Quantitative Literacy Report: 2021/22

Closing the loop on Quantitative Literacy: Analysis

Contents

Section 1: Previous Results and need for Closing the Loop

For the 2020-2021 school year, Quantitative Literacy (QL) was in a data collection and analysis phase for the Assessment Cycle at Oregon Tech. The Quantitative Literacy Committee collected data in three ways: a survey of students on their understanding of financial literacy focused on student loans, a survey of faculty departments on their perceptions of student quantitative literacy for their field, and an analysis of QL within the Math 361 (Statistical Methods 1) courses at Oregon Tech. While the results for the first two were informative and sufficient for analysis, the direct course assessment was poorly implemented in 2020/21 and more significant analysis would be helpful.

Section 2: Question Analysis

Two sections of Math 361 were selected in the fall term (both taught by Prof. Joseph Reid). These courses had similar questions (different numbers but similar context) on each of three exams which tested concepts of sampling designs, exploration of data, probability and distributions, null hypothesis testing, and regression methods. A total of 89 questions from these three exams were recorded and 88 of them were tagged with one of the five categories of quantitative literacy within the rubric for the institution.

Calculate: Perform mathematical calculations correctly and evaluate/confirm that they have done so.

Questions associated with calculation appeared on all three artifacts. The first exam calculation questions involved summarizing data in standard statistics. For the second exam, these questions involved probability and expected value calculations. In the third exam, these calculations were of test statistics, expected values in regression, p-values, and degrees of freedom. 33 of the 88 assessed questions were associated with calculation.

Interpret: Extract and interpret quantitative information presented in various commonly used forms.

14 questions were tagged with the criteria of "Interpret." These questions typically involved reading histograms, boxplots, scatterplots, tables, and pulling quantitative information out of the description of a problem. This quantitative information may also include identifying the types of variables, role of variables within the context of the problem, and finding patterns and deviations from those patterns.

Construct Representations: Convert relevant quantitative information and data into different forms as appropriate.

For the criteria of "Construct Representations," a total of 6 questions were assessed. These questions involved such tasks as constructing a box plot associated with statistics for the data,

constructing a probability tree to calculate conditional and reverse-conditional probabilities, and constructing confidence intervals associated with statistical measures of the data.

Apply in Context: Apply appropriate quantitative methods, draw justified conclusions, evaluate claims, and make decisions based on quantitative information. Make and evaluate key assumptions in estimation, modeling, and data analysis.

The "Apply in Context" criteria constituted 24 questions in the analysis. These questions encompassed identifying what statistical tests to use, forming hypotheses, identifying sources of bias and confounding, assessing underlying assumptions, identifying conclusions based on p-values and effect size, interpreting what role random chance plays within the context of experiments and their conclusions, etc.

Communicate: In writing and (where appropriate) in speaking, effectively communicate accurate quantitative information in support of conclusions. In doing so, use representations of quantitative evidence appropriate to both audiences and purpose.

Communication of results involved 9 assessment items. These were all written responses in regard to the context and conclusions of a given scenario. Typically, these came in three forms: an assessment for a given sampling strategy in terms of potential sources of bias and strength of the design; secondly, as conclusions to a null hypothesis testing framework; and finally as an interpretation of the statistical output for a regression problem.

Section 3: Methodology

A total of 47 students within two sections of these courses were analyzed. Students who only participated in the first exam and did not complete the course were not included in the results due to a lack of data. Furthermore, missing data for exam 1, exam 2, or the final exam was present within some student data resulting in partial information. For one student, this caused insufficient information in the "Apply in Context", and "Communicate" sections.

All 88 scores as well as the final course grade were recorded for each of the students in the data set (when available). Students were then assigned a random number (based on the Gaussian normal distribution), sorted by this random number and assigned a student number from 1 to 47. Names were then removed from the data prior to analysis. Finally, as not all questions were out of the same number of points, scores were standardized based on the number of earned points divided by the number of total points such that, for any given item, a perfect score of 1 and minimal score of 0 is available. Question information is assumed to be reasonably additive and univocal, and thus the distribution of averages associated with the items is assumed to be representative. Each participant is thus scored in each of the criteria by the average score (from 0 to 1) associated with the collection of items constituting that criterion.

