7-2019

Beyond linear regression: A reference for analyzing common data types in discipline based education research

Elli J. Theobald
University of Washington

Melissa Aikens
University of New Hampshire

Sarah L. Eddy
Biology Departmentand STEM Transformation Institute, Florida International University, seddy@fiu.edu

Hannah Jordt
University of Washington

Beyond linear regression: A reference for analyzing common data types in discipline based education research

Elli J. Theobald,[1,*] Melissa Aikens,[2] Sarah Eddy,[3] and Hannah Jordt[1]

[1]*Department of Biology, University of Washington, Seattle, Washington, 98195, USA*
[2]*Department of Biological Sciences, University of New Hampshire, Durham, New Hampshire, 03824, USA*
[3]*Department of Biological Sciences and the STEM Transformation Institute, Florida International University, Miami, Florida, 33199, USA*

(Received 24 August 2018; published 3 July 2019)

[This paper is part of the Focused Collection on Quantitative Methods in PER: A Critical Examination.] A common goal in discipline-based education research (DBER) is to determine how to improve student outcomes. Linear regression is a common technique used to test hypotheses about the effects of interventions on continuous outcomes (such as exam score) as well as control for student nonequivalence in quasirandom experimental designs. (In quasirandom designs, subjects are not randomly assigned to treatments. For example, when treatment is assigned by classroom, and observations are made on students, the design is quasirandom because treatment is assigned to classroom, not subject (students).) However, many types of outcome data cannot be appropriately analyzed with linear regression. In these instances, researchers must move beyond linear regression and implement alternative regression techniques. For example, student outcomes can be measured on binary scales (e.g., pass or fail), tightly bound scales (e.g., strongly agree to strongly disagree), or nominal scales (i.e., different discrete choices for example multiple tracks within a physics major), each necessitating alternative regression techniques. Here, we review extensions of linear modeling—generalized linear models (glms)—and specifically compare five glms that are useful for analyzing DBER data: logistic, binomial, proportional odds (also called ordinal; including censored regression), multinomial, and Poisson (including negative binomial, hurdle, and zero-inflated) regression. We introduce a diagnostic tool to facilitate a researcher's identification of the most appropriate glm for their own data. For each model type, we explain when, why, and how to implement the regression approach. When: we provide examples of the types of research questions and outcome data that would motivate this regression approach, including citations to articles in the DBER literature. Why: we name which linear regression assumption is violated by the data type. How: we detail implementation and interpretation of this modeling approach in R, including R syntax and code, and how to discuss the regression output in research papers. Code accompanying each analysis can be found in the online github repository that is associated with this paper (https://github.com/ejtheobald/BeyondLinearRegression). This paper is not an exhaustive review of regression techniques, nor does it review nonregression-based analyses. Rather, it aims to compile and summarize regression techniques useful for the most common types of DBER data and provide examples, citations, and heavily annotated R code so that researchers can easily implement the technique in their work.

DOI: 10.1103/PhysRevPhysEducRes.15.020110

I. INTRODUCTION

Undergraduate education is undergoing a transformation: traditional lecture-based teaching is being replaced with active learning in which students engage with course content, and each other, in class. This transformation is progressing because active learning increases student academic outcomes [1]. At the heart of this progress are instructors who want to help students maximize their learning, and discipline-based education researchers who assess the success of instructional methods and course innovations.

Linear regression has been established as a best practice in testing hypotheses in discipline-based education research [2]. Linear regression is a common statistical technique in which a continuous outcome variable is modeled as a linear function of one, or multiple, predictor variables. Physics education researchers have used linear regression models to

*Corresponding author.
ellij@uw.edu

test many hypotheses related to educational outcomes such as student performance, e.g., Refs. [3–8] and student evaluation scores [9].

Linear regression, however, is not appropriate for all the diverse types of data that education researchers collect. For example, in addition to being interested in student performance, researchers also may investigate student retention, the number of times a student participates in class, student affect (e.g., attitudes), or the roles students assume in group work. These outcome variables, for example, are binary, count, or categorical, rather than continuous (Fig. 1), and violate assumptions of linear regression models (see below for more details). Thus, researchers with these types of data must move beyond linear regression models and employ generalized linear models.

Generalized linear models are a group of regression models that are often suitable alternatives to linear regression. These models mathematically convert noncontinuous data into data that can be modeled linearly, thus the name generalized *linear* models. Generalized linear models were first formally introduced in the 1970s by statisticians Nelder and Wedderburn. At this time, statistical methods to analyze binary and some forms of count data had been developed [10,11]. However, in their seminal paper *Generalized Linear Models* [12], Nelder and Wedderburn discussed a unified framework for analyzing binomial and count data using linear models, by incorporating a link function into the analysis that transforms the relationship between predictor and outcome variables into an additive, linear model. At about the same time that generalized linear models were

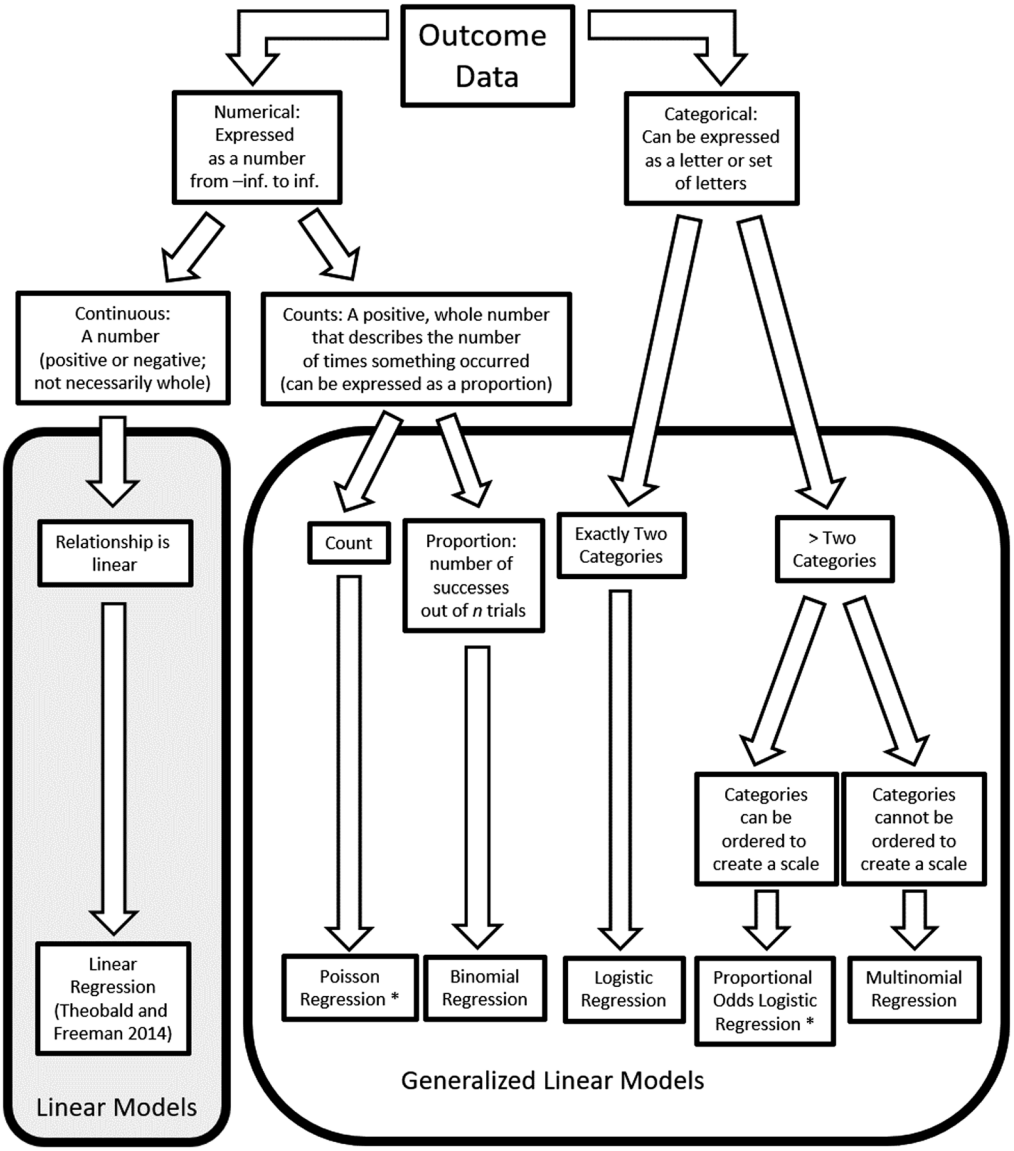

FIG. 1. Diagnostic tool useful for determining which type of generalized linear model is appropriate for the data collected. This tool is used in an analogous manner as a dichotomous key. The * indicates that there are special cases of these models that are discussed in this paper.

formalized, extensions of binary models for analyzing categorical data were being developed, primarily in the field of econometrics [13]. Since categorical data analysis was based on binary models, the generalized linear modeling framework could also be applied to these data types. Data analysis using generalized linear models grew in popularity, partly owing to the development of computer programs that conducted this type of analysis. In 1983, the first book on generalized linear models was published by McCullagh and Nelder [14].

Generalized linear models have been employed in the fields of sociology, psychology, and biomedical science at least since the 1990s. For example, an article in the journal *Psychological Bulletin* in 1995 advocates for the use of a generalized linear modeling method for count data [15], a paper that has been cited over 1200 times. In 2001, Sage

TABLE I. Glossary of common terms.

Term	Definition
[Model] Assumptions	Criteria that have to be met for a model to produce valid results. If assumptions are violated, model output could be misleading. Most assumptions for glms relate to the nature of the outcome variable, the relationship between predictor variables and the outcome variable, and the distribution of error. Linearity is the primary model assumption that generalized linear models overcome.
[Residual] Error	The difference between the observed value of the outcome and the value of the outcome predicted from the model. Each modeled data point has residual error. Residual error is visualized in a plot between predicted values and actual values
Heteroscedastic (opposite: homoscedastic)	Unequal residual error across the range of fitted values or across the range of predictor values. This occurs when a model fits better for some values of a predictor than others. Diagnosed by visualizing a plot of residuals versus fitted values, or residuals versus any predictor in the model; there should not be a strong pattern in the residuals across the fitted (or predictor) values.
Interaction	The multiplicative effect of two or more predictor variables on the outcome variable. This means that the effect of one predictor on the outcome changes depending on the value of the second predictor.
Link or linking function	A function that allows a non-linearly distributed outcome variable to vary linearly with predictor variables. For example, the logit link (log odds) and log link are used in glms discussed in this paper.
Log odds (logit transformation of probabilities)	The logarithm of the odds. The output of most glm models is log odds; log odds can be back transformed into odds using $e^{(\log \text{odds})}$.
Main effects	The direct effect of a predictor variable on the outcome variable. Main effects are outputted as "all else equal"—at the average value of all centered predictor variables, or the 0 value (or reference) for any noncentered predictor variable in the model.
Model selection	A technique that allows users to identify a subset of predictor variables that best describe their outcome variable. Methods can include stepwise decisions where one predictor is included or omitted from subsequent models. It is often applied by comparing models using likelihood ratio tests or Information Criterion (Akaike Information Criterion, AIC; Bayesian Information Criterion, BIC, etc.).
Multicollinearity	An occurrence wherein one predictor variable can be predicted from one or more predictor variables in the model. This results in a high correlation between predictor variables. This should be avoided in regression models. Synonym: correlated predictors.
Odds	The number of successes relative to the number of failures. Note: Odds are different from a proportion, which documents the number of successes out of the total number of trials. Odds can be converted to probability with the formula: $\text{probability} = \text{odds}/(1 + \text{odds})$
Odds ratio	The ratio of the odds for one group relative to another group. For example, if the odds for males is 0.2 and the odds for females is 0.3, the odds ratio is $0.2/0.3 = 0.667$.
Outcome variable	The variable that is being predicted in a regression model. Also called a response variable or dependent variable.
Predictor variable	The variable(s) that are being used to predict the outcome variable. Also called an explanatory variable or independent variable.
Reference level	The value of a categorical predictor variable that all other values are compared to. In 0/1 variables, defaults to 0; can be releveled in R (with function `relevel`).
Regression coefficient	An estimation of the change in the outcome variable accompanying a one unit increase in the predictor variable. Each predictor variable has a regression coefficient. Abbreviated β; also called estimate (*noun*).
Variance	The square of the standard deviation of a sample; describes how far each value in the data set is from the mean of the data. Note: variation is a general term describing the amount of variability; it is measured with various quantities, including variance.

publishing house released a book on the use of generalized linear models, in their series Quantitative Applications in the Social Sciences [16], demonstrating the popularity and utility of glms. Generalized linear models are now widely implemented across statistical platforms including SPSS, STATA, R, and SAS. The wide use, the ease of implementation, as well as the suitability of generalized linear models for data commonly collected by DBER researchers, suggest the utility of these models for our community.

