FERRIS STATE UNIVERSITY

General Education Report*

Quantitative Literacy Outcome QNT3, Fall 2017

Dr. Clifton Franklund
General Education Coordinator

Fall 2017

Contents

Abstract

"Assessment is not a spreadsheet; it's a conversation." — Irmeli Halinen

This report is a proof-of-concept for the proposed General Education assessment strategy at Ferris State University. Course-level student assessment data was gathered using TracDat and de-identified using a custom script. The clean and tidy data set was used to generate this report in both PDF and HTML formats with the bookdown package for the R statistical programming language. A total of 21 semesters of student performance on a lecture exam were used to evaluate student competency on Ferris Learning Outcome (FLO) SCI1. A meta-analysis of these data demonstrated that performance was essentially at the criterion of success. There was substantial variation in enrollment and course performance over the time span examined The utility of reports like these to analyze, distribute, and act up General Education assessment data will be investigated using faculty focus groups in the fall of 2016.

*Report number 1710, DOI 10.17605/OSF.IO/35GSR

Table 1: Conversion of percentages to rubric scores

Percent Correct	Rubric	Interpretation
0.0 to 49.0%	0	Unsatisfactory
50.0 to 59.9%	1	Beginning
60.0 to 69.9%	2	Developing
70.0 to 84.9%	3	Proficient
85.0 to 100.0%	4	Advanced

Introduction

This report is an actual analysis of real course-level assessment data from a 200-level Biology course. However, its primary purpose is to serve as a proof-of-concept for the new General Education assessment process at Ferris State University. Assessment is perhaps best viewed as a scholarly activity that is focused upon programmatic improvement. Such scholarly work should be built upon, and contribute to, the relevant professional literature (Weimer, 2015). To emphasize that reality, this report is formatted in the form of a journal article. This report, and ones like it, will be authored, published, and cited in future work to support the development and improvement of the General Education program at Ferris.

Methods

Collection of assessment data

Student performance on the first lecture exam in a 200-level Biology course was analyzed. The content assessed in all exams was biological diversity. However, the number and format of the questions used varied by semester. Individual student scores were collected using the new General Education Natural Sciences "scores" data workbook for 21 semesters. Student scores were automatically converted to a rubric score by the workbook using the equivalencies shown in Table 1.

These workbook files contain personally identifiable information (PII) and are, therefore, subject to FERPA regulations. For this reason, they are not directly shared. Instead, they are permanently housed within the Proof_of_Concept folder under Core Competency: Natural Sciences in TracDat.

De-identification of student data

Copies of the 21 data files were downloaded from TracDat. An R aggregator script was used to read the data from these data sheets and concatenate it into one data set in a destructive process – the downloaded copies were deleted in the process. Student names and identification numbers were redacted and each student's entry was given a unique eight-digit identifier - the Record.Key. These keys may be used for longitudinal studies in the future. The algorithm used is kept in an encrypted site and shared with no one. The de-identified data set contains 748 student entries and is formatted as a comma-delimited text file (mathData.csv).

Data provenance

Data provenance refers to a system that permits tracking of the origin, movement, modification, and utilization of data sets (Buneman et al., 2001). The provenance of General Education data will be explicitly declared to facilitate the reproducibility and extensibility of these studies.

Location of public website files

All files related to this report can be found online at the Open Science Framework (Nosek, 2012). This site contains all of the files needed to reproduce this report from the de-identified data set. The site's url is https://osf.io/t6u8m/.

Session information

This report was written using RStudio (RStudio Team, 2015) and the R statistical programming language (R Core Team, 2013). These products are free to download for PC, Macintosh, and Linux operating systems. The following information pertains to the session parameters used to generate this report. If you have trouble reproducing this report, it may be due to different session parameters. You may contact Dr. Franklund if you need assistance.

R version 3.5.1 (2018-07-02)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

locale: en_US.UTF-8||en_US.UTF-8||en_US.UTF-8||C||en_US.UTF-8||en_US.UTF-8

attached base packages: grid, stats, graphics, grDevices, utils, datasets, methods and base

other attached packages: forestplot(v.1.7.2), checkmate(v.1.8.5), magrittr(v.1.5), weights(v.1.0), mice(v.3.3.0), gdata(v.2.18.0), Hmisc(v.4.1-1), Formula(v.1.2-3), survival(v.2.42-6), lattice(v.0.20-35), moments(v.0.14), modeest(v.2.1), pander(v.0.6.2), forcats(v.0.3.0), stringr(v.1.3.1), dplyr(v.0.7.6), purrr(v.0.2.5), readr(v.1.1.1), tidyr(v.0.8.1), tibble(v.1.4.2), ggplot2(v.3.0.0) and tidyverse(v.1.2.1)