For example, in the area of Apply in Context, there are 24 associated questions. Each question is scaled between 0 and 1 for a score and then the average of these 24 scores is taken to represent the

student's score in this area. In context, we will assume that a score of 0.7 demonstrates minimal acceptable proficiency within this area and as a goal, we wish for >70% of students to achieve proficiency in each area. As an indirect measure of student effectiveness, we will use final grade (composed of exam scores, in class exercises, and homework's) as an indication of proficiency (or not) within QL. Here, an acceptable grade of "C" or higher is considered proficient.

Section 4: Summary of Results and Conclusions (TLDR)

	Apply in Context	Calculate	Communicate	Construct	Interpret	Final Grade
Total Proficient	**33**	**36**	**33**	**39**	**38**	**45**
Total Proficient %	**70.21%**	**75.00%**	**70.21%**	**81.25%**	**80.85%**	**93.75%**

- Proficient is determined by an average score of >= 0.7 in each category
- Proficient in "Final Grade" is associated with a grade of "C" or better

Descriptive Statistics for each Criteria

	Apply in Context	Calculate	Communicate	Construct	Interpret
N	46	47	46	47	46
Missing	1	0	1	0	1
Mean	0.763	0.804	0.754	0.827	0.821
Median	0.764	0.832	0.778	0.819	0.817
Standard deviation	0.106	0.165	0.159	0.133	0.104
Minimum	0.549	0.296	0.417	0.521	0.624
Maximum	0.964	0.994	1.000	1.000	0.986

All criteria reached the proficiency score of 0.7 by more than 70% of the students with average attainment significantly above this desired average. With 93.75% of the students assessed (those who completed the course) having earned a grade of C or better, we have met the basic goals for this round of QL assessment.

The committee will discuss the appropriateness of the 70% achievement goals and sufficiency criteria for proficiency within each criterion. In terms of methodology, the technique of tagging multiple questions under each criterion, calculating scores as percentages, and calculating averages is extendable through multiple instructors and instruction methodologies for comparable results of assessing

institutional learning outcomes; however, it does not account for item information criteria. A discussion of norming in assessment would be a reasonable follow-up discussion.

We note that one weakness in our data is that there was no effort to establish inter-rater reliability or do training to calibrate the evaluation of student responses by individual faculty members; these are areas for future improvement. We believe that there is room to improve the quality of the data by improving the calibration of the faculty doing the evaluation; there might also be room to better leverage the learning management system (LMS) to facilitate the collection, aggregation, and evaluation of student data. The results of this study will be passed to the assessment commission and a determination of the dissemination of results and further steps in closing the loop for the university as a whole will be planned for the following year.

Section 5: In depth Analysis and Results

	Apply in Context	Calculate	Communicate	Construct	Interpret	Final Grade
Total Proficient	**33**	**36**	**33**	**39**	**38**	**45**
Total Proficient %	**70.21%**	**75.00%**	**70.21%**	**81.25%**	**80.85%**	**93.75%**

- Proficient is determined by an average score of >= 0.7 in each category
- Proficient in "Final Grade" is associated with a grade of "C" or better

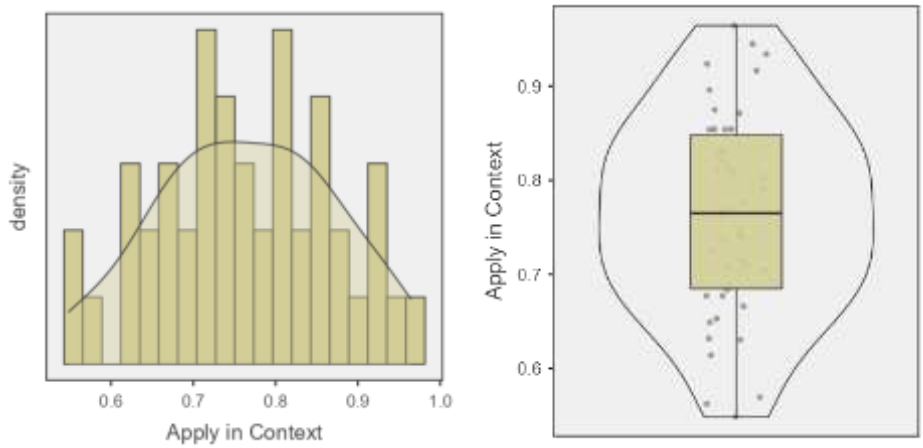
Descriptive Statistics for each Criteria

