

A Customizable Inquiry-Based Statistics Teaching Application for Introductory Biology Students

Mikus Abolins-Abols^{1,2*}, Natalie Christian¹, Jeffery A. Masters¹, and Rachel M. Pigg¹

¹Department of Biology, University of Louisville

²Center for Integrative Environmental Health Sciences, University of Louisville

Abstract

Building strong quantitative skills prepares undergraduate biology students for successful careers in science and medicine. While math and statistics anxiety can negatively impact student learning within biology classrooms, instructors may reduce this anxiety by steadily building student competency in quantitative reasoning through instructional scaffolding, application-based approaches, and simple computer program interfaces. However, few statistical programs exist that meet all needs of an inclusive, inquiry-based laboratory course. These needs include an open-source program, a simple interface, little required background knowledge in statistics for student users, and customizability to minimize cognitive load, align with course learning outcomes, and create desirable difficulty. To address these needs, we used the Shiny package in R to develop a custom statistical analysis application. Our “BioStats” app provides students with scaffolded learning experiences in applied statistics that promotes student agency and is customizable by the instructor. It introduces students to the strengths of the R interface, while eliminating the need for complex coding in the R programming language. It also prioritizes practical implementation of statistical analyses over learning statistical theory. To our knowledge, this is the first statistics teaching tool where students are presented basic statistics initially, more complex analyses as they advance, and includes an option to learn R statistical coding. The BioStats app interface yields a simplified introduction to applied statistics that is adaptable to many biology laboratory courses.

Citation: Abolins-Abols M, Christian N, Masters JA, Pigg RM. 2024. A Customizable Inquiry-Based Statistics Teaching Application for Introductory Biology Students. CourseSource 11. <https://doi.org/10.24918/cs.2024.6>

Editor: Katie Burnette, University of California, Riverside

Received: 6/2/2023; **Accepted:** 1/16/2024; **Published:** 4/5/2024

Copyright: © 2024 Abolins-Abols, Christian, Masters, and Pigg. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

Conflict of Interest and Funding Statement: Authors were supported by a Kentucky IDeA Networks of Biomedical Research Excellence Course-based Undergraduate Research Experiences Award (to MAA, NC, JAM, RMP), funded by NIH NIGMS Grant #P20GM103436. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. None of the authors have a financial, personal, or professional conflict of interest related to this work.

***Correspondence to:** 139 Life Sciences Bldg., University of Louisville, Louisville, Kentucky 40292 m.abolins-abols@louisville.edu

INTRODUCTION

Students in the biological sciences need strong quantitative skills to prepare them for careers in science and medicine (1). While biology students can gain experience with quantitative skills in other STEM courses and departments, it is increasingly evident that more emphasis is needed on quantitative learning outcomes within the biology curriculum itself (2, 3). Skills such as statistical analysis, data visualization, and data interpretation are integral to modern biology (4). Developing these skills within biology coursework can help students form a strong interdisciplinary foundation in the life sciences (3). However, students in biology courses often suffer from math and statistics anxiety that can hinder their learning (5). Reducing anxiety and building competency in quantitative reasoning could be achieved by providing students with scaffolded guidance during learning experiences (6, 7). Scaffolded instruction uses high structure, formative assignments, and frequent feedback to guide students from introductory to more advanced concepts and skills. This approach has previously been shown to promote student learning in laboratory courses (8).

Some of the most important tools that instructors use to teach quantitative skills are statistical programs and applications. Instructors can choose from a variety of conventional statistical programs for their introductory biology and introductory

statistics courses (Table 1; see [9] for additional programs). These applications vary in their target audience, licensing, ease of use, and customizability. However, instructors may struggle to find an application that meets all the needs of their classroom simultaneously. For instance, an inclusive learning experience in an inquiry-based introductory biology laboratory would require a statistics tool that is free for students to use, is easily accessible, can be used by students with no background in statistics, contains the most common tests needed for the analysis of biological data, and, optimally, can be customized to scaffold learning throughout the course. Existing commercial and open-source programs fit some, but not all, of these criteria. For example, Microsoft Excel is often used in introductory classes (10–12). It allows for data management and can run some common statistical tests. However, statistical analyses and data manipulation in Excel can be difficult for beginners. Furthermore, Excel lacks common tests used in biological research (such as non-parametric tests). Another popular statistics tool in introductory laboratories is R, a free statistical programming language that is commonly used in the natural sciences. However, despite its customizability and power, coding in R can be difficult for beginners (4) and may induce anxiety in some students (13), making it unsuitable for many first-year biology students. Perhaps among the most student-friendly analysis tool is the [HHMI Data Explorer](#), an online application where students can use an intuitive graphical user

