Tomkin and West
International Journal of STEM Education (2022) 9:27
https://doi.org/10.1186/s40594-022-00343-1

International Journal of STEM Education

RESEARCH Open Access

STEM courses are harder: evaluating inter-course grading disparities with a calibrated GPA model

Jonathan H. Tomkin[1*] and Matthew West[2]

Abstract

Background: Grades in college and university STEM courses are an important determinant of student persistence in STEM fields. Recent studies have used the grade offset/grade penalty method to explore why students have lower grades in STEM courses than their GPAs would predict. The results of these studies are in doubt; however, as they use GPA as a reliable measure of academic performance, which is a disputed assumption. Using a predictive model of student performance, it is possible to produce a more accurate measure of academic performance than the observed GPA and discover if STEM courses are graded more stringently, and under which circumstances.

Results: A weighted logistic model of GPA better predicts academic performance than the observed GPA. Using this calibrated GPA it is found that the grade offset method predicts that STEM courses, departments, and programs grade significantly more stringently than non-STEM courses. The average grade difference between STEM and non-STEM course grades and GPAs is around four tenths of a grade point. An exception is general education courses offered by STEM departments, which are graded with the same leniency as non-STEM courses. Grade offset calculations that use the observed GPA systematically underestimate the negative offset in STEM grading relative to calculations that use the calibrated GPA. The calibrated GPA is much more highly correlated with standardized tests such as the ACT ($r = 0.49$) than the observed GPA is ($r = 0.25$).

Conclusion: Observed GPA is a systematically biased measure of academic performance, and should not be used as a basis for determining the presence of grading inequity. Logistic models of GPA provide a more reliable measure of academic performance. When comparing otherwise academically similar students, we find that STEM students have substantially lower grades and GPAs, and that this is the consequence of harder (more stringent) grading in STEM courses.

Keywords: STEM education, Standardized tests, GPA, Grade offset, Grade penalty

Introduction

Grades matter. Good grades in college can lead to course credit, academic scholarships, access to graduate degree programs, and, of course, the awarding of a degree (Cohen, 2000; Rosovsky & Hartley, 2002). Poor grades might mean no course credit, the removal from courses of study, and lowered employment prospects. Furthermore, since grades have large persistence effects (Ost, 2010; Rask, 2010; Stinebrickner & Stinebrickner, 2011) it is reasonable to wonder if unfair grading might be a contributor to systematic group differences in STEM enrollment (Cromley et al., 2016). We might ask ourselves if grades in STEM courses are biased in some way, and, if so, how that bias impacts students.

How do we know if (and how) a course is graded differently than other courses? Ordinarily, we might compare the grading in different courses by looking at the average

*Correspondence: tomkin@illinois.edu
[1] School of Earth, Society, and the Environment, University of Illinois at Urbana-Champaign, 1301 W. Green St., Urbana, IL 61801, USA
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

grade of the course. However, this neglects the usual grade of the students in the course—the grades in the course might reflect the type of student who takes it. One way around this is to determine the difference between the grade students average in a particular course versus their average grade in other courses. This difference is known variously as the "grade offset" (Vanderbei et al., 2014) and the "grade penalty" (Koester et al., 2016; Matz et al., 2017), and is a new way to quantitatively compare grading. A course in which most students do worse than their GPA (Grade Point Average) has a negative offset—and is "harder" (or at any rate graded more stringently) than other courses. If, for example, "B" students (overall GPA = 3.0) have an average grade of B − (score equivalent to 2.67) in a given course, then that course has a grade offset of − 0.33 and, equivalently, a grade penalty of 0.33. Conversely, courses in which most students do better than their GPA are "easier"—they have positive grade offsets and (somewhat confusingly) negative grade penalties. For clarity, we use the term "grade offset" here (positive offsets mean "easier" and negative offsets mean "harder").

The grade offset does not, of course, capture many important aspects of courses (including how much is taught and learned!) but it does directly address the question of whether or not there is grading disparity between courses. In our work we include all courses (including the examined one) when calculating grade offsets, which will tend to moderate the size of the offset. This enables us to calculate a single calibrated GPA value for each student.

Unfortunately, the grade offset method has a built-in flaw: GPA and course grades are not independent. The grade offset is the difference between the course grade and the GPA, but the GPA itself is the sum of all course grades. This biases grade offsets towards zero. In the extreme case, where students take only one course, there are no grade offsets recorded at all (as the course grade is equal to the GPA). Taking more courses doesn't solve this problem if those courses have grade offsets that are systematically related.

Tomkin et al. (2016) illustrated this with an example (updated here to use grade offsets):

> *Imagine a student who, if she took all courses at an institution, averaged a "B"; her GPA should be 3.0. But no student takes all courses at an institution. If a "B" student exclusively enrolled in grade-penalizing courses that have an average grade offset of −0.33, she would have an observed GPA of 2.67 — not 3.0, as specified. This in turn would impact the observed grade offset of these lower-grade courses: this student records no grade offset for taking these courses. Her observed GPA is 2.67, her average grade in these courses is 2.67, and so the apparent grade offset is zero — even though the actual grade offset was defined as −0.33.*

As this example shows, GPA itself is a biased measure of academic ability, and that the bias arises from the grade offset that we are trying to measure.

Fortunately, there is a solution: construct an unbiased measure of student academic performance. If students took every single course in the sample then we would no longer have course selection effects, and GPAs would be a stable measure of student achievement. Although we can't make students take all the same courses, we can extrapolate from the courses that they have taken to predict their performance (Tomkin et al., 2016, 2018). This method uses all of the observed grades (from all students in all courses) to predict how each student would do in courses that they did not take. Since every student now has an assigned grade for every course in the sample, we can now give each student a GPA (a "calibrated GPA") that is free of course selection bias. We can then use this calibrated GPA and the observed course grades to calculate a grade offset that removes course selection effects.

In this paper, we make a methodological advance on this method by weighting by size of courses in the calculation of the calibrated GPA. We show that the new weighted calibrated GPA is a significantly better predictor of student performance than the "observed" GPA (that is, the transcript GPA that is calculated from course grades), and has lower error and is better calibrated than previous modeling approaches. Importantly, the calibrated GPA produces large, and systematically different, grade offset predictions than approaches that use the observed GPA. The change to the grade offset is significant when used to examine individual courses, categories of courses, programs of study, and groups of students. We determine that any grade offset method that uses observed GPA without correction is systematically biased. By both describing the improved method and including the code for its implementation we hope to encourage the research community to move to accurate methods.

To this end we investigate five Research Questions. In each case, we are interested in determining whether these hypotheses are supported, and, perhaps more importantly, the magnitude and significance of any effect.

1 Is a weighted logistic model of GPA better than the observed GPA in predicting academic performance?
2 Are STEM majors graded more stringently than non-STEM majors?
3 Are STEM general education courses graded less stringently than STEM major courses?

4 Are gateway courses in STEM graded less stringently that STEM major courses?
5 To what extent do standardized test scores predict academic performance?

These research questions are specified in the "Methods". We note here that academic performance is measured with course grades, we specify stringency as the relative grade offset, and that the standardized test scores referred to are the ACT and SAT.

Background

It has been a long-standing finding that STEM and non-STEM grades (and thus GPAs) are not equivalent (Goldman & Widawski, 1976). There are a number of studies that seek to improve upon the validity of observed grades using statistical methods. Using student data from Carnegie Mellon University, Caulkins et al. (1996) found that incorporating a course difficulty adjustment into grades produces a GPA that better correlated with high school GPA and standardized tests. Comparing item-response and Bayesian approaches, Johnson et al. (1997) used a student ranking approach to adjust grades, and also found adjusted grades to be more predictive than observed grades. Johnson went on to write a monograph on grade adjustment methods (Johnson, 2003), concluding that they are superior to the current approach and that the heterogeneity in current grading practices undermines academic standards and the assessment of student learning. More recently, Vanderbei et al. (2014) regressed grade data from Princeton University to calculate course offsets so as to create a model of student aptitude. They also found smaller predictive errors when using these adjusted grades.