	Apply in Context	Calculate	Communicate	Construct	Interpret
N	46	47	46	47	46
Missing	1	0	1	0	1
Mean	0.763	0.804	0.754	0.827	0.821
Median	0.764	0.832	0.778	0.819	0.817
Standard deviation	0.106	0.165	0.159	0.133	0.104
Minimum	0.549	0.296	0.417	0.521	0.624
Maximum	0.964	0.994	1.000	1.000	0.986

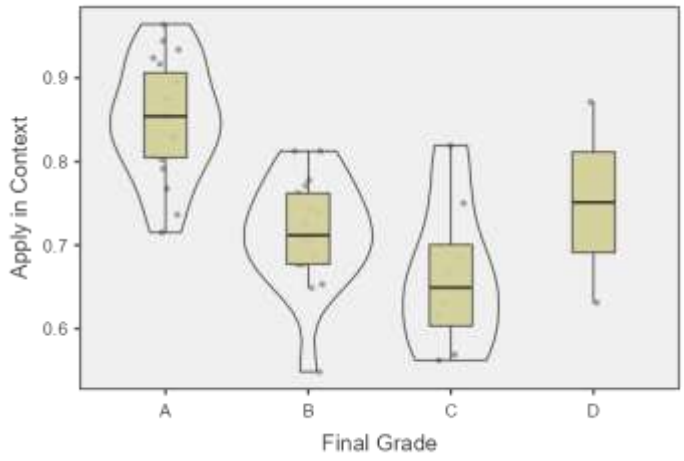
The preceding results are strong in support of the claims that our students were, on average, significantly above the minimal proficiency goal of 0.7. In regard to proportional claims, there is insufficient data to test a that more than 70% of our students achieve this goal (assuming this was a random sample… which it was not); however, within the sample, the goal was met in all 6 areas including the indirect measurement of final grade.

Apply in Context:

The "Apply in Context" criteria had 70.21% of the students who earned a 0.7 or higher average score with the mean score being 0.76328 and standard deviation of 0.10592. The range of scores was between 0.549 and 0.96429. A one sample t-test performed on this average indicates (with p = 0.00001, Cohen's d = 0.59749) that this is moderately higher than the goal of 0.7 on average and is statistically significant at the 95% confidence level. The distribution of these scores appears moderately normal with no significant patterns.



When considering the scores disaggregated by Final Grade, the expected pattern appears to be present (ignoring a grade of D as there are only two data points).



When assessing this with ANOVA, all of the Welch's (unequal variances), Fisher's ANOVA, and Kruskal Wallis suggest that we reject the Omnibus hypothesis of equal means between groups. With further post-hoc analysis, we find differences between those who earned A's and those

who earned B's (Tukey p-value = 0.00002, DSCF p-value = 0.00009.) Similarly, and expectedly, the difference between A's and C's was also significant. Interestingly, the difference between the B and C groups is not significantly different (Tukey p-value = 0.3727, DSCF p-value = 0.39414).

