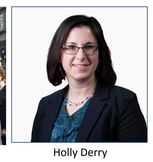
Reducing Gendered Performance Differences in Intro STEM Courses

Tim McKay, Galina Grom, Ben Koester: LSA Physics
Holly Derry, Ben Hayward: Digital Innovation Greenhouse


Tim McKay


Ben Koester (he's the big one)


Holly Derry


Ben Hayward

Average Grade Anomalies and Gendered Performance Differences

Variations in grading practice among courses and across disciplines have been widespread since letter grades were adopted. These variations affect the student experience in important ways, imposing grades anomalies: grades which differ from expectations in ways which are experienced as penalties or bonuses. To explore the nature of these signals, we define the average grade anomaly (AGA) of a course as the mean difference between grades received in this course and the grades received in others:

$$AGA = \langle Grade - GPAO \rangle_{\text{All students}}$$

When this quantity is positive, students on average received grades higher than they're used to: a grade bonus. When it is negative, students on average received grades lower than they're used to, experiencing a grade penalty. We compare AGAs for male and female students to form a measure of gendered performance difference (GPD):

$$GPD = AGA_{\text{female}} - AGA_{\text{male}}$$

When this quantity is positive, the relative performance of female students is better than that of male students. When it is negative, the relative performance of female students is worse than that of male students.

As part of the University of Michigan's Provost's Learning Analytics Task Force we have investigated patterns of AGA and GPD across thousands of courses. Focusing on large STEM courses (Figure 1), we note that first year 'prerequisite' STEM lecture courses have large AGAs, as large as -0.54 and averaging -0.41 letter grades. They also impose substantial GPDs, as large as -0.28 and averaging -0.18. Labs associated with these same fields show quite a different pattern, with AGAs averaging 0.12 (a small bonus) and GPDs which are small (averaging 0.02) and various, ranging from -0.1 to 0.06.

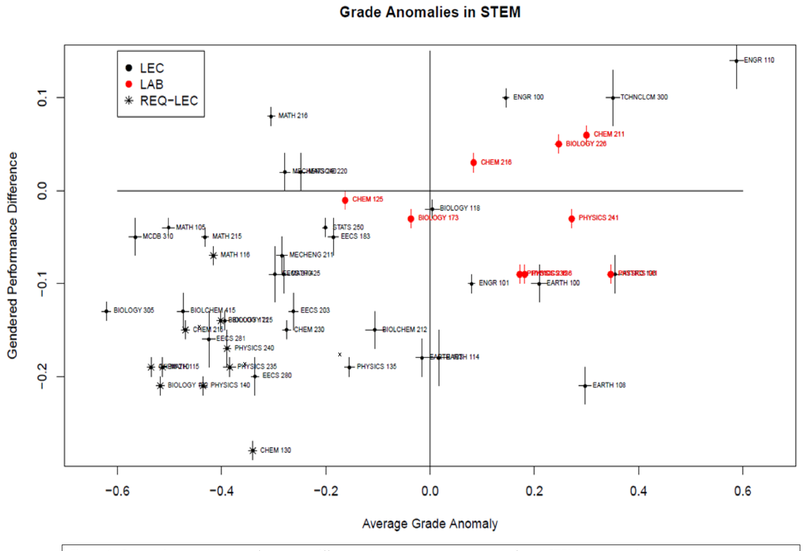


Figure 1: Comparing gendered performance difference to average grade anomaly for all STEM courses with average enrollments over 200. First year STEM lecture courses are starred – all show substantial grade anomalies and gendered performance differences. The most severe is in Chem 130, the first STEM course taken by 60% of students. First year labs are red – they show smaller grade anomalies and no average gender performance differences. AGAs, GPDs, and their errors are determined via bootstrap resampling. This is a seven year sample – errors in single year data for these classes are approximately 3x larger.

Despite generations of gradual progress, women remain underrepresented in the leadership of all STEM disciplines. The causes of this disparity are certainly various, but one important factor is the existence of gendered performance differences (GPDs) in introductory STEM courses. These GPDs persist even when accounting for various measures of prior performance, including high school GPA, standardized tests, and prior college performance. In recent years, learning analytics efforts which began at the University of Michigan have revealed a consistent pattern in these GPDs: while they are ubiquitous and substantial in lecture courses evaluated by timed examinations, they are absent in lab courses evaluated through more authentic means. The pattern observed at Michigan has now been confirmed in data from four other R1 universities.

Does the evaluative style of these courses evoke stereotype threat?

This pattern suggests that evaluative style might be responsible for these substantial gendered performance differences, rather than subject matter or intrinsic ability. We hypothesize that stereotype threat (ST) plays a central role. Stereotype threat is a well established social-psychological phenomenon. When an individual is placed in an evaluative environment in which they know others might expect them to confirm a negative stereotype, they expend some cognitive resources on this concern, modestly reducing their ability to perform. In the US, widespread gender schemas, familiar to both male and female students, express an expectation that female students will be less successful in STEM than male students. These schemas may trigger stereotype threat, reducing the performance of female students by ~10%.