loaded via a namespace (and not attached): httr(v.1.3.1), jsonlite(v.1.5), splines(v.3.5.1), modelr(v.0.1.2), gtools(v.3.8.1), assertthat(v.0.2.0), latticeExtra(v.0.6-28), cellranger(v.1.1.0), yaml(v.2.2.0), pillar(v.1.3.0), backports(v.1.1.2), glue(v.1.3.0), digest(v.0.6.17), RColorBrewer(v.1.1-2), minqa(v.1.2.4), rvest(v.0.3.2), colorspace(v.1.3-2), htmltools(v.0.3.6), Matrix(v.1.2-14), plyr(v.1.8.4), pkgconfig(v.2.0.2), broom(v.0.5.0), haven(v.1.1.2), bookdown(v.0.7), scales(v.1.0.0), lme4(v.1.1-18-1), htmlTable(v.1.12), withr(v.2.1.2), pan(v.1.6), nnet(v.7.3-12), lazyeval(v.0.2.1), cli(v.1.0.0), crayon(v.1.3.4), readxl(v.1.1.0), mitml(v.0.3-6), evaluate(v.0.11), nlme(v.3.1-137), MASS(v.7.3-50), xml2(v.1.2.0), foreign(v.0.8-71), tools(v.3.5.1), data.table(v.1.11.8), hms(v.0.4.2), munsell(v.0.5.0), cluster(v.2.0.7-1), bindrcpp(v.0.2.2), compiler(v.3.5.1), rlang(v.0.2.2), nloptr(v.1.2.1), rstudioapi(v.0.7), htmlwidgets(v.1.3), base64enc(v.0.1-3), rmarkdown(v.1.10), gtable(v.0.2.0), R6(v.2.2.2), gridExtra(v.2.3), lubridate(v.1.7.4), knitr(v.1.20), bindr(v.0.1.1), jomo(v.2.6-4), rprojroot(v.1.3-2), stringi(v.1.2.4), parallel(v.3.5.1), Rcpp(v.0.12.18), rpart(v.4.1-13), acepack(v.1.4.1), tidyselect(v.0.2.4) and xfun(v.0.3)

Processing instructions

This project produced a computationally reproducible assessment report (this document). Anyone wishing to recreate this report from the source document will need to install the following on their computer:

1. An installation of the R programming language
2. An installation of the RStudio IDE
3. An installation of LaTeX

The necessary source files include the de-identified data set (BIOL200Data.csv), Rmarkdown code files (index.Rmd, 01-Introduction.Rmd, 02-Methods.Rmd, 03-Results.Rmd, 04-Discussion.Rmd, and 05-References.Rmd), bibtex reference file (references.bib), and custom art file in the /art folder.

To process the files, you must first open the project in RStudio. Click on the "Build Book" button in the Build menu. Bookdown allows you to build this project as git_book (html site), pdf_book (via LaTeX), or epub_book (compatible with iBooks and other e-book readers).

Citation of this work

Results

This document itself is the primary result of the project. It will be shared with members of the General Education Committee, Academic Senate, and the Department of Mathematics at Ferris State University. Their comments and suggestions will be included in the Discussion.

Summary statistics

A total of 748 student performances on Ferris Learning Outcome 3 were collected over 21 different math courses. Student scores were converted to rubric scores as described above. The overall average rubric score for all students and semesters was 2.69. The mode and median scores were 4 and 3, respectively. The average was not statistically different from the threshold score for competence (2.6) as evaluated with a one-value, two-tailed t-test (t=1.75, df=747, p=0.081). The effect size for the difference between the average and the threshold was tiny (d=0.06). We can infer from this that the overall average rubric score is not practically different than the threshold score. A total of 31 courses were initially registered to submit data for the semester, so we achieved a 67.7% completion rate. According to Banner records, 1,997 students were enrolled in Quantitative Literacy courses in the fall semester. We captured 37.5% of this population with our census.

> Do you think that this level of performance on the outcome is sufficient? Or, do you think that there is a need to increase this level?

The distribution of all rubric scores is shown in Figure 1. This distribution exhibited a moderate negative skew (skew = -0.86). This result may simply indicate that the teaching, materials, and student learning are all functioning well when the scores are viewed in aggregate. A total of 502 students (67.1%) met or exceeded the competence threshold over the semesters investigated.

> Do you think that this distribution of rubric scores seems right? Are there too many ones and fours, or is this distribution expected in these courses?

The distribution of rubric scores by course is shown in Figure 2. There are rather obvious differences in both the distribution of rubric scores and class sizes between semesters.

> What do you think accounts for the variability in scoring from course to course. Is this a problem in measurement, or does the variation reflect real differences in the course populations?

Meta-analysis

Meta-analysis of the student performance was performed using R (Del Re, 2015). This analysis resulted in a weighted average of rubric scores. This value was calculated using formula (1). The value X_i average rubric scores for the semesters, while P_i is the weighting factor (student enrollment).