One-Way ANOVA

		F	df1	df2	p
Apply in Context	Welch's	12.41109	3	4.41076	0.01324
	Fisher's	15.33043	3	42	< .00001

Tukey Post-Hoc Test – Apply in Context

		A	B	C	D
A	Mean difference	—	0.13471 ***	0.18768 ***	0.09876
	p-value	—	0.00002	< .00001	0.30957
B	Mean difference		—	0.05297	-0.03595
	p-value		—	0.37274	0.92017
C	Mean difference			—	-0.08892
	p-value			—	0.45548
D	Mean difference				—
	p-value				—

Note. * p < .05, ** p < .01, *** p < .001

Kruskal-Wallis

	χ^2	df	p
Apply in Context	23.76139	3	0.00003

Pairwise comparisons - Apply in Context

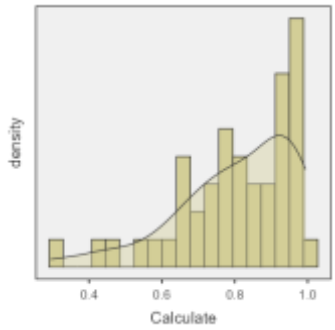
	W	p

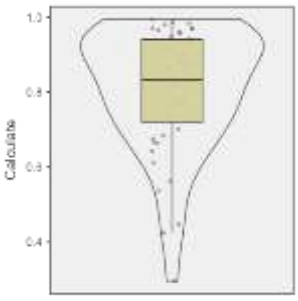
Tukey Post-Hoc Test – Apply in Context

		A	B	C	D
A	B	-6.12209	0.00009		
A	C	-5.11498	0.00170		
A	D	-1.19030	0.83483		
B	C	-2.22526	0.39414		
B	D	0.18803	0.99917		
C	D	1.10782	0.86216		

Calculations:

The "Calculations" criteria had 75% of the students who earned a 0.7 or higher average score with the mean score being 0.80433 and standard deviation of 0.16513. The range of scores was between 0.29605 and 0.99394. The distribution of calculation scores indicates data that is heavily skewed to the left. A Wilcoxon-W test (with p = 0.00007, Cohen's RBC = 0.64) indicates that this is moderately to substantially higher than the goal of 0.7 on average and is statistically significant at the 95% confidence level.



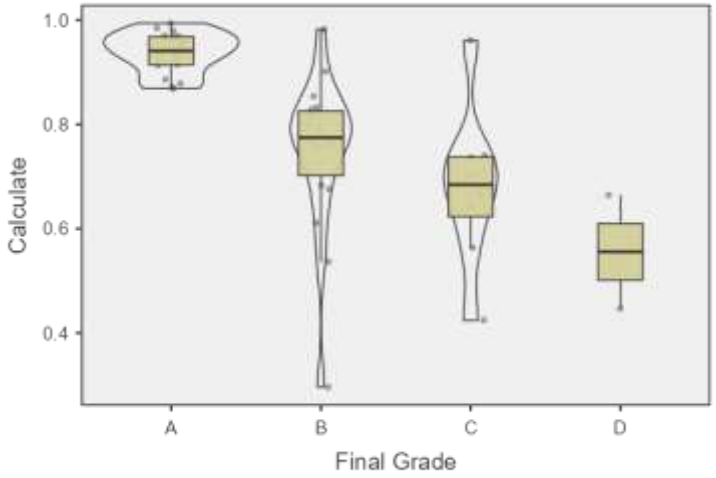


One Sample T-Test

		Statistic	df	p		Effect Size
Calculate	Student's t	4.33137	46.00000	0.00004	Cohen's d	0.63180
	Wilcoxon W	925.00000		0.00007	Rank biserial correlation	0.64007

Note. H_a $\mu > 0.7$

When disaggregated by final grade, there is a wide range of behavior from the groups with almost all of the A students very high on the spectrum of average scores.



When analyzing the data again with ANOVA (Kruskal Wallis due to outliers), the omnibus test is again rejected ($p < 0.00001$) indicating a difference between groups on average with A being different from B and C (p-value = 0.00003, p-value = 0.00378 respectively) but B and C again not being statistically significantly different in medians ($p = 0.31204$).

