexts when presenting a position.	Questions some assumptions. Identifies several relevant contexts when presenting a position. May be more aware of others' assumptions than one's own (or vice versa).	Shows an emerging awareness of present assumptions (sometimes labels assertions as assumptions). Begins to identify some contexts when presenting a position.
Student's position (perspective, thesis/hypothesis)	Specific position (perspective, thesis/hypothesis) is imaginative, taking into account the complexities of an issue. Limits of position (perspective, thesis/hypothesis) are acknowledged. Others' points of view are synthesized within position (perspective, thesis/hypothesis).	Specific position (perspective, thesis/hypothesis) takes into account the complexities of an issue. Others' points of view are acknowledged within position (perspective, thesis/hypothesis).	Specific position (perspective, thesis/hypothesis) acknowledges different sides of an issue.	Specific position (perspective, thesis/hypothesis) is stated, but is simplistic and obvious.
Conclusions and related outcomes (implications and consequences)	Conclusions and related outcomes (consequences and implications) are logical and reflect student's informed evaluation and ability to place evidence and perspectives discussed in priority order.	Conclusion is logically tied to a range of information, including opposing viewpoints; related outcomes (consequences and implications) are identified clearly.	Conclusion is logically tied to information (because information is chosen to fit the desired conclusion); some related outcomes (consequences and implications) are identified clearly.	Conclusion is inconsistently tied to some of the information discussed; related outcomes (consequences and implications) are oversimplified.

Appendix B: Campus Representativeness

Our sample size is too small to make any claims about representativeness of the data, and we urge caution in using these results to generalize about learning on campus. This appendix is primarily to demonstrate the distribution of our sample across Colleges and Departments, as well as to compare the demographics of our sample to those of the undergraduate population.

Table 1. Select Comparison Demographic ≥ 75% of Credits for Graduation Earned: Comparison of student with submitted work sample and UMass Amherst students as a whole.

	Sample 2017	UMass 2017
School/College		
Education	0%	0.3%
PHHS	23%	10%
CNS	17%	28%
HFA	29%	8%
Engineering	6%	11%
CICS	0%	5%
SBS	20%	17%
ISOM	3%	18%
Nursing	0%	2%
Other-BDIC	4%	2%
Gender		
Female	50%	47%
Male	50%	53%
URM		
URM	14%	9%
Asian	6%	10%
White	67%	71%
Unknown	12%	10%
International	2%	4%
PELL Eligible		
Yes	23%	22%
No	77%	78%
Transfer		
Transfer	25%	24%
Frosh	75%	76%
First Gen		
Yes	32%	27%
No	68%	73%

Appendix C. UMass Amherst Revised Critical Thinking Rubric, adapted from AAC&U VALUE rubric

	Capstone 4	Milestone 3	Milestone 2	Benchmark 1
Explanation of Issues	Issue/problem to be considered critically is stated clearly and described comprehensively, delivering all relevant information necessary for full understanding	Issue/problem to be considered critically is stated, described, and clarified so that understanding is not seriously impeded by omissions	Issue/problem to be considered critically is stated, but description leaves some terms undefined, ambiguities unexplored, boundaries undetermined, and/or backgrounds unknown	Issue/problem to be considered critically is stated without clarification or description
Evidence	Information is taken from primary and/or secondary source(s) with enough interpretation/ evaluation to develop a comprehensive analysis or synthesis	Information is taken from primary and/or secondary source(s) with enough interpretation/ evaluation to develop a coherent analysis or synthesis	Information is taken from primary and/or secondary source(s) with some interpretation/ evaluation, but not enough to develop a coherent analysis or synthesis, or it is rather obvious, too generalized, or not backed by evidence	Information is taken from primary and/or secondary source(s) without any interpretation/ evaluation
Influence of Contexts (See note)	Identifies one or more contexts and thoughtfully evaluates their relevance when presenting a position	Identifies one or more contexts and adequately evaluates their relevance	Identifies one or more relevant contexts but the evaluation is rather shallow	Begins to identify some contexts when presenting an opinion
Student's Position (perspective, thesis, hypothesis, conclusion)	Specific position is insightful, taking into account the complexities of an issue/topic. Limits of position are acknowledged; and/or others' points of view are synthesized within position	Specific position takes into account the complexities of an issue/topic. Limits of position and/or others' points of view are acknowledged within position	Position is stated and goes beyond being a simplistic assertion, but is overly generalized and/or doesn't acknowledge limits of position or others' points of view or does so only minimally	Position is stated, but is simplistic and obvious.
Logical Coherence	Student's position and supporting points are logical	Student's position and supporting points	Student's position is logically tied to information for the	Student's position is inconsistently tied to the information

	Capstone 4	Milestone 3	Milestone 2	Benchmark 1
	and reflect student's informed evaluation and ability to place evidence and perspectives discussed in priority order	are logically tied to range of information (including others' points of view as relevant)	most part but with some breakdown in the logic of the line of development either overall or within paragraphs	discussed and/or the overall work lacks a logical line of development to develop the position.