In this paper, we identify when to move beyond linear regression, and how to apply generalized linear models to education data. We begin by reviewing the assumptions of linear regression. We then detail five types of generalized linear regression models: logistic, binomial, proportional odds (also called ordinal), multinomial, and Poisson. We also detail several special cases of these glms, including censored regression, overdispersion, and excess zeros. For each model type, we explain when, why, and how to implement the regression approach.

- When: we provide examples of the types of research questions and outcome data that would motivate this regression approach.
- Why: we explain which linear regression assumptions are violated by the data type.
- How: we detail implementation and interpretation of this modeling approach in R, including R syntax and annotated code, and how to discuss the regression output in research papers. All code associated with analyses can be found in the online github repository associated with this paper [17].

We ground our explanations in a diagnostic tool (Fig. 1) that allows researchers to identify the most appropriate model type for their own data. A glossary of common terms can be found in Table I.

We intend for this paper to be used as a reference, not necessarily read from start to finish. For this reason, each section can stand alone and readers may notice repetition throughout. We hope this allows researchers to easily implement each of these glm techniques in their work. This paper is not an exhaustive review of regression techniques, nor does it review nonregression-based analyses. There are many ways to test hypotheses in DBER—regression analysis, discussed here, is one. In contrast, hypotheses regarding social interactions between groups of students can be tested with network analysis [18]; hypotheses about how items on assessments relate to each other, differentiate respondents, and show bias can be tested with factor analysis [19], Rasch analysis [20], and differential item functioning [21], and some hypotheses are best developed and tested with mixed-methods research [22]. All of these methods are beyond the scope of this paper but are not to be overlooked when considering the best method to analytically test hypotheses in DBER.

Finally, there are several steps a researcher must take before fitting regression models including, but not limited to, data cleaning, checking for outliers, and checking for multicollinearity. These topics are intentionally avoided in this paper. We refer readers to two reference books, Gelman and Hill [23] and Fox and Weisberg [24], for more information on preparing data for use in regression analyses. Additionally, we have found these books to have particularly clear explanations, examples, and R code for implementation of regression models. Throughout this paper we have relied heavily on these books, as well as online tutorials (e.g., Introduction to R. UCLA: Statistical Consulting Group [25]) and the primary literature.

II. THE LIMITS OF LINEAR REGRESSION

This paper assumes a basic familiarity with linear regression. For readers who are new to statistical modeling and linear regression, we recommend the following sources: Moore and Notz [26] is a nicely written introductory text for developing understanding about why and how researchers should (and should not) use statistical models. Theobald and Freeman [2] extend this discussion by demonstrating the fallacy of applying nonregression based techniques to DBER data. Finally, Gelman and Hill [23] culminate with 25 chapters dedicated to applying all types of regression in social science research. The latter uses R to analyze the data in all of their examples. Researchers new to R may turn to helpful resources learning how to code, including online tutorials such as Try R [27] or STAT 545 [28], and the Field, Miles, and Field book *Discovering Statistics Using R* [29].

A. Linear regression assumptions

To understand why generalized linear models are most appropriate for the data types discussed above, it is helpful to understand the assumptions of linear regression that limit its broader applicability. In order of importance (according to Gelman and Hill [23]), the six assumptions of linear regression are (i) validity of the model, (ii) linearity, (iii) additivity, (iv) independence of errors, (v) equal variance of errors, and (vi) normality of errors. The majority of this paper focuses on the assumption of linearity, as it is *mathematically* the most important assumption to avoid violating [23]. However, before discussing linearity and how to remedy violations, we will briefly describe the other five assumptions which are worth considering.

Validity of the model.—All modeling rests on the assumption that the model being fit is a valid model. By validity, we mean that three conditions are met: (i) the outcome that is being modeled represents the phenomenon that is being predicted (i.e., that a researcher can answer their question with the data being collected), (ii) the appropriate predictors are included in the model, and (iii) the population from which the data is collected represents the population that predictions are being made about [23]. For example, a researcher interested in

understanding the factors that influence first-generation college student persistence in the physics major would need to (i) appropriately measure persistence in the physics major (e.g., by longitudinally tracking declaration of physics and nonphysics majors), (ii) choose a set of predictors they hypothesize influence persistence in the major based on theory or empirical studies (e.g., gender, grades in introductory courses, attitudes toward physics, etc.), and (iii) collect data from first-generation physics majors to make predictions about this specific population. It is primarily up to the researcher to assess the violation of this assumption based on their knowledge and the consensus in the primary literature. However, readers and reviewers are also responsible for assessing the validity of the regression models used in any paper before drawing conclusions from that paper. Validity is arguably the most important assumption to avoid violating [23] (even though it is not a mathematical assumption). Succinctly put, statistics cannot help you if your data do not match to your research question.

Additivity.—Additivity assumes that the relationship between the outcome and predictors is additive; in other words that

$$y = x_1 + x_2 + x_3 + \in,$$

where y is the outcome, x_{1-3} are predictors, and $\in$ is the error. If the relationship is not additive, but is actually multiplicative, such that

$$y = x_1 + x_2 * x_3 + \in$$

then the researcher can either add interactions (for example, $x_2 * x_3$), or transform the predictors (for example, $\log x_2 + \log x_3$) to make the relationship additive. The assumption of additivity is tested by plotting the outcome variable with each predictor variable to check for a linear relationship and also testing models that include interactions or transformed predictors [23].

Independence of errors.—This assumption states that observations [and the error ($\in$), or uncertainty, associated with these observations] are independent. Nonindependence can arise from nested, or clustered, designs or in repeated measures designs. For example, observations made on students within sections are nested within the section. Students in the same section may have outcomes that are more highly correlated to each other than to students in different sections due to their shared environment. Alternatively if a study is conducted across several years, the observations within year not independent. Similarly, in repeated measures designs, the same outcome is measured on the same subjects multiple times (e.g., over time) and those observations are not independent. Insufficiently correcting nonindependence (i.e., violations of the assumption of independence of errors) can lead to spurious conclusions. This nonindependence can be accounted for with a statistical method called multilevel modeling, as detailed in Gelman and Hill [23] and applied to DBER in Theobald [30]. Adding a random effect term of section, year, or student (respectively, from the examples above) accounts for the nonindependent nature of the observations. Nonindependence is a relatively common problem in education studies because of the frequent use of quasirandom study designs. Multilevel modeling can accommodate many types of the glms we discuss in this paper [23,30].

Equal variance of errors.—Also known as homoscedasticity (as opposed to heteroscedasticity), equal variance of errors assumes that the deviation of the modeled value of the outcome versus the actual value of the outcome is the same for all the modeled values. One well-known example of heteroscedasticity in educational data is in the relationship between SAT scores and college grades [31]. The relationship between these two variables tends to be stronger for students with high SAT scores and weaker for students with low SAT scores. This means that college grades for students with low SAT scores will not be predicted as well by the model as college grades for students with high SAT scores. The assumption of equal variance of errors is tested by visually examining plots of residuals versus fitted values. Ideally, there should be no strong pattern observed in these plots. If the assumption is grossly violated, Gelman and Hill [23] recommend using weighted least squares, where each point is weighted inversely proportional to its variance. Alternatively, one can use robust standard errors when data are heteroscedastic [23]. Heteroscedasticity has to be quite strong before it violates the model assumptions, so in most cases in DBER, including the example described above, it does not have to be corrected.

Normality of errors.—The errors of a linear regression model are assumed to have a normal distribution. This assumption is generally the least important, and according to Gelman and Hill [23], it is not necessary (or even recommended) to diagnose the normality of errors, particularly if the goal of the model is to test a hypothesis, as opposed to make predictions.

Linearity.—For the rest of the paper, we focus almost exclusively on this final assumption of linearity. Linearity means that the relationship between the outcome and each of the predictors is a straight line. In other words, a one unit change in the predictor translates to a specific amount of change in the outcome. The exact amount of change in the outcome is the same for every unit change in the predictor; this change in the outcome is the regression coefficient in the regression model. This assumption is violated when the outcome is not linear—generalized linear models were developed to overcome violations of linearity [12].

III. GENERALIZED LINEAR MODELS

The assumption of linearity is violated when the outcome variable is categorical, tightly bounded (such as a Likert-scale items with few response options), or expressed as counts. To illustrate this, let us consider a simple case of a count

outcome: how many times students raise their hand to volunteer in class. If an intervention is designed to increase the number of times each student raises their hand, a researcher might compare the number of times each student raises their hand in an intervention class to the number of times each student raises their hand in a non-intervention class. As all teachers know, getting a student who never raises her hand to raise her hand once is much harder than getting a student who raises her hand five times to raise her hand a sixth time. Thus, the outcome from 0 to 6 is not linear because the "distance" between 0 and 1 is much larger than the "distance" between 5 and 6. In a Poisson regression (discussed at length below), the outcome is log-transformed to become linearized.

The other common transformation that glms utilize is the logit transformation. Logit transformations are used when outcome data are categorical, and the probability of being in one category versus other categories is the outcome of interest. Specifically, a logit transformation is computed as

$$\text{logit} = \log\left(\frac{x}{1-x}\right).$$

When x is a probability (e.g., the probability of passing or failing a class, or the probability of majoring in one of several fields, etc.), $(\frac{\text{probability}}{1-\text{probability}})$ represents the odds of the event. Thus, the logit transformation when x is a probability is the logarithm of the odds, or the log odds. In these ways, generalized linear models move beyond the assumption of linearity that is a central tenet of linear regression. In fact, this is why these models are called generalized *linear* models: instead of modeling the outcome itself, these models relate predictor variables to outcomes using a link function, like the log or logit link in the examples above.

A. Interpreting log odds: Challenges and solutions

The output of glm models (i.e., the regression coefficients) that use a logit transformation are on the log odds scale, and log odds are difficult to interpret. By exponentiating the coefficient [e^{β}; $\exp(\beta)$ in R], the log odds are converted to odds ratios. Odds ratios are relative, meaning one group is always being compared to another group (i.e., odds ratio: odds for category 1/odds for category 2). Despite being easier to interpret than log odds [32], odds ratios are also notoriously challenging to interpret [33] because of the tendency to conflate the interpretation of odds with the interpretation of proportions [32,33].