There have been several recent studies using grade offsets (or, equivalently, grade penalties) to examine grading disparities between students in STEM programs of study. These studies used the observed GPA to calculate the course grade penalty, and then used additional covariates (such as standardized tests scores and programs of study) to determine differences in group outcomes. Koester et al. (2016) used grade point penalties to compare the performance of male and female students in large courses at the University of Michigan. They found that female students had larger grade point penalties than male students in STEM lecture courses, but not in lab courses. A follow-up study performed across five large research universities (Matz et al., 2017) found the same result. They called for interventions to reduce the material and statistically significant differences between the STEM grade penalties for male and female students. Witteveen and Attewell (2020) compared the grading penalty between STEM and non-STEM courses, They also found that STEM courses have higher average grade penalties, with grades between 0.25 and 0.4 points lower (on a 4.0 scale) in STEM courses, and that women had higher STEM grade penalties, but that this did not impact women's graduation rates.

Another group (Tomkin et al., 2016, 2018) studied a similar data set (from the University of Illinois) but came to a different conclusion. Women appeared to have larger STEM penalties than men when observed GPA was used to calculate the penalty. When a logistic grade model was used (to account for gender heterogeneity in course choice), they found that gender did not predict grade penalties in STEM courses: women and men had similar STEM grade penalties. They described the spurious relationship observed by other workers as an example of "Simpson's Paradox".

Predictive models of student grades have been used previously by multiple authors to estimate true student aptitude. Vanderbei et al. (2014) used two-parameter linear models of student grades for this estimation and then used these models to quantify course grade inflation. Tomkin et al., (2016, 2018) used two-parameter logistic models in a similar way and showed that two-parameter logistic models had slightly better performance on real-world data than two-parameter linear models. Logistic models are widely used in other psychometrics and educational settings, including item response or latent trait models (Nering & Ostini, 2010), and they are special cases of generalized additive models (Hastie et al., 2009). For the purpose of modeling student grades, logistic models have the advantage of always predicting grades that fall within the grading scale (i.e., they cannot predict negative grades or grades above a 4.0).

Methods

Student grade data and observed GPA

Our data set consists of 64,860 students, 3606 courses, and a total of 1,984,111 student grade records from the College of Engineering and the College of Liberal Arts and Sciences at the University of Illinois at Urbana-Champaign over a period of 10 years (2006 to 2015 inclusive). These two colleges were chosen as they contain a well-balanced population of both STEM and non-STEM students with substantial rates of cross-college and cross-department course selection. The sample consists of 35,034 STEM majors and 29,826 non-STEM majors. The students had grades from an average of 30.6 courses. STEM majors took 8.2 (or 27.7%) non-STEM courses, and non-STEM majors took 5.5 (or 18.4%) STEM courses. Courses had an average of 550.2 students over the 10-year period. The data set includes only those courses with a total enrollment of at

least 30 students over the 10-year period, and students with at least 10 courses, to ensure model identifiability.

To define grade point averages mathematically, we denote by $N=1{,}984{,}111$ the number of enrollment records, where record (i_n, k_n, g_n) indicates that student i_n took course k_n and received grade g_n, for $n=1,...,N$. It is possible that the same student took a given course multiple times and received either the same or different grades each time. Different offerings of a course, either in the same term or in different terms, are considered to be the same course. There are a total of $I=64{,}860$ students and $K=3606$ courses. Grades are measured on a standard four-point scale with $g=0.0$ being the lowest grade (F) and $g=4.0$ being the highest grade (A or A+). We denote by K_i the number of courses records for student i, so that the observed GPA is

$$\text{observedGPA}_i = \frac{1}{K_i} \sum_{\substack{n=1,\ldots,N \\ such that \\ i_n = i}} g_n. \tag{1}$$

Logistic grade models

We follow Tomkin et al. (2018) and use a two-parameter logistic model for the predicted grade $\widehat{g}_{ik}$ of student i in course k, which gives

$$\widehat{g}_{ik} = \frac{4}{1+\exp(-a_k(\theta_i - b_k))} \tag{2}$$

where there is one student parameter and two course parameters given by

$$\theta_i = \text{"ability of student" } i \tag{3a}$$

$$b_k = \text{"difficulty of course" } k, \tag{3b}$$

$$a_k = \text{"discrimination of course" } k. \tag{3c}$$

The "difficulty" of a course is the ability level θ at which the logistic model crosses 2.0 (i.e., a grade of C). That is, a student with an ability equal to the course difficulty will be predicted to receive exactly a grade of C in the course. The "discrimination" of a course indicates how strongly a course distinguishes between students of different ability levels (higher discrimination courses provide more information about a students' ability level).

For a set of parameters (3) the root mean square error (RMSE) of the predicted grades is

$$e = \sqrt{\sum_{n=1}^{N} (g_n - \widehat{g}_{i_n,k_n})^2} \tag{4}$$

The optimal parameters θ_i^*, b^*_k, and a^*_k are determined by minimizing the error e over all parameter values. This can be done, for example, by an iterative procedure that begins by initializing student abilities θ_i to observed GPA scores and then alternates between finding the optimal course parameters, while the student abilities are held fixed, and finding the optimal student abilities while fixing the course parameters. That is, we alternate between

$$\left(a_k^*, b_k^*\right) = \text{argmin}_{a_k,b_k} \sum_{\substack{n=1,\ldots,N \\ \text{such that } k_n = k}} \left(g_n - \hat{g}_{i_n,k}\right)^2 \quad \text{for } k = 1,\ldots,K \tag{5a}$$

and

$$\theta_i^* = \text{argmin}_{\theta_i} \sum_{\substack{n=1,\ldots,N \\ \text{such that} i_n = i}} \left(g_n - \hat{g}_{i,k_n}\right)^2 \quad \text{for } i = 1,\ldots,I. \tag{5b}$$

To remove the effect of course choice on GPA for student i, we first compute their predicted grade $\widehat{g}_{ik}$ in all courses $k=1,...,K$. Following Tomkin et al. (2018), we then define their (unweighted) calibrated GPA to be their GPA as if they had taken every course in the university, using these predicted grades:

$$\text{unweighted - calibrated GPA}_i = \frac{1}{K}\sum_{k=1}^{K} \hat{g}_{ik} \tag{6}$$

In this paper we introduce the weighted calibrated GPA for student i, which is defined by:

$$\text{weighted - calibrated GPA}_i = \frac{1}{\sum_{k=1}^{K} N_k} \sum_{k=1}^{K} N_k \hat{g}_{ik} \tag{7}$$

where N_k is the number of student records for course k. The weighted calibrated GPA for a student is their predicted GPA if they took all courses at the university, where courses are weighted proportional to enrollment. Equivalently, this is weighting each course by the probability that a random student takes that course. This corrects for the problem that the unweighted calibrated

GPA in Equation (6) is overly influenced by the large number of small courses, which tend to have higher average grades. Upper level (300 and 400 level) courses have lower total average enrollments (146) and higher average scores (3.28) than lower level courses (668 and 3.19, respectively). In the analysis below for Research Questions 2–5 we will always use the weighted model and we will refer to the corresponding predictions as simply "calibrated grades" and "calibrated GPA".

We have made the Python code available (https://github.com/jtomkin/calibrated-GPA) so that other workers can use this calibrated GPA when analyzing their own data.

Methods for Research Question 1: Is a weighted logistic model of GPA better than the observed GPA in predicting academic performance?

Different courses have different course offsets. The observed GPA of a student is the average of course grades, which includes the average of these offsets. If course offsets are systematic then the observed GPA will be heavily influenced by the set of courses taken and observed GPA will not be a reliable measure of academic ability.

This suggests that we can build a more accurate model of student ability by computing each student's expected grade over all courses, and then finding the weighted average of all of these grades to determine each student's GPA. Since this calibrated GPA compares students with the same set of courses (i.e., all of them) there would be no bias due to course choice.

If the calibrated GPA is a better measure of student ability then the observed GPA, it would be a better predictor of the discrete course grades actually observed in students (i.e., it would have a lower error).

Methods for Research Question 2: Are STEM majors graded more stringently than non-STEM majors?

One way to test this hypothesis would be to compare the observed GPAs of students in different majors. If the hypothesis is true, we would expect GPAs to be lower in the more difficult STEM majors. However, this approach does not account for the possibility of program-dependent variability in academic ability, which would produce different baseline GPAs: if STEM students were more academically able on average then this would raise grades and hide the effect of more stringent grading. Average GPA is, therefore, not a reliable measure of program difficulty as it is dependent on the average ability of students in the major.

We solve this by measuring the difference between the STEM program's observed grades and the student's calibrated GPA—the grade point offset. The grade offset measures the difference in grades relative to the student's ability, and so is a measure of course challenge that is independent of the student's observed academic ability. The lower the average grade offset of a major overall, the more stringent the grading. If STEM programs have systematically more negative grade offsets then non-STEM programs this supports the hypothesis that STEM majors are graded more stringently.