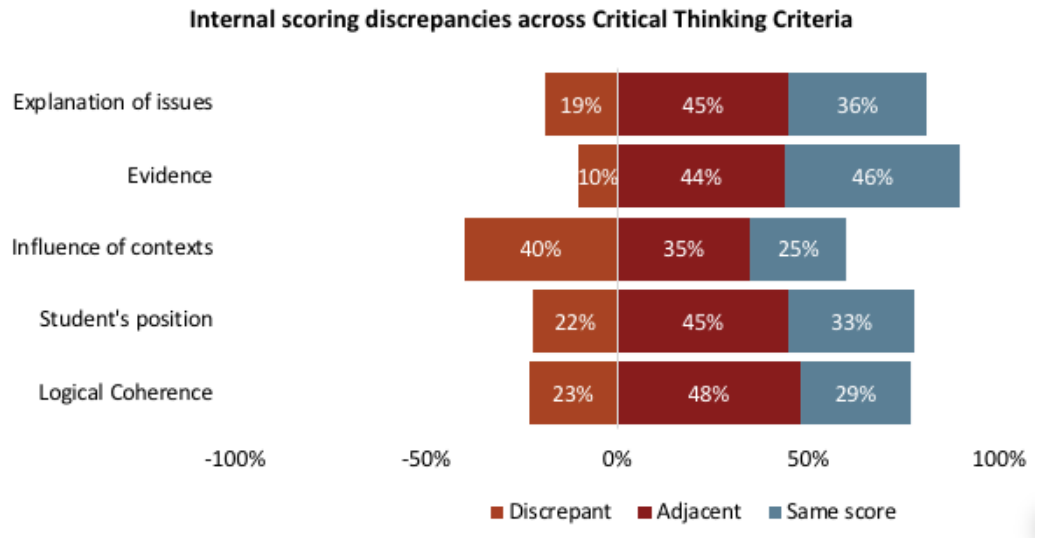
Appendix D. Distribution of the sample across School/College and Department 2017

School/College	School/College Represented in sample*	# Departments in School/College	# Departments in sample
College of Education		3	0
School of Public Health and Health Sciences	*	7	2
College of Natural Sciences	*	15	3
College of Humanities and Fine Arts	*	14	6
College of Engineering	*	5	1
College of Information and Computer Sciences		1	0
College of Social and Behavioral Sciences	*	15	3
Isenberg School of Management	*	7	0
College of Nursing		1	0
Other--ROTC	*	1	0

Appendix E: UMass Internal Scoring Reliability

There were two UMass Amherst readers for each paper. Scores on each criterion were considered acceptable if they were the same or differed by only one level (for example, a score of 2 and 3). When the two readers differed by two or more levels on the rubric (for example, a 2 and a 4), their scores were considered "discrepant." When a paper received discrepant scores on two or more criteria for a given paper, a third reader was used. In all cases, whether a paper was scored by two or three readers, the reader scores were averaged to yield the final score for the paper.

As the graph below illustrates, the percentage of scoring discrepancy is similar for four of the five criteria, with influence of context being the exception (with substantially more discrepant scores). For influence of context, only one-quarter of all scores were the same and more than one-third had a discrepancy of more than one. We attribute this low reliability to the particular challenge of scoring the context criterion.



Of the 114 papers scored internally, 34 (30%) required a third reading; that is, the initial two readers' scores differed by two or more score points on at least two of the traits.

Appendix F: Correlational Analyses

The scoring process revealed that two paper characteristics had a relationship with national average score: length and use of external sources.

Table 1: 2016-2017 Descriptive Information on Number of Pages and References for 115 Papers.

	Overall Average	Overall Median	Overall Range
Number of Pages	9.20	8	3-30
Number of References	9.67	7	0-73

Table 2 shows that there is a weak positive correlation between a paper's page length and external scores, and the same type of correlation for number of references and external scores. Longer papers and those using more external sources tended to score higher on the rubric's Critical Thinking criteria than shorter papers and those using fewer references. The correlation with internal scores was not significant.

Table 2. Correlation values for number of pages and references by year

	2016		2017	
	External Average Score	Internal Average Score	External Average Score	Internal Average Score
Number of Pages	0.60**	0.58**	0.21*	0.11
Number of References	0.55**	0.61**	0.27**	0.15

**Correlation is significant at the .05 level*
*** Correlation is significant at the 0.01 level*

Table 3. Correlation values for Math SAT, Verbal SAT, Combined SAT, and GPA by external and external scorers for 2017

	Math SAT scores	Verbal SAT scores	SAT Combined	GPA
External Average 2017	0.15	0.20*	0.22	0.16
Internal Average 2017	0.11	0.18	0.17	0.21*

**Correlation is significant at the .05 level*
*** Correlation is significant at the 0.01 level*