Odds are a measure of the number of successes relative to the number of failures (successes versus failures) rather than the number of successes out of the number of trials (successes versus trials; i.e., proportions). For example, say we have a class of 200 students. Sixty-five (65) of the men pass the class and 35 fail; 90 women pass the class and 10 fail. The odds of a man passing the class are 65/35 or 1.8. The proportion of men who pass the class are 65/100 or 0.65. (The odds of a woman passing the class are 90/10 or 9.) The odds ratio for men versus women passing the class is the odds of men passing the class divided by the odds for women or $1.8/9 = 0.2$: the odds for men passing the class are 0.2 that of the odds for women. From this odds ratio we can get a sense of the direction and magnitude of the effect: an odds ratio of 1 indicates that there are equal odds for each group, an odds ratio greater than 1 indicates there are greater odds for the nonreference group, and an odds ratio less than 1 indicates there are greater odds for the reference group. Thus, here, women have better odds of passing than men, but the exact impact of gender on passing is still unclear because odds are not intuitive. Unfortunately, we cannot simply say something straightforward such as "men are 0.2 times as likely to pass the class as women"—this is not representing odds accurately. Instead, because the odds ratio is relative, it is actually saying something more convoluted and harder to understand: for every man not passing, 0.2 times as many men will pass than the number of women passing for every woman not passing.

An additional complication (for reasons nicely explained by Osborne [32]) of interpreting odds ratios occurs when the odds ratio is less than one. In these cases, it is preferable to take the reciprocal of the odds ratio and either reverse the relationship or the reference category. For example, when the odds ratio for men compared to women is 0.2, it is best to describe this relationship as the odds for women are 5 times (1/0.2) that of the odds of men.

An even more natural way to present these relationships is by describing changes in the *probability* of the outcome for each group. The conversion of odds to probabilities is relatively easy: probability is equal to the odds divided by the odds plus one. The function `plogis` does this in R. So here, the probability that a man will pass is $1.8/(1+1.8) = 64\%$ and the probability that a woman will pass is $9/(1+9) = 90\%$. However, converting *odds ratios* (the back-transformed glm regression coefficients) to probabilities is not as straightforward, particularly with continuous predictor variables. The *effects* package in R [34] or the *sjPlot* package in R [35], on the other hand, does this seamlessly. The *effects* package can model the predicted probability of being in a particular outcome category, given changes in the predictor variables, using the log odds generated from a glm model [36,37]. Specifically, it isolates the influence of each predictor individually in an "all else equal" context: for example, it calculates the probability of each outcome as the predictor of interest varies, while setting all other predictors in the model equal to their average value. These effects can be tabulated or plotted, making them easy to visualize.

IV. GENERALIZED LINEAR MODEL TYPES

A. Choosing between generalized linear model types

In the following sections we describe five common generalized linear regression models that are appropriate

to use with different types of outcome variables and how to interpret the results. Knowing which generalized linear model type to implement can be a daunting task, but Fig. 1 can help. Specifically, Fig. 1 is to be used as a decision tree: it is a guide to choosing the most appropriate generalized linear model based on the outcome data being analyzed.

B. Logistic regression

1. When to use logistic regression

Logistic regression is appropriate when the outcome variable is categorical with only two possible categories (Fig. 1). There are many examples of outcome data of this nature in DBER. For example, researchers have tested whether student characteristics influence: passing or failing a test [38] or a course [39], majoring or not majoring in physics [40], matriculating or not matriculating into medical or graduate school [41], or completing or not completing a science, technology, engineering, and math (STEM) Ph.D. [42].

In logistic regression, the binary outcome data are represented as 0 and 1, where 1 represents the outcome that the researcher is interested in modeling (i.e., "success" as defined by the researcher) and 0 is the reference level (i.e., the "failure" as defined by the researcher). The model estimates the probability that the outcome equals 1, given a set of predictor variables. For example, Kost and colleagues [38] examined how gender influenced passing the Force and Motion Concept Evaluation as a post-test. Since passing the test was the outcome of interest, each student in the sample was assigned a 1 if they passed the post-test and a 0 if they failed the post-test. Kost and colleagues [38] then used logistic regression to determine whether gender influenced the probability of passing the post-test.

2. Why use logistic regression

Because binary outcome data are categorical, the probability of being in a particular category is typically the outcome of interest. However, probability values are bounded between 0 and 1, resulting in a nonlinear relationship with predictor variables, and thus violating the linear regression assumption of linearity [43,44]. Logistic regression models estimate binary outcome data by using the logarithm of odds, also called logit transformation, to transform the outcome probability values into continuous, unbounded log odds that can be predicted with a linear function [45].

3. How to implement logistic regression in R

Logistic regression can be implemented in R using the `glm` function in the *base* package [46]. With logistic regression, `family=binomial` and `link="logit"` must be specified within the `glm` function (Table II; Appendix 1 [47]). If the binary outcome data are coded as 0 and 1 in the data set, then R may interpret these as integer data, when in fact they are categorical. Thus, it is important to change this data structure in R to be factor data, and set the reference level to 0. If the binary response data are coded as text in the data set (e.g., coded as "pass" or "fail"), then R should recognize it as a factor (always double check to be sure), but it is important to set the group that represents "0" as the reference level.

Here, we use modified data from a study that sought to understand life science majors' task values for using mathematics in the context of biology [63]. Specifically, we examine how several characteristics predict whether students report being likely to take an elective mathematical modeling biology course. In this model, the outcome is either "unlikely" to take the course (outcome = 0) or "likely" to take the course (outcome = 1). The predictor variables include: interest in using mathematics to understand biology (abbreviated: interest), perceptions of the usefulness of mathematics for their life science career (abbreviated: utility value), perceptions of the cost of incorporating mathematics into biology courses (abbreviated: cost), as well as students' gender, year in school, and highest mathematics course taken in high school. Annotated R code to conduct this analysis is provided in the logistic example in the github repository in Ref. [17].

4. How to interpret the output from logistic regression

Standard output from logistic regression models include coefficients for the intercept and each predictor variable, along with standard errors and p values, based on a z statistic (Wald test statistic), for each coefficient to assess whether the coefficient is significantly different from 0 (Appendix 1 [47]). An AIC value is also included in the output and can be used in model selection [64]. Because the logit transformation converts the outcome variable into the log odds of the outcome variable equaling 1, the regression coefficients indicate an increase or decrease in the log odds of the outcome variable equaling 1 for a one-unit increase in a continuous variable. Or they indicate an increase or decrease in the log odds of the outcome variable equaling 1 for a group in comparison to a reference group in a categorical variable. The intercept represents the log odds of the outcome variable equaling 1 when all continuous predictor variables equal 0 and when all categorical predictor variables are at their reference level.

Again, it is not intuitive to consider changes in log odds due to a predictor variable, so each regression coefficient should be back-transformed into an odds ratio by using the equation e^{β}, where β is the regression coefficient (as explained above and as is implemented in the R code). Interpreting the logistic regression in our example data, increased interest in using mathematics to understand biology significantly increased the log odds of reporting being likely to take a modeling course ($\beta = 0.71$, $p < 0.0001$; Appendix 1 [47]). The odds of reporting being likely to take a modeling course compared to reporting

TABLE II. Summary of how to use generalized linear models for data types frequently encountered in discipline-based education literature.

Data type		Outcome data example(s)	Regression type	R package	R model syntax	Example paper
Continuous		Exam performance	Linear model	*Base*[a]	Mod←lm(outcome∼predictor, data = data)	[2]
Binary		DFW/passing; 0/1; yes/no	Logistic	*Base*[a]	Mod ← glm(cbind(numerator, denominator)predictor,	[48]
Proportion		Proportion of classes attended; proportion of assignments completed; proportion of activities that employ active learning	Binomial	*Base*[a]	family = binomial(link = "logit"), data = data)	[49]
Categorical ordinal		Strongly Disagree to Strongly Agree, any number of categories	Proportional odds; Ordinal	*MASS*[b]	Mod ← polr(as.factor(outcome)predictor, data = data)	[50]
	Special case: Censored regression	Combining responses from multiple survey questions, resulting in a ceiling (or floor) effect	Censored regression	*censReg*[d]	Mod←censReg(outcome∼predictor, data = data, right = y_1, left = y_2), where right is the highest value possible (the ceiling) and left is the lowest possible value (the floor)	[51]
Multinomial		Choice of concentration within a major (e.g., ecology, cell, physiology, general)	Multinomial regression	*nnet*[e]	Mod←multinom(outcome∼predictor, data = data)	[52,53]
Count		Number of hand raises (e.g., per student)	Poisson	*Base*[a]	Mod←glm(outcome∼predictor, family = Poisson, data = data)	[54]
	Special case: Overdispersed count	Number of teaching resources provided to faculty	Negative binomial	*MASS*[b]	Mod←glm.nb(outcome∼predictor, data = data)	[55]
	Special Case: Count with excess zeros	Number of hand raises per student in a large class	Hurdle or Zero-inflated	*pscl*[c] [56]	Mod←hurdle(outcome∼predictor, data = data, dist = "poisson" or "negbin") Mod←zeroinfl(outcome∼predictor, data = data)	[57,58]

[a][46]
[b][59]
[c][60]
[d][61]
[e][62]

