FERRIS STATE UNIVERSITY

General Education Report*

Meta-Analysis of Student Achievement on Natural Sciences FLO SCI1 in a 200-Level Biology Course

Dr. Clifton Franklund
General Education Coordinator

Fall 2016

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 13 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.

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

*Report number 201601, DOI 10.17605/OSF.IO/35GSR

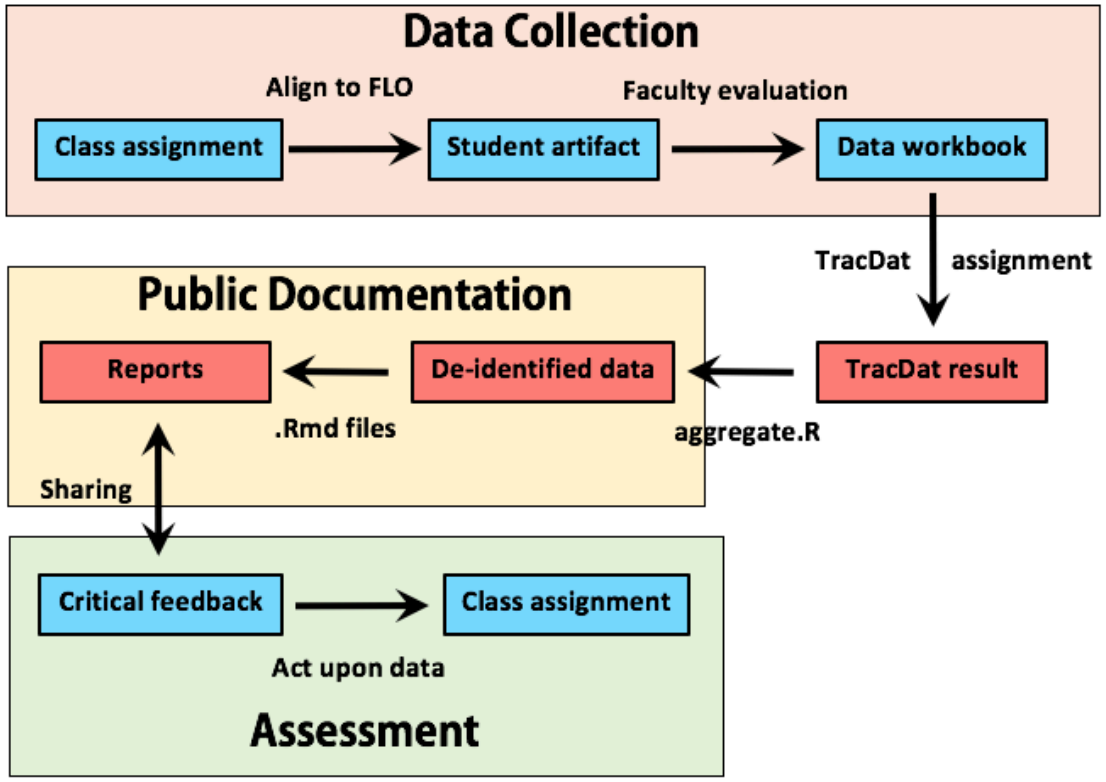

Figure 1: A diagram illustrating the flow of data from initial collection to storage, access, and use. This process constitutes "closing the loop" on assessing General Education competencies. The blue boxes highlight steps with direct faculty involvement; the red boxes indicate processes carried out by the General Education Coordinator. The golden region indicates the files that are publicly available on the Open Science Framework.

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.

Many different approaches can be used to assess a General Education program; direct and indirect assessment can take place at the course, program, or institutional levels. The structure, strengths and weaknesses of each of these are highlighted elsewhere (Allen, 2006). Regardless of the approach used, a quality program evaluation must possess five key attributes: utility, feasibility, propriety, accuracy, and accountability (Yarbrough et al., 2011). Both this report, and the assessment processes underlying it, are designed to satisfy these requirements.