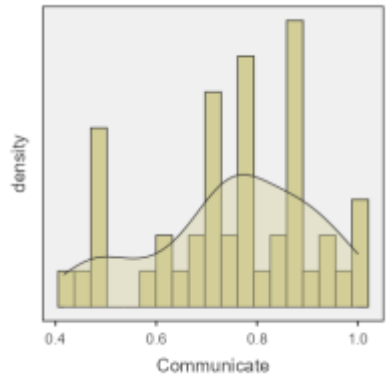
Communicate and Construct outcomes:

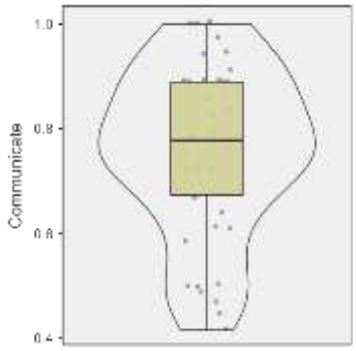
The "Communicate" and "Construct" follow a similar pattern to those present in Calculations. The "Communicate" criteria had 70.21277% of the students who earned a 0.7 or higher average score with the mean score being 0.75403 and standard deviation of 0.15867. The "Construct" Criteria had 81.25% of the students who earned a 0.7 or higher average with the mean score being 0.82639 and standard deviation 0.13316. The distribution of calculation scores indicates data that is heavily skewed to the left for each of these criteria. A Wilcoxon-W test (with $p = 0.01478$, Cohen's RBC = 0.36910) indicates that Communicate is slightly higher than the goal of 0.7 on average and is statistically significant at the 95% confidence level. Similarly, a Wilcoxon-W for Construct (with $p < 0.00001$, RBC = 0.79610) indicates a median that is substantially higher than 0.7 on average.

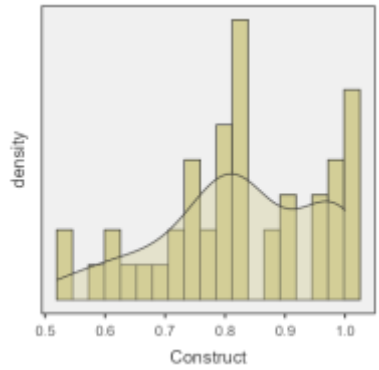
One Sample T-Test

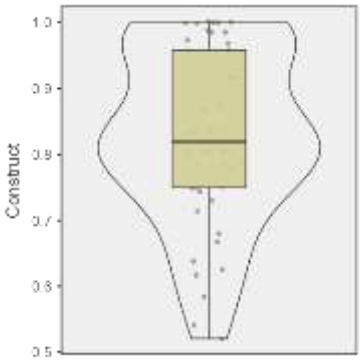
		Statistic	p		Effect Size
Communicate	Wilcoxon W	740.00000	0.01478	Rank biserial correlation	0.36910
Construct	Wilcoxon W	1013.00000	< .00001	Rank biserial correlation	0.79610

Note. H_a $\mu > 0.7$

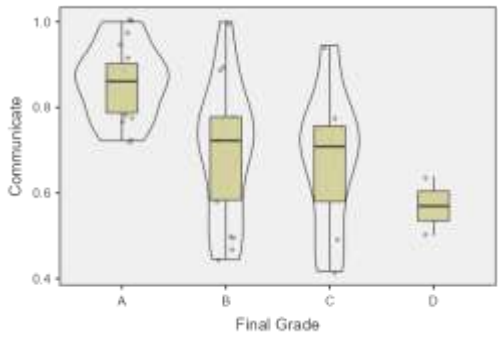


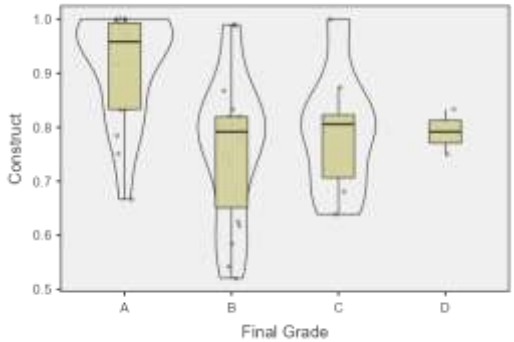






Plots and ANOVA associated with these criteria follow the same pattern from “Calculate.”