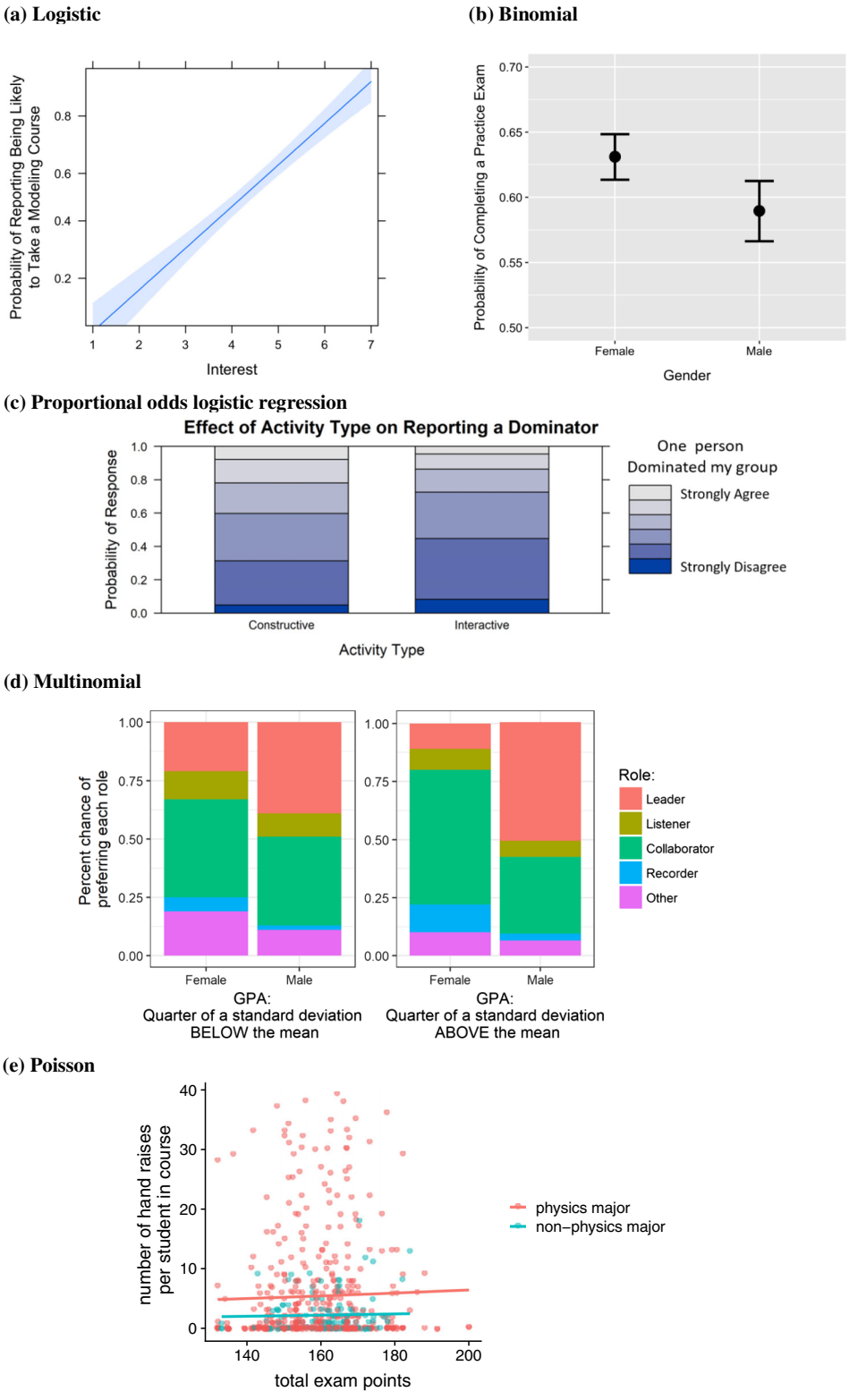

FIG. 2. Examples of publication-style figures for each type of generalized linear model. These figures were made from the data presented in each extended example. (a) Logistic: increasing interest in using mathematics to understand biology is related to an increase in the probability of reporting being likely to take a mathematical modeling in biology course. (b) Binomial: female students were more likely to complete optional practice exams than male students. (c) Proportional odds logistic regression: students were less likely to agree that someone dominated their group after they completed the Interactive Activity. (d) Multinomial: GPA moderates the influence of gender on the roles students prefer during group work. (e) Poisson: Physics majors (red line) are predicted to raise their hands more than nonphysics majors (blue line) across all values of exam point totals (dots are data points, not estimates).

being unlikely to take a modeling course increases by a factor of 2 for every unit increase in students' interest in using mathematics to understand biology (odds ratio $= e^{\beta} = e^{0.71} = 2.04$; Appendix 1 [47]).

It is important to remember that odds ratios are different than probabilities [32]. In the example here, an odds ratio of 2 does not mean that students with a one unit increase in interest are twice as likely to report being likely to take a modeling course. Rather, it means that for a one unit increase in interest, for every student that reports *not* being likely to take a modeling course, twice as many students will report being likely to take a modeling course. Odds ratios can be converted into probabilities for each category of the predictor, which are the most intuitive way to understand the effects of a predictor on an outcome. The *effects* package does this easily and can be used to tabulate or plot how each predictor, holding all others at their average values, affects the probability of reporting being likely to take a mathematical modeling course [e.g., Fig. 2(a)].

When odds ratios are less than one, they become more difficult to intuitively interpret [65]. Again, in our example, we found a significantly lower log odds of a fourth-year student reporting being likely to take a modeling course than a first-year student ($\beta = -0.52$, $p = 0.03$), corresponding to an odds ratio of 0.60. This can be interpreted as the odds of a fourth-year student reporting being likely to take a modeling course is 0.60 that of the odds of a first-year student reporting being likely to take a modeling course. Although clearly this is lower than the odds of a first-year student, it is difficult to conceptualize the magnitude of this difference. Again, the best approach is to calculate the odds ratio of the opposite comparison, by taking the reciprocal of the odds ratio ($1/\text{odds ratio}$), so that an odds ratio greater than 1 can be interpreted. For example, instead of calculating an odds ratio for fourth-year students compared to first-year students, the odds ratio of first-year students compared to fourth-year students can be calculated: $1/0.60 = 1.67$. This would be interpreted as the odds of a first-year student reporting being likely to take a modeling course is 1.67 times that of the odds of a fourth-year student reporting being likely to take a modeling course.

Thus far, we have focused on interpreting the coefficients for predictor variables, but the coefficient for the intercept can also be useful for interpreting the results. The intercept is particularly useful when the researcher is interested in visualizing the odds or the probability of an outcome for each level of a categorical predictor. As a reminder, the intercept represents the log odds of the outcome at the reference level, and the regression coefficients show change in the log odds as the parameters change (i.e., away from the reference). Therefore, the intercept and regression coefficient for a categorical variable can be added together to obtain the log odds of the outcome for each categorical level. For example, we could compute the odds of a student in each year of school (first through fourth year) reporting being likely to take a mathematical modeling course, holding all other variables constant at the reference level (if categorical) or at 0 (if continuous). In our model, the intercept of −2.02 represents the log odds of male, first-year students with a calculus background and scores of 0 on interest, utility value, and cost reporting being likely to take a mathematical modeling course (Appendix 1 [47]). To compare this to second-year students of the same gender, background, and task values, we would add the regression coefficient for second year (−0.16) to the intercept (−2.02), obtaining a log odds value of −2.18. Similarly, we would obtain values for third-year students (−2.21) and fourth-year students (−2.54). Exponentiating these values will give you the odds of students in each year of school (again, holding all other variables constant at the reference level or at 0) reporting being likely to take a mathematical modeling course (0.13, 0.11, 0.11, 0.08). More intuitively, these numbers can be converted to probabilities using the equation: $\text{odds}/(1 + \text{odds})$, or the `plogis` function in R, leading to probabilities of 0.12, 0.10, 0.10, and 0.07, which can be easily graphed and interpreted.

An important consideration when doing this is whether the reference values represented by the intercept make sense. In our example, the continuous variables of interest, utility value, and cost take a range of values from 1–7, so it does not make sense to compare students at a value of 0 for these variables. An alternative is to compare different levels of a categorical predictor at the mean value of each continuous predictor in the model. This can be done by centering the continuous variables before running the regression (i.e., subtract mean from each value). A different example: if SAT score were included in a model as a predictor, the intercept would indicate the expected value of the outcome when the SAT score is 0. This value is nonsensical, so centering SAT score by taking each student's score and subtracting the class mean centers SAT. In this case, the intercept is interpreted as average SAT score. The *effects* package can also be used to determine the odds or probability of an outcome for each level of a categorical variable (though the *effects* package will also hold categorical variables constant at the mean rather than a reference level; see Ref. [36] for an explanation of this).

5. Assumptions and how to test the assumptions of logistic regression

After a logistic model has been fit to the data, it is important to check the assumptions of the model. Logistic regression assumes that (i) the outcome variable is binary, (ii) the observations are independent, and (iii) the continuous predictor variables are linearly related to the log odds (logit) of the outcome variable. The first two assumptions are based on knowledge of the data and experimental

design. The third assumption can be tested graphically by creating a smooth scatterplot of the log odds of the predicted probabilities (also called fitted values) as a function of each continuous predictor in the model [44]. A smooth scatterplot does not assume a linear relationship between variables, but rather uses a smoothed curve to depict the relationship between variables. A linear relationship would indicate the assumption is met, whereas a nonlinear relationship would suggest either a transformation of the predictor variable is necessary or the logistic regression model is not an appropriate model for the data. For example, instead of the log odds (logit) transformation, a probit or complementary log-log model might be more appropriate for the data. We refer the reader to other authorities on the topic of when to use alternative transformations [13,23,24,45]. In our example data, the scatterplot of the log odds of the predicted probabilities as a function of interest looks linear. However, the scatterplots of the log odds of the predicted probabilities as a function of utility value and cost appear to have some curvature. Therefore, exploring probit and complementary log-log models would be worthwhile with this data set. All of these options are shown in the extended example in the code in the online repository [17].

C. Binomial regression

1. When to use binomial regression

Binomial regression, also called binomial logistic regression, is a more general form of logistic regression in which the outcome data are proportions. These proportions come from counts of the number of successful outcomes out of a given number of trials (Fig. 1). For example, a researcher interested in the probability of persistence in a STEM major may count the number of students that remain in a STEM major ("success" outcome) out of the total number of students who enroll in an introductory STEM class. The outcome is represented as the proportion of the total number of students that remain in a STEM major. This proportion, when measured across multiple introductory classes, could be used as an outcome variable. An example of the use of binomial data in education research can be found in Desjardins [49]: researchers calculated the proportion of students with at least one day of suspension to determine the effects of gender, ethnicity, poverty status, and whether the student was in a special education program on the probability of being suspended. Binomial data may also be generated by counting the number of "successes" in n independent trials (if each trial has a binary outcome), and representing that as a proportion. For example, a researcher may count the number of times a student turned in a homework assignment throughout the semester to calculate the proportion of homework assignments turned in by each student. Thus, logistic regression (as described above) is actually a special case of binomial logistic regression in which the number of trials equals one.

As in binary logistic regression, in binomial logistic regression, the researcher is modeling the probability of an outcome of interest, designated as a "1" or a "success." Even though there are multiple trials represented in the binomial data, the binomial regression still models the probability of success in one trial as a function of a set of predictor variables, and assumes that this probability is the same for all trials. For example, if the outcome data are the proportion of homework assignments turned in by each student, binomial regression will model the probability of a student turning in a homework assignment based on a set of predictors and assume that probability is the same for turning in each of the rest of the homework assignments.

There are two additional considerations for binomial regression. First, in binomial regression, the total number of trials can differ among variables or individuals in the data set. In these cases, it is important to input the total number of trials into the analysis (described below). When the total number of trials is not known, beta regression is an alternative technique that may be useful [66]. Second, binomial outcome data can resemble Poisson outcome data as they are both counts. If the count is tightly bounded, a binomial regression is more appropriate; on the other hand, as n becomes larger and the probability of success becomes smaller, the Poisson is a good approximation and thus Poisson regression can be used [67].

2. Why use binomial regression

Using proportional count data in a linear regression model is problematic for the same reasons as using binary data in a linear model. Primarily, because the data are proportions, they are bounded between 0 and 1, which results in a nonlinear relationship with predictor variables. (Particularly in the case of modeling "rare events," i.e., if probability of success or failure is small [68].) Furthermore, with proportional count data, the distribution of residuals is not always normally distributed or of equal variance [23]. As in logistic regression, a log odds transformation (also called logit transformation because the outcome is a probability) is applied to transform the bounded outcome into a continuous outcome that can be modeled as a linear function of predictor variables.

3. How to use binomial regression in R

To demonstrate how to implement and interpret the output of a binomial regression, we will use a data set we modified from Jackson and colleagues [69]. Here, students were given the opportunity to complete nine practice exams over the course of the semester, but completing practice exams was optional. In order to get a better sense of which students complete these optional practice exams, we examined whether gender, first-generation status (i.e., if they were the first in their family to attend college), and GPA influenced the

probability of completing a practice exam. The data and the R code can be found in the binomial regression example in the online repository [17].

