Equilibrium Grade Inflation with Implications for Female Interest in STEM Majors*

Tom Ahn	Peter Arcidiacono	Amy Hopson	James Thomas
Naval Postgraduate School	Duke University NBER & IZA	Bureau of Labor Statistics	Federal Trade Commission

February 4, 2022

Abstract

Substantial earnings differences exist across majors with those that pay well also assigning lower grades and higher workloads. We show that the stricter grading policies in STEM and economics courses reduce STEM enrollment, especially for women. To show this, we estimate a model of student demand for courses and optimal effort choices of students given professor grading policies. Professor grading policies are treated as equilibrium objects that in part depend on student demand for courses. Restrictions on grading policies that equalize average grades across classes reduce the STEM gender gap and increase overall enrollment in STEM classes.

1 Introduction

Even after accounting for selection, significant earnings differences exist across majors. Majors in engineering and the sciences, as well as in economics and business, pay substantially more than other fields, with earning disparities increasing over time (Gemici & Wiswall (2014)) and with these majors being more robust to recessions (Altonji, Kahn & Speer (2016)).[1] Despite their value in the marketplace, STEM (Science, Technology, Engineering, and Mathematics) fields are perceived to be under-subscribed, with significant government activity centered on increasing the supply of

*The views expressed in this article are those of the authors and do not necessarily reflect those of the Bureau of Labor Statistics or the Federal Trade Commission.

[1]See Altonji et al. (2012) and Altonji, Arcidiacono & Maurel (2016) for reviews of the literature on the returns to major. For international evidence, see Kirkeboen et al. (2016) and Hastings et al. (2013).

STEM workers.[2] Of particular concern is the lack of female representation in STEM occupations; women represent 47% of the workforce but 30% of STEM workers (U.S. Bureau of Labor Statistics (2019)).

Universities may be contributing to lower STEM enrollment and the gender gap in STEM by being *laissez-faire* with regard to differences in grading and study time across fields. The same majors that pay well also give significantly lower grades (Freeman (1999), Bagues et al. (2008), Rask (2010)) and are associated with more study time (Brint et al. (2012), Stinebrickner & Stinebrickner (2014)), suggesting universities are actually subsidizing students to go into low-paying majors. Lower grades and higher study times deter enrollment. Sabot & Wakeman-Linn (1991) show that the absolute level of grades was a far more important indicator of taking further courses in the subject than their ranking within the class. To the extent that women may value grades more than men, lower STEM grades may have a compounding effect for women (Rask & Bailey (2002), Rask & Tiefenthaler (2008)).

In this paper, we model both how students choose courses in different fields and how professors choose grading policies and instructor effort, paying particular attention to the role of gender. The demand-side has students choosing courses and exerting effort based on their preferences for classes and departments, costs of studying, and expected grades. Expected grades in turn depend on the optimal choice of study time given their abilities and the professors' grading policies. The supply-side has professors make choices based on their own tastes but also demand for their classes, where their grading policies consist of choosing average grades for their course as well as the returns to studying and course-specific abilities. Modeling both sides of the market allows us to examine the equilibrium effects of policies such as restricting average grades across classes, where professors may respond to such policies by changing the returns to studying and the effort they make to enhance the attractiveness of their courses.

To estimate both the demand and supply sides of the model, we use academic transcript and course evaluation data from the University of Kentucky (UK). The transcript data contains student and professor characteristics, course choices, and grades. The course evaluation data provides

[2]State legislatures have introduced bills to waive or reduce tuition or forgive student loans for STEM students who agree to teach in the state (Hinz (2019), Latek (2019), Chapman (2014)). A bill to expand STEM education, especially increasing participation of girls in computer science was signed into law (Building Blocks of STEM Act (S. 737)) (Stevens 2021). The GI Bill was expanded to support veterans seeking a STEM degree (Veteran STEM Scholarship Improvement Act) (Gross 2019). Legislation has been introduced to retain international STEM students by easing restrictions on green card issuance (Keep STEM Talent Act, STEM Jobs Act (H.R. 6429)) (Graeml 2019).

student evaluations of professors and self-reported information on hours studied. The raw data show that STEM classes are associated with grades over 0.3 points lower and average study time almost 40% higher. Women have higher grades in both STEM and non-STEM classes but make up a much smaller share of STEM enrollment.[3]

The estimates of the demand-side model help explain the lack of female representation in STEM courses. One component is comparative advantage. Females have characteristics associated with relatively higher grades and non-grade utility in non-STEM classes. Another component is that our model estimates suggest women value grades 27% more than men, amplifying the effects of comparative advantage. Identification of preferences for grades—and how they differ between men and women—comes from how students sort into classes *within* a field of study. Some classes within a particular field will reward student ability in that field more than others. To the extent that students of higher field-specific ability sort into those classes—above and beyond the fact that those with higher field-specific ability may find classes in that department attractive—reveals the value of grades. Other channels such as gender-specific study costs and preferences for departments as well as female preferences for female professors also affect the gender gap but only by modest amounts.

The estimates of the supply-side model account for the differences in grading policies and instructor effort decisions across departments. The professor model posits preferences over enrollments, grades, workloads, and effort. Professors who face low demand for their courses then have incentives to raise their grades and own effort while also lowering student workloads to attract more students; professors with high demand face the opposite incentives. Our estimates imply that a substantial share of the differences in grades between STEM and non-STEM classes is due to differences in student demand.

Finally, we use the both sides of the model to simulate a counterfactual scenario where all courses must have an average grade of a B. Holding fixed professor choices of effort and workloads, the policy increases STEM enrollment by almost 25% and shrinks the STEM gender gap by 22%. In equilibrium, however, professors respond by changing student workloads and the effort they put into instruction. Incorporating professor responses still results in a 21% increase in STEM enrollment and an 18% reduction in the gender gap. As the course choice model takes the declared major of junior and seniors as given, the long run effects of the policy would be even larger as

[3] We include economics as part of STEM as grading practices and classes sizes are closer to traditional STEM fields as is demand for the major in the labor market.

students adjust major choices. As such, a standard curve could be a low-cost and effective way to both increase overall STEM participation and decrease the gender gap in STEM.

That changing average grades across fields would have such large effects may seem surprising. If students and employers have complete information, conditional on major choice, differences in grades, especially across fