UNIVERSITY OF
TEXAS
ARLINGTON

INSTITUTIONAL EFFECTIVENESS AND REPORTING

Measuring Critical Thinking

SPRING 2018 REPORT

The University of Texas at Arlington

Measuring Critical Thinking, Spring 2018 Report

This report summarizes evidence of *Critical Thinking* in embedded assignments from students enrolled in undergraduate TCC courses at The University of Texas at Arlington (UTA). The outcome was measured using the *Critical Thinking* VALUE rubric developed by the Association of American Colleges and Universities ([AAC&U]; AAC&U, 2013; Rhodes, 2010). The purpose of this report is to present findings from the assessment of *Critical Thinking* during the 2018 spring semester at UTA

The university sustains assessment using a multi-year schedule that covers the six TCC objectives within each of the eight Foundational Component Areas (FCA). It represents an effort to reduce the burden of assessment on departments while maintaining consistent data collection. This report contains a summary of the findings from three FCAs: Communication, Creative Arts, and Social and Behavioral Sciences.

Method

The project gathered evidence of *Critical Thinking* within a representative sample of students enrolled in Texas Core Curriculum (TCC) courses at UTA and recruited qualified and engaged raters read/score each written student artifact. More than half of the students were female (59%; $n = 118$), the rest were male (41%, $n = 83$). Students primarily represented four ethnic groups: White (46%, $n = 92$), Hispanic (22%, $n = 44$), Black/African American (15%, $n = 31$) and Asian (7%, $n = 14$). Many (42%) perceived themselves to be first-generation college students and almost half (49%) were Pell Grant eligible (see Table 1). In terms of UTA college and school representation, a majority of the students were from the College of Nursing and Health Innovation (30%), the College of Liberal Arts (29%) or the College of Science (15%), however the student artifacts were collected from a variety of approved TCC general education courses.

Course types included both on-line and on-campus modalities of traditional semester and dynamically-dated duration. Most students in this sample were enrolled in a traditional on-campus

Table **1**
Student Demographics

Categorical Information	Number of Students	Percent
Gender		
Female	118	59
Male	83	41
Ethnic Description		
American Indian or Alaskan Native	1	<1
Asian	14	7
Black, African American	31	15
Foreign, Non-Resident Alien	10	5
Hispanic, All races	44	22
Multiple Ethnicities	8	4
Unknown or Not Specified	1	<1
White, Caucasian	92	46
Level		
Freshman	43	21
Sophomore	82	41
Junior	56	28
Senior	20	10
Type of course		
Traditional 16-week on-campus	127	63
Traditional 16-week on-line	36	18
Accelerated, dynamic-dated on-line	38	19
First generation college student (self-report)		
Yes	84	42
No	117	58
Pell Grant eligible upon admission (self-report)		
Yes	98	49
No	103	51
Transfer Student		
Yes	103	51
No	98	49
College or School		
College of Nursing and Health Innovation	61	30
College of Liberal Arts	58	29
College of Science	30	15
College of Business	20	10
Division of Student Success	15	8
College of Engineering	8	4
College of Architecture, Planning, & Public Affairs	5	3
School of Social Work	3	2
College of Education	1	<1

setting in which they met in a classroom face-to-face with their instructor several times a week for 16 weeks (63%). Others interacted with the course instructor and course materials (37%) via a

curriculum-management system over the internet. Some of the on-line students (*n* = 38) were enrolled in accelerated, dynamically dated course sections, typically eight weeks in duration. The remainder of the on-line students (*n* = 36) followed the traditional semester schedule.

Student essays were collected from TCC courses to measure attainment of *Critical Thinking*. Typically freshmen and sophomore-level students enroll in these courses. That said, upper division and transfer students who need to meet graduation criteria for the TCC also enroll. Because the research study examines attainment of *Critical Thinking,* an ideal sample would have a higher ratio of sophomores and juniors than freshmen, because they likely completed more TCC courses at UTA.