Because binomial regression is a form of logistic regression, it is also implemented using `glm` and specifying `family=binomial (link="logit")`, as shown in Table II. However, instead of inputting a single column as the binary outcome of the model, a two-column matrix containing the number of "successes" and the number of "failures" is inputted as the outcome (Table II, Appendix 2 [47]). Alternatively, a column containing the proportion of successes (bounded between 0 and 1) can be input as the outcome, but "weight" must be set to a column containing the total number of trials from which each proportion of successes was calculated.

4. How to interpret the output from binomial regression

As in binary logistic regression, the output from binomial regression reports the regression coefficients as log odds, and their associated standard errors, z-values, and p-values. Exponentiating the regression coefficients (e^{β}) will yield odds ratios that describe the odds of success to failure for the reference and non-reference group. Here, both GPA and gender significantly predict the probability a student completed a practice exam. Exponentiating the regression coefficient for gender ($e^{0.17}$) gives an odds ratio of 1.19: the odds that a female completed a practice exam are 1.19 times the odds of a male student completing a practice exam. We can use the *effects* package to get a better understanding of the impact of gender on completing practice exams: a female student with the average score across all other variables is predicted (by the model) to have a 63% chance of completing a practice exam and an equivalent male student a 59% chance [Fig. 2(b)].

When the odds ratio is less than one, as is the case for GPA (Appendix 2 [47]), we take the reciprocal of the odds ratio: for each 1-unit increase in GPA the odds of completing a practice exam decreases by a factor of 4.1. Using the *effects* package, we see that all else equal, a student with a 2.4 GPA (for example) has an 85% chance of completing a practice exam, and a student with a GPA of 3.5 has a 55% chance (Appendix 2 [47]).

5. Assumptions and how to test the assumptions of binomial regression

Binomial data are proportions based on counts (or frequencies) of successes from independent trials, each with the same probability of success. Thus, binomial regression has two critical assumptions: (i) that the trials are independent, and (ii) that there is the same probability of success for each trial [24]. Thus, binomial regression is not appropriate for all proportion data, notably proportion data from trials that are not independent, or trials that do not have the same probability of success. For example, considering the sample data on practice exams, if students had a higher probability of completing the first practice exam than a later practice exam, perhaps because students were too busy to complete optional practice exams later in the semester, the probability of success (completing a practice exam) would differ between some of the trials (each instance of completing a practice exam), violating the assumptions of binomial regression. Furthermore, proportions should not be confused with ratios, which are quotients often used with the purpose of relativizing numbers (e.g., surface area to volume ratio). Ratios are not appropriately modeled with binomial regression, rather beta regression may be a more appropriate regression method. Cribari-Neto and Zeileis [66] provide more information on beta regression, including implementing it in R through the *betareg* package [70].

One final note about the binomial distribution: when there are a large number of trials, the proportion may be better estimated with a linear regression because as n trials increases, the binomial distribution approaches a normal distribution [67]. For example, if a researcher were interested in the number of days a student attended class, a binomial regression would be best if there were few class sessions—for example, a seminar class that met once per week (outcome $=$ n attended/n weeks). On the other hand, in K-12, most school districts require a 180-day school calendar. With this many days (i.e., trials), a linear regression is more appropriate as it will closely approximate a binomial regression. In these cases, a linear regression is preferred because the output of a linear regression (days) is easier to interpret than the output from a binomial regression (log odds), and may be more meaningful.

Similar to binary logistic regression, binomial regression also assumes that the predictor variables are linearly related to the log odds of the binomial outcome variable. This assumption can be tested by examining a smooth scatterplot of the log odds of the proportion data or the predicted probabilities as a function of each continuous predictor variable [45]. In our example, a scatterplot shows that GPA, the only continuous variable in the data set, is linearly related to the log odds of the proportion data.

Overdispersion, in which the observed variance is greater than that expected for a binomial distribution, occurs frequently in binomial regression [23]. Although there are multiple ways to test for overdispersion in a data set, in our example we examine the sum of the squared Pearson residuals divided by the residual degrees of freedom. This value should approximate 1 [71], and it can be used in a more formal test of overdispersion by comparing it to a χ^2 distribution [23]. To do this in R, we direct the readers to an overdispersion function presented for generalized linear mixed models [72]. For conducting and interpreting this test, see the R code for binomial regression found in the github repository [17].

D. Proportional odds logistic regression

1. When to use proportional odds logistic regression

Proportional odds logistic regression models (also called proportional odds models, ordered logistic regression models, ordinal logistic regression models, ordered logit regression models, or cumulative link models) are a type of glm that models categorical ordinal outcome data [73]. The outcome for this model can have any number of categories (greater than two), but the categories must be ordered (Fig. 1). For example, Likert-scale data from a survey where respondents report agreement on a scale from "strongly agree" to "strongly disagree" are categorical ordinal; the levels (i.e., "agree") are categories, and the categories are ordered in that strongly disagree is "lower" than disagree, which is "lower" than agree, etc. Categorical ordinal data are distinguished from binary data by having more than two categories and are distinguished from multinomial data in that multinomial data are not ordered (see below for more details about multinomial regression).

Categorical ordinal data are commonly generated via surveys. Surveys are a convenient way to quantitatively assess student attitudes, affect, or experience in a class. One common example used in physics DBER is the Colorado Learning Attitudes about Science Survey (CLASS), which measures students' beliefs about physics and attitudes about learning physics [74]. Proportional odds logistic regression would be ideal for analyzing results from this and other surveys with Likert-scale responses. For example, Theobald and colleagues [75] assessed whether students' affect changes after studying local or global examples of the biological impacts of climate change. Through an identical pre- and post-test, they asked three affect questions: (i) How likely is it that climate change will impact your life? (ii) How willing are you to change your behavior to reduce the impacts of climate change? And (iii) How much do you support government action to reduce greenhouse-gas emissions? Students answered these questions on a 5-point scale, where 1 = relatively little, 3 = a moderate amount, and 5 = a great deal. The researchers fit three proportional odds logistic regression models, one for each question, to test the hypothesis that the intervention is correlated with how students answered these questions. A thorough explanation of their methods details can be found in their paper [75].

Another common type of categorical ordinal data come from quizzes, concept inventories, or tests that have few questions (for few possible points). For example, Wiggins, Eddy, Grunspan, and Crowe [76] tested whether an intervention improved student learning on an 8-item post-test. The distribution of test scores was both highly skewed and tightly bounded, so treating them as linear and analyzing them with a linear regression would violate the assumption of linearity. Instead, the authors considered the outcome, measured as the number of questions correct, from 0-8, as categorical ordinal with each additional question answered correct as a "higher" category, and analyzed the data with a proportional odds model.

2. Why use proportional odds logistic regression

Categorical ordinal data violate the assumption of linearity because the linear distance between each level is not necessarily equivalent. For example, on a ruler, the distance between 1 and 2 in. is equivalent to the distance between 4 and 5 in. In Likert-scale survey data, however, the difference between strongly agree and agree may be smaller than the distance between disagree and neutral. Additionally, categorical ordinal data may approach a ceiling, or a floor, wherein students would have answered higher (or lower) if the outcome had allowed. For these reasons, modeling categorical ordinal data as continuous does not accurately reflect the structure of the data and thus can lead to inaccurate conclusions. Furthermore, when Likert-scale responses are reduced to percent agreement, or groups of respondents are compared simply using *t* tests, as is common with many surveys, a tremendous amount of information is lost. For example, averages can be the same but the variance much different, and it is impossible to account for differences between group compositions that might drive those differences in responses. Rather, the outcome should be modeled using proportional odds logistic regression: Instead of treating the outcome as linear, the outcome is translated to log odds (see logistic regression) and a logistic regression is fit for each level (i.e., category) compared to the others.

3. How to fit proportional odds logistic regression in R

Proportional odds logistic regression models can be fit in R with the `polr` function in the *MASS* package [59], as shown in Table II. If it is necessary (see proportional odds assumptions below) to have more advanced control of the linking function, thresholds or cutoffs (reported as Intercepts in the R output; Appendix 3 [47]), or to fit multilevel proportional odds models, the *ordinal* package [77] can help.

Theobald and colleagues [50] fit proportional odds models to test hypotheses about how students experience group work. They used both an eight-item pre-post-test as well as a single survey question (that was answered on a Likert-scale) to better understand the impact of a "dominator" in group work. Here, we will work through one of their examples in detail. For illustrative purposes, we have truncated their data to avoid the need to fit multilevel models, and will use the `polr` function in the *MASS* package [59]. The truncated data as well as the R code to analyze it can be found in the proportional odds logistic regression section in the online repository [17].

Theobald and colleagues [50] hypothesized that different kinds of group work would be more (or less) conducive to a single student dominating a group. To assess this question, they implemented two activities that required different

types of group work; they called these activities "interactive group work" and "constructive group work." After each activity, they asked students a series of questions, including to what extent each student thought a single student dominated their group. Students answered this "dominator" question on a 6-point Likert scale, from strongly disagree to strongly agree. For analysis, this scale was converted to 1 to 6 with higher numbers indicating higher agreement that there was a dominator in their group.

4. How to interpret the output of proportional odds logistic regression

Proportional odds models fit successive logistic regressions, comparing one level of the outcome to any level higher. The output is presented as a single coefficient and it represents the log odds of responding one or more levels higher on the outcome scale for every unit change in the predictor. For example, if GPA describes a survey score, the regression coefficient for GPA quantifies the log odds of a student agreeing one or more additional levels on the survey for each unit increase in GPA. Only one coefficient is reported because it is assumed that the transition between each level of the outcome is equivalent for all levels of the outcome (see below).

The output from proportional odds models does not include p values so classical hypothesis testing with p values is not straightforward. While there are ways to calculate a p value, these p values become increasingly biased as the sample size decreases [13,59]. Thus, it is instead recommended that model selection is used to test hypotheses [64]. The complete details of model selection are too extensive to cover here, but Burnham and Anderson [64] are considered an authority on the topic. Instead, we will provide a brief overview of backward selection using AIC. The key difference between hypothesis testing with model selection and hypothesis testing with p values is that when conducting model selection, each model is considered a distinct hypothesis. Thus, the relative fit of the models (as measured by AIC in this case) is how each hypothesis is tested.

When employing backward selection, the researcher fits a complex model and singularly removes parameters in subsequent models [30,64]. Models are compared using AIC and the model with the lowest AIC is selected as best fit, with the important caveat that models within 2 AIC are considered to have equivalent fit [64]. In these cases, rules of parsimony dictate that the simplest model is preferred [64]. Once the best fitting model is selected, it should be compared to a null model, without any predictor variables. The model with the fewest number of parameters with the lowest AIC is selected as best fitting and interpreted.

After employing backward selection, Theobald and colleagues [50] found that the best fitting model controlled for course grade and ethnicity and included activity type; indeed, activity type impacted the extent to which students agreed that a single student dominated their group [Fig. 2(c)]. The output of the model includes the regression coefficients ("values" Appendix 3 [47]) for each predictor variable, the standard errors, and t values for the regression coefficients. The coefficients are reported on the log odds scale, so similar to logistic and binomial regression, the coefficients should be translated to odds ratios or probabilities for interpretation and publication [Fig. 2(c); Appendix 3 [47]].