interface to explore and analyze data using common statistical tests. However, there is currently no means for the instructor to remove options from the HHMI Data Explorer interface to scaffold the introduction of increasingly complex analyses and visualizations for learners across several assignments. Such customizability would help instructors further reduce the cognitive load of program interfaces for their students (14), thereby facilitating the use of statistical programs for beginners.

Faced with a lack of suitable student-focused statistical applications, some instructors have developed their own custom web-based apps (also called applets; for examples see [Rossman and Chance Applet Collection](#) and [Statpages](#)) to meet their course's and students' unique needs. The development of custom apps has been accelerated by the Shiny package in R, which allows anyone trained in R programming to create interactive web-based applications. Indeed, many apps developed using Shiny are already used to teach introductory statistics (Table 2). However, these apps either do not have a data file upload option or do not contain statistical tests that are relevant for biological studies, making them poorly suited for our students' needs. Additionally, many focus more on teaching statistical theory rather than practical application, which may be more suitable for an introductory statistics class rather than an introductory biology class. We therefore used the Shiny platform to develop our own interactive statistical application, BioStats, which satisfied our criteria by allowing us to create scaffolded learning experiences for our students, provide a simple, open-source user interface, and promote student agency in an inquiry-based, introductory biology laboratory course. Fully annotated open-source R code for our app is available on [GitHub](#) and a link to the web-based BioStats app is available upon request.

IMPLEMENTATION

Our BioStats application is built using the Shiny platform (15), an open-source R package that enables custom web application development. Applications built using Shiny can be implemented for free online either through one's own servers or through the [shinyapps.io website](#), which offers both free and paid subscriptions that differ in maximum usage time. Alternatively, Shiny-based applications can be run locally through RStudio on any computer free of charge.

We developed the BioStats application as part of a Course-Based Undergraduate Research Experience (CURE) that is embedded in the introductory biology curriculum across two laboratory courses. The primary learning goal of our introductory laboratory courses is to train students in the scientific method. In the CURE, students generate large datasets as a class, and are given freedom to propose and test their own hypotheses. Specifically, students conduct an ecological study to assess predicted relationships between soil microbial diversity, microbial abundance, soil chemistry, and various landscape or habitat variables (such as plant diversity or road density).

While designing this CURE, we looked for a statistical program that would allow students to use their own data to run simple statistical tests, without burdening students with statistical language, theory, or complexity that, at this point

in their learning, is unnecessary and may hinder them from achieving our courses' learning outcomes. Rather than learn the mathematics behind statistical tests, we wanted our students to learn how to **use and interpret** statistical tests in biology (16). To achieve this objective, we capitalized on the customizability of Shiny to create an application that differs from standard statistical software, including other Shiny statistics applications, in three main ways.

First, in keeping with best pedagogical practices, one of our main goals was to scaffold student learning about the practical uses of statistics, so that students can build on important concepts throughout the semester. We therefore created three, successive BioStats application "levels," where the higher levels added more options and information compared to the previous level (Table 3). For example, *t* tests are available in all levels, but the option to conduct an ANOVA is only available in higher BioStats levels, after students have familiarized themselves with how to use more basic statistical tests. This design allows us to minimize the cognitive load for students using the software during earlier laboratory assignments and introduce desirable difficulty in later assignments (17).

