UNIVERSITY OF TEXAS ARLINGTON

INSTITUTIONAL EFFECTIVENESS AND REPORTING

EMPIRICAL & QUANTITATIVE SKILLS ASSESSMENT USING AAC&U VALUE RUBRICS AT THE UNIVERSITY OF TEXAS AT ARLINGTON

Summer 2023 Report

Empirical & Quantitative Core Objective Assessment at UT Arlington

Empirical and Quantitative Skill (EQS) is one of six core objectives identified by the Texas Higher Education Coordinating Board (THECB) when the current core curriculum was established in 2011 (THECB, 2019). The assessment of the EQS Objective is required in three of the eight Foundational Component Areas listed by THECB; thus, EQS is implemented within core curriculum coursework at the undergraduate level in Life and Physical Sciences, Mathematics, and Social and Behavioral Sciences. The University of Texas at Arlington (UT Arlington) assesses the six THECB core objectives on a multi-year cycle to examine the extent of student achievement.

At UT Arlington, the EQS Objective is assessed using written or computational samples of undergraduate student work from approved Signature Assignments embedded in the existing core courses. The quality of EQS in student work was rated by UT Arlington faculty and staff using a rubric developed by the Association of American Colleges and Universities (AAC&U, 2009). This report aims to present EQS ratings and information gleaned from undergraduate student work samples collected during the fall 2022 semester at UT Arlington for use by faculty to make recommendations as needed regarding student learning within the core curriculum.

Method

Participants

Written student work samples were obtained from undergraduates enrolled in the fall 2022 semester. Data were collected from two high-enrollment core curriculum courses in mathematics that are offered every fall and spring semesters: MATH 1315 – College Algebra for Economics & Business Analysis and MATH 1402 – College Algebra. These were high-enrollment courses, with 250-300 students with different majors enrolled in these courses every semester. Both algebra courses introduce algebraic and numerical functions and focus on the real-world applications of those functions, making them suitable sample courses for evaluating EQS.

The following demographic information describes 210 students for which data were available. Over half of the students were male (56.19%; n = 118), and the remainder were female (43.81%; n = 92). Regarding race and ethnicity, the sample also reflected the rich diversity of students at UT Arlington. Less than half of the student participants identified as Hispanic/Latino (46.67%; n = 98), while less than one-fifth identified as White (16.19%; n = 34) and a similar number of students who identified as Asian (18.57%; n = 39) were included in this sample. Black/African American, American Indian/Alaskan Native, Foreign, and mixed-race individuals comprised the remainder. Most of these students were either freshmen or sophomores (93.33%; n = 196). Slightly under half (47.62%; n = 100) identified themselves as first-generation students, and slightly over half (51.90%; n = 109) were Pell Grant recipients (see Table 1).

Table 1. Student Demographics

Categorical Information	N	%
Gender		
Female	92	43.81%
Male	118	56.19%
Racial/Ethnic Description		
Asian	39	18.58%
Black, African American	19	9.05%
Foreign, Non-Resident Alien	6	2.87%
Hispanic, All Races	98	46.64%
Two or More Races/Ethnicities	12	5.72%
Unknown, Not Specified	2	0.95%
White, Caucasian	34	16.19%
Level		
Freshman	150	71.43%
Sophomore	46	21.90%
Junior	13	6.19%
Senior	1	0.48%
First Generation Student		
Yes	100	47.62%
No	110	52.38%
Pell Grant Recipient		
Yes	109	51.90%
No	101	48.10%
UTA Enrollment Year		
2022 – 2023	187	89.05%
2021 – 2022	20	9.52%
2020 – 2021 and prior	2	1.43%
Transfer Student		
Yes	10	4.76%
No	200	95.24%

For most students who provided work samples (89.05%; n = 187), the initial enrollment year at UTA was the 2022-2023 academic year. More than 75% of the participating students represented the College of Business (43%; n = 91) and the College of Engineering (35.71%; n = 75). Student participants from each college are listed in Table 2 below.

Table 2: Students by Colleges/Schools

College/School	Number of Students	Percentage
College of Nursing and Health Innovation	2	0.95%
College of Business	91	43.34%
College of Education	4	1.90%
College of Engineering	75	35.72%
College of Liberal Arts	4	1.90%
College of Science	30	14.29%
Division of Student Success	4	1.90%

Procedure

Faculty currently teaching undergraduate courses in the Mathematics Foundational Component Area agreed to submit course Signature Assignment(s) for this report. The syllabus for each core curriculum class at UT Arlington describes the Signature Assignment(s). Students enrolled in core courses complete the Signature Assignment(s) as they would complete other required coursework and assignments. The samples submitted for this assessment process were ungraded, and the Office of Institutional Effectiveness and Reporting de-identified copies of student work before using them for the core curriculum assessment to mitigate rater bias and to protect the confidentiality of student information.