Replicating the conclusions from Theobald *et al.* [50], we see that controlling for course grade and ethnicity, students are less likely to report that one person dominated their group after completing an interactive group activity (odds ratio for the interactive activity compared to the constructivist activity $= 0.56$). The odds ratio is less than one, thus the reciprocal and the opposite relationship is reported ($1/0.56 = 1.79$): the odds that a student who worked on the constructivist activity more strongly endorsed that one person dominated their group were 1.79 that of students who worked on an interactive activity. By "more strongly endorse," we specifically mean reported one level or more higher on the question asking them if there was a dominator in their group. (Note that the size of the effect reported here differs slightly from that reported in Theobald *et al.* [50], but that is because the data were truncated here for illustrative purposes.) From the figure created with the *effects* package [Fig. 2(c)] we see that there is a higher probability that students disagree (and lower probability that they agree) that there was a dominator when they worked on the interactive activity than on the constructivist activity.

5. Assumptions and how to test the assumptions of proportional odds models

There are two key assumptions to proportional odds models: first, that the levels of the categorical outcome are ordered, and second that the relationship between each pair of levels is the same (hence the name proportional). To test whether the outcome is best modeled as ordinal, compare the model fit of the proportional odds model to the model fit of the same model that is fit as a multinomial model. A multinomial regression model models a categorical outcome that has more than two categories which are not ordered (see below for extensive details on how to fit multinomial models). This is easily done with the function `multinom` in the *nnet* package in R [62]. Once both the proportional odds and multinomial models are fit, comparing the AIC of the models will determine if the outcome is ordered (i.e., if the AIC of the `polr` model is lower) or not (i.e., if the AIC of the `multinom` model is lower).

Second, the assumption that the relationship between each pair of outcome groups is the same is an underlying assumption of proportional odds logistic regression. In other words, the model assumes that the relationship between the lowest versus all higher categories is the same as the relationship between the second lowest versus all

higher categories: the odds of choosing one level over another is proportional for each level. (Two side notes: 1) this assumption is where the name *proportional odds models* comes from; 2) it is because of this assumption that there is only one set of coefficients that gets reported when the model is fit. Otherwise, there would be a coefficient for each level of the outcome variable.) This assumption is not straightforward to test in R, although an online example [78] has been modified in the supplemental R code for proportional odds logistic regression found at github [17]. The assumption is tested by first fitting separate logistic models for each level compared to all higher levels. For example, model one would predict level one compared to all higher levels, model two would predict level two compared to all higher levels, and so forth. Then, compare the output from each logistic model to the output of the proportional odds model. The difference between the coefficients for one higher level of the outcome should be the same (i.e., roughly the same) for each level of the outcome.

Special cases.—Proportional odds logistic regression models are useful for understanding student responses to single Likert-scale questions. However, surveys often have multiple items that comprise a construct. The mean or the sum of the items in the construct is used to determine a construct score that is on a continuous scale. Using linear regression to model these scores is intuitive, but has a critical flaw. Typically with Likert-scale data, student responses approach a ceiling (or floor, although this is much less common in survey data). A ceiling effect occurs when some respondents who responded the highest value (e.g., of agreement) would have responded even higher if the scale had permitted. Conversely, a floor effect occurs when responses are bounded at the lower end of a scale. Ceiling and floor effects result from censoring: the true value is unknown because it occurs outside the range of the measurement instrument. Ceiling and floor effects (i.e., the effects of censoring) can be diagnosed by visually inspecting the residuals of a linear regression model; data appearing artificially bounded indicate a ceiling or floor effect. (See the supplement of Ref. [51] for an example.)

When a ceiling or floor effect is suspected, a censored regression should be fit [79]. A censored regression is not unique to proportional odds logistic regression, however it is particularly useful when modeling responses to survey constructs, as these types of data often experience a strong ceiling or floor [79]. A censored regression accounts for the ceiling or floor effect by actually modeling an uncensored latent outcome instead of the censored observed outcome. In this way, the estimates as well as the standard error of the estimates are more accurate. In their analysis of survey data that grouped into constructs, Wiggins and colleagues [51] fit a censored regression to confirm that their results did not qualitatively differ from the results of their linear regression. That paper, including the Supplemental Material, provide details of how and why to fit a censored regression. Censored regression models can be fit using the *censReg* package in R [61].

E. Multinomial regression

1. When to use multinomial regression

Multinomial regression is employed when the outcome variable has more than two categories that cannot be meaningfully ordered (Fig. 1). To illustrate this, consider an example from the education literature where the outcome variable is the students' choice of major [80,81]. College majors are categorical and there are more than two options, so simple logistic regression cannot be applied. In addition, majors cannot be meaningfully ranked to create levels, because one major is not necessarily better than another, thus proportional odds logistic regression is not supported. Multinomial regression was developed to address these types of outcome variables.

Multinomial regression is most commonly used in education to take a more nuanced look at student persistence (for example, beyond the simplistic binary of graduate to dropout) and other student decisions, e.g., Refs. [43–45, 82]. For example, Jones-White and colleagues [83] recognized that many students leave their home institution, but go on to graduate from a different institution. These students are in fact college graduates, even though from the lens of the home institution, they appear to have dropped out. Thus, Jones-White and colleagues [83] chose to expand the classic binary outcome for graduation success to a four level outcome variable to understand what predictors impacted a student's trajectory: (i) graduating with a Bachelor's degree at a student's initial institution, (ii) graduating with a Bachelor's degree from another institution, (iii) graduating with an associate's degree, or (iv) not graduating in six years.

Multinomial regression is also useful in other contexts. For example, Eddy and colleagues [53] used multinomial regression to explore the factors related to the role students preferred to assume during groupwork—i.e., whether students prefer to be a listener, talker, collaborator, or recorder. Similarly, Weerts and Cabrera [84] documented factors that influenced the type of civic engagement college students engaged in (superengager, apolitical-engager, social-cultural engager, and nonengager). Finally, Prevost and colleagues [85] used multinomial regression to analyze student understanding of biology concepts by moving beyond simply categorizing student responses as right or wrong and allowing for different types of incomplete understanding.

2. Why use multinomial regression

Multinomial regression is used when the outcome variable is categorical, not continuous, but when the categories cannot be ordered. Multinomial regression is

similar to logistic regression in process and interpretation: in fact, it involves the simultaneous calculation of multiple sets of logistic regression models. The simultaneous calculation of these logistic regressions has the advantage that it produces smaller standard errors around the estimates than when each model is run separately [13]. The number of logistic regressions fit in a multinomial regression depends on the number of categories in the outcome variable; specifically, the number of models is "1 − the number of categories" due to one category being selected as the reference level and each of the other levels being modeled in comparison to that reference.

3. How to fit multinomial regression in R

To illustrate how to fit a multinomial regression in R we will replicate the results in Eddy *et al.* [53]. In their paper, Eddy and colleagues [53] asked students to describe the role they preferred to assume when completing in-class groupwork. Student responses were categorized into five mutually exclusive bins: listener, talker, collaborator, recorder, and other. Here, we use a subset of their data to explore the relationship between gender, class standing, and college GPA on the roles students assume during group work. Note that college GPA is centered so that the mean GPA in the sample is 0. Code to run this analysis in R as well as the data set are provided in the multinomial regression section in the online repository [17].

The first step of multinomial regression is selecting the reference level for the outcome variable. The regression coefficients from a multinomial regression are always relative: specifically, the coefficient describes the log odds of being in each other category relative to a reference category. The reference outcome category will be the same across all the individual logistic regression models. The choice of reference level is up to the researcher, but it is best to set the reference level to something that creates "clinically meaningful" odds ratios [44]. If there is one category that is the most important to the research to understand why participants are in or out of it, then set that as the reference level. In the case of group work, theory from cognitive science suggests that the collaborator role is the ideal role for students to assume during group work, because groups where all participants listen and speak are most likely to reach interactive engagement which can lead to the deepest learning [86]. Thus, all comparisons to the collaborator role will be clinically meaningful, so this is designated as the reference.

Next, we specify the model and run a multinomial regression. In R, multinomial regression can be implemented with the `multinom` function in the *nnet* package [62], as shown in Table II. In this example, the multinomial regression involves simultaneously running four logistic regressions with the following contrasts of outcome categories: Leader versus collaborator, listener versus collaborator, recorder versus collaborator, other versus collaborator. The initial model includes interactions between gender and class standing and gender and college GPA to test whether the influence of these two variables on the role students assume in groups is different for students of different genders.

Once the initial model is identified, we can evaluate which predictor variables contribute to a better prediction of the outcome. Like logistic regression and proportional odds logistic regression, multinomial regression models the log odds of being in a particular category of the outcome variable relative to being in the reference category. What complicates interpretation of predictors is that for each predictor variable, R reports a different regression coefficient, standard error, and p-value based on the z statistic from the Wald test for each comparison (i.e., for each logistic regression model). In this way, the researcher models a full picture of the relationship between the predictors and the possible categories, but it can be challenging to interpret. To evaluate the contribution of the predictors across all the simultaneous logistic regressions, Hosmer and colleagues [44] recommend comparing multinomial models with and without each predictor variable using a likelihood ratio test (as opposed to the Wald tests). This approach is necessary for two big reasons: First, the p values derived from the Wald test statistic are specific to each logistic regression, and the Wald statistic varies in significance depending on the outcome contrast being modeled by the logistic regression (collaborator versus talker or collaborator versus listener)—together, this makes the p values unuseful as a global measure of the importance of a predictor [44]. For example, college GPA might be a significant predictor for the odds of preferring to be a leader relative to a collaborator, but not a listener relative to a collaborator, making it difficult to assess how important college GPA is in predicting which role a student assumes. In contrast, a likelihood ratio test is calculated for the predictor as a whole and across all the individual logistic regressions, thus globally assesses each predictor. Second, the Wald statistic is generally less reliable than the likelihood ratio test with small sample sizes [13].

To conduct a likelihood ratio test to evaluate the contribution of each predictor variable on the role a student assumes, we test the contribution of each term to the model by comparing models with and without that term, using backwards selection. In this example we have five possible predictors: gender, college GPA at the start of the course, class standing, gender×GPA, and class standing×GPA. As is standard practice with backward selection [30,64], we start by testing the inclusion of each interaction by comparing models, using likelihood ratio tests, with and without the interaction. Next, we test the main effects of gender, college GPA, and class standing. Appendix 4 [47], documents the outcomes of the likelihood ratio tests. We do not see substantial change in the fit of the model by

including an interaction term for gender and class standing ($p = 0.195$) or class standing as a main effect ($p = 0.736$). There is a loss of fit when gender×college GPA is removed ($p = 0.048$). Thus, model selection via likelihood ratio tests suggests the critical variables to consider when describing the role students prefer (from among the initial variables listed) are gender, college GPA, and a gender×college GPA interaction.

4. How to interpret the output of multinomial regression

The regression coefficients in log odds of the final model are represented in Appendix 4 [47]. For each comparison of outcome level to collaborator (i.e., each row in the table) there is a different regression coefficient for each predictor variable. The p value in each cell is from a Wald test, which tests whether the coefficient is significantly different from zero. We can see from the third column in the table that the coefficient for gender is significant in the comparison of the talker versus collaborator role, but does not significantly influence the log odds of not being a collaborator in any of the other comparisons.

To develop a sense of the magnitude of the gender impact we convert the log odds to an odds ratio by exponentiating the coefficient for gender ($e^{1.4}$). Here, the significant main effect of gender indicates that the odds of males with the average university GPA preferring the role of talker over collaborator are 4.1 times that of a similar female preferring the role of talker over collaborator. But, this is not the whole story: there is also a significant interaction between gender and GPA. Understanding the impact of this interaction is substantially easier using the predicted probabilities of the *effects* package. The probabilities from the *effects* package are global, meaning that they incorporate the results from all the comparisons. In this example, as GPA increases above the mean, men become less likely to prefer being collaborators and more likely to prefer being a leader. Women with the same GPA, on the other hand, become more likely to be collaborators [Fig. 2(d), Appendix 4 [47]].