Second, we wanted to give students autonomy with respect to the tools that they use to learn. The R programming environment is becoming increasingly commonplace in modern natural sciences, and we strongly feel that students should be introduced to R as early as possible. However, studies have shown that if coding in R is required, it can be intimidating and stressful to students, especially students that are underrepresented in STEM or undertrained in quantitative approaches (13). The anxiety of coding may be reduced by providing the opportunity for students to optionally work with computer code in a risk-free environment. We therefore provided optional R code and instructions next to statistical tests on our BioStats application for students to run tests on their own computers if they chose to. To further encourage student agency and inquiry-based learning, our Shiny app asks students to make choices about the most appropriate statistical test for their data. For example, one of our learning objectives asks students to decide whether their data should be analyzed using a parametric or non-parametric test. To address this learning outcome, we first ask students to inspect the distribution of the residuals from the parametric test without showing students the *p* value or the test statistic—based on this distribution, students then proceed with the appropriate test.

Third, the customizability of the Shiny platform allows us to modify statistical outputs in BioStats. Traditional outputs of some statistical tests can be confusing for an introductory student. The output of statistical tests in BioStats is therefore simplified to provide a more intuitive interpretation of the analyses for beginners. For example, one of the major learning outcomes of our course is to interpret and contrast *p* values and effect sizes. In our BioStats tool, the output of linear regression includes only the *p* value and the effect size, R^2 . However, we intentionally leave out the degrees of freedom and alternative test statistics that are commonly shown in the output of conventional statistics programs. Other instructors may choose to customize the app differently to include other types of output to align with their course's learning objectives (see *Customization* below).

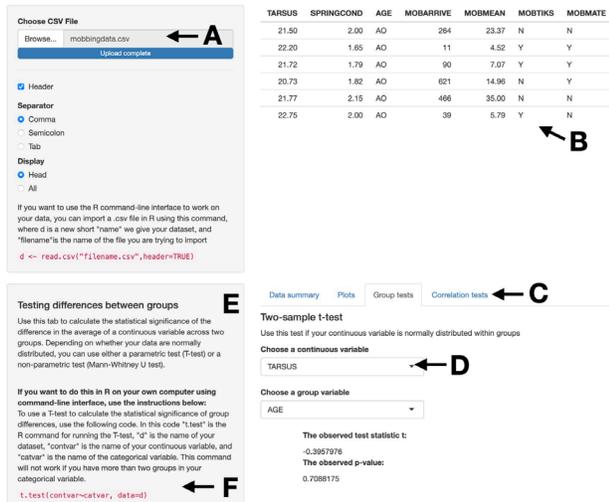


Figure 1. BioStats application layout. (A) Students can browse their computer directories for CSV files. (B) By choosing “Head” or “All” under the “Display,” students can view the uploaded file in its entirety or only the first 6 rows. (C) After the file has been uploaded, students can choose between different tasks from the menu (Data summary, Plots, Group tests, Correlation tests). Levels 2 and 3 add additional options to this menu. (D) Students select variables for running statistical tests using drop-down menus. (E) Upon opening a particular menu item, the shaded sidebar shows relevant information about the statistical test. (F) The sidebar also provides instructions about how to run these tests in R.

The BioStats app consists of a file upload section, a menu for statistical tests, and a side bar that explains the function of the statistical tests and provides an optional code for students to run these tests in R (Figure 1). Students can upload and view any CSV file in the app. After uploading a CSV file, students can choose to run summary statistics, generate plots (histograms, box plots, and scatter plots), conduct correlation tests, conduct parametric and non-parametric two-sample group tests, conduct ANOVA (levels 2 and 3 only), run chi-squared or Fisher’s exact tests (level 3 only), or conduct paired two-sample tests (level 3 only, see Table 3 for the full list of capabilities of each BioStats level). To run each test, students choose variables from drop-down menus, which read the column names from the imported CSV file. Upon choosing the variables, BioStats provides the test statistic and *p* value.

Our students primarily access BioStats applications online through the shinyapps.io website using their personal computers, which forgoes the barrier of novel software installation. Additionally, we have made the software locally available on laboratory computers to provide backup access to the app during Internet outages and to make it available to students that do not own a laptop. Because of the large number of students enrolled in our introductory biology sequence (up to 23 sections of 24 students each semester), our online app is used almost constantly by hundreds of students throughout the academic year. Our institution therefore subscribes to the Basic shinyapps.io plan that provides more usage time. Courses with fewer students may be able to use the free shinyapps.io subscription option. Alternatively, institutions with laboratory computers can run the application for free locally.

Example Activity with BIOSTATS

Students use BioStats for four, two-hour laboratories throughout our introductory biology lab sequence (Figure 2).