Some TCC courses assigned students a research position paper that examined an issue from different viewpoints. Work samples from the Social and Behavioral Sciences FCA consisted of the student's reflections after interviewing a person who was born outside our country. A third assignment from the Creative Arts FCA courses directed the students to attend and analyze performance art. Preparation of the work samples for rating involved assigning the papers a coded tracking number and then removing all personal identification information (e.g., the student's name, the faculty instructor's name) to prevent rater bias during the planned group "Scoring Day" activities.

Assessment Instrument

The AAC&U's Critical Thinking Rubric (AAC&U, 2015) was used as the assessment instrument. It was developed by a multi-disciplinary team of faculty experts gathered by the AAC&U with funding from the Lumina Foundation. The rubric is conceptually divided into dimensions that represent *Critical Thinking:* 1) *Explanation of issues,* 2) *Evidence,* 3) *Influence of context & assumptions,* 4) *Student's position (perspective, thesis/hypothesis),* and 5) *Conclusions & related outcomes (implications and consequences).* The rubric contained a narrative description of the expected quality for each essay and the corresponding point values for rating the five dimensions. Rating values ranged from 1 - 4, with 4 representing the highest observed levels of *Critical Thinking*.

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.

Figure 1. Critical Thinking VALUE Rubric.

Raters, best practices for "Scoring Day" and inter-rater reliability goals

Providing training in the use of rubrics as a professional development opportunity, among other things, seemed to enhance recruitment efforts to gather a multi-disciplinary group of raters from within the UTA faculty. All had earned masters or doctoral degrees in their respective fields and three had professional certifications (see Table 2). On average, they had eight years of teaching experience at the university level *(M* = 8.06, *SD*= 7.27, *Range*= 0 - 22). The multi-disciplinary group of raters represented the College of Business (6%), College of Liberal Arts (56%), Division of Student Success, (13%), and Other (Center for Distance Education, Office of Institutional Effectiveness and Reporting; 25%).

Table 2
Rater Demographics

Categorical Information	Number of Raters	Percent
Gender		
Female	10	62.5
Male	6	37.5
Ethnic Description		
Asian	0	0
Black, African American	1	6.3
Hispanic, All races	3	18.8
White, Caucasian	13	81.3
Classification		
Faculty	5	31.3
Graduate Teaching Assistant	6	37.4
Staff	5	31.3
Highest Degree Received		
Masters	10	37.5
Doctoral	6	62.5
Additional Certifications		
IEEE	1	6.3
Licensed Mediator	1	6.3
Texas Teaching Certificate	1	6.3

On scoring day, a faculty expert from the English department led 16 raters in a guided a group discussion about the distinctions of between rating and grading and the use of the rubric. For example, the facilitator described identifiable features for each level of the rubric and then all the raters read a student work sample chosen by the facilitator for discussion. During this step in the calibration process, each rater read the essay and assigned ratings for each rubric dimension. After the facilitator tallied the dimension ratings using a simple show of hands, she led a discussion aimed at reaching a common understanding of each measure of *Critical Thinking* and the group discussed the elements that a paper must contain for awarding a score at each level. After sufficient consensus was reached, the scoring process began. A minimum of two raters individually read each paper and scored it independently using the rubric. After rating, Rater A placed an adhesive "post-it"-type note as a covering over their ratings on the score sheet to avoid biasing Rater B with their scores. Then Rater A passed the paper to Rater B to read and score.

Achievement of inter-rater agreement was a high priority. If the values awarded by the two

raters were identical or within two points, then scoring was completed and during analyses the scores were averaged. For example, if Rater 1 scored the *Explanation of issues* measure with a value of 2 and Rater B scored *Explanation of issues* with a value of 4, then the average of the two scores for *Explanation of issues* was 3. If the scores from the two raters differed by more than two points, then a third rater would read the paper and an average of the three scores would be calculated. For this group of essays and raters, a third rater never became unnecessary. Figure 3 displays an image of the rater score sheet.