The only other regression coefficient that the Wald test indicates is significantly different than zero, is college GPA on the odds ratio of being in the category of other versus collaborator. With a one-point increase in GPA, the odds that a student prefers being a collaborator are 7.6 times ($1/e^{-2}$) that of a student preferring assuming an "other" role.

Odds ratios can still be confusing because the interpretation is always relative to the reference level. The *effects* package can provide additional clarity by converting the odds ratios produced by multinomial regression into probability (percent chance) of being in each outcome category. This measure is not relative and is much easier to compare. Figure 2(d) and Appendix 4 [47] represent one way to summarize these relationships for a publication.

5. Assumptions and how to test the assumptions of multinomial regression

Multinomial regression relies on several assumptions. First, the data have to be case specific; i.e., each individual can only be in one outcome category not several. This means that this type of analysis may work better with forced choice type questions versus questions where students are allowed to check as many categories as they want. Alternatively, the researchers have to look at student choices and bin them into single categories. The second assumption is the independence of irrelevant alternatives. This assumption states that adding other outcome categories to the analysis will not change the relationship between the other outcome variables [83]. In our example this would mean that if we added a 6th possible group role, there would be no impact on the log odds of being a leader versus collaborator. If a new outcome could impact the relationships between the existing outcome categories, then this assumption is violated and coefficients may be miscalculated [87]. When researchers perceive this assumption could be violated, it is recommended to use the probit linking function in the multinomial regression, although, this has not been emphasized. Multinomial probit models can be run in R using the *MNP* package [88].

Multinomial regression is a useful tool when an outcome variable is categorical with more than two levels, but it has several limitations. First, because so many tests are being calculated and these tests require the use of maximum likelihood estimates, multinomial requires a larger sample size than ordinal or logistic regression [44]. Second, even though there is no limit to how many categories the outcome variable can have, interpretations become increasingly difficult with more outcomes. Thus, it is wise to try to keep the number of possible outcomes small.

F. Poisson regression

1. When to use Poisson regression

As with binomial regression, the outcome of Poisson regression models are count data. However, it differs in that the count data must be unbounded, i.e., not a proportion of some specified number of occurrences (Fig. 1; Ref. [67]). In discipline-based education research, Poisson regression might be used to determine how student, classroom, or program characteristics influence the number of students from a particular program who persist in a discipline [89] or the number of undergraduates completing research projects within a department [90]. Auerbach and colleagues [54] used Poisson regression to model the number of times an instructional practice, such as promoting metacognition, was noticed by experts compared to novices as they analyzed videos of active-learning classrooms. Andrews and colleagues [55] examined how faculty gender, academic rank, DBER or not DBER academic position, and participation in teaching professional development influenced change of teaching practices. They measured change

by counting the number of times a faculty member was reported in a survey as the cause of a change, thus necessitating a Poisson model of their count data.

2. Why use Poisson regression

Linear regression is not appropriate for count data for two reasons. First, a linear regression model may predict values that are negative, which is beyond the range of possible values for count data. Second, count data are generally not normally distributed, but rather are right-skewed, and thus better modeled with a Poisson distribution. It is especially important to use a Poisson regression model when the mean of the outcome variable is low (below 10) as the differences in the Poisson and normal distributions typically become distinct within this range [91]. Poisson regression accounts for these violations by log-transforming the outcome variable, which no longer bounds the data at 0 [13].

3. How to fit Poisson regression in R

Poisson regression is fit in `R` using the `glm` function (`family=Poisson`) in the *base* package [46], as shown in Table II. To illustrate the implementation of Poisson regression, we created a sample data set loosely based on ecological data that had count outcomes [92]. We use this toy data set to "test" the hypotheses that (i) students' major (physics or not) and (ii) course exam performance affect the number of times a student raises their hand in the classroom (outcome variable). The data and R code can be found in the Poisson example in the github repository [17].

4. How to interpret the output of Poisson regression

Similar to the output from the logistic and binomial regression models, the output from a Poisson regression includes regression coefficients, their standard errors, and p values, based on a z statistic, for each of the coefficients to determine if each is significantly different than zero (Appendix 5 [47]). However, unlike the output on a log odds scale, the output of a Poisson regression is on the log scale. Exponentiating the coefficients transforms the output to be a multiplicative effect as we demonstrate below. For more details, see Gelman and Hill [23].

In our hand-raising example, the number of times a student raises their hand can be predicted by whether or not a student is a physics major ($\beta = -0.872$, $p < 0.0001$; Appendix 5 [47]), as well as their total exam points ($\beta = 0.004$, $p < 0.01$; Appendix 5 [47]). To facilitate interpretation of these coefficients, we exponentiate them (e^{β}; where β is the coefficient) and interpret them as multiplicative effects. For example, controlling for the total exam points received by a student, the expected number of hand raises for a nonphysics major is $e^{(-0.87)} = 0.42$ times the expected number of hand-raises for a physics major [Fig. 2(e)]. In other words, the expected number of hand raises in class is reduced by 58% for nonphysics majors compared to physics majors. Furthermore, when controlling for major status, every 1-point increase in exam points translates into a 0.4% increase in the expected number of hand raises in class for physics majors [$e^{(0.004)} = 1.004$; Fig. 2(e)].

In addition to the standard regression output, the AIC value of the model is also reported. AIC can be helpful in model selection and can be used to determine if a special case of Poisson regression (as described below) better explains the data.

5. Assumptions and testing the assumptions of Poisson regression

Poisson regression models rely on several assumptions. The first, and most fundamental assumption, is that the variance in the outcome is equal to the mean of the outcome. The easiest way to test this assumption is by summarizing the data [i.e., `summary(data$outcome)`]. With count data it is common for the variance to be greater than the mean, a phenomenon known as overdispersion. In the code online [17], we detail how to determine whether overdispersion is occurring in a data set. We first fit a negative binomial model (not to be confused with a binomial regression) with the *MASS* package [59]. These models are commonly used to account for overdispersed count data. Second, we test for overdispersion with `odTest` in the *pscl* package [60]. With both of these tests, we find that our data are highly overdispersed.

Another assumption of Poisson regression that is commonly violated includes having an overabundance of zeros in the data set (i.e., the case when many subjects never display the behavior you are recording). In this case, the data can be fit with either a zero-inflated Poisson or Hurdle model, both of which can be run in R using the *pscl* package [56,60]. The zero-inflated model is most appropriate for our hand-raising example as it is used when there is more than one reason we might expect a value of zero to be observed. In our case, students may choose not to raise their hand, or instead, they may be absent from class thus unable to raise their hand. More details describing these special cases can be found in Appendix 6 [47].

V. CONCLUSIONS

Regression models are essential in discipline-based education research [2] in that they both allow for rigorously testing hypotheses and control for differences among students in quasirandom experimental designs. However, linear regression models have several assumptions that are violated with various types of outcome data in education. Thus, researchers must move beyond linear regression and consider generalized linear models, or glms, for appropriately analyzing their data.

In this paper we provide a diagnostic tool for identifying the most appropriate regression model (Fig. 1) and describe

how to fit and interpret five glm types in R. For readers who want more details, there are many great resources on how to analyze data in a regression framework and an equal number of resources on how to implement regression in R. We have cited Fox and Weisberg [24] and Gelman and Hill [23] extensively throughout this paper. Additionally, there are other resources which may be helpful for specific kinds of analyses, including: Hosmer *et al.* [44] for logistic, ordinal, and multinomial regressions; Agresti [13] for logistic, ordinal, and multinomial regressions; and Faraway [67] for logistic, binomial, and Poisson regressions.

ACKNOWLEDGMENTS

The authors would like to thank several people and groups for their support and improvement of this paper: the authors who graciously shared their data for the extended examples, Scott Freeman for the financial support and encouragement, and A. Clemmons, M. Mack, R. Theobald, and members of the Biology Education Research Group for their friendly reviews. Funding for the project came from a grant from the University of Washington, Office of the Dean of Arts and Sciences.

[1] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).

[2] R. Theobald and S. Freeman, Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research, CBE Life Sci. Educ. **13**, 41 (2014).

[3] J. Stewart, G. Stewart, and J. Taylor, Using time-on-task measurements to understand student performance in a physics class: A four-year study, Phys. Rev. ST Phys. Educ. Res. **8**, 010114 (2012).

[4] N. Hall and D. Webb, Instructors' support of student autonomy in an introductory physics course, Phys. Rev. ST Phys. Educ. Res. **10**, 020116 (2014).

[5] A. L. Rudolph, B. Lamine, M. Joyce, H. Vignolles, and D. Consiglio, Introduction of interactive learning into French university physics classrooms, Phys. Rev. ST Phys. Educ. Res. **10**, 010103 (2014).

[6] P. R. L. Heron, Effect of lecture instruction on student performance on qualitative questions, Phys. Rev. ST Phys. Educ. Res. **11**, 010102 (2015).

[7] C. S. Stevens, M. Marder, and S. R. Nagel, Patterns in Illinois educational school data, Phys. Rev. ST Phys. Educ. Res. **11**, 010113 (2015).

[8] B. D. Mikula and A. F. Heckler, Framework and implementation for improving physics essential skills via computer-based practice: Vector math, Phys. Rev. Phys. Educ. Res. **13**, 010122 (2017).

[9] G. Potvin and Z. Hazari, Student evaluations of physics teachers: On the stability and persistence of gender bias, Phys. Rev. Phys. Educ. Res. **12**, 020107 (2016).

[10] W. G. Cochran, Some methods for strengthening the common $\chi 2$ tests, Biometrics **10**, 417 (1954).

[11] D. R. Cox, The regression analysis of binary sequences, J. R. Stat. Soc. Ser. B **20**, 215 (1958).

[12] A. J. A. Nelder and R. W. M. Wedderburn, Generalized linear models, J. R. Stat. Soc. Ser. A **135**, 370 (1972).

[13] A. Agresti, *Categorical Data Analysis*, 3rd ed. (John Wiley & Sons, Hoboken, 2013).

[14] P. McCullagh and J. A. Nelder, *Generalized Linear Models* (Chapman and Hall, London, 1983).

[15] W. Gardner, E. P. Mulvey, and E. C. Shaw, Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models, Psychol. Bull. **118**, 392 (1995).

[16] J. Gill, *Generalized linear models: A unified approach*, Sage Unive (Sage, Thousand Oaks, CA, 2001).

[17] https://github.com/ejtheobald/BeyondLinearRegression.

[18] D. Z. Grunspan, B. L. Wiggins, and S. M. Goodreau, Understanding classrooms through social network analysis: A primer for social network analysis in education research, CBE Life Sci. Educ. **13**, 167 (2014).

[19] E. Knekta, C. Runyon, and S. L. Eddy, One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research, CBE Life Sci. Educ. **18**, ar1 (2019).

[20] W. J. Boone, Rasch analysis for instrument development: Why, when, and how?, CBE Life Sci. Educ. **15**, ar51 (2016).

[21] P. Martinková, A. Drabinová, Y. L. Liaw, E. A. Sanders, J. L. McFarland, and R. M. Price, Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments, CBE Life Sci. Educ. **16**, ar19 (2017).

[22] A. R. M. Warfa, Mixed-methods design in biology education research: Approach and uses, CBE Life Sci. Educ. **15**, ar51 (2016).