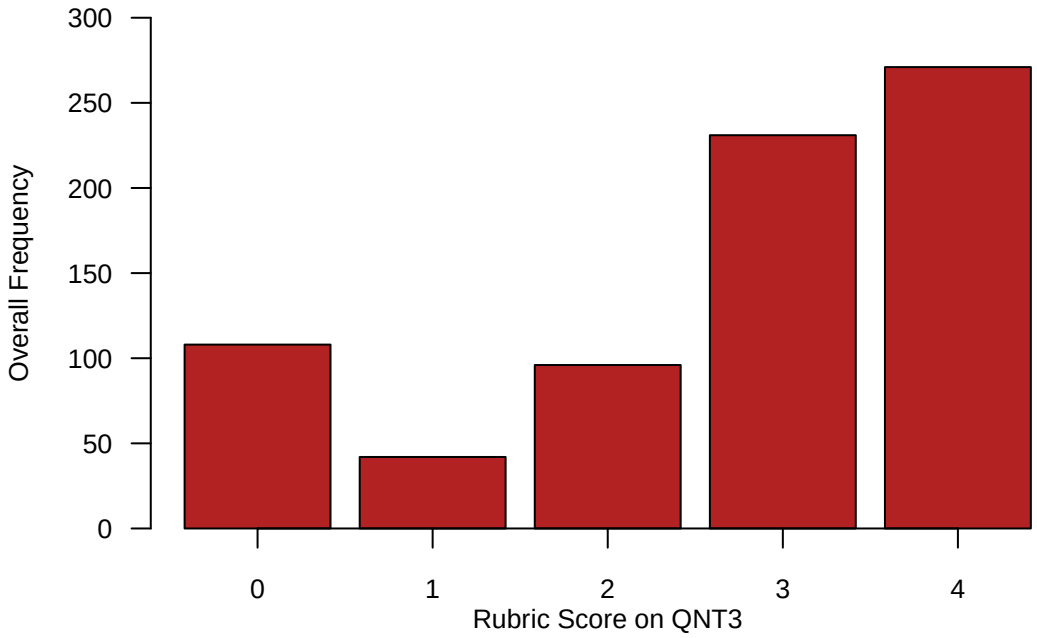

Figure 1: A histogram of the distribution of individual rubric score frequencies over all thirteen semesters.

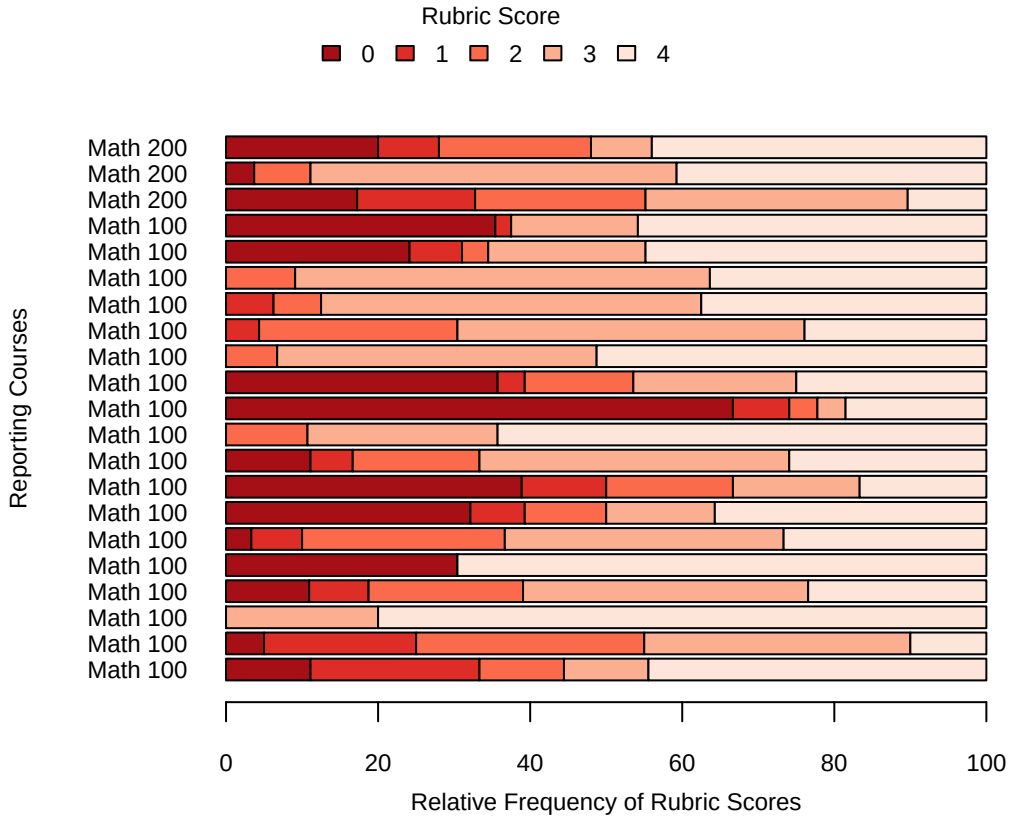

Figure 2: A barplot showing the distribution of rubric scores broken down by semester.