Assessment Instrument

The Signature Assignments were assessed using the Valid Assessment of Learning in Undergraduate Education (VALUE) Rubric for Quantitative Literacy (AAC&U 2009) developed by the Association of American Colleges and Universities (AAC&U). The rubric categorizes EQS into six dimensions: Interpretation, Representation, Calculation, Application/Analysis, Assumptions, and Communication. The rubric describes each dimension and uses a four-point scoring scale (see Appendix A). The rubric functions as a matrix that provides narrative descriptions of expected work quality and corresponding point values for scoring the six measures. The point values range from 1 to 4, with 1 indicating baseline performance (Benchmark-1), 2 indicating approaching milestone (Milestone-2), 3 indicating achieved milestone (Milestone-3), and 4 indicating the highest mastery (Capstone-4) of Empirical and Quantitative skills. AAC&U, the authors of the rubric, permit zero ratings if the paper does not meet the minimum content or quality standards defined in the rubric. The attainment target (numerical ratings) was set at a score of 2 (Milestone-2). The attainment target was set above the benchmark following recommendations from AAC&U research (Greenhoot & Bernstein, 2012) and standard acceptance criteria in the assessment community.

All raters assigned a score to each of the six dimensions in the rubric for each student work sample. Higher values indicate more evidence of EQS in student work and vice versa. Beginning this year, raters were advised to use zero per AAC&U recommendations if any dimension is absent in student work. Typically, in student samples, the six dimensions are adequately represented in the narrative. It is important to note that the EQS samples with visual components in the form of charts, graphs, and figures enhanced the identification of the Representation and Communication dimensions. Since written and visual communications are required across all eight FCAs, a visual component was expected in students' work in foundational mathematics courses.

Raters, Rater Calibration, and Scoring

Raters scored the student writing samples during a scheduled scoring day, and each paper was reviewed twice (two separate raters) in a group setting. A third "tiebreaker" rating was obtained when ratings diverged by more than one rating interval on a single dimension. In these cases, the two sets of most closely aligned ratings were used in calculating the mean scores. The rater group included thirteen faculty members and professional staff with advanced degrees.

The scoring day began with an orientation and description of the rating process. A qualified UTA facilitator led the raters through reviewing the rubric and discussing the rating dimensions and scale designed to calibrate the rater's understanding and use of the rubric in the rating process. Then, the entire group read and rated one practice anchor paper, which was chosen beforehand by the facilitator. Following the sample paper review, the facilitator led a discussion among all raters using the anchor paper to reach a common understanding of the EQS dimensions and to find exemplar indicators within the paper for the rubric levels of mastery. Following completion of the calibration activity, formal review and rating of the de-identified student papers began. During the formal review and rating of papers, raters read each paper and assigned scores for each dimension on the rubric using the four-point scale (plus the available "zero" rating). If the values of the skill measure scores for a paper from the two raters were identical or within one point difference, then the two scores were considered in agreement and averaged. For example, if Rater A scored the Calculation measure with a value of 2 and Rater B scored the same measure with a value of 3, then the rating was considered in agreement, and scores for that dimension were averaged, resulting in a score value of 2.5. If the scores from the two raters differed by over two points, a third rater was assigned the paper. In such cases, three scores were averaged together to determine the final score. For example, if Rater A scored the Calculation measure with a value of 1 and Rater B scored the same measure with a value of 3, the rating was not in agreement, and a third rater was asked to read and score the paper.

Analysis and Results

Inter-rater Agreement

Once each paper had been rated twice, the IER staff collected the rating sheets, entered the rating scores into a spreadsheet, and analyzed them to determine agreement. Each score was calculated as the average of the two rater scores if the values assigned by the raters differed by one point or less. The percentage of agreement among raters for each dimension remained between 82% and 87% for the six EQS dimensions (see Table 3). Regarding rating differences that exceeded two points, a third rater read and scored the paper, and then the average of the two most similar scores was used as the dimension score. In this report, seventeen (8%) student artifacts were rated by a third rater.