The case for indicting evaluative scheme

The first year STEM courses with substantial GPDs differ in many ways. Most are taught in large lecture sections with more or less active engagement, but some (Math 115/116) are taught in small studio sections. Some are highly quantitative, others are not. The one feature they share is their use of timed, relatively mechanical examinations to determine the bulk of student grades. Labs are evaluated quite differently, focusing on more scientifically authentic work, conducted without rigid time constraints. We speculate that this difference in evaluative scheme plays a role in generating the striking pattern of gendered performance differences we observe.

Two decades of social psychological research have shown that individuals at risk of confirming negative group stereotypes in evaluative environments often underperform. They do not fall apart, but the cognitive load associated with stereotype threat impedes performance at the 5-10% level. Many factors affect the perception of threat which activates this effect. Widespread societal schemas about who succeeds in science and in high stakes testing clearly place women and some minorities at risk. Ironically, strong desire to succeed and identification with the field in question elevates risk . All of these factors are present for female students in prerequisite STEM courses.

Extensive stereotype threat research would lead us to expect group performance differences in these courses. Our analysis clearly shows that they are present. Eliminating these GPDs is the central goal of this project.

DIG Digital Innovation Greenhouse

Alleviating stereotype threat: Intervening to elicit the best performance from all students

E Coach
E² COACH
Personalized Messaging to Students

M-Write
M
Writing-to-Learn Pedagogies at Scale

Delivering interventions through ECoach and Mwrite

Since the discovery of stereotype threat in the 1990s, social psychologists have developed a variety of interventions which reduce its effects during evaluations. These include reframing attributions about struggle, affirming essential values, and emphasizing a growth mindset. In recent years, an important body of theory has emerged to explain why such brief interventions can create lasting impacts, along with early demonstrations that these interventions can be meaningfully scaled. Beginning in Fall 2016, we will conduct a large scale test of the ability of these interventions to reduce STEM gendered performance differences in field settings over a multi-year period.

To conduct these tests, we will use two tools being developed as part of the **Digital Innovation Greenhouse**: ECoach and MWrite. ECoach is a well established computer-tailored communication system, already delivering personalized feedback, encouragement, and advice to thousands of students per term. MWrite is a new toolkit, being developed in collaboration with the Sweetland Center for Writing, which supports the use of writing-to-learn methods at scale.

In collaboration with leading social psychologists, we will iteratively develop online writing interventions delivered by the combined ECoach/MWrite system, applying each to more than 15,000 students per term in an array of physics, chemistry, biology, and engineering courses. Using tailored communication, the ECoach framework will allow us to design micro-randomized trials, exploring the potentially large intervention space in real-time. By the end of this trial, we will have aggressively explored the possibilities for ameliorating the impact of stereotype threat on gendered performance differences in introductory STEM lecture courses.

SUBJECT	NBR	N/YR	M	F	AGA	GPD	GPD 1 yr err	MATCHED
BIOLOGY	171	1322	531	791	-0.40	-0.14	0.03	-0.14
BIOLOGY	172	1174	497	677	-0.52	-0.21	0.03	-0.19
BIOLOGY	305	807	378	429	-0.62	-0.13	0.03	-0.11
CHEM	130	1965	1002	963	-0.34	-0.28	0.02	-0.17
CHEM	210	1934	931	1003	-0.54	-0.19	0.02	-0.14
CHEM	215	1408	686	722	-0.47	-0.15	0.02	-0.14
CHEM	230	678	313	364	-0.28	-0.15	0.04	-0.11
EECS	280	1346	1116	230	-0.34	-0.20	0.02	-0.23
EECS	281	1048	902	146	-0.43	-0.16	0.03	-0.15
MATH	115	2217	1249	968	-0.51	-0.19	0.02	-0.11
MATH	116	1491	983	508	-0.42	-0.07	0.02	-0.05
PHYSICS	140	1252	909	343	-0.44	-0.21	0.03	-0.23
PHYSICS	240	935	741	194	-0.39	-0.17	0.03	-0.20
PHYSICS	135	613	273	340	-0.15	-0.19	0.04	-0.16
PHYSICS	235	476	230	246	-0.38	-0.19	0.04	-0.20
All		18665	10741	7924	-0.42	-0.18	0.005	-0.16

Table 1: Details for the 1st year STEM lecture courses which trigger participation in Welcome2STEM, including subject, course number, annual enrollment (and by gender), average grade anomaly, gendered performance difference, *annual* GPD uncertainty, and GPD from the optimal matching procedure described in the text. Individual course GPD's are detected in *single year data* at the 3.5σ to 14σ level. Averaging over all courses measures this GPD at more than 40σ. A 50% change in GPD would be clearly detectable in every individual course. A 10% change could be easily detected in the aggregate.