	Rater 1				**Rater 2**				**Rater 3** (only if needed)			
Explanation of issues	4	3	2	1	4	3	2	1	4	3	2	1
Evidence	4	3	2		4	3	2		4	3	2	
Influence of context & assumptions	4	3	2		4	3	2		4	3	2	
Student's position	4	3	2		4	3	2		4	3	2	
Conclusions & related outcomes	4	3	2		4	3	2		4	3	2	

Figure 3. Rater Score Sheet used on scoring day with the Critical Thinking VALUE Rubric

Analysis and Results

Inter-rater reliability

Inter-rater agreement analyses assessed whether the rater scores corresponded to each other for a particular student paper. Levels of agreement were determined by calculating the intraclass correlation coefficient (ICC). High ICC values (Fleiss Kappa) indicate more agreement between rater scores (Fleiss, 1986; Shrout & Fleiss, 1979). For this sample, values indicated a trend of good to excellent agreement (see Table 3). These high values give confidence to proceed with analyses involving student attainment.

Table 3
Intraclass Correlation Coefficient (Fleiss 'Kappa) for Critical Thinking dimensions

Critical Thinking VALUE Rubric Dimension	*n* = 201
Explanation of issues	0.80
Evidence	0.71
Influence of context & assumptions	0.69
Student's position	0.68
Conclusions & related outcomes	0.69

Note 1: *less than 0.40 = poor agreement; between .40 and.74 = fair to good agreement; greater than . 74 = excellent agreement.*
Note 2: *the intra-class correlation coefficient (ICC) was calculated as a one-way random effects model. Values in this type of model with random rater pairings are typically expected to be lower than models where rater pairings are fixed throughout rating day.*

Scores from Signature Assignment ratings

The distributions of score frequencies for each of the dimensions closely followed standard normal curves with more student scores along the mean (rated values between 2 and 3) and fewer scores at the two tales of the curve (rated values between 1 and 4). Table 4 contains the score frequencies of all the ratings. Because each paper was rated twice there are twice as many ratings *(n* = 402) as papers *(n* = 201). The means for each dimension (see Table 5) show that one of the five dimensions, *Explanation of Issues,* had an average score of 2.5. Importantly, the rest of the average scores attained the standard targeted threshold recommended by the AAC&U, a score of 2. Our institution follows their recommendation and targets 2 as the targeted outcome. These results indicate that, on average, UTA undergraduates exceeded the target in all five targeted dimensions.

Table 4
Frequencies for Critical Thinking Dimension Rating Scores

		Rubric Values (Percent of Student papers)							
Measurement dimensions	N	1		2		3		4	
Explanation of issues	402	33	(8%)	156	(39%)	191	(48%)	22	(5%)
Evidence	402	54	(13%)	193	(48%)	145	(36%)	5	(2%)
Influence of context	402	57	(14%)	186	(46%)	147	(37%)	12	(3%)
Students position	402	57	(14%)	150	(37%)	180	(45%)	15	(4%)
Conclusions	402	70	(17%)	192	(48%)	124	(31%)	16	(4%)

Note: Each paper was rated twice, therefore the number of ratings contained in this table is double the number a/papers (N=201).

Table 5
Means for Critical Thinking Measure Scores

Measurement Dimensions	N	Mean	SD	Percent> *μ-1rJ*
Explanation of issues	201	2.50	0.66	91.8
Evidence	201	2.28	0.64	86.6
Influence of context	201	2.28	0.64	85.8
Students position	201	2.38	0.67	85.8
Conclusions	201	2.22	0.68	82.6

Analyses probed the student scores further using standardized scores and the Empirical Rule (e.g., 68-95-99.7 Rule, first described by de Moivre in 1733) in order to answer the question "what percent of students score within one standard deviation of the mean or better?" These analyses assume a standard normal curve (e.g., bell-shaped) and analyses found that these data were skewed negatively with more rating values of 1 than rating values of 4. That said, the Empirical Rule drills deeper into the data to count the student scores that are above the mean or not statistically different from the mean. This step adds to the evidence by examining meaningful target thresholds for student attainment. The targeted threshold proposed from the Empirical Rule determines whether 84% of students would have a score that was greater than negative 1 standard deviation from the mean (84% > *μ-1rJ).* For this sample, students exceeded that goal in four of the five dimensions and more than eighty percent of the students scored greater than negative 1 standard deviation of the mean (see Table 5) across all measures. This indicates that a majority of undergraduates enrolled in these TCC courses scored above the mean or statistically no different than the mean.