Students first use BioStats level 1 to analyze a simple dataset on songbird antipredatory behavior in the Introduction to Statistics laboratory. In the following lab, students use BioStats level 2 to test their own hypotheses about more complex data that the class generated on soil microbes, soil chemistry, and site properties. Finally, students use BioStats level 3 to analyze data that they collect as part of two laboratories that explore concepts related to respiration and animal behavior. These latter laboratories can include paired or multi-group designs, therefore level 3 is the most advanced BioStats level.

Before students arrive for the in-person Introduction to Statistics laboratory, they complete a formative 20-question pre-lab assignment online (available upon request) where students learn to distinguish between independent and dependent variables, describe the roles of sample size and random sampling in statistical testing, interpret the meaning of *p* values, and describe the difference between the *p* value and the effect size. To appreciate that statistics is not a pre-determined list of procedural steps, but that different data require different statistical tests, students also learn to distinguish between different data distributions, and choose tests that are best suited for analyzing different types of data. The pre-lab assignment is due before the in-person laboratory.

During the in-person Introduction to Statistics laboratory, students are provided with a fillable electronic worksheet (Supporting Files S1, S2) and a CSV file (Supporting File S3) containing a dataset from (18). Working in groups of four, students first read a summary of the study, explore the structure of the dataset, and propose two hypotheses about the relationship between variables in the dataset. Students then use BioStats level 1, in conjunction with the worksheet, to plot data relevant to their hypotheses and choose the statistical tests most appropriate for testing their hypotheses. For example, students are asked to choose between correlation and group tests, and between parametric and non-parametric tests. Students then report the outcomes of the statistical tests and paste the figures into their worksheet. At the end of the lab, students write their statistical methods and results in the format of a scientific paper. Optionally, interested students can use code displayed in the app next to each test to conduct the tests in RStudio using the command-line interface.

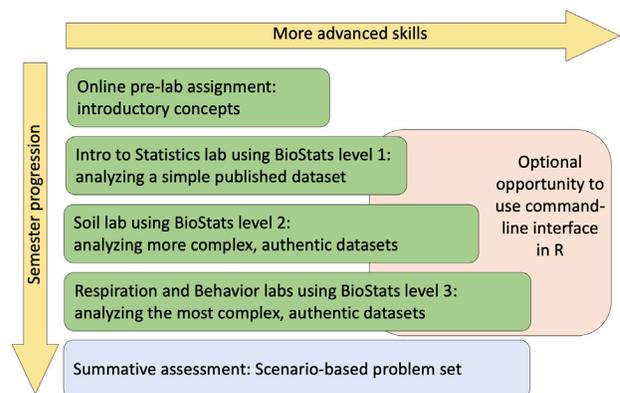


Figure 2. Instructional scaffolding of BioStats app usage throughout the semester. Green boxes indicate formative assignments, blue box indicates the summative assessment.

We use the BioStats application in three other laboratories later in the semester, adding more advanced tools and capabilities in almost every iteration (worksheets for laboratories associated with more advanced BioStats levels are available upon request). Students report their findings in full-length reports in the format of a scientific paper. We provide rubrics for student lab reports, including specific guidance on conducting, visualizing, and reporting the outcome of statistical tests. The rubric provides clear and consistent expectations for the students and graduate teaching assistants to assess student learning.

By scaffolding our instruction, we thus guide students through the Bloom's taxonomy hierarchy of learning outcomes. Pre-lab assignments focus on **understanding** foundational statistical concepts, while in the laboratories students successively progress towards higher-order learning outcomes. Specifically, students **evaluate** which statistical tests are appropriate for the data and **apply** their understanding about statistical test results to determine whether their hypotheses have been supported or refuted. Our setup allows students to actively participate in the process of science—designing experiments, analyzing and interpreting data, and communicating their findings—thus developing a sense of ownership of their work and agency as a developing scientist. At the end of the semester, students are tested on their mastery of the learning outcomes related to statistics (Supporting File S4) using a summative, graded online problem set. This assignment consists of 20 scenario-based questions (19) that assess students' ability to select appropriate statistical tests and evaluate test results correctly.