To determine the grade offset of a major, we first compute the grade offset for each student in each of their courses as the difference between their course grade and their calibrated GPA. Second, we average these values over all courses taken by a student to find the average grade offset for each student. Finally, we average over all students in the major. This method captures the total course choices of students in each program of study, and so takes into account factors that might not be apparent in a catalog description of the major, including hidden requirements, electives, general education requirements, concentrations, minors, previous changes of major, and advanced courses.

Methods for Research Question 3: Are STEM general education courses graded less stringently than STEM major courses?

Colleges and universities in the United States require students to take courses in "general education" which are outside of their major course of study. Courses that are taken by non-STEM students to fulfill STEM general education requirements are usually not the same as those taken by STEM students to fulfill major requirements, and are sometimes disparaged with diminutives, such as "physics for poets" or "rocks for jocks" (Gilbert et al., 2012). These diminutives may arise from a folk belief that STEM courses that are required of major students have higher academic standards than STEM courses taken to fulfill a general education requirement.

Our method is to compare the average offsets of the required major courses and general education courses, and determine if they are statistically significantly different. We define STEM general education courses as those satisfying the following two conditions: (a) they provide Quantitative Reasoning and/or Natural Science general education credit in the University, and (b) they do not count towards the major of the Department that offers the course. For each department that offers such a course we compare this STEM general education course with the first required major course that the department offers (the "gateway" course). If a department offers more than one general education course we use the course with the largest enrollment. We compare courses from the

same department with one another so as to remove the effect of inter-departmental variability in grading (which is shown, in the results for Research Question 2, to be potentially considerable).

Methods for Research Question 4: are gateway courses in STEM graded less stringently that STEM major courses?

All STEM majors are required to take gateway courses that are foundational to STEM programs of study. Poor performance in these courses will dissuade many students from continuing in STEM (Ost, 2010; Rask, 2010), and is seen as being part of the "leaky pipeline" problem in STEM education (Seymour & Hewitt, 1997). Stringent grading in these courses is sometimes perceived by students as being part of a "weed-out" strategy, which may disproportionately dissuade underrepresented groups of students from continuing in a STEM major (Sanabria & Penner, 2017).

If this weed-out hypothesis is true, then we would necessarily find evidence of large, negative grade point offsets in gateway STEM courses. On its own, this is not sufficient evidence of a weed-out course; however, as this grade offset may merely reflect the rigorous grading found in all STEM disciplines (Table 1). In other words, gateway STEM courses may be hard, because all STEM courses are hard. If the weed-out hypothesis is true then gateway courses will be inconsistently graded; we will observe lower grade point offsets in these courses than in others offered in the department. We, therefore, compare the grade offsets of the gateway courses and their host departments.

For the purposes of this study, we consider weed-out courses to be the gateway courses that all STEM majors are required to take for their course of study. General Chemistry 1, University Physics 1, and Calculus 1 are required courses for all physical science and engineering programs. Life science and bio-engineering students are all required to take the introductory microbiology course.

Methods for Research Question 5: to what extent do standardized test scores predict academic performance?

As part of the University's admission process, students submit standardized test scores. As the institution is in the Midwest, the standard score submitted is the ACT, but students also submit SAT scores. The institution is selective; average composite ACT scores are 28.3 ± 3.9 (mean and standard deviation) and average SAT total scores are 1316 ± 132.

To determine if the standardized test scores predict grades at the university, we perform a linear regression analysis with the tests and GPA. If standardized tests are a predictor of course performance, there should be a correlation between test scores and GPA. This means that if the calibrated GPA is a better predictor of course performance than the observed GPA (Research Question 1), we would then expect that the correlation between the calibrated GPA and test scores to be higher than the correlation between the observed GPA and test scores.

Results

Results for Research Question 1: Is a weighted logistic model of GPA better than the observed GPA in predicting academic performance?

We fitted the logistic model (2) to the student grade data set using the iterative procedure (5) and we computed the root mean square error (RMSE) of the predictions. Figure 1 shows the difference between the error when using the observed GPA and the error when using the model, as a predictor of student grades. This shows that the model is a better predictor of student course grades than observed GPA for 83.6% of students, with a mean improvement of 0.11 grade points.

Figure 2 shows the calibrated GPA in both unweighted and weighted versions, plotted against the observed GPA of each student. As observed in Tomkin et al. (2018), the unweighted calibrated GPA is systematically higher than the observed GPA, especially for students with low GPAs. The use of the weighted calibrated GPA corrects this bias and we see that the observed and weighted calibrated GPAs are similar on average. The unweighted calibrated GPA has a linear regression slope of 0.62 against observed GPA, while the weighted calibrated GPA has a slope of 0.89, showing that weighting produces a model which is much better calibrated to the observed GPA on average. Note, however, that for any individual student the weighted calibrated GPA may be substantially higher or lower than the observed GPA, reflecting the fact that observed GPAs are not accurate measurements of an individual student's ability.

Although we are primarily interested in the predicted grades, $\widehat{g}_{ik}$ from the model and the resulting calibrated GPAs, we can also consider the model parameters themselves. The student ability parameters, θ_i, are monotonically predictive of the (weighted) calibrated GPA, so these have a Spearman rank correlation of $r_s = 1$. The course difficulty, b_k, and discrimination, a_k, parameters are quite highly correlated ($r_s = 0.79$) so more-stringent courses tend to have both higher difficulty and higher discrimination (and lower average grades). However, the course

Table 1 Average grade offset of students in different majors, as measured by the difference in the average observed and calibrated GPAs of each major

Degree	*n*	Avg. Obs. GPA	Avg. Cal. GPA	Avg. Offset
Computer Engineering	1959	3.09	3.38	− 0.29
Electrical Engineering	2960	3.18	3.44	− 0.26
Chemical Engineering	1457	3.18	3.42	− 0.24
Engineering Mechanics	310	3.08	3.32	− 0.24
Engineering Physics	519	3.29	3.48	− 0.20
Physics	687	3.13	3.33	− 0.20
Math and Computer Science	374	3.05	3.24	− 0.19
Computer Science	2790	3.19	3.38	− 0.19
Materials Science and Engr	1004	3.23	3.40	− 0.18
Computer Sci and Astronomy	9	3.00	3.18	− 0.18
Aerospace Engineering	1174	3.12	3.30	− 0.18
Civil Engineering	2275	3.18	3.35	− 0.17
General Engineering	1197	3.11	3.27	− 0.16
Astronomy	120	2.77	2.92	− 0.16
Industrial Engineering	553	3.16	3.32	− 0.16
Statistics and Computer Science	107	3.15	3.30	− 0.15
Nuclear, Plasma, Radiolgc Engr	322	3.04	3.20	− 0.15
Computer Sci and Chemistry	13	3.17	3.31	− 0.14
Mechanical Engineering	2508	3.26	3.40	− 0.14
Nuclear Engineering	127	3.25	3.34	− 0.09
Computer Sci and Anthropology	9	2.95	3.03	− 0.08
Statistics	504	3.10	3.17	− 0.07
Actuarial Science	1112	3.34	3.41	− 0.07
Biochemistry	345	3.37	3.44	− 0.07
Agricultural Engineering	108	3.19	3.25	− 0.06
Bioengineering	467	3.47	3.53	− 0.06
Agricultural and Biological Engr	326	3.14	3.20	− 0.06
Chemistry	1707	3.14	3.20	− 0.06
Molecular and Cellular Biology	4654	3.38	3.41	− 0.03
Mathematics	1480	3.33	3.36	− 0.02
Teaching of Latin	2	3.47	3.49	− 0.01
Engineering Undeclared	10	2.90	2.87	0.03
Computer Sci and Linguistics	16	3.36	3.31	0.05
Integrative Biology	1713	3.22	3.16	0.06
Economics	3463	3.19	3.12	0.07
Geology	241	2.98	2.90	0.08
Finance	185	3.36	3.28	0.09
Agr Engineering and Agr Science	23	2.75	2.66	0.09
Atmospheric Sciences	205	3.07	2.96	0.11
Technical Systems Management	694	2.87	2.73	0.14
Biology	927	3.23	3.06	0.17
Philosophy	324	3.01	2.83	0.18
Classics	80	3.30	3.10	0.20
Earth, Soc, Env Sustainability	493	3.03	2.83	0.20
Russian Lang and Literature	8	3.21	3.00	0.21
Geography	141	3.03	2.81	0.22
Linguistics	205	3.27	3.05	0.22
Germanic Lang and Lit	74	3.18	2.96	0.22