Further examination of the relationships between the student characteristics and the five written *Critical Thinking* dimensions used analysis of variance (ANOVA). Significant effects for gender, Pell eligibility, GPA, if they transferred from another institution (yes or no), if they live on campus (yes or no), if course was taught on-line 8-week, on-line 16-week or on-campus 16-week, and first generation (yes or no) for the five rubric dimensions were not found. As expected there was an effect for upper division students (juniors and seniors) versus lower division students (freshmen and sophomores) for

Explanation of issues F(l) = 5.81, $p = 0.02$; *Evidence* F(l) = 5.36, $p = 0.02$; *and Conclusion s,* F(l) = 5.38, $p = .02$; where upper students attained higher scores on average. That said, linear regression revealed that higher numbers of completed semester credit hours (SCH) did not significantly predict higher scores on any of the five rubric dimensions.

Summary

The current assessment of signature assignments used an adapted AAC&U Critical Thinking VALUE rubric. Results revealed *Critical Thinking* strengths and weaknesses in a sample of undergraduate students. In addition , analyses included an examination of student characteristics in order to identify trends and comparisons by groups.

In this sample of papers scored in the spring of 2018, average student scores were strongest for the *Explanation of issues* dimension from the rubric. The means for the other four dimensions exceeded the threshold value. Importantly, for all dimensions, the student's average scores met previous threshold criteria established by the university and standard use criteria set by the AAC&U (rubric values of two or better).

In addition, this *Critical Thinking* report includes analyses that examine additional attainment criteria using standardized scores and the Empirical Rule. In doing so, this report continued the inquiry into a new target of having 84% of the students attain scores above or within one standard deviation of the mean for each dimension. Used in conjunction with the AAC&U threshold, which indicated attainment for all dimensions, this additional analysis drilled down a bit further to show that students met the threshold of 84% for all but one dimension of the *Critical Thinking* Core Curriculum Objective. While these analyses continue to be exploratory in nature, they suggest that future studies continue this analytical approach to examine trends in student performance and improvement because they further differentiate strengths and weaknesses beyond a

simple look at the mean score.

An examination of student characteristics indicated that the sample was generally descriptive of the university. Continued evidence of the student attainment of *Critical Thinking* supports that there is the same quality of instruction in the dynamically dated on-line courses as in the traditional length semester. This is encouraging, as students in these courses interact within an accelerated schedule outside of the traditional brick-and-mortar institution. That said, this evidence was limited by the size of the sample, and plans to continue this line of inquiry should span all six TCC objectives.

This report contains evidence from three of the eight Foundational Component Areas (Communication, Creative Arts, and Social and Behavioral Sciences). Authentic performance-based student work samples were collected for this measurement as part of the multi-year plan to assess *Critical Thinking.* This report presents positive evidence of student attainment for *Critical Thinking* in the five AAC&U Critical Thinking VALUE Rubric dimensions using the student essays rated in the spring 2018. All of the reports developed by UTA to meet the THECB requirements are available from the Office of Institutional Effectiveness and Reporting.

References

Association of American Colleges and Universities (2015). *VALUE Rubrics.* Retrieved from https://www.aacu.org/value-rubrics/

Fleiss J. L. (1986). *The design and analysis of clinical experiments.* New York: John Wiley & Sons.

National Association of Colleges and Employers. (2016). *Job Outlook 2016.* Bethlehem, PA.

Rhodes, T. (Ed.). (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics.* Washington, DC: Association of American Colleges and Universities.

Shrout, P., & Fleiss, J. L. (1979). Intraclass correlation: uses in assessing rater reliability. *Psychological Bulletin, 86(2),* 420 - 428.