UNIVERSITY OF
TEXAS
ARLINGTON

INSTITUTIONAL EFFECTIVENESS AND REPORTING

Measuring Critical Thinking

JUNE 2019 REPORT

The University of Texas at Arlington

Measuring Critical Thinking, June 2019 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* that occurred in June 2019 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 the Social and Behavioral Sciences FCA.

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 (63.3%; n = 143), the rest were male (36.7%, n = 83). Students primarily represented four ethnic groups: White (28.3%, n = 64), Hispanic (36.7%, n = 83), Black/African American (13.3%, n = 30) and Asian (11.9%, n = 27). Many (45.1%) perceived themselves to be first-generation college students and almost half (48.7%) were Pell Grant eligible (see Table 1).

Table 1
Student Demographics

Categorical Information	N	%
Gender		
Female	143	63.3%
Male	83	36.7%
Racial/Ethnic Description		
American Indian or Alaskan Native	0	0.0%
Asian	27	11.9%
Black, African American	30	13.3%
Foreign, Non-Resident Alien	13	5.8%
Hispanic, All Races	83	36.7%
Two or More Races/Ethnicities	8	3.5%
Unknown, Not Specified	1	0.4%
White, Caucasian	64	28.3%
Level		
Freshman	24	10.6%
Sophomore	63	27.9%
Junior	74	32.7%
Senior	65	28.8%
First Generation Student		
Yes	102	45.1%
No	124	54.9%
Pell Grant Eligible*		
Yes	110	48.7%
No	116	51.3%
Transfer Student		
Yes	70	31.0%
No	156	69.0%

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, 2019) 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 raters had earned masters or doctoral degrees in their respective fields. The multi-disciplinary group of raters represented the College of Liberal Arts (70.6%), College of Education, (11.8%), College of Nursing and Health Innovation (5.9%), and the College of Science (5.9%).

Table 2
Rater Demographics

Categorical Information	**N**	**%**
Gender		
Female	10	58.8%
Male	7	41.2%
Ethnic Description		
Hispanic, All Races,	1	5.9%
White, Caucasian	16	94.1%
Classification		
Faculty	14	82.4%
Graduate Teaching Assistant	2	11.8%
Staff	1	5.9%
Highest Degree Received		
Masters	7	41.2%
Doctoral	10	58.8%

On scoring day, a faculty expert from the department of Curriculum and Instruction led 17 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	1	4	3	2	1	4	3	2	1
Influence of context & assumptions	4	3	2	1	4	3	2	1	4	3	2	1
Student's position	4	3	2	1	4	3	2	1	4	3	2	1
Conclusions & related outcomes	4	3	2	1	4	3	2	1	4	3	2	1

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 (Cronbach's Alpha) indicate more agreement between

rater scores (Fleiss, 1986; Shrout & Fleiss, 1979). For this sample, values indicated a trend of good agreement (see Table 3). These high values give confidence to proceed with analyses involving student attainment.

Table 3

Intraclass Correlation Coefficient for Critical Thinking dimensions

Critical Thinking VALUE Rubric Dimension	*n* = 226
Explanation of issues	0.74
Evidence	0.77
Influence of context & assumptions	0.76
Student's position	0.73
Conclusions & related outcomes	0.74

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.*

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* = 452) as papers (*n* = 226). The means for each dimension (see Table 5) show that one of the five dimensions, *Explanation of Issues,* had an average score of 2.48. 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 Communication Dimension Rating Scores

		Rubric Values (Percent of Student Papers)							
	Total	1		2		3		4	
Measurement dimensions	N	N	%	N	%	N	%	N	%
Explanation of issues	452	60	13.3%	156	34.5%	192	42.5%	44	9.7%
Evidence	452	55	12.2%	199	44.0%	169	37.4%	29	6.4%
Influence of context and assumptions	452	80	17.7%	200	44.3%	151	33.4%	21	4.7%
Student's position	452	90	19.9%	204	45.1%	133	29.4%	25	5.5%
Conclusions and related outcomes	452	118	26.1%	209	46.2%	109	24.1%	16	3.5%

Table 5

Means for Communication Measure Scores

Measurement Dimensions	N	Mean	SD	Percent> μ-1σ
Explanation of issues	226	2.48	0.75	83.6%
Evidence	226	2.38	0.70	83.1%
Influence of context and assumptions	226	2.25	0.72	76.5%
Student's position	226	2.20	0.73	85.8%
Conclusions and related outcomes	226	2.05	0.72	82.7%

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 > -1 standard deviation from the mean (84% > μ-1σ). For this sample, students exceeded that goal in two of the five dimensions and more than eighty percent of the students scored > -1 standard deviation of the mean (see Table 5) across all measures. However, it should be noted that all scores were close to the 84% mark (with the exception of *Influence of*

context and assumptions). 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 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.. 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 one of the eight Foundational Component Areas (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 June 2019. 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 (2019). *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. (2018). *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.