Table 3. Scoring Agreement Percentage Among Raters for Empirical and Quantitative Skills Dimensions

Dimension (EQS VALUE Rubric)	Percentages
Interpretation	84%
Representation	86%
Calculation	84%
Application/Analysis	87%
Assumptions	80%
Communication	82%

Note: If values assigned by the raters differed by the rating interval of one point or less, it was counted as agreement. The agreement percentage was computed by dividing the number of agreements by the total number of ratings.

Apart from the simple percentage agreements, researchers widely measure the reliability of rating agreements between different raters to eliminate chance agreements. All raters who participated in the scoring process had advanced degrees and work experience and attended the same training just before the scoring session. Hence, the probability of chance agreement was very low, but inter-rater agreement was computed to follow best research practices. Inter-rater reliability is the consistency among raters when scoring the same subjects independently. The extent to which different raters agree on their judgments establishes the validity and credibility of measurements or ratings.

The inter-rater agreement was determined to check the consistency level of the rating by calculating the Intraclass Correlation Coefficient (ICC). High ICC values indicate more reliability between rater scores. Commonly accepted guidelines were used to interpret the ICC results. These suggest that the range of 0.40 to 0.74 is considered fair to good inter-rater agreement, with results above 0.74 classified as excellent inter-rater agreement and results lower than 0.40 considered poor inter-rater agreement (Fleiss, 1986; Shrout & Fleiss, 1979). ICC values for the Empirical and Quantitative Reasoning scoring session are presented above in Table 4. The ICC values for all dimensions except Assumptions were above 0.60, with Assumptions IC value at 0.50. Since ICC values for all dimensions indicate good inter-rater agreement, mean scopes for each dimension were determined to be reliable.

Table 4. Intraclass Correlation Coefficient for Empirical and Quantitative Skills Dimensions

Dimension (EQS VALUE Rubric)	Coefficient
Interpretation	0.65
Representation	0.66
Calculation	0.68
Application/Analysis	0.64
Assumptions	0.50
Communication	0.66

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 two-way random effects model. Values in this model type with random rater pairings are typically expected to be lower than those where rater pairings are fixed throughout the rating day.*

Students Performance

The final data set contains rating scores on the six dimensions, and all student papers (n = 210) were rated on the scoring day. Across the six dimensions, students scored highest (mean = 2.10) in the *Representation* category, *and* the *Assumptions* category had the lowest scores (mean = 1.26). A rating score of two indicates that dimension milestones were met, and a rating above 1 means the skill is developing at the benchmark level. The means and standard deviations of the analyzed data for each dimension are presented in Table 5.

Table 5. Means for Empirical and Quantitative Skills Measure Scores

Measurement Dimensions	N	Mean	SD
Interpretation	210	2.09	0.94
Representation	210	2.10	0.96
Calculation	210	2.07	1.03
Application/Analysis	210	1.78	0.95
Assumptions	210	1.26	0.96
Communication	210	1.92	0.97

A pattern of strengths and weaknesses in the written work produced by this sample of undergraduates emerged from assessing the student work samples. According to the rating scores, student work exhibited strength in three areas: Interpretation, Representation, and Calculation. However, the student work was rated lower in the Application/Analysis.

Observations and Limitations

During the analysis, some key observations were prominent and worth mentioning here. The majority of UTA student papers included in the rating activity (93.33%) were written by freshmen (71.43%) and sophomore (21.90%) students. Therefore, an average score of 2 indicates that these students are above the benchmark and developing through the milestone level, which is notable given that most participants were first-year students. Additionally, most participants (89.05%) were enrolled at UTA in Fall 2022-2023. These students were new in the university settings, which may have impacted their performance in core courses.

Certain limitations in our data analysis were beyond our control. Firstly, student work samples were only available from one discipline. To enhance representation, future samples should include Life and Physical Sciences courses and Social and Behavioral Sciences. Secondly, aligning the Signature Assignments with the VALUE rubrics is crucial for accurately rating student work. While the rater calibration activity and subsequent discussions of the anchor paper improved interrater reliability, achieving alignment between the rubric and the Signature Assignments was occasionally challenging.

Overall, this assessment of the EQS Core Objective built on previous studies that reported on the use of Signature Assignments as measures of student mastery at UT Arlington. The multi-year plan of assessing the six THECB Core Objectives continues through 2023. Thus far, evidence suggests adequate mastery at the level of students assessed in four out of six EQS dimensions at UT Arlington.