Table 1 (continued)

Degree	*n*	Avg. Obs. GPA	Avg. Cal. GPA	Avg. Offset
Psychology	5211	3.34	3.10	0.23
Religious Studies	38	3.17	2.92	0.25
Teaching of German	10	3.46	3.21	0.25
Political Science	3081	3.16	2.90	0.26
Individual Plans of Study	85	3.39	3.13	0.26
Religion	30	3.21	2.94	0.27
Anthropology	558	3.15	2.86	0.28
E Asian Languages and Cultures	228	3.18	2.90	0.28
Italian	19	3.06	2.77	0.29
Global Studies	525	3.38	3.07	0.31
History	1552	3.28	2.97	0.31
Latin American Studies	19	3.09	2.78	0.31
International Studies	537	3.34	3.03	0.31
Spanish	658	3.22	2.92	0.31
Russian and E European Studies	9	3.21	2.89	0.32
Teaching of Spanish	127	3.69	3.36	0.33
French	155	3.31	2.98	0.33
Russian, E Eur, Eurasian St	11	3.44	3.11	0.33
Teaching of French	21	3.60	3.27	0.33
Comparative Literature	54	3.33	2.99	0.34
History of Art	124	3.24	2.90	0.34
Portuguese	6	2.61	2.25	0.36
English	2353	3.30	2.92	0.38
Interdisciplinary	98	3.01	2.60	0.41
Communication	2553	2.99	2.58	0.41
Slavic Studies	2	3.40	2.97	0.43
Sociology	1493	2.95	2.52	0.43
Speech Communication	1071	3.00	2.53	0.47
Creative Writing	80	3.20	2.72	0.48
Gender and Women's Studies	47	2.93	2.43	0.50
Rhetoric	351	3.01	2.50	0.51
African American Studies	30	2.80	2.21	0.58
LAS—Undeclared	288	2.95	2.27	0.68
Latina/Latino Studies	18	2.47	1.79	0.68

stringency is better determined by the difficulty parameter ($r_s = -0.88$) than by the discrimination parameter ($r_s = -0.43$).

Results for Research Question 2: are STEM majors graded more stringently than non-STEM majors?

The average course offset for students in each program is shown in Fig. 3. The list of observed and calibrated GPA values for each major is given in Table 1. The average GPA offset is the difference between the average observed and calibrated GPAs. The grade offset for each major is highly correlated with the average calibrated GPA ($r = 0.80$). The slope is negative (-0.58): the higher the academic performance of students in a major the more negative the grade offset.

Figure 3 indicates that STEM majors are the only majors in which students are penalized on net. Figure 4 compares the average observed and calibrated GPA of all STEM majors (those majors in which a student earns a B.S. or B.Eng., $n = 35{,}034$) with non-STEM majors ($n = 29{,}826$). The observed GPA is 3.21 in STEM majors, similar to the value of 3.19 for non-STEM students. The calibrated GPA is 3.32 for STEM majors, and 2.90 for non-STEM majors. The average STEM offset is -0.12 (STEM majors have observed GPAs that are about a tenth of a GPA point lower than their predicted score), while non-STEM students have

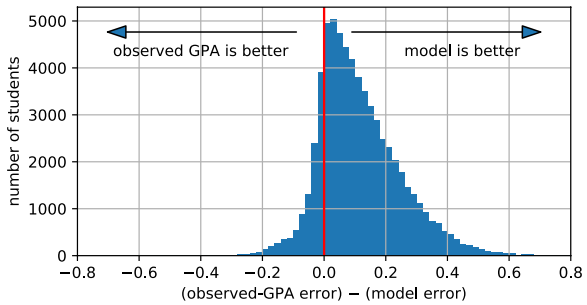

Fig. 1 The difference between the RMSE (root mean square error) of observed GPA and the logistic model (2) as a predictor of student grades. Values to the right of center indicate that the model (2) is a better predictor than observed GPA, while points to the left indicate the reverse. This shows that the model (2) is a better overall predictor of student grades than observed GPA

an offset of 0.29. The effect size (for both Cohen's d and Hedges' g) of belonging to a STEM major on a student's GPA is 0.61.

Results for Research Question 3: are STEM general education courses graded less stringently than STEM major courses?

In our sample there are 10 departments that offer general education courses that fit the criteria described in the method section. For example, the Department of Geology offers GEOL 100 "Planet Earth" to non-science majors, and GEOL 107 "Physical Geology" to Geology majors. Table 2 shows the characteristics for the two types of courses from each of the departments, their offsets, the p value of the likelihood of the difference in those offsets, and the results of an unpaired t test for significance in the magnitude of the difference in offsets.

Using a fixed effects model (Borenstein et al., 2011) we determine the average effect size (Cohen's d) of replacing the STEM major course with the STEM general education course to be 0.40 (95% CI [0.39, 0.42]), as shown in Fig. 5. The result is statistically significant, with $z = 45.3$ and $p < 0.0001$.

If we use observed GPA to calculate the grade offset, the observation that general education courses are graded more stringently largely disappears (Table 2). The average difference in offsets is reduced from 0.42 (for the calibrated GPA case) to 0.11. The individual departments now show mixed results: although 5 departments show statistically significant (to the $p < 0.05$ level) offset differences in which the gateway course is graded more stringently, there are two courses that are not statistically different, and three courses in which the gateway course is grade more leniently. In all

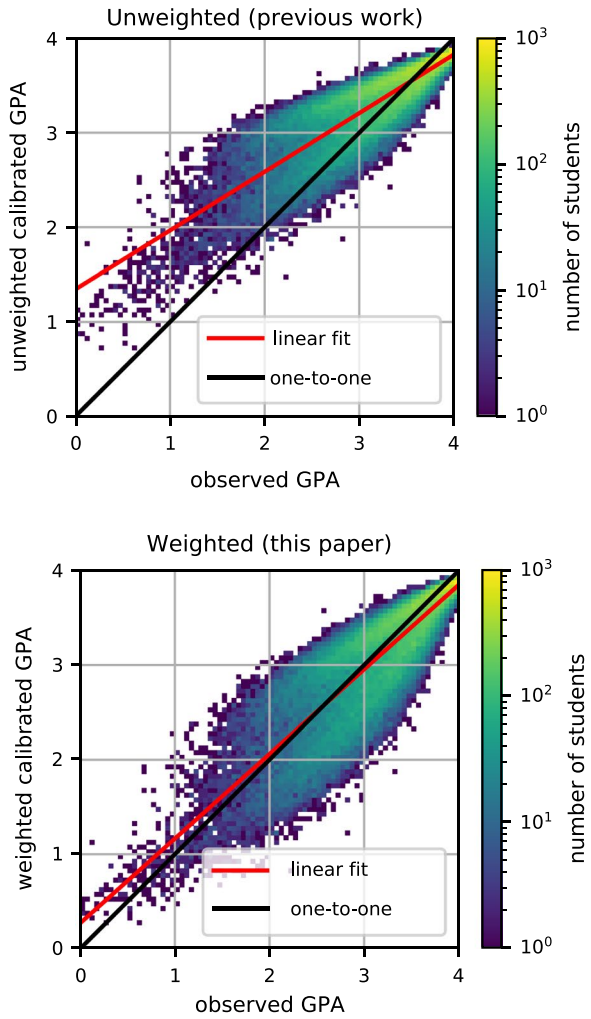

Fig. 2 Observed GPA plotted against unweighted calibrated GPA (left panel, Equation (6) and weighted calibrated GPA (right panel, Equation (7)) for each student. The right panel shows that using the weighted calibrated GPA almost completely eliminates the overestimation that is visible in the left panel for the original unweighted calibrated GPA

cases the trend is for the general education course to appear to be graded more leniently when observed GPA is used in the calculation.

Results for Research Question 4: are gateway courses in STEM graded less stringently that STEM major courses?