Kruskal-Wallis

	χ^2	df	p	ε^2
Communicate	16.33494	3	0.00097	0.36300
Construct	13.53428	3	0.00361	0.29422

Dwass-Steel-Critchlow-Fligner pairwise comparisons

Pairwise comparisons - Communicate

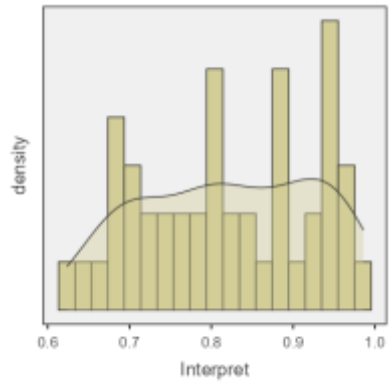
		W	p
A	B	-4.61776	0.00603
A	C	-4.03108	0.02269
A	D	-3.23505	0.10099
B	C	-0.49661	0.98516
B	D	-1.51158	0.70859
C	D	-1.10782	0.86216

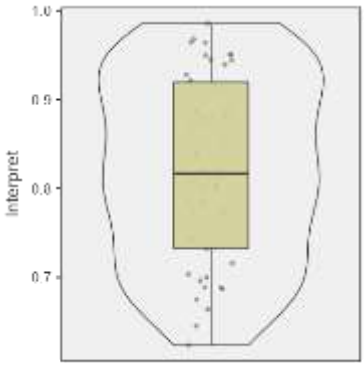
Pairwise comparisons - Construct

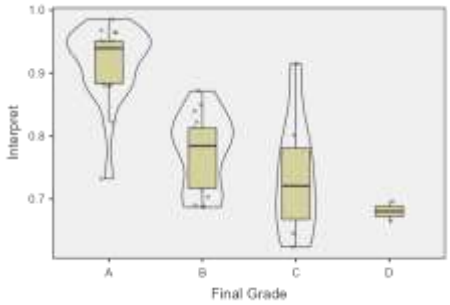
		W	p
A	B	-4.84660	0.00342
A	C	-3.21513	0.10436
A	D	-2.22886	0.39250
B	C	0.55224	0.97987
B	D	0.71458	0.95787
C	D	0.37383	0.99355

Interpret:

The “Interpret” criteria had 80.851% of the students who earned a 0.7 or higher average score with the mean score being 0.82131 and standard deviation of 0.10447. The range of scores was between 0.62381 and 0.98571. A one sample t-test performed on this average indicates (with $p < 0.00001$, Cohen’s $d = 1.16$) that this is substantially higher than the goal of 0.7 on average and is statistically significant at the 95% confidence level. The distribution of these scores appears moderately uniform with slight upward skew.







One-Way ANOVA (Non-parametric)

Kruskal-Wallis

	χ^2	df	p	ε^2
Interpret	28.38512	3	< .00001	0.63078

Dwass-Steel-Critchlow-Fligner pairwise comparisons

Pairwise comparisons - Interpret

		W	p
A	B	-6.54437	0.00002
A	C	-4.95847	0.00257
A	D	-3.22240	0.10317
B	C	-2.18363	0.41113
B	D	-2.63014	0.24577
C	D	-1.10782	0.86216

ANOVA (again, Kruskal Wallis due to outliers and skewness) indicates a difference between group averages with a $p < 0.00001$. This difference is primarily between those with grades of A versus the other groups ($p = 0.00002$ and $p = 0.00257$), but again, no difference is found between those with grades of B and C ($p = 0.41113$)

The differences between B and C in criteria for Apply in Context, Calculate, and Interpret may be due to a lack of statistical power and need a larger sample size to detect these differences. Furthermore, the number of hypotheses tested within the results above is substantial and begs the question of p-hacking. With regards to this, there are 5 ANOVA and 5 one sample tests run above. The largest p-value was more than a factor of 10 less than 0.05 indicating that, even with a conservative Bonferroni correction, few of these (if any) are likely to be the result of a type 1 error.

The results are strong in support of the claims that our students were, on average, significantly above the minimal proficiency goal of 0.7. In regard to proportional claims, there is insufficient data to test a that more than 70% of our students achieve this goal (assuming this was a random sample… which it was not); however, within the sample, the goal was met in all 6 areas including the indirect measurement of final grade.