The utility of the assessment process is a measure of how useful it is to the relevant stakeholders. A broad sampling of our faculty are engaged in the section of assessment outcomes, collection of data, and interpretation of assessment findings. This involvement ensures that any results are viewed within an appropriate context, and increases their value for program evaluation. The automated nature of the data collection, aggregation, and analysis increases the feasibility of this approach. Much of the reports must still be authored by the General Education Coordinator. However, having the data manipulations and analysis done automatically greatly simplifies the task. Propriety speaks to the ethical use of the data and results. Every effort has been made to ensure that the identities of all students and faculty involved in these studies is protected. No personally identifiable information will ever be included in these results. Furthermore, the General Education assessment results exist solely for the improvement of the General Education program – the results will never be used for the evaluation of specific courses or instructional personnel. The accuracy

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

of these reports is improved by the very nature of the analysis and reporting used. Meta-analyses (Borenstein et al., 2011) are used to compare groups of related assessment results. This approach can account for variation in scoring and student ability between courses and provide an a more realistic overview of student competencies. The range of meta-data collected in addition to student evaluations will permit the testing of a variety of research hypotheses. This report is also a form of reproducible research (Stodden et al., 2014). This report is computationally reproducible because the code needed to manipulate the de-identified data, perform the analyses, and create the figures are included within the Rmarkdown (.Rmd) files themselves. This approach was first described as "literate programming" in the 1980's (Knuth, 1984). The principle advantage to this approach is that anyone (at any time) can reproduce, critique, and extend these studies without needing to track down multiple documents, graphics files, and data sets. Finally, the accountability of reports such as this one is safeguarded by the involvement of faculty in contextualizing the results. All reports will be shared with appropriate focus groups for their input. Their comments and recommendations for future actions will be summarized and included within the discussion section of each document.

The overall process employed in this assessment strategy is illustrated in Figure 1 and described in the Methods. As a proof-of-concept, 13 semesters of student results from a 200-level Biology course are analyzed. A more typical analysis would be from a variety of courses (say from Biology, Chemistry, Physics, Geology, and Geography) to evaluate a specific FLO over a specific period of time.

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 13 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 13 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 973 student entries and is formatted as a comma-delimited text file (BIOL200Data.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 3.4.1 (2017-06-30)

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), checkmate(v.1.8.3), magrittr(v.1.5), dplyr(v.0.7.2), weights(v.0.85), mice(v.2.30), gdata(v.2.18.0), Hmisc(v.4.0-3), ggplot2(v.2.2.1), Formula(v.1.2-2), survival(v.2.41-3), lattice(v.0.20-35), moments(v.0.14), modeest(v.2.1) and pander(v.0.6.1)

loaded via a namespace (and not attached): gtools(v.3.5.0), splines(v.3.4.1), colorspace(v.1.3-2), htmltools(v.0.3.6), yaml(v.2.1.14), base64enc(v.0.1-3), rlang(v.0.1.1), foreign(v.0.8-69), glue(v.1.1.1), RColorBrewer(v.1.1-2), bindrcpp(v.0.2), plyr(v.1.8.4), bindr(v.0.1), stringr(v.1.2.0), munsell(v.0.4.3), gtable(v.0.2.0), htmlwidgets(v.0.9), evaluate(v.0.10.1), latticeExtra(v.0.6-28), knitr(v.1.16), htmlTable(v.1.9), Rcpp(v.0.12.12), acepack(v.1.4.1), scales(v.0.4.1), backports(v.1.1.0), gridExtra(v.2.2.1), digest(v.0.6.12), stringi(v.1.1.5), bookdown(v.0.4), rprojroot(v.1.2), tools(v.3.4.1), lazyeval(v.0.2.0), tibble(v.1.3.3), cluster(v.2.0.6), pkgconfig(v.2.0.1), MASS(v.7.3-47), Matrix(v.1.2-10), data.table(v.1.10.4), assertthat(v.0.2.0), rmarkdown(v.1.6), rstudioapi(v.0.6), R6(v.2.2.2), rpart(v.4.1-11), nnet(v.7.3-12) and compiler(v.3.4.1)

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.