The grade offsets of gateway courses are shown in Table 3, along with the weighted average of grade offsets in their host departments. The n for all courses is large. In all cases, the gateway courses have negative grade offsets (they are graded more stringently than most courses). The average grade offset of the gateway course departments is also negative (Table 1). In general, the gateway

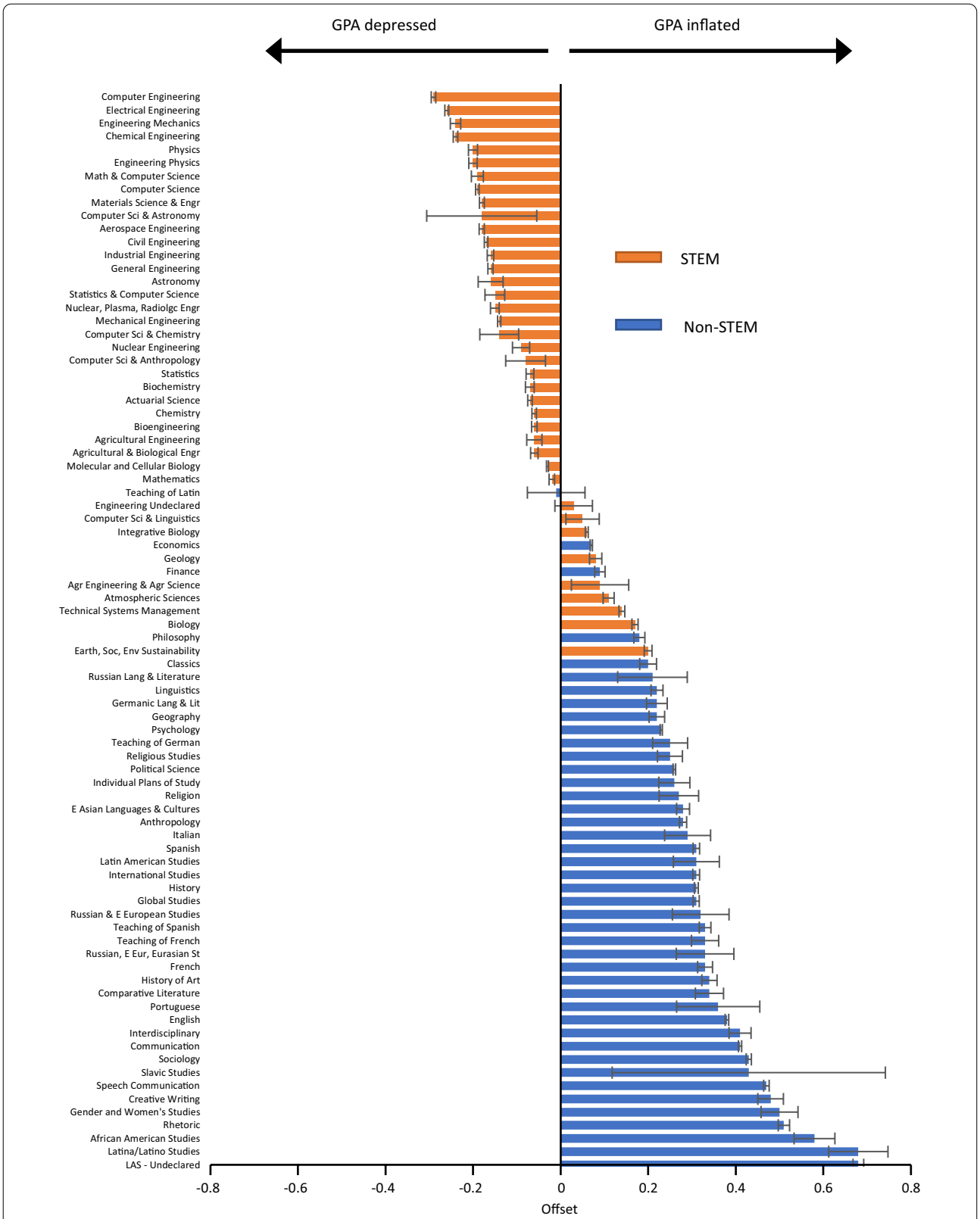

Fig. 3 Average GPA offset for students in each major. A grade difference of 1.0 is equivalent to a letter grade. Error bars display standard error of the mean (± 1 S.E.). The differences in the S.E. are largely determined by differences in the number of students in each program of study

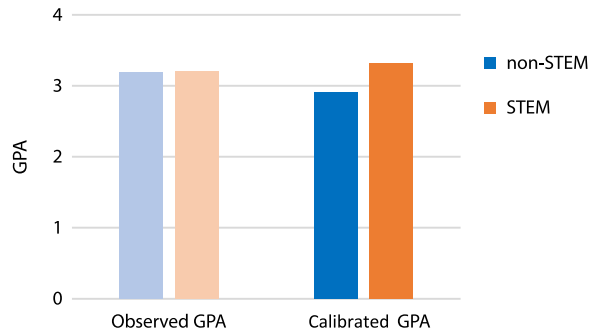

Fig. 4 Average observed and calibrated GPA for STEM students (those who are in programs that result in B.S., B.Eng, and mathematics degrees) and non-STEM (all other programs). The observed GPA is 3.19 for non-STEM and 3.21 for STEM. The calibrated GPA is 2.90 for non-STEM and 3.32 for STEM. This implies that the observed non-STEM GPA is inflated by 0.29 points, and the observed STEM GPA is depressed by 0.12 points. The standard deviation of the GPA is 0.67. The Standard Error (S.E.) for all bars is 0.03 and is omitted for clarity

courses have similar course offsets to the department averages and so appear to have representative offsets.

Results for Research Question 5: to what extent do standardized test scores predict academic performance?

The Pearson correlation between the composite ACT score and observed GPA is $r=0.25$ ($n=54{,}436$). The Pearson correlation for the same students between the composite ACT score and calibrated GPA is $r=0.49$. The data set also includes the ACT math subscore, which is as predictive as the composite score ($r=0.20$ for the observed GPA and $r=0.49$ for the calibrated GPA).

The slope of the regression is also higher for the calibrated GPA. Every additional point gained in the composite ACT score predicts an additional 0.071 for the calibrated GPA versus an increase of 0.035 for the observed GPA. Every additional point gained in the SAT score predicts an additional 0.0014 for the calibrated GPA versus an increase of 0.0008 for the observed GPA.

This means that students with low standardized test scores have higher observed than calibrated GPAs, and that students with high standardized test scores have lower observed than calibrated GPAs. Students with composite ACT scores between 12 (the lowest in the sample) and 30 all have positive average offsets, while students with scores of 31 to 36 all have negative average offsets. The largest positive average offset is 0.69, for students with composite ACT scores of 15 (average GPA 2.66), while the most negative average offset, of −0.10, was for students with composite ACT scores of 34 (average GPA 3.34).

STEM majors have higher average tests scores than non-STEM majors in this sample. For STEM majors, the average composite ACT scores are 29.7 ± 3.9 and average SAT total scores are 1347 ± 115. For non-STEM majors, the average composite ACT scores are 26.7 ± 3.3 and average SAT total scores are 1249 ± 142.

Discussion

Discussion for Research Question 1: is a weighted logistic model of GPA better than the observed GPA in predicting academic performance?

GPA is a biased measure of academic achievement. Grade offset and grade penalty studies that use observed GPAs as baselines of student ability are likely to be systematically biased. Logistic models are better at predicting student performance than the observed GPA. Previous work (Tomkin et al., 2016, 2018) has shown that using observed GPA overestimates STEM gender and racial grading disparities in this data set. In the examples shown in this study, using observed GPA to calculate course offsets (Koester et al., 2016; Matz et al., 2017) substantially

Table 2 Comparisons of largest general education course ("Gen-ed") offered by ten STEM departments and that department's first course for majors ("Gateway")

Department	Gen-ed	Offset	*n*	Gateway	Offset	*n*	Difference	*p* value	*t* value
Atmospheric Sci	ATMS 120	0.64	7243	ATMS 201	0.68	506	− 0.03	0.2126	1.25
Astronomy	ASTR 100	0.26	6597	ASTR 210	0.27	811	0.53	< 0.0001	19.67
Chemistry	CHEM 108	0.92	199	CHEM 102	− 0.62	22,356	1.54	< 0.0001	31.92
Computer Sci	CS 105	− 0.02	6448	CS 125	− 0.04	5373	0.01	0.3803	0.88
El. & Co. Eng	ECE 101	0.32	243	ECE 110	− 0.33	6635	0.65	< 0.0001	13.93
Geology	GEOL 100	0.13	4164	GEOL 107	0.10	1154	0.02	0.2180	1.23
Integrative Bio	IB 105	0.42	3729	IB 150	0.00	8037	0.42	< 0.0001	35.09
Mat. Sci. Eng	MSE 101	0.16	116	MSE 201	− 0.27	868	0.43	< 0.0001	7.90
Physics	PHYS 140	0.04	3471	PHYS 211	− 0.40	17,329	0.44	< 0.0001	30.59
Statistics	STAT 100	0.44	17,384	STAT 200	0.23	1814	0.21	< 0.0001	15.07