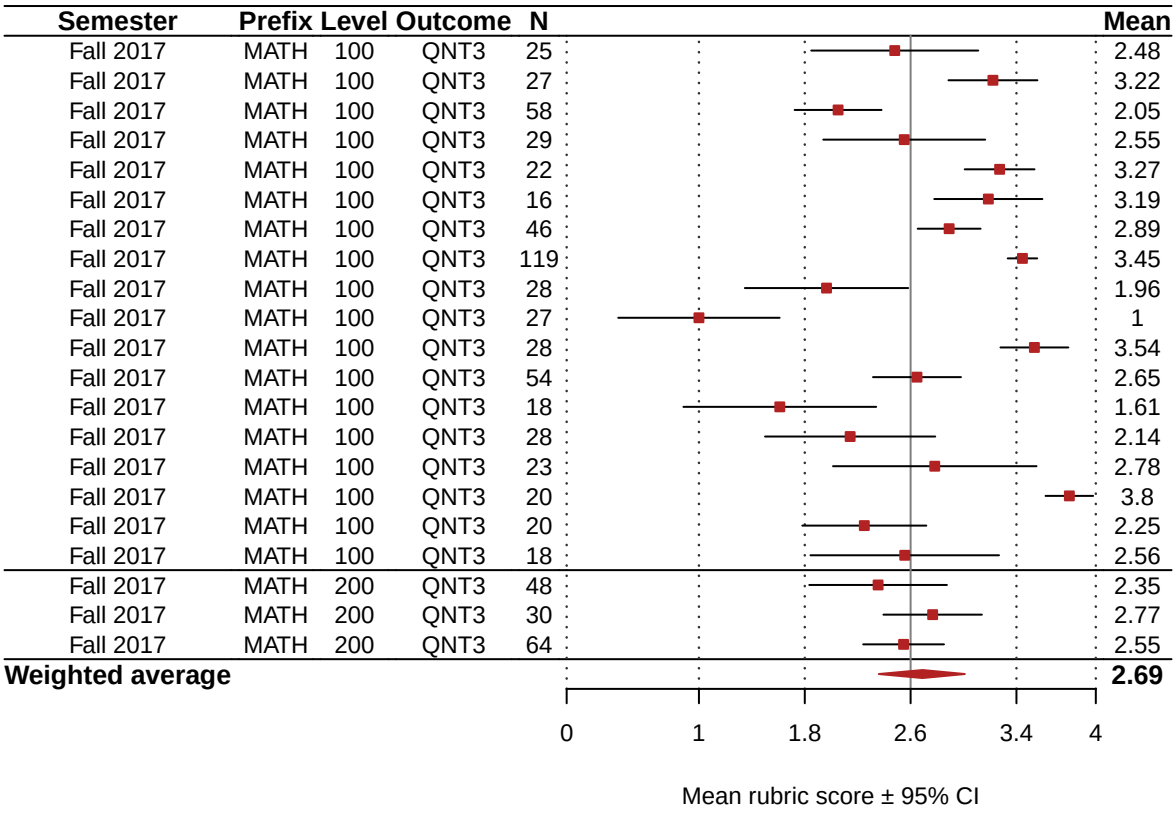

Figure 3: A forest plot of the average scores for each semester with a weighted mean estimate for the entire period investigated. Error bars indicate the 95% confidence intervals.

$$\bar{X}_w = \frac{\sum X_i P_i}{\sum P_i} \tag{1}$$

The confidence interval for the weighted mean was calculated using the weighted variance. However, the weighted variance is actually not simple to calculate. Several different methods have been compared to bootstrapping (Gatz and Smith, 1995). The most accurate method was initially described by Cochran (Cochran, 1977) and that one was used in this study. The calculation to obtain the weighted variance is shown in formula (2).

$$\begin{aligned}(SEM_w)^2 = \frac{n}{(n-1)(\sum P_i)^2}\big[\sum (P_i X_i - \bar{P}\bar{X}_w)^2 \\ -2\bar{X}_w \sum (P_i - \bar{P})(P_i X_i - \bar{P}\bar{X}_w) + \bar{X}_w^2 \sum (P_i - \bar{P})^2\big]\end{aligned} \tag{2}$$

A forest plot of the meta-analysis is shown in Figure 3. In this representation, each semester is illustrated as a separate line. The mean and 95% confidence intervals for each semester are plotted in the right panel and their associated meta-data are given in the table to the left. The weighted average of all the data is plotted at the bottom of the figure. The width of the diamond indicates the 95% confidence interval.

> Do you think that this reflects the actual status of student performance in these courses?

The rubric scale can be conceptually divided into five areas as shown in Table 2. Of the 21 semesters, 10 fell in the proficient range, 9 fell in the developing range, and 2 fell in the beginning range. The weighted mean

Table 2: Interpretation of average rubric scores

Average Score	Interpretation
0.00 to 0.99	Unsatisfactory
1.00 to 1.79	Beginning
1.80 to 2.59	Developing
2.60 to 3.39	Proficient
3.40 to 4.00	Advanced

Table 3: One-way ANOVA analysis of scores by standard measure used

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Standard.Measure	3	10.70376	3.567919	1.864508	0.1341432
Residuals	744	1423.71736	1.913599	NA	NA

score was not significantly different from the threshold of competence as judged by a weighted, one-factor, two-tailed t-test (t=0.63, df=20, p=0.53). We can conclude that the weighted average score is practically equivalent to the competency threshold score.

Performance by Course Level

The average rubric scores for 100-level and 200-level courses was 2.73 and 2.53, respectively . Both of these averages are near the threshold value of 2.6. The difference between these scores was not statistically different as evaluated with a two-tailed t-test (t=1.49, df=205.41, p=0.14). The effect size for the difference between was tiny (d=0.1). We can infer from this that there was no evidence of a measurable difference in student performance by course level.

> The performance in 200-level students is no better than that of 100-level students. Is this a concern, or are these results expected for these courses? Why or why not?

Performance by Standard Measure

A one-way ANOVA was used to compare the rubric scores by student gender (Table 3). There was not a statistically significant difference between the student scores. So, there is currently no evidence to suggest that there is a difference between students of different standard measures on the Quantitative Literacy learning outcome.

> Do you think that the standard measures used in this cycle are the best ones for the outcome being addressed? Are the results surprising to you?

Performance by Student Gender

A one-way ANOVA was used to compare the rubric scores by student gender (Table 4). There was not a statistically significant difference between the student scores. So, there is currently no evidence to suggest that there is a difference between students of different genders on the Quantitative Literacy learning outcome.