[23] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Heirarchical Models* (Cambridge University Press, New York, 2007).

[24] J. Fox and S. Weisberg, *A R Companion to Applied Regression*, 2nd ed. (Sage, Thousand Oaks, CA, 2011).

[25] UCLA: Statistical Consulting Group, Introduction to R, 2018.

[26] D. S. Moore and W. I. Notz, *Statistics: Concepts and Controversies*, 9th ed. (W.H. Freeman and Company, New York, 2017).

[27] http://tryr.codeschool.com.

[28] http://stat545.com/.

[29] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R* (Sage, Thousand Oaks, CA, 2012).

[30] E. Theobald, Students are rarely independent: When, why, and how to use random effects in discipline-based education research, CBE Life Sci. Educ. **17**, rm2 (2018).

[31] J. L. Kobrin, S. Sinharay, S. J. Haberman, and M. Chajewski, An investigation of the fit of linear regression models to data from an SAT validity study, College board Research Report 2011-3; ETS Research Report RR-11-19, 2011, https://eric.ed.gov/?id=ED521177.

[32] J. W. Osborne, Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses, Pract. Assess. Res. Eval. **11** (2006); http://pareonline.net/getvn.asp?v=11&n=7.

[33] W. L. Holcomb, T. Chaiworapongsa, D. A. Luke, and K. D. Burgdorf, Obstet. Gynecol. **98**, 685 (2001).

[34] J. Fox and S. Weisberg, Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals, J. Stat. Software **89**, 1 (2018).

[35] D. Lüdecke, sjPlot: Data Visualization for Statistics in Social Science, http://dx.doi.org/10.5281/zenodo.1308157 (2018).

[36] J. Fox, Effect displays in R for generalised linear models, J. Stat. Software **8**, 1 (2003).

[37] J. Fox and J. Hong, Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package, J. Stat. Software **32**, 1 (2009).

[38] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).

[39] S. L. Eddy and K. A. Hogan, Getting under the hood: How and for whom does increasing course structure work?, CBE Life Sci. Educ. **13**, 453 (2014).

[40] L. J. Sax, K. J. Lehman, R. S. Barthelemy, and G. Lim, Women in physics: A comparison to science, technology, engineering, and math education over four decades, Phys. Rev. Phys. Educ. Res. **12**, 020108 (2016).

[41] R. H. Tai, X. Kong, C. E. Mitchell, K. P. Dabney, D. M. Read, D. B. Jeffe, D. A. Andriole, and H. D. Wathington, Examining summer laboratory research apprenticeships for high school Students as a factor in entry to MD/PhD programs at matriculation, CBE Life Sci. Educ. **16**, ar37 (2017).

[42] K. I. Maton, T. S. Beason, S. Godsay, M. R. Mariano, T. S. C. Bailey, S. Sun, and F. A. Hrabowski, Outcomes and processes in the meyerhoff scholars program: STEM PhD completion, sense of community, perceived program benefit, science identity, and research self-efficacy, CBE Life Sci. Educ. **15**, ar48 (2016).

[43] A. DeMaris, Logistic regression: Basic foundations and new directions, in *Handbook of Psychology Vol. 2: Research Methods in Psychology*, edited by J. A. Schinka, W. F. Velicer, and I. B. Weiner (John Wiley & Sons, Hoboken, NJ, 2013), pp. 543–570.

[44] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. (John Wiley & Sons, Hoboken, 2013).

[45] D. Collett, *Modeling Binary Data*, 2nd ed. (Chapman and Hall, New York, 2003).

[46] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

[47] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.020110 for annotated output for logistic regression, binomial regression, proportional odds logistic regression, multinomial regression, and Poisson regression, as well as descriptions of special cases of Poisson regression.

[48] R. L. Forrest, D. W. Stokes, A. B. Burridge, and C. D. Voight, Math remediation intervention for student success in the algebra-based introductory physics course, Phys. Rev. Phys. Educ. Res. **13**, 020137 (2017).

[49] C. D. Desjardins, Modeling zero-inflated and overdispersed count data: An empirical study of school suspensions, J. Exp. Educ. **84**, 449 (2016).

[50] E. J. Theobald, S. L. Eddy, D. Z. Grunspan, B. L. Wiggins, and A. J. Crowe, Student perception of group dynamics predicts individual performance: Comfort and equity matter, PLoS One **12**, e0181336 (2017).

[51] B. L. Wiggins, S. L. Eddy, L. Wener-Fligner, K. Freisem, D. Z. Grunspan, E. J. Theobald, J. Timbrook, and A. J. Crowe, ASPECT: A survey to assess student perspective of engagement in an active-learning classroom, CBE Life Sci. Educ. **16** (2017).

[52] R. Ivie, S. White, and R. Y. Chu, Women's and men's career choices in astronomy and astrophysics, Phys. Rev. Phys. Educ. Res. **12**, 020109 (2016).

[53] S. L. Eddy, S. E. Brownell, P. Thummaphan, M.-C. Lan, and M. P. Wenderoth, Caution, student experience may vary: Social identities impact a student's experience in peer discussions, CBE Life Sci. Educ. **14**, ar45 (2015).

[54] A. J. Auerbach, M. Higgins, P. Brickman, and T. C. Andrews, Teacher knowledge for active-learning instruction: Expert–novice comparison reveals differences, CBE Life Sci. Educ. **17**, ar12 (2018).

[55] T. C. Andrews, E. P. Conaway, J. Zhao, and E. L. Dolan, Colleagues as change agents: How department networks and opinion leaders influence teaching at a single research university, CBE Life Sci. Educ. **15**, ar15 (2016).

[56] A. Zeileis, C. Kleiber, and S. Jackman, Regression Models for Count Data in R, J. Stat. Software **27** (2008).

[57] A. Baccini, L. Barabesi, M. Cioni, and C. Pisani, Crossing the hurdle: The determinants of individual scientific performance, Scientometrics **101**, 2035 (2014).

[58] C. F. Kot, The impact of centralized advising on first-year academic performance and second-year enrollment behavior, Res. High. Educ. **55**, 527 (2014).

[59] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S. Fourth Edition* (Springer, New York, 2002), ISBN 0-387-95457-0.

[60] S. Jackman, pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. United States Studies Centre, University of Sydney. Sydney, New South Wales, Australia. R package version 1.5.2, https://github.com/atahk/pscl/.

[61] A. Henningsen, censReg: Censored Regression (Tobit) Models. R package version 0.5-26, https://CRAN.R-project.org/package=censReg.

[62] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S. Fourth Edition* (Springer, New York, 2002), ISBN 0-387-95457-0.

[63] S. E. Andrews and M. L. Aikens, Life Science Majors' Math-Biology Task Values Relate to Student Characteristics and Predict the Likelihood of Taking Quantitative Biology Courses, J. Microbiol. Biol. Educ. **19**, 1 (2018).

[64] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. (Springer, New York, 2002).

[65] H. T. O. Davies, I. K. Crombie, and M. Tavakoli, When can odds ratios mislead?, Br. Med. J. **316**, 989 (1998).

[66] F. Cribari-Neto and A. Zeileis, Beta regression in R, J. Stat. Software **34**, 1 (2010).

[67] J. J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (CRC Press, Boca Raton, 2016).

[68] C. J. Peng, K. U. K. L. Lee, and G. M. Ingersoll, An introduction to logistic regression analysis and reporting, J. Educ. Res. **96** (2002).

[69] M. A. Jackson, A. Tran, M. P. Wenderoth, and J. H. Doherty, Peer vs. self-grading of practice exams: Which is better?, CBE Life Sci. Educ. **17**, es44 (2018).

[70] F. Cribari-Neto and A. Zeileis, Beta Regression in R, J. Stat. Software **34**, 1 (2016).

[71] W. Venables and B. Ripley, *Modern Applied Statistics with S* (Springer, New York, 2002).

[72] https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#overdispersion.

[73] A. Agresti and J. B. Lang, A proportional odds model with subject-specific effects for repeated ordered categorical responses, Biometrika **80**, 527 (1993).

[74] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[75] E. J. Theobald, A. Crowe, J. Hillerislambers, M. P. Wenderoth, and S. Freeman, Front. Ecol. Environ. **13**, 132 (2015).

[76] B. L. Wiggins, S. L. Eddy, D. Z. Grunspan, and A. J. Crowe, AERA Open **3**, 1 (2017).

[77] R. H. B. Christensen, ordinal—Regression Models for Ordinal Data. R package version 2019.3-9, http://www.cran.r-project.org/package=ordinal/.

[78] https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/.

[79] A. Henningsen, Estimating censored regression models in R using the censReg Package, R package vignettes **5**, 12 (2010).

[80] M. Pinxten, B. De Fraine, W. Van Den Noortgate, J. Van Damme, T. Boonen, and G. Vanlaar, 'I choose so I am': a logistic analysis of major selection in university and successful completion of the first year, Stud. Higher Educ. **40**, 1919 (2015).

[81] J. Jorstad, S. S. Starobin, Y. Chen, and A. Kollasch, STEM aspiration: The influence of social capital and chilly climate on female community college students, Community Coll. J. Res. Pract. **41**, 253 (2017).

[82] S. G. A. van Herpen, M. Meeuwisse, W. H. A. Hofman, S. E. Severiens, and L. R. Arends, Early predictors of first-year academic success at university: pre-university effort, pre-university self-efficacy, and pre-university reasons for attending university, Educ. Res. Eval. **23**, 52 (2017).

[83] D. R. Jones-White, P. M. Radcliffe, R. L. Huesman, and J. P. Kellogg, Redefining student success: Applying different multinomial regression techniques for the study of student graduation across institutions of higher education, Res. High. Educ. **51**, 154 (2010).

[84] D. J. Weerts and A. F. Cabrera, Understanding civic identity in college, J. Coll. Character **16**, 22 (2015).

[85] L. B. Prevost, M. K. Smith, and J. K. Knight, Using student writing and lexical analysis to reveal student thinking about the role of stop codons in the central dogma, CBE Life Sci. Educ. **15**, ar65 (2016).

[86] M. T. H. Chi and R. Wylie, The ICAP framework: Linking cognitive engagement to active learning outcomes, Educ. Psychol. **49**, 219 (2014).

[87] S. Washington, M. G. Karlaftis, and F. L. Mannering, *Statistical and Econometric Methods for Transportation Data Analysis* (CRC Press, Boca Raton, 2003).

[88] K. Imai and D. van Dyk, MNP: R Package for Fitting the Multinomial Probit Model. R package version 3.1-0, https://CRAN.R-project.org/package=MNP.

[89] K. Richardson, A. Tarr, S. Miller, N. Sibanda, L. Richardson, and K. Mikaere, in *Māori Pasifika High. Educ. Horizons* (Emerald Group Publishing Limited, Bingley, UK, 2014), pp. 179–200.

[90] B. Mellis, P. Soto, C. D. Bruce, G. Lacueva, A. M. Wilson, and R. Jayasekare, Factors affecting the number and type of student research products for chemistry and physics students at primarily undergraduate institutions: A case study, PLoS One **13**, e0196338 (2018).

[91] S. Coxe, S. G. West, and L. S. Aiken, The analysis of count data: A gentle introduction to poisson regression and its alternatives, J. Pers. Assess. **91**, 121 (2009).

[92] E. J. Theobald, H. Gabrielyan, and J. HilleRisLambers, Lilies at the limit: Variation in plant-pollinator interactions across an elevational range, Am. J. Bot. **103**, 189 (2016).