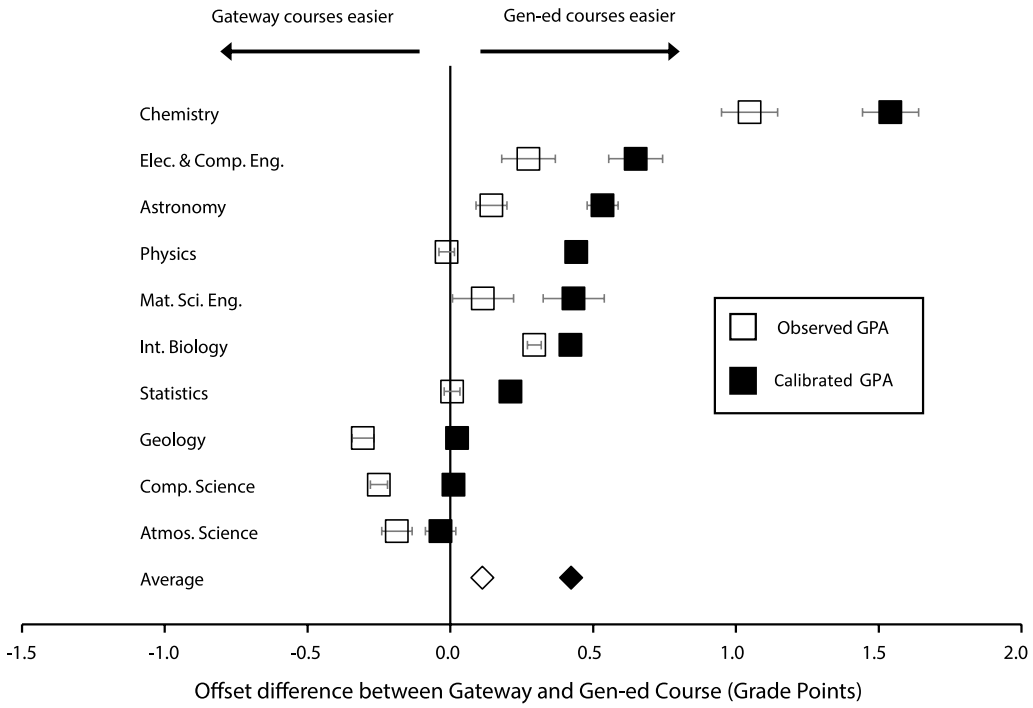

Fig. 5 Average grading difference between general education and gateway courses within ten STEM departments. A grade difference of 1.0 is equivalent to a letter grade. Positive values indicate that the gateway course was graded more stringently than the general education course. Open squares represent the grade offset calculated using the observed GPA. Black squares represent the grade offset calculated using the calibrated GPA. Error bars display ± 1 S.E. Diamonds are the weighted average of the ten departments

underestimates the stringency of grading of STEM students (relative to non-STEM students), substantially underestimates the stringency of grading in STEM major courses relative to STEM general education courses, and substantially underestimates the ability of standardized tests to predict university academic performance.

Table 3 Comparisons of grade offsets of gateway STEM courses ("Offset") with the weighted average grade point penalty of that department ("Dept. Average Offset")

Department	Gateway	n	Offset	Dept. Average Offset	r_{GPA}
Chemistry	Chemistry 1	22,356	− 0.62	− 0.20	0.73
Mathematics	Calculus 1a	8464	− 0.47	− 0.47	0.65
Mathematics	Calculus 1b	8137	− 0.27	− 0.47	0.66
Molec. and Cell. Bio	Biology 1	10,201	− 0.40	− 0.36	0.78
Physics	Physics 1	17,329	− 0.40	− 0.46	0.72

In all cases, the Standard Error of the Mean is equal or less than 0.01 of a grade point. Note that there are two first courses in calculus 1 ("Calculus 1a" is MATH 220, for those who have not been exposed to calculus previously, and "Calculus 1b" is MATH 221, for those who have). Chemistry 1 has the rubric CHEM 102, Physics 1 has PHYS 211, and Biology 1 has MCB 150 at the institution. We write "r_{GPA}" for the correlation coefficient *r* between a student's grade in the gateway course listed and their calibrated GPA

All our results are consistent with STEM programs having more stringent grading than non-STEM programs. Individual programs within STEM also have considerable grading variation. Studies that sort students by degree program (Herman et al., 2018) will control for the major grading biases, but as individual students still have course choice (in electives and in general education) individual variation is not entirely accounted for by aggregating grades at the program level. Individual students with a strong preference for high grades will likely have higher course offsets than other students in the same program, due to their course choices.

Accounting for the heterogeneity of student course choice is, therefore, necessary in any study that uses course grades as a proxy for ability, equity, or bias, as it substantially impacts the interpretation of the data. Recent work using observed GPA to measure grade penalties (Koester et al., 2016; Matz et al., 2017) erroneously concluded that women suffer higher grade penalties in STEM courses than men, for example. A similar study that used a calibrated GPA (Tomkin et al., 2018) found that there was no gender disparity: the expected performances of women and men were the same. Course choice was not the same, however. Male students (on average) took courses with higher penalties, which depressed

their observed GPA, which made it appear as if they were overachieving in STEM courses. We, therefore, encourage all workers to use a calibrated GPA when performing grade penalty/grade offset research, and have made the Python code freely available at (repository URL to be posted at publication).

Discussion for Research Question 2: are STEM majors graded more stringently than non-STEM majors?

The observed GPA is almost the same between STEM and non-STEM students, but this hides a large and systematic difference in grading practices between majors (Figs. 3, 4). The average difference in grade offset between STEM and non-STEM students is 0.41 of a grade point. In other words, non-STEM students have GPAs that are more than a third of a letter grade higher than their observed GPAs indicate, compared to STEM students. This is large in both absolute and relative terms. The effect size on GPA of belonging to a STEM major is $d = 0.61$. Therefore, while the observed GPAs suggest that STEM and non-STEM students have the same average GPA, this analysis suggests that there is a very large difference in actual academic performance, with STEM students penalized relative to their peers. Our results support earlier observations of disparities in STEM grading (Johnson, 2003).

Different STEM programs do not have equal offsets (Fig. 3). Treating STEM programs as equivalent (Sonnert & Fox, 2012; Witteveen & Attewell, 2020) will remove some, but not all, of the bias in observed GPA.

The high correlation ($r = 0.80$) between GPA offsets and calibrated GPAs of majors acts to reduce variation in the observed GPAs: in general, the better your performance, the more your grades are depressed, and the worse your performance, the more your grade is inflated. As the slope of the relationship is less than one, but greater than one-half, most, but not all, of the actual difference in academic performance of students is not present in the observed GPA.

Some researchers have proposed (Cotner and Ballen (2017); Matz et al., 2017) that assessment-style effects (e.g., low stakes versus high-stakes assessments) can explain disparities between otherwise comparable students. Because STEM courses are typically considered to have more high-stakes assessments, this could help explain the findings here. If this is true, and non-STEM students are relatively better at low-stakes assessments, then we would expect the model to be better at predicting the grades of non-STEM students in non-STEM (relative to STEM) courses. We find the opposite: the model predicts non-STEM major's observed grades better in STEM courses ($r = 0.78$) than in non-STEM courses ($r = 0.72$). Similarly, if STEM majors were (relatively) better in high-stakes exam settings than non-STEM majors, we would expect that ACT scores, which are the result of high-stakes exams, to better predict the course grades of STEM majors. We again find the opposite: the calibrated grades of non-STEM majors are better predicted by ACT scores ($r = 0.44$) than they are for STEM majors ($r = 0.31$). Although the difference in correlations is at least partly due to range restriction (Sackett et al., 2009) (as more STEM majors are limited by the upper bound on ACT scores), the high value of the observed correlation is sufficient in itself to argue against a large assessment-style effect. Taken as a whole, the difference between STEM and non-STEM students does not appear to be driven by different reactions to high-stakes exams.

Although the reasons for the STEM/non-STEM grading disparity go beyond the scope of this study, we speculate that a simple explanation for this result is that professors tend to give the same average grade as one another, regardless of the average ability of their students. Given that the STEM students in our study have higher average academic ability (according to both standardized test scores and grades in individual courses), and that STEM students take more similar courses with each other than with non-STEM students, this would predict that STEM students would get the same average grades as non-STEM students, which is observed.