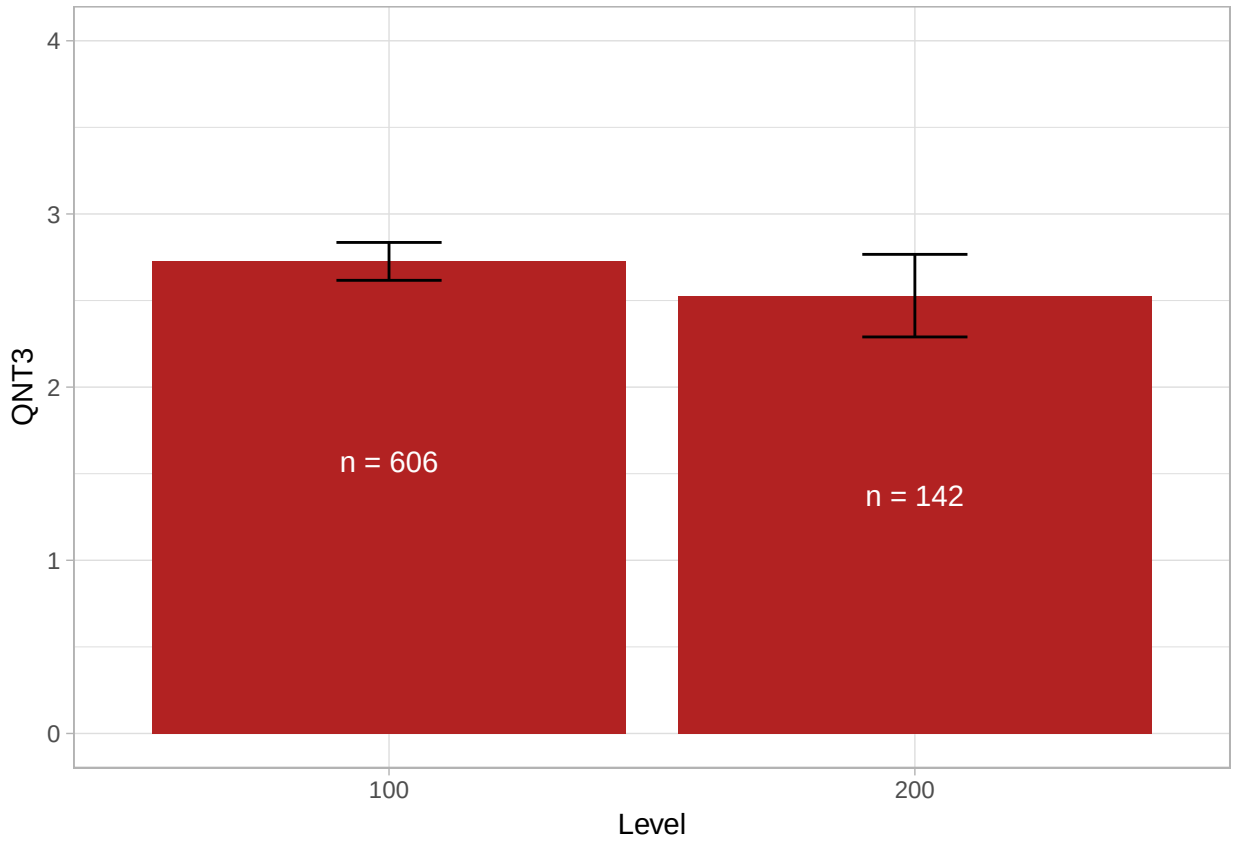

Figure 4: A comparison of student performance on QNT3 by course level. Error bars indicate the 95% confidence intervals.

Table 4: One-way ANOVA analysis of scores by self-identified student gender

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	2	8.681903	4.340952	2.268303	0.1042019
Residuals	745	1425.739220	1.913744	NA	NA

