UNIVERSITY OF
TEXAS
ARLINGTON

INSTITUTIONAL EFFECTIVENESS AND REPORTING

Measuring Critical Thinking

MAY 2018 REPORT

The University of Texas at Arlington

Measuring Critical Thinking, May 2018 Report

Evidence of *Critical Thinking* was examined in embedded assignments from students enrolled in undergraduate Texas Core Curriculum (TCC) courses at The University of Texas at Arlington (UTA). The tool used for measuring this outcome was the *Critical Thinking* VALUE rubric developed by the Association of American Colleges and Universities ([AAC&U]; AAC&U, 2013; Rhodes, 2010). This report presents findings from the assessment of *Critical Thinking* during the 2018 spring semester at UTA.

Assessment is sustained at the university using a multi-year schedule that covers the six TCC objectives within each of the eight Foundational Component Areas (FCA). This schedule was developed 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; History; Language, Philosophy and Culture; and Life and Physical Sciences.

Method

The project gathered evidence of *Critical Thinking* within a representative sample of students enrolled in TCC courses at UTA. Qualified and engaged raters were recruited to read/score each written student artifact. Demographic data for most of the sample *(n* = 250) was obtained to describe the students represented. More than half of the students were female (56%; *n* = 141), the rest were male (44%, *n* = l 09). Ethnicity varied as expected for a diverse student body such as UTA: Hispanic/Latino (34%, *n* = 85), White (30%, n = 72), Asian (17%, *n* = 43), and Black/African American (8%, *n* = 21). The average age of the students who prepared these work samples was 20 ½ years of age *(M*= 20.5, *SD*= 2.1, *Range*= 18 - 33). More than a third (36%) perceived themselves to be first-generation college students and many (44%) were Pell Grant eligible (see Table 1). All UTA college and school representation were represented, however, all the student artifacts were collected from a variety of approved TCC-approved general education courses in four FCAs.

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, as they likely completed more TCC courses at UTA.

Table **1**
Student Demographics

Categorical Information	Number of Students	Percent
Gender		
Female	141	56
Male	109	44
Ethnic Description		
American Indian or Alaskan Native	2	<1
Asian	43	17
Black, African American	21	8
Foreign, Non-Resident Alien	16	6
Hispanic, All races	85	34
Multiple Ethnicities	7	3
Unknown or Not Specified	2	<l
White, Caucasian	74	30
Level		
Freshman	22	9
Sophomore	82	33
Junior	74	30
Senior	72	29
First generation college student (self-report)		
Yes	89	33
No	161	60
Pell Grant eligible upon admission (self-report)		
Yes	111	42
No	139	52
Transfer Student		
Yes	41	15
No	209	78
Not reported	**1**7	6
Housing		
On campus	11	4
Off campus	239	89

Note: Student demographics variables were available for 250 of the 267 essays, but not availabl e for six percent of the sample (n = I 7).

Some TCC courses assigned students a research position paper that examined an issue from different viewpoints. Work samples from Language, Philosophy and Culture FCA consisted of the student's application of ethical frameworks to an issue or decision. An assignment from History FCA courses directed the students analyze historical perspectives in scholarly papers ranging from the influence of international partners on United States economic policy to labor union policies in the United States. The last type of assignments were lab reports that reported on results of experiments; these represented the Life and Physical Sciences FCA. It is important to note that all were embedded assignments. 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 assessment instrument used in this report was the AAC&U's Critical Thinking Rubric (AAC&U, 2015), developed by a multi-disciplinary team of faculty experts. 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.* Raters read the student papers and rated each measure.

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 raters had earned master's or doctoral degrees in their respective fields and six had professional certifications (see Table 2). On average, they had over nine years of

teaching experience at the university level *(M* = 9.33, *SD*= 6.98, *Range*= 0 - 20). The multi-disciplinary group of raters represented the College of Liberal Arts (47%), College of Education (20%), College of Nursing and Health Innovation, (2%), College of Science, (7%), School of Social Work, (7%), and Other (7%).

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.

Table 2
Rater Demographics

Categorical Information	Number of Raters	Percent
Gender		
Female	11	73
Male	4	27
Ethnic Description		
Hispanic, All races	0	0
White, Caucasian	15	100
Classification		
Faculty	9	60
Graduate Teaching Assistant	5	33
Staff	1	7
Highest Degree Received		
Masters	7	47
Doctoral	8	53
Additional Certifications		
IEEE	1	7
K-12 Teaching Certificate	1	7
Licensed Social Worker	1	7
Registered Nurse (RN) or RN, Clinical Nurse Specialist	2	14
Texas Principal Certificate	1	7

On scoring day, a faculty expert from the English department led 15 raters in a guided group discussion about the distinctions 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. Using a simple show of hands to tally the dimension ratings, the facilitator 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, assignment of the paper to a third rater was 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	4	3	2		4	3	2	1
Evidence	4	3	2	4	3	2	1	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

As the name implies, "inter-rater agreement" analyses assess whether raters that read the same paper agreed on the values to award. 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 good to excellent agreement (see Table 3). These high values give confidence to proceed with analyses involving student attainment.

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

Critical Thinking VALUE Rubric Dimension	*n* = 267
Explanation of issues	0.79
Evidence	0.71
Influence of context & assumptions	0.72
Student' s position	0.74
Conclusions & related outcomes	0.76

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 coef ficient (ICC) was calculated as a one-wa y random effe cts model. Values in this type of model with random rater pairings are typi cally expected to be lower than models where rater pairings are fix ed 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* = 534) as papers (*N* = 267). The means for each dimension (see Table 5) show that scores were highest for *Explanation of Issues,* (*M* = 2.7). Importantly, the rest of the average scores attained the standard targeted threshold recommended by the AAC&U, a score of 2 or above. 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	534	30	(6%)	181	(34%)	245	(46%)	78	(15%)
Evidence	534	48	(9%)	206	(39%)	239	(45%)	41	(8%)
Influence of context	534	73	(14%)	206	(39%)	216	(40%)	39	(7%)
Students position	534	51	(10%)	188	(35%)	247	(46%)	48	(9%)
Conclusions	534	65	(12%)	239	(45%)	173	(32%)	57	(11%)

Note: Each paper was rated twice, therefore the number of ratings contained in this table is double the number of papers (N=267).

Table 5
Means for Critical Thinking Measure Scores

Measurement Dimensions	N	Mean	SD	Percent> μ-1a
Explanation of issues	267	2.69	0.71	94.4
Evidence	267	2.51	0.67	91.0
Influence of context	267	2.41	0.72	86.3
Students position	267	2.55	0.70	90.4
Conclusions	267	2.42	0.75	87.8

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 tally 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-1a). For this sample, students exceeded that goal (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.

Summary

The current assessment of signature assignments used an adapted AAC&U Critical Thinking VALUE rubric. Results revealed *Critical Thinking* strengths in a sample of undergraduate students. 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 also exceeded the threshold of 84% for all dimensions of the *Critical Thinking* Core Curriculum Objective. More students exceeded the threshold for *Explanation of the issues, Influence of Context,* and *Conclusions* than the other two dimensions. Results 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. Student demographics were available for most of the sample (94%) but not for the remainder. That said, this evidence may be 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 four of the eight Foundational Component Areas (Communication; History; Language Philosophy and Culture; Life and Physical Sciences). Authentic 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.