Customization

We encourage instructors to use BioStats and adapt it to their own curricula. To facilitate this, we have created a simplified version of BioStats (level 0), which can be used as a customizable starting point in other courses. We have published fully annotated open-source R code for creating all four levels (0, 1, 2, 3) of the BioStats app on [GitHub](#). Links to the web-based BioStats levels are available upon request. Shiny provides free video and internet [tutorials](#) about creating Shiny apps that can be used to supplement the annotations in our code. We will continue posting updates, improvements, and additions to BioStats on GitHub.

SUPPORTING MATERIALS

- S1. Statistics Application – Worksheet
- S2. Statistics Application – Worksheet Key
- S3. Statistics Application – Dataset
- S4. Statistics Application – Learning Outcomes

ACKNOWLEDGMENTS

Authors were supported by a Kentucky IDeA Networks of Biomedical Research Excellence Course-based Undergraduate Research Experiences Award (to MAA, NC, JAM, RMP), funded by NIH NIGMS Grant #P20GM103436. We additionally thank our graduate teaching assistants for their feedback during the creation of the BioStats app and associated course materials and our department for their support while creating the CURE (in particular, Drs. Perri Eason, Deborah Yoder-Himes, and Linda Fuselier).

Table 1. Characteristics of commonly used commercial and open-source software used in introductory statistics instruction. GUI: Graphical user interface; \$\$: cost to the student or institution.

Software	Type	Cost	Interface	Level	Example Citations
Excel	web or desktop	free /\$\$	spreadsheet	introductory-intermediate	(10–12)
HHMI Data Explorer	web	free	GUI	introductory	(20)
Minitab	web or desktop	\$\$	GUI	introductory-advanced	(11, 21)
R	desktop	free	command-line	intermediate-advanced	(9)
R Studio	web or desktop	free	command-line	intermediate-advanced	(22)
SAS	web or desktop	free	GUI	introductory-advanced	(23)
SPSS	desktop	\$\$	GUI	introductory-advanced	(9, 11)

Table 2. Examples of statistics applications developed using the Shiny platform.

App(s)	Institution	Capability	Data Upload Possible	Level
Cal Poly Shiny App collection	Cal Poly	t test	Y	introductory
Happy Apps	Grand Valley State University	t test, regression, chi-squared	Y/N	introductory-intermediate
Little Apps	Macalester College	t test, regression	Y	introductory-intermediate
The Book of Apps	Penn State	t test, regression, ANOVA	N	introductory-advanced
Duke ShinyEd	Duke University	ANOVA	N	introductory-intermediate
Biostats	University of Louisville	t test, Mann-Whitney U , Spearman and Pearson correlations, ANOVA, Kruskal-Wallis, chi-squared, Fisher's exact	Y	introductory

Table 3. Statistical capability of each BioStats application level.

Capability	Level 1	Level 2	Level 3
Summary statistics (mean, sample size, range, variance)	✓	✓	✓
Histogram	✓	✓	✓
Scatter plot	✓	✓	✓
Box plot	✓	✓	✓
Pearson's product moment correlation	✓	✓	✓
Spearman's rank-order correlation	✓	✓	✓
Two-sample t test (unpaired)	✓	✓	✓
Two-sample t test (paired)			✓
Mann-Whitney U test (unpaired)	✓	✓	✓
Mann-Whitney U test (paired)			✓
ANOVA		✓	✓
Tukey's multiple comparisons post hoc test for ANOVA		✓	✓
Kruskal-Wallis test		✓	✓
Dunn's multiple comparisons post hoc test for Kruskal-Wallis test		✓	✓
Chi-squared test			✓
Fisher's exact test			✓