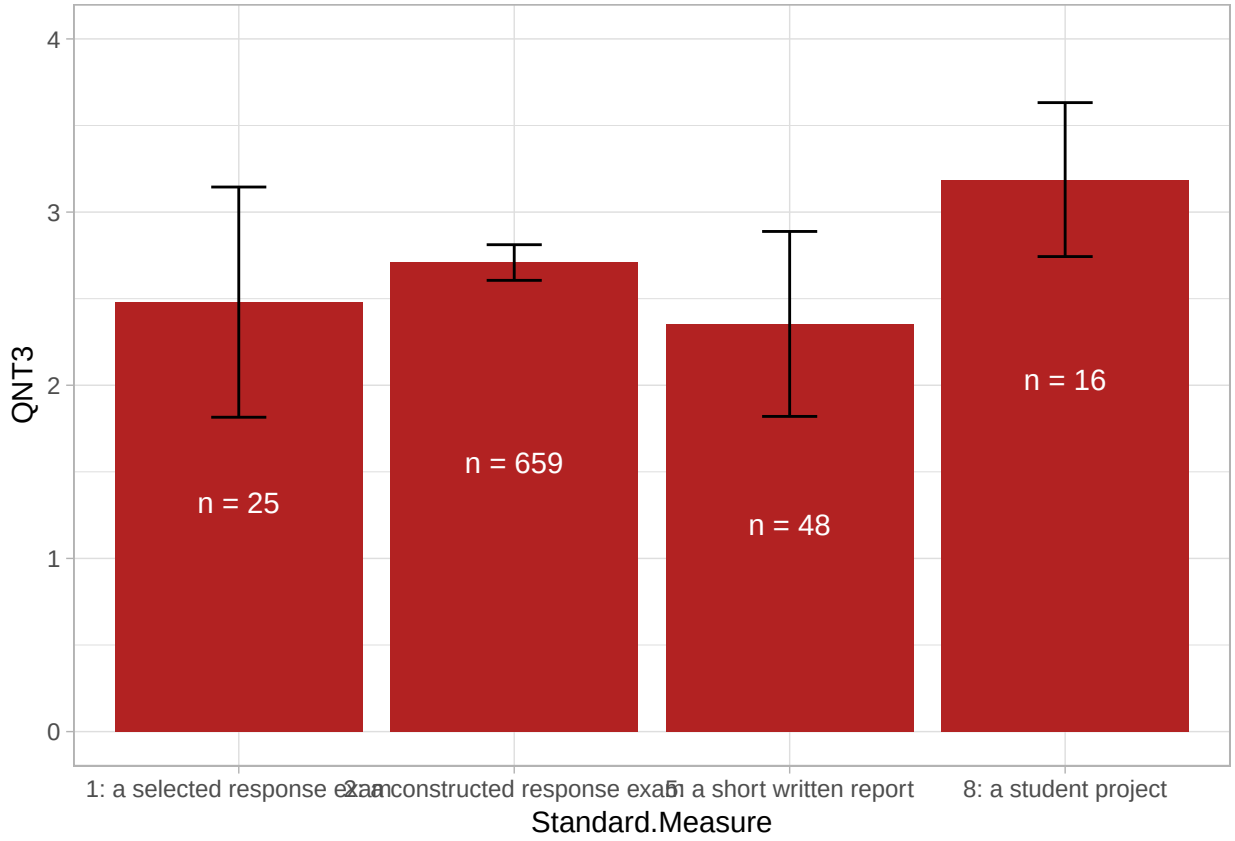

Figure 5: A comparison of student performance on QNT3 by standard meeasure used. Error bars indicate the 95% confidence intervals.

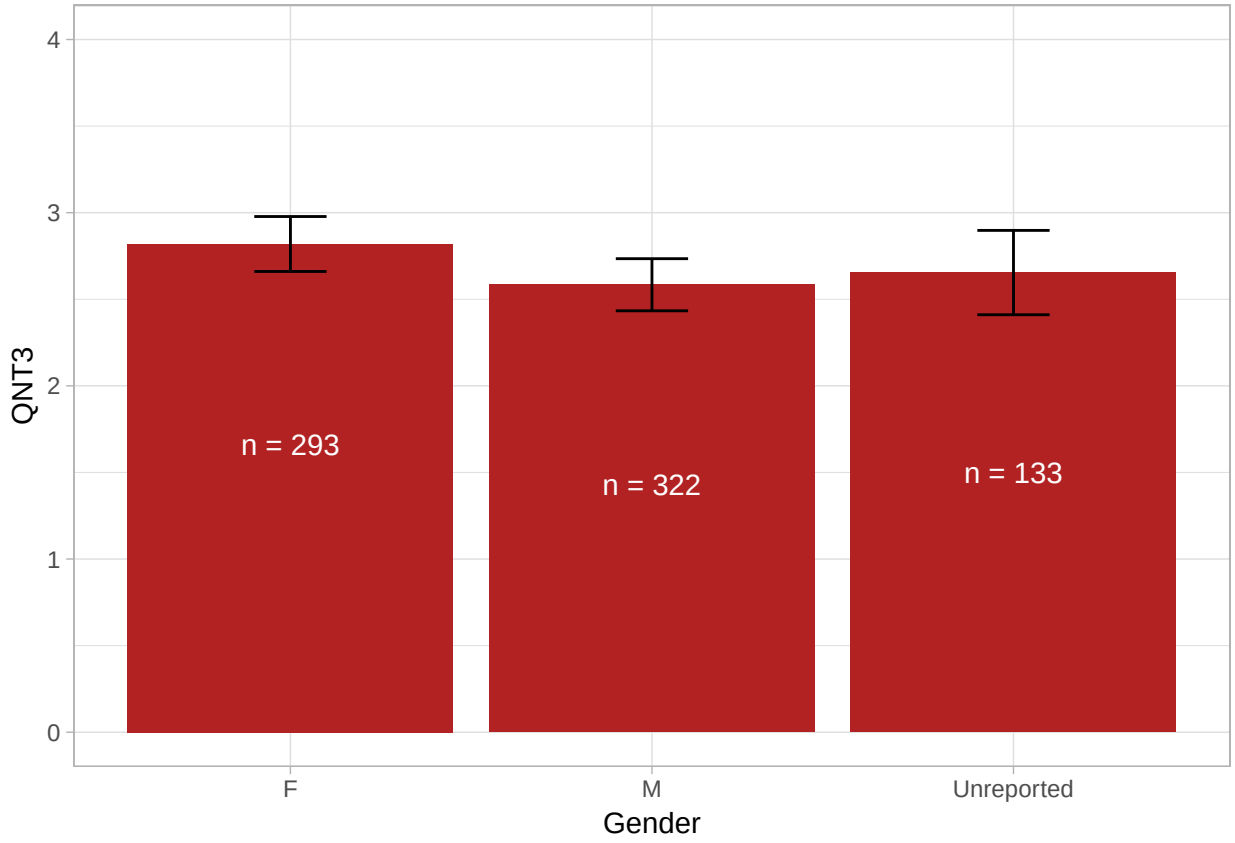

Figure 6: A comparison of student performance on QNT3 by student gender. Error bars indicate the 95% confidence intervals.