References

Andersen, S., Leon, G., Patel, D. et al. The Impact of COVID-19 on Academic Performance and Personal Experience Among First-Year Medical Students. Med.Sci.Educ. 32, 389–397 (2022). https://doi.org/10.1007/s40670-022-01537-6

Association of American Colleges and Universities. (2009). Quantitative Literacy VALUE rubric. https://www.aacu.org/initiatives/value-initiative/value-rubrics/value-rubrics-quantitative-literacy

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

Georgesen, J. (2015). Evolving from big data to smart data: New ways CX researchers predict customer behavior. Retrieved on 12/08/15 from http://mrweek.com/content

Landers, R.N. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower, 2*:e143518.81744. DOI: 10.15200/winn.143518.81744.

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.

Texas Higher Education Coordinating Board (2019). Texas Core Curriculum. Retrieved on 08/22/2019 from http://www.thecb.state.tx.us/apps/tcc/

Appendix A: Quantitative Literacy VALUE Rubric

QUANTITATIVE LITERACY VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

Definition

Quantitative Literacy (QL) – also known as Numeracy or Quantitative Reasoning (QR) – is a "habit of mind," competency, and comfort in working with numerical data. Individuals with strong QL skills possess the ability to reason and solve quantitative problems from a wide array of authentic contexts and everyday life situations. They understand and can create sophisticated arguments supported by quantitative evidence and they can clearly communicate those arguments in a variety of formats (using words, tables, graphs, mathematical equations, etc., as appropriate).

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 4	**Milestones** 3	2	**Benchmark** 1
Interpretation *Ability to explain information presented in mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Provides accurate explanations of information presented in mathematical forms. Makes appropriate inferences based on that information. *For example, accurately explains the trend data shown in a graph and makes reasonable predictions regarding what the data suggest about future events.*	Provides accurate explanations of information presented in mathematical forms. *For instance, accurately explains the trend data shown in a graph.*	Provides somewhat accurate explanations of information presented in mathematical forms, but occasionally makes minor errors related to computations or units. *For instance, accurately explains trend data shown in a graph, but may miscalculate the slope of the trend line.*	Attempts to explain information presented in mathematical forms, but draws incorrect conclusions about what the information means. *For example, attempts to explain the trend data shown in a graph, but will frequently misinterpret the nature of that trend, perhaps by confusing positive and negative trends.*
Representation *Ability to convert relevant information into various mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Skillfully converts relevant information into an insightful mathematical portrayal in a way that contributes to a further or deeper understanding.	Competently converts relevant information into an appropriate and desired mathematical portrayal.	Completes conversion of information but resulting mathematical portrayal is only partially appropriate or accurate.	Completes conversion of information but resulting mathematical portrayal is inappropriate or inaccurate.
Calculation	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem. Calculations are also presented elegantly (clearly, concisely, etc.)	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem.	Calculations attempted are either unsuccessful or represent only a portion of the calculations required to comprehensively solve the problem.	Calculations are attempted but are both unsuccessful and are not comprehensive.
Application / Analysis *Ability to make judgments and draw appropriate conclusions based on the quantitative analysis of data, while recognizing the limits of this analysis*	Uses the quantitative analysis of data as the basis for deep and thoughtful judgments, drawing insightful, carefully qualified conclusions from this work.	Uses the quantitative analysis of data as the basis for competent judgments, drawing reasonable and appropriately qualified conclusions from this work.	Uses the quantitative analysis of data as the basis for workmanlike (without inspiration or nuance, ordinary) judgments, drawing plausible conclusions from this work.	Uses the quantitative analysis of data as the basis for tentative, basic judgments, although is hesitant or uncertain about drawing conclusions from this work.
Assumptions *Ability to make and evaluate important assumptions in estimation, modeling, and data analysis*	Explicitly describes assumptions and provides compelling rationale for why each assumption is appropriate. Shows awareness that confidence in final conclusions is limited by the accuracy of the assumptions.	Explicitly describes assumptions and provides compelling rationale for why assumptions are appropriate.	Explicitly describes assumptions.	Attempts to describe assumptions.
Communication *Expressing quantitative evidence in support of the argument or purpose of the work (in terms of what evidence is used and how it is formatted, presented, and contextualized)*	Uses quantitative information in connection with the argument or purpose of the work, presents it in an effective format, and explicates it with consistently high quality.	Uses quantitative information in connection with the argument or purpose of the work, though data may be presented in a less than completely effective format or some parts of the explication may be uneven.	Uses quantitative information, but does not effectively connect it to the argument or purpose of the work.	Presents an argument for which quantitative evidence is pertinent, but does not provide adequate explicit numerical support. (May use quasi-quantitative words such as "many," "few," "increasing," "small," and the like in place of actual quantities.)