REFERENCES

- American Association for the Advancement of Science (AAAS). 2011. Vision and change in undergraduate biology education: A call to action. AAAS, Washington, DC.
- Hoffman K, Leupen S, Dowell K, Kephart K, Leips J. 2016. Development and assessment of modules to integrate quantitative skills in introductory biology courses. *CBE Life Sci Educ* 15:ar14. doi:10.1187/cbe.15-09-0186.
- Bray SR, Duffin PM, Wagner JD. 2016. Thinking deeply about quantitative analysis: Building a biologist's toolkit. *CourseSource* 3. Source. doi:10.24918/cs.2016.4.
- Clark AD, Stevison LS. 2023. Learning R for biologists: A mini course grab-bag for instructors. *CourseSource* 10. doi:10.24918/cs.2023.12.
- Yang S, Hazlehurst J, Taniguchi DAA. 2021. Cats teach stats: An interactive module to help reduce anxiety when learning statistics in biology. *Am Biol Teach* 83:542–544. doi:10.1525/abt.2021.83.8.542.
- Kirschner PA, Sweller J, Clark RE. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ Psychol* 41:75–86. doi:10.1207/s15326985ep4102_1.
- Eddy SL, Hogan KA. 2014. Getting under the hood: How and for whom does increasing course structure work? *CBE Life Sci Educ* 13:453–468. doi:10.1187/cbe.14-03-0050.
- Killpack TL, Fulmer SM, Roden JA, Dolce JL, Skow CD. 2020. Increased scaffolding and inquiry in an introductory biology lab enhance experimental design skills and sense of scientific ability. *J Microbiol Biol Educ* 21:21.2.56. doi:10.1128/jmbe.v21i2.2143.
- Stemock B, Kerns L. 2019. Use of commercial and free software for teaching statistics. *Stat Educ Res J* 18:54–67. doi:10.52041/serj.v18i2.140.
- Popolizio TR, Killpack TL. 2021. A remote-learning framework for student research projects: Using datasets to teach experimental design, data analysis and science communication. *CourseSource* 8. doi:10.24918/cs.2021.27.
- Prvan T, Reid A, Petocz P. 2002. Statistical laboratories using Minitab, SPSS and Excel: A practical comparison. *Teach Stat* 24:68–75. doi:10.1111/1467-9639.00089.
- Warner CB, Meehan AM. 2001. Microsoft Excel as a tool for teaching basic statistics. *Teach Psychol* 28:295–98. doi:10.1207/S15328023TOP2804_11.
- Forrester C, Schwikert S, Foster J, Corwin L. 2022. Undergraduate R programming anxiety in ecology: Persistent gender gaps and coping strategies. *CBE Life Sci Educ* 21:ar29. doi:10.1187/cbe.21-05-0133.
- de Jong T. 2010. Cognitive load theory, educational research, and instructional design: Some food for thought. *Instr Sci* 38:105–134. doi:10.1007/s11251-009-9110-0.
- Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B. 2023. shiny: Web application framework for R. (R package version 1.7.4.9002), <<https://CRAN.R-project.org/package=shiny>>.
- Bradstreet TE. 1996. Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *Am Stat* 50:69–78. doi:10.1080/00031305.1996.10473545.
- Bjork EL, Bjork RA. 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning, p 59–68. *In* Gernsbacher MA, Pew RW, Hough LM, Pomerantz JR (ed), *Psychology and the real world: Essays illustrating fundamental contributions to society*. Worth Publishers, New York, NY.
- Abolins-Abols M, Ketterson ED. 2017. Condition explains individual variation in mobbing behavior. *Ethology* 123:495–502. doi:10.1111/eth.12625.
- Pigg RM, Rauschert ESJ, Yang S. 2023. Long story short: Using current primary literature to generate scenario-based question sets. *Am Biol Teach* 85:169–171. doi:10.1525/abt.2023.85.3.169.
- Park PJ, Gorsen DM, Girgenti CD, Del Bello IG, Tobitsch CM, Wu JR, Hong AJ. 2022. From statistics to ecological analyses: An application of the New York City East River Ichthyology Database (ERID) using spreadsheets. *Adv Biol Lab Educ* 42. doi:10.37590/able.v42.art15.
- Metz AM. 2008. Teaching statistics in biology: Using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. *CBE Life Sci Educ* 7:317–326. doi:10.1187/cbe.07-07-0046.
- Del Toro I, Dickson K, Hakes AS, Newman SL. 2022. Early undergraduate biostatistics & data science introduction using R, R Studio & the Tidyverse. *Am Biol Teach* 84:124–129. doi:10.1525/abt.2022.84.3.124.
- Lawton S, Taylor L. 2020. Student perceptions of engagement in an introductory statistics course. *J Educ Stat* 28:45–55. doi:10.1080/10691898.2019.1704201.