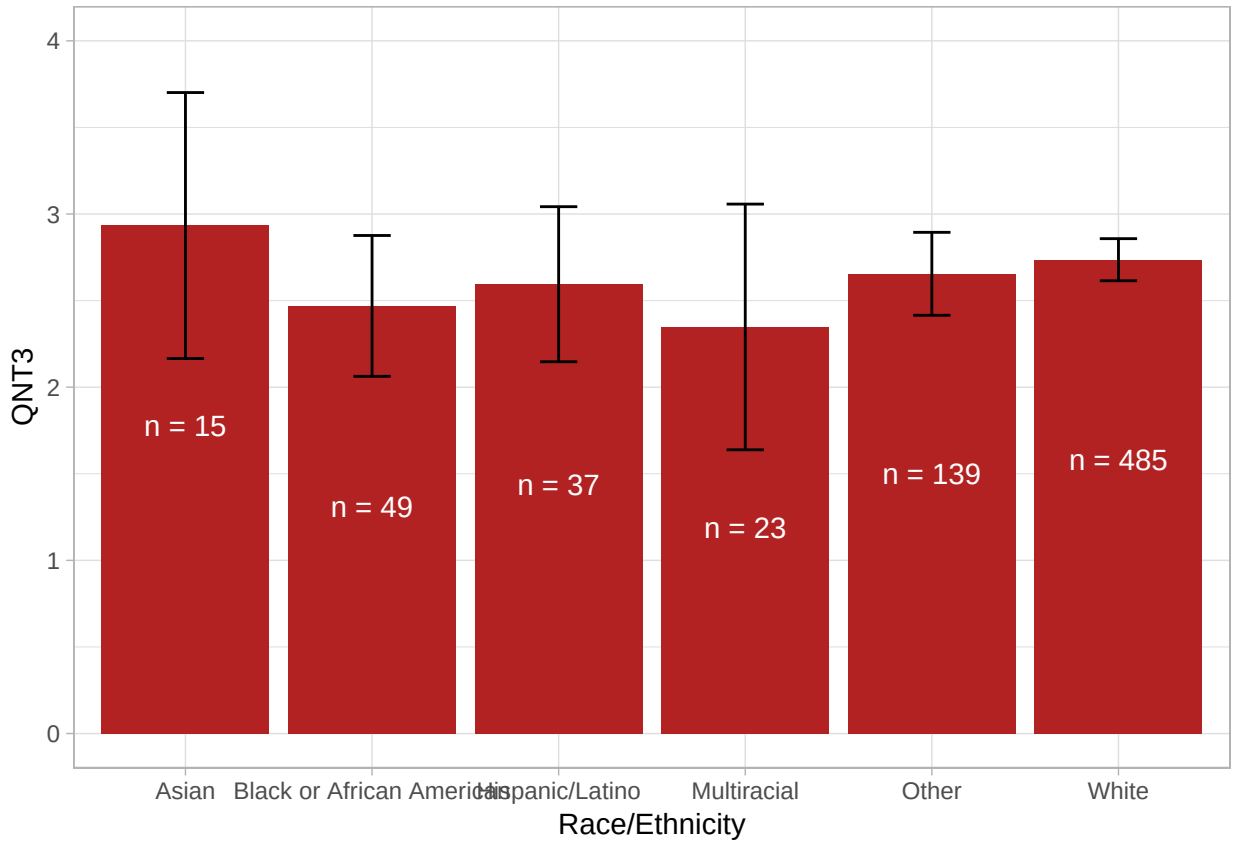

Figure 7: A comparison of student performance on QNT3 by student race. Error bars indicate the 95% confidence intervals.

Table 5: One-way ANOVA analysis of scores by student race

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
'Race/Ethnicity'	5	7.504381	1.500876	0.7804591	0.5639121
Residuals	742	1426.916742	1.923068	NA	NA

Do these results seem right to you? Are we reaching all gender groups as well as we ought? Or, do you think that these scores are an anomaly due sampling bias?

Performance by Student Race

A one-way ANOVA was used to compare the rubric scores by student race (Table 5). There was not a statistically significant difference between the student scores. So, there is currently no evidence to suggest that there is a difference between students of different races on the Quantitative Literacy learning outcome. Note however, that the sample size for most races other than "white" are rather small yet.

Do these results seem right to you? Are we reaching all minority groups as well as we ought? Or, do you think that these scores are an anomaly due to the small sample sizes?