If professors tend to give the same average grade in their courses, regardless of student performance, we would predict the observed grades to show less dispersion than the calibrated grades. This is, in fact, observed. The standard deviation of the observed average course grades is 0.26, while the standard deviation of the calibrated course grades is 0.36. This is consistent with the hypothesis that professors assign similar grades regardless of the average academic ability of students in their courses.

Discussion for Research Question 3: are STEM general education courses graded less stringently than STEM major courses?

STEM general education courses are usually graded more leniently than other STEM courses. When we compare the major and general education offerings of ten STEM departments, the majority (7 out of the 10) have a significantly lower offset in the major course. The remaining three are statistically indistinguishable. The average difference is statistically significant ($p < 0.0001$).

The grade differences themselves can also be very large. The unweighted average difference in offsets between the general education and major courses is 0.423. The largest difference is seen in Chemistry, with an average offset of 0.922 for the general education course and -0.619 for the major course. A student with a GPA of 3.0—a B

average—would average an A in the general education CHEM 108 ("Chemistry Everyday Phenomena") and only a C+ in the major gateway CHEM 102 ("General Chemistry I"). We hypothesize that this difference arises due to the incentive structures that face departments and instructors. If the college or university rewards departments with higher enrollments, and students desire courses in which they are most likely to get good grades (and possibly even courses that do not require much work), then departments will have a compelling reason to offer courses that are not graded stringently.

There is evidence that higher grades are correlated with higher student assessment scores (Eiszler, 2002), and courses with a favorable reputation are liable to attract more students in the future. Previous workers have associated the growing number of contingent faculty (Kezim et al., 2005; Sonner, 2000) with this trend, as these instructors are most susceptible to departmental pressure to increase enrollments at the expense of rigor. This is consistent with our sample, in which the majority of the general education courses were taught by non-tenure-track instructors.

This mechanism would impact general education courses the most, as these courses are not designed to be foundational to the discipline: students will not need to use any knowledge gained in general education coursework for future courses. The academic standard of major courses need not suffer from the same incentive to inflate grades, as these courses need to be held to a consistent standard so that majors are adequately prepared for upper-level coursework, and careers, in their discipline.

Although we believe that the differing grading practices in general education courses and major courses that we observed is consistent with this incentive hypothesis, further study is required to show that this is the causal mechanism.

Discussion for Research Question 4: are gateway courses in STEM graded less stringently that STEM major courses?

If the weed-out hypothesis is correct, stringent grading in gateway STEM courses is a strategy enacted by departments to discourage academically weaker students from majoring in their discipline. The educational rationale for such an approach is dubious at best (would it not be better to presume that all the students can be taught? There is considerable evidence (Freeman et al., 2014; Handelsman et al., 2004) that student-centric teaching produce better outcomes than traditional methods, for example), so it is no surprise that departments do not, generally, openly state that their gateway courses are designed to prevent students from further study.

It is clear that gateway STEM courses are rigorously graded, with negative grade offsets (Table 3). However, it is also apparent that the other courses in these departments are also rigorously graded, and also have negative grade offsets. There is no consistent differences between the gateway courses and weighted department averages; in one case (Chemistry) the gateway course has a grade offset that is more than 0.1 points lower than the department average, three courses were within 0.1 of a grade point of the department average, and one gateway course (in calculus) had a offset more than 0.1 points higher than the department average. Given this mixed result, our evidence does not support the hypothesis that gateway STEM courses are designed to weed out students. Gateway STEM courses are stringently graded not, because they are gateway courses, but because they are representative STEM courses.

Furthermore, success in gateway STEM courses is highly predictive of overall academic success, as measured by calibrated GPA (Table 3). The correlation between the two ranges between $r=0.65$ and $r=0.78$. This also argues against a weed out interpretation, as it implies that grading in gateway courses is consistent with other grades received, and that the grades that a student receives in gateway courses are honest signals of future undergraduate academic success.

Despite this finding, it may very well be true that other factors associated with gateway courses are disproportionately dissuading underrepresented groups from continuing in STEM. Gasiewski et al. (2012) found that gatekeeping courses that socially distanced their students reduced engagement, and Sanabria and Penner (2017) found that women react more negatively to poor grades then men. Gateway STEM courses at the university studied here are large, and historically taught in a traditional way. Since the data period of this study the authors have been part of an effort to introduce more active learning and student engagement in these foundational STEM courses (Herman et al., 2018; Mestre et al., 2019).

The reader will note that Introductory Chemistry is a gateway course with a highly negative grade offset, and this may be worthy of further study. One of the authors asked the Director of General Chemistry about their grading practices. Final exams are very important in determining the final grade in CHEM 102, and this exam is centralized across all sections. Section instructors are not tenure-track faculty, and they do not set the exam. The final exam has not changed in any significant way in several decades, and the grading expectation is that the median student will be graded in the C range. This implies that there is an expectation, if not a mandate, for a grading curve. As the average grade cannot change, the implication is that Chemistry does not believe that innovations in instructional practices can improve student learning.

The highly negative grade offset in chemistry is consistent with a model in which chemistry has held its grades constant, while other departments have changed. The average GPA at the University of Illinois has been increasing at about 0.1 grade points per decade, which is similar to other universities in the US (this is the so called "grade inflation" (Rojstaczer & Healy, 2016)). If cultural practices in the department (i.e., having a set expectation of student success tied to a final exam, and not changing the final exam for several decades) mean that grading is "stuck" in the 1980s, then we could explain the apparent grade offset in chemistry courses as a result of grade drift. Chemistry has stayed consistent, while many other units have grades that have consistently trended upwards, so Chemistry has become relatively harder. If true, this model underlines the importance of departmental culture in instructional practice in this sample (Ma et al., 2019). Testing the validity of this model would require further work.

Discussion for Research Question 5: to what extent do standardized test scores predict academic performance?

The use of standardized tests in university admissions is a topic of long-running debate (Atkinson & Geiser, 2009). The tests are criticized for being poor predictors of student outcomes that contribute to disparities (Buchmann et al., 2010; Soares, 2015), while defenders (Sackett & Kuncel, 2018) argue that they are neutral measures of academic ability. This debate is beyond the scope of this article; we restrict ourselves to noting how predictive of academic performance these tests are in our data.

Both the ACT and SAT tests are much better at predicting student performance than comparisons with observed GPA would suggest. By incorporating the grading difficulty of individual student's course of study, we find that the observed correlation between the ACT and grades increases from 0.25 to 0.49, and between SATs and grades increases from 0.18 to 0.37. This means that standardized tests are better predictors of how well students do in individual courses than suggested by studies that use observed GPAs. Our hypothesis is that students with high ACT scores tend to take more challenging courses of study, and so suffer a penalty to their GPA, which in turn lowers the observed correlation. This is consistent with the observation that the STEM students in the sample both have lower average grade offsets (-0.41) and higher average standardized test scores (about 3 points higher in the ACT and 98 points higher in the SAT).

It should also be pointed out that our sample likely suffers from range restriction (Cohen et al., 2013); the University is a selective institution and applicants with low scores either do not apply or are not admitted. Range restriction reduces the observed correlations between variables. Our correlation is, therefore, likely a modest underestimate: the actual correlation between standardized tests and academic performance is even larger than what we present here.

Standardized tests are, therefore, a surprisingly accurate tool for predicting academic success for undergraduates. Our results are consistent with previous work that has attempted to determine how predictive SAT scores are after correcting for institutional type and range restriction (Sackett et al., 2009). We do not have access to other relevant information about student applications (such as high school grades or socioeconomic status), but, taken in isolation, the use of the ACT and SAT as tools for admission are consistent with a goal of admitting students who will perform well in university courses.

Conclusions

In this paper we proposed a proposed a methodological enhancement to the calculation of calibrated or predicted grades for students, and we used these calibrated grades to investigate the stringency of grading in STEM versus non-STEM courses at the University of Illinois. Our analysis strategy followed earlier work in using calibrated grades to compute grade offsets (actual grades minus predicted grades) to understand the extent to which an individual course grades leniently versus stringently, relative to other courses.