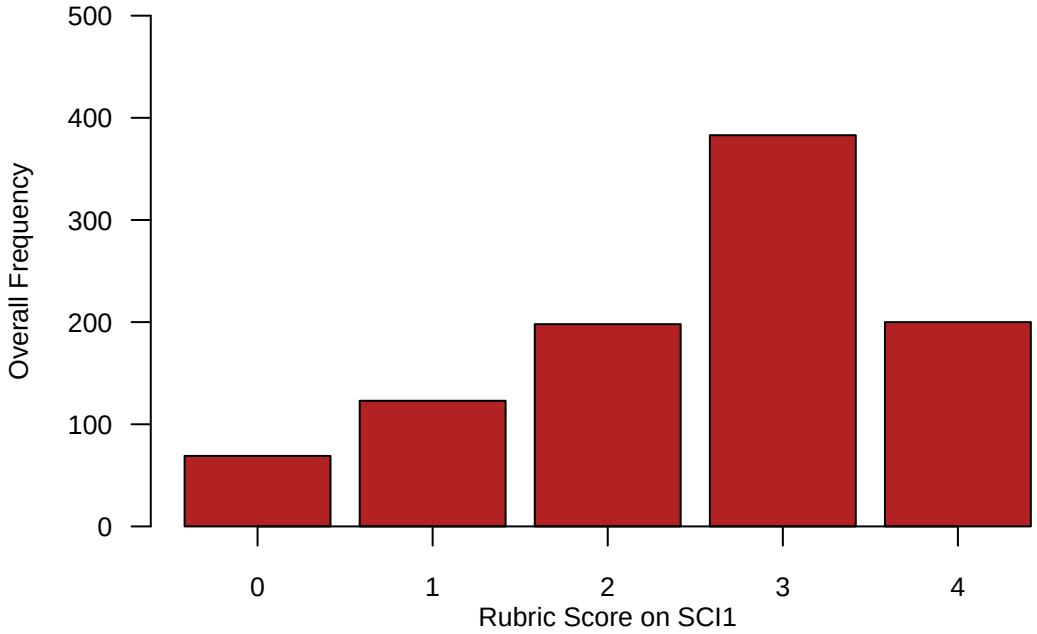

Figure 2: A histogram of the distribution of individual rubric score frequencies over all thirteen semesters.

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

All of the de-identified data, analysis code, and documentation that constitute this report project may be freely used, modified, and shared. The de-identified data set, BIOL200Data.csv, is released under the Creative Commons CC0 license. All documentation, including README.md, Codebook.md, and this report, are released under the Creative Commons CC-BY licence. Any questions, comments, or suggestions may be sent to Dr. Franklund.

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 Biological Sciences at Ferris State University. Their comments and suggestions will be included in the Discussion.

Summary statistics

A total of 973 student performances on exam 1 were collected over 13 semesters of instruction. Student scores were converted to rubric scores as described above. The overall average rubric score for all students and semesters was 2.54. The mode and median scores were 3 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=-1.71, df=972, p=0.087). The effect size for the difference between the average and the threshold was tiny (d=-0.05). We can infer from this that the overall average rubric score is not practically different than the threshold score.

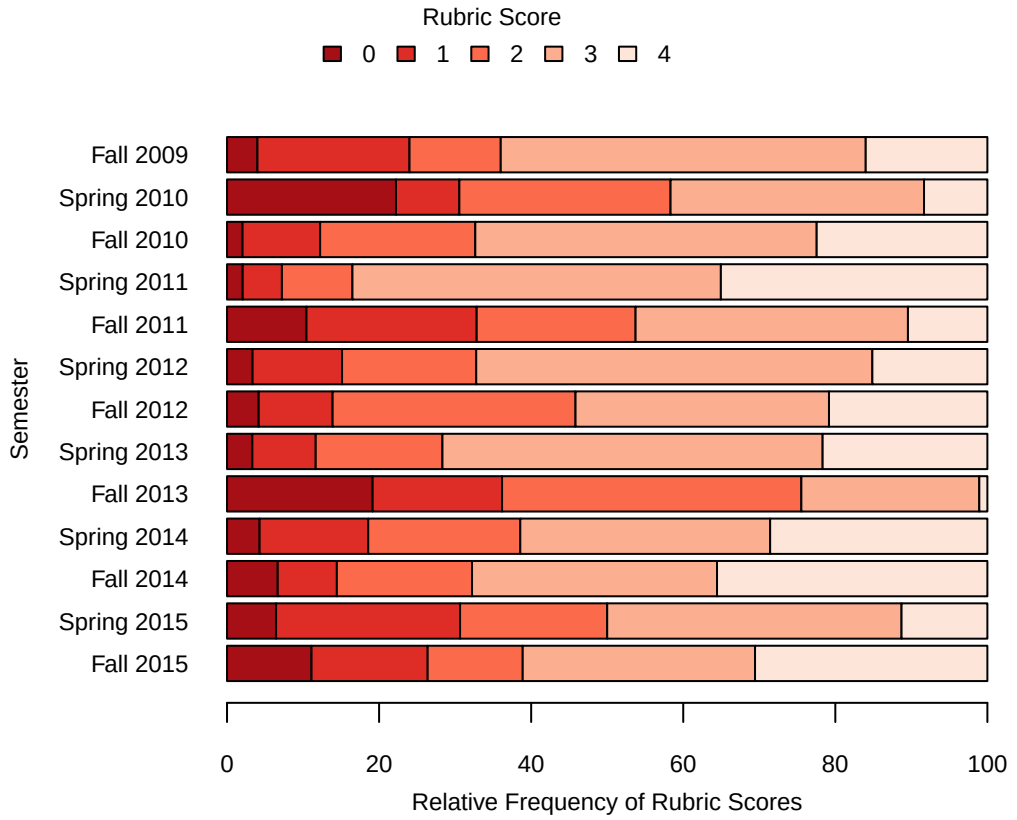

Figure 3: A barplot showing the distribution of rubric scores broken down by semester.

Table 2: One-way ANOVA analysis of scores by semester

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Semester	12	142.7398	11.894980	9.850787	0
Residuals	960	1159.2150	1.207516	NA	NA

The distribution of all rubric scores is shown in Figure 2. This distribution exhibited a moderate negative skew (skew = -0.62). 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 583 students (59.9%) met or exceeded the competence threshold over the semesters investigated.

The distribution of rubric scores by semester is shown in Figure 3. There are rather obvious differences in both the distribution of rubric scores and class sizes between semesters. A one-way ANOVA was used to compare the rubric scores by semester (Table 2). Unsurprisingly, there were statistically significant differences between semester scores. Semester of instruction, however, explained a relatively small amount of the overall variance ($\eta^2 = 0.11$).

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).

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

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).

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

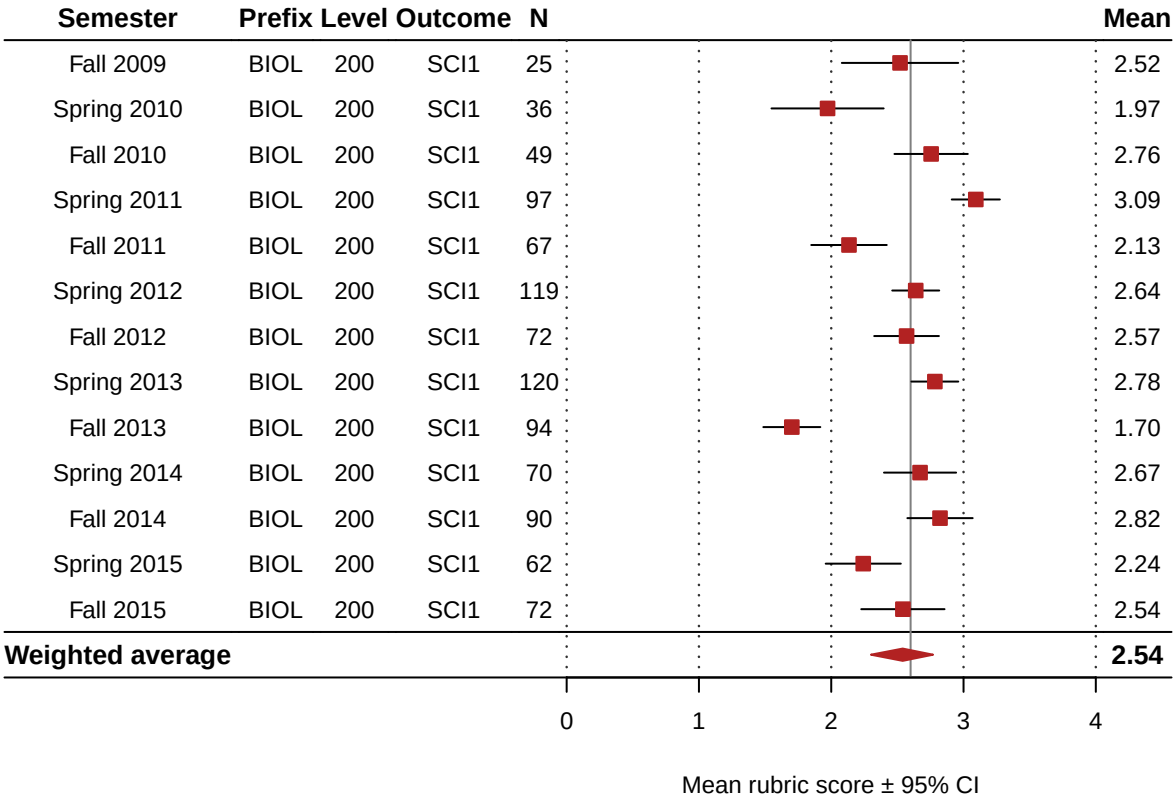

Figure 4: A forest plot of the average scores for each semester with a weighted mean estimate for the entire period investigated. Error bars indicate the 95% confidence intervals.

A forest plot of the meta-analysis is shown in Figure 4. 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.

The rubric scale can be conceptually divided into five areas as shown in Table 3. Of the 13 semesters, 6 fell in the proficient range, 6 fell in the developing range, and 1 fell in the beginning range. The weighted mean score, 2.54, was not significantly different from the threshold of competence as judged by a weighted, one-factor, two-tailed t-test (t=-0.57, df=12, p=0.58). We can conclude that the weighted average score is practically equivalent to the competency threshold score.

Table 3: 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

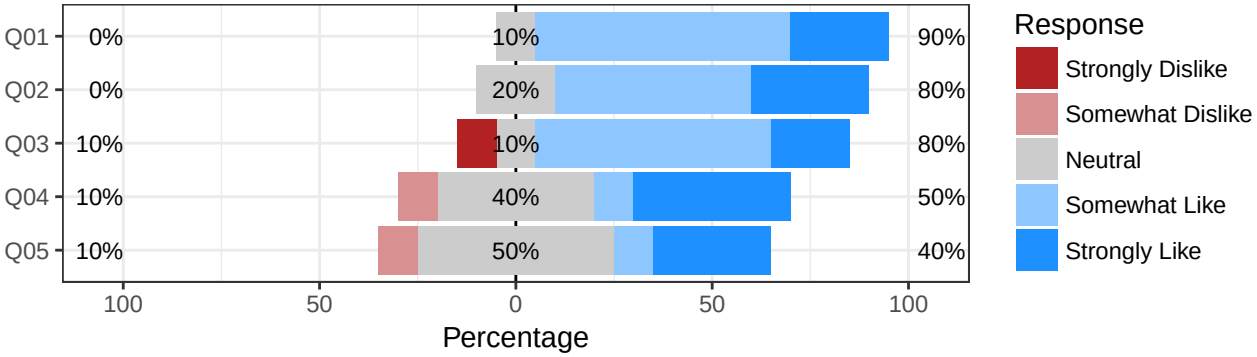

Figure 5: Results of the faculty survey (n=12)

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 from a 200-level Biology course was used as a proof-of-concept.

Over a span of 13 semesters, 46.2% of courses had mean scores considered to be proficient. Of all students in all semesters, 59.9% met or exceeded the competence threshold. From these data it is inferred that the students meet (just barely) the threshold of competence.

Faculty feedback

This report was distributed to members of the General Education Committee, Academic Senate, and the Department of Biological Sciences. These individuals were asked to provide their comments, suggestions, and concerns about this report and the processes involved in its creation. A simple SurveyMonkey instrument was developed to gather some feedback. The instrument consisted of the following six questions. The first five questions were five-level Likert scale items (Strongly dislike, Somewhat dislike, Neutral, Somewhat like, and Strongly Like). The format of the final item was free response (short answer).

- Q01 - What is your opinion of the format of the General Education assessment report?
- Q02 - What is your opinion of the content of the General Education assessment report?
- Q03 - What is your opinion of the data provenance plan for General Education?
- Q04 - What is your opinion of the use of meta-analysis in the General Education assessment report?
- Q05 - What is your opinion of the plan to publicly share assessment data, analyses, and reports?
- Q06 - What suggestions do you have to improve the General Education Report?

A total of 12 responses were obtained from the faculty. The results for the first five questions are summarized in Figure 5.

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

Allen, M. J. (2006). Assessing general education programs. Anker Pub. Co, San Fransisco, CA.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2011). Introduction to Meta-Analysis. John Wiley & Sons, West Sussex, UK.

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.

Knuth, D. E. (1984). Literate Programming. Comput. J., 27(2):97–111.

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.

Stodden, V., Leisch, F., and Peng, R. D. (2014). Implementing reproducible research. CRC Press/Taylor and Francis, Boca Raton, FL.

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

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., and Caruthers, F. A. (2011). The program evaluation standards : a guide for evaluators and evaluation users. SAGE, Thousand Oaks, CA.