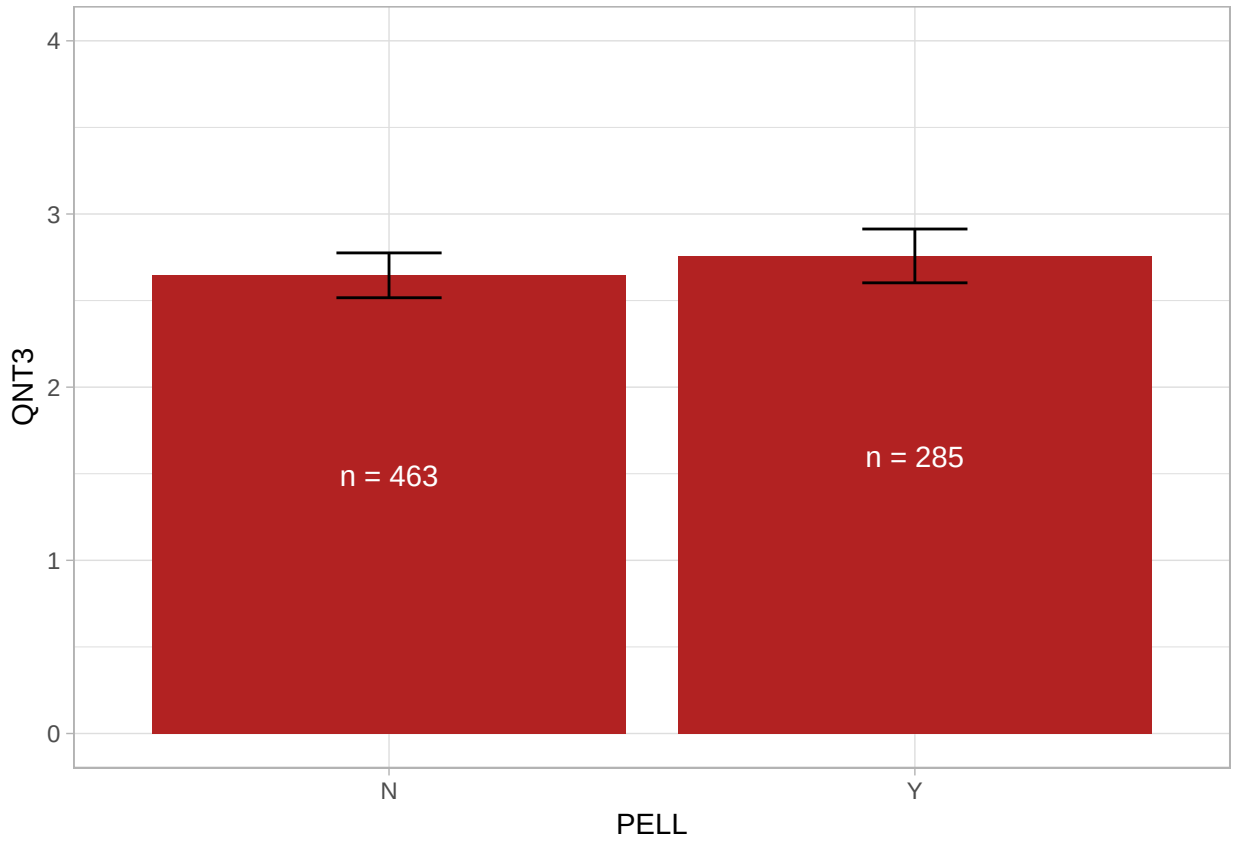

Figure 8: A comparison of student performance on QNT3 by student PELL eligibility. Error bars indicate the 95% confidence intervals.

Performance by PELL Eligibility

The average rubric scores for Pell-eligible and non-Pell-eligible students was 2.76 and 2.65, respectively. Both of these averages are slightly above the threshold value of 2.6. The difference between these scores was not statistically different as evaluated with a two-tailed t-test (t=-1.09, df=629.54, p=0.28). The effect size for the difference between was tiny (d=-0.04). We can infer from this that there was no evidence of a measurable difference in student performance by Pell eligibility.

> The performance of Pell-eligible students is no worse than that of those that are not. Is this a surprise, or are these results expected for these courses? Why or why not?

Discussion

A novel approach for the collection, aggregation, analysis, and reporting of General Education assessment data has been developed. Computationally reproducible reports can easily be generated and distributed to improve the program over time. A meta-analysis of data collected for the first time from a selection of Mathematics courses. Over a total of 21 courses, 47.6% had mean scores considered to be proficient. Of all students in all semesters, 67.1% met or exceeded the competence threshold. From these data it is inferred that the students are about at the threshold of competence. Disaggregation of scores by standard measure employed, course level, student race, student gender, and Pell eligibility all failed to find any significant differences in student performance.

Faculty feedback

This report has been distributed to members of the General Education Committee, Academic Senate, and the Department of Mathematics at Ferris State Univerity. These individuals were asked to provide their comments, suggestions, and concerns about this report and the processes involved in its creation. Their reflections are captured in the Disqus comments within this report. Summary of their ideas and action steps will eventually be included in an Addendum section to this report.

Plan of action

After analyzing the data and considering the comments provided in the faculty feedback, the relevant General Education sub-committee members will make one or more recommendations for future work. Some of the possible actions could include:

- No modifications – continue to gather data
- Convene a training session to get better inter-course reliability
- Suggest modifications to the types of assignments that are used
- Suggest modifications to which data workbooks are used
- Suggest that instructors consider modifying the scope or sequence of instruction
- Modify the learning outcomes themselves
- Modify the competency as a whole

Acknowledgments

This report was built using Rmarkdown and the bookdown R package. The valuable contributions made by the members of the General Education Committee, Academic Senate, and Department of Biological Sciences are also greatly appreciated.

References

Buneman, P., Khanna, S., and Wang-Chiew, T. (2001). Why and Where: A Characterization of Data Provenance, pages 316–330. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cochran, W. G. (1977). Sampling techniques. Wiley, New York, NY, 3rd editio edition.

Del Re, A. C. (2015). A Practical Tutorial on Conducting Meta-Analysis in R. Quant. Methods Psychol., 11(1):37–50.

Gatz, D. F. and Smith, L. (1995). The standard error of a weighted mean concentration—I. Bootstrapping vs other methods. Atmos. Environ., 29(11):1185–1193.

Nosek, B. (2012). An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. Perspect. Psychol. Sci., 7(6):657–660.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team (2015). RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA.

Weimer, M. (2015). Enhancing scholarly work on teaching and learning. John Wiley & Sons, San Fransisco, CA.