Our proposed methodological improvement was to use course-size weighting when computing calibrated GPAs with a logistic model. Consistent with prior work, in Research Question 1 we found that using a logistic model resulted in much better predictions of academic performance than the observed GPA. This finding confirms that observed GPA should not be used as a basis for determining whether there are biases in courses grades unless individual student course selection effects are taken into account. We further found that using course-size weighting resulted in better-calibrated calibrated GPAs, with a calibration slope of 0.89 rather than 0.62, where ideal calibration would have a slope of 1.

A promising avenue of future work would be to use calibrated grades to help students succeed in college. Currently we use observed grades to monitor student progress, but as we have demonstrated, observed grades are flawed. Incorporating grade offsets would enable academic advisors to give more accurate advice to students, and could improve systems designed to help struggling students get back on track.

In Research Question 2 we found that students in non-STEM majors are graded more leniently than students in STEM majors at this institution, with an average difference in our sample of 0.41 grade points. This

means that students choosing to major in STEM fields or choosing to take many optional STEM courses will, on average, be penalized by almost half a letter grade on their GPA for their STEM courses.

We next investigated two important sub-categories of STEM courses, namely, STEM general education courses, and "gateway" STEM courses that are the first required courses in STEM majors. Research Question 3 revealed that STEM general education courses are usually graded more leniently than STEM gateway courses at this institution, with an average grading difference of 0.42 grade points, making them essentially equivalent to non-STEM courses in their average grading stringency. On the other hand, gateway STEM courses, investigated in Research Question 4, were found to be graded as stringently as other STEM courses, but not more so. This suggests that there is no weed-out effect apparent in the grade distributions.

Finally, in Research Question 5 we found that standardized tests (ACT and SAT) have a much higher correlation with calibrated grades than with observed GPAs. This shows both that calibrated grades are better indicators of student performance than observed GPAs, and that the ACT and SAT are good predictors of course grades at this institution ($r = 0.49$ and $r = 0.37$).

This paper raises a number of immediate questions. First, to what extent are these results reflective of other institutions within the U.S. and across the world? It is important to replicate this analysis for other schools. Second, our analysis did not disaggregate students by demographic or other features. We believe that the calibrated GPA could create a more accurate predictive model of student performance, which could in turn more accurately locate factors that contribute to disparities in STEM courses and programs. Previous work has investigated the effects of gender using the "grade offset/grade penalty" approach, and it would be interesting to understand how other demographic information interacts with STEM course difficulty.

Acknowledgements
Thanks to Lin Fan and Debbie Dilman for their invaluable help in gathering and formatting the data for this analysis. Thanks to the three anonymous reviewers for their insightful comments which greatly improved this article.

Authors' contributions
J.T. and M.W. jointly generated the core idea of this work. M.W. developed the logistic grade model and wrote the software. Both J.T. and M.W. produced modeling results. J.T. performed the statistical analysis. Both J.T. and M.W. discussed the results, produced the figures, and contributed to the final manuscript. All authors read and approved the final manuscript.

Funding
This material is based on work supported by the National Science Foundation under Grant No. DUE-1347722.

Availability of data and materials
The URL for the source code is at https://github.com/jtomkin/calibrated-GPA.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Earth, Society, and the Environment, University of Illinois at Urbana-Champaign, 1301 W. Green St., Urbana, IL 61801, USA. [2]Department of Mechanical Engineering, University of Illinois at Urbana-Champaign, 1206 W. Green St., Urbana, IL 61801, USA.

Received: 28 July 2021 Accepted: 2 March 2022
Published online: 17 March 2022

References

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher, 38*(9), 665–676.
Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.
Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow education, american style: Test preparation, the sat and college enrollment. *Social Forces, 89*(2), 435–461.
Caulkins, J. P., Larkey, P. D., & Wei, J. (1996). Adjusting gpa to reflect course difficulty.
Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.
Cohen, W. D. (2000). The grade point average (gpa): An exercise in academic absurdity. *National Teaching & Learning Forum, 9*(5), 1–4.
Cotner, S., & Ballen, C. J. (2017). Can mixed assessment methods make biology classes more equitable? *PLoS ONE, 12*(12), e0189610.
Cromley, J. G., Perez, T., & Kaplan, A. (2016). Undergraduate stem achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 4–11.
Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*(4), 483–501.
Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415.
Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory stem courses. *Research in Higher Education, 53*(2), 229–261.
Gilbert, L. A., Stempien, J., McConnell, D. A., Budd, D. A., van der Hoeven Kraft, K. J., Bykerk-Kauffman, A., & Wirth, K. R. (2012). Not just "rocks for jocks": Who are introductory geology students and why are they here? *Journal of Geoscience Education, 60*(4), 360–371.
Goldman, R. D., & Widawski, M. H. (1976). A within-subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement, 36*(2), 381–390.
Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., De- Haan, R., & Wood, W. B. (2004). Scientific teaching. *Science, 304*(5670), 521–522.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
Herman, G. L., Greene, J. C., Hahn, L. D., Mestre, J. P., Tomkin, J. H., & West, M. (2018). Changing the teaching culture in introductory stem courses at a large research university. *Journal of College Science Teaching*, *47*(6), 32–38.
Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York, NY: Springer-Verlag.
Johnson, V. E., et al. (1997). An alternative to traditional gpa for evaluating student performance. *Statistical Science*, *12*(4), 251–278.
Kezim, B., Pariseau, S. E., & Quinn, F. (2005). Is grade inflation related to faculty status? *Journal of Education for Business*, *80*(6), 358–364.

Koester, B. P., Galina, B. G., & McKay., T. A. (2016). Patterns of gendered performance difference in introductory stem courses. arXiv preprint.
Ma, S., Herman, G. L., West, M., Tomkin, J., & Mestre, J. (2019). Studying stem faculty communities of practice through social network analysis. *The Journal of Higher Education*, *90*(5), 773–799.
Matz, R. L., Koester, B. P., Fiorini, S., Grom, G., Shepard, L., Stangor, C. G., et al. (2017). Patterns of gendered performance differences in large introductory courses at five research universities. *AERA Open*, *3*(4), 2332858417743754.
Mestre, J. P., Herman, G. L., Tomkin, J. H., & West, M. (2019). Keep your friends close and your colleagues nearby: The hidden ties that improve stem education. *Change: the Magazine of Higher Learning, 51*(1), 42–49.
Nering, M. L., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models*. Routledge.
Ost, B. (2010). The role of peers and grades in determining major persistence in sciences. *Economics of Education Review, 29*(6), 923–934.
Rask, K. (2010). Attrition in stem fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review, 29*(6), 892–900.
Rojstaczer, S., & Healy, C. (2016). *Gradeinflation.com: Grade inflation at american colleges and universities.* Retrieved from https://gradeinflation.com/.
Rosovsky, H., & Hartley, M. (2002). *Evaluation and the academy: Are we doing the right thing? grade inflation and letters of recommendation*. American Academy of Arts & Sciences.
Sackett, P. R., & Kuncel, N. R. (2018). Eight myths about standardized admissions testing. *Measuring success: Testing, grades, and the future of college admissions*. Johns Hopkins University Press, 13–38.
Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin, 135*(1), 1.
Sanabria, T., & Penner, A. (2017). Weeded out? Gendered responses to failing calculus. *Social Sciences, 6*(2), 47.
Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving*. Westview Press.
Soares, J. A. (2015). *Sat wars: The case for test-optional college admissions*. Teachers College Press.
Sonner, B. S. (2000). A is for "adjunct": Examining grade inflation in higher education. *Journal of Education for Business, 76*(1), 5–8.
Sonnert, G., & Fox, M. F. (2012). Women, men, and academic performance in science and engineering: The gender difference in undergraduate grade point averages. *The Journal of Higher Education, 83*(1), 73–101.
Stinebrickner, T. R., & Stinebrickner, R. (2011). *Math or science? Using longitudinal expectations data to examine the process of choosing a college major*. National Bureau of Economic Research.
Tomkin, J., West, M., & Herman, G. L. (2016). A methodological refinement for studying the STEM grade-point penalty. In *46th annual frontiers IEEE frontiers in education conference (fie 2016)*.
Tomkin, J. H., West, M., & Herman, G. L. (2018). An improved grade point average, with applications to cs undergraduate education analytics. *ACM Transactions on Computing Education (TOCE), 18*(4), 1–16.
Vanderbei, R. J., Scharf, G., & Marlow, D. (2014). A regression approach to fairer grading. *SIAM Review, 56*(2), 337–352.
Witteveen, D., & Attewell, P. (2020). The stem grading penalty: An alternative to the "leaky pipeline" hypothesis. *Science Education, 104*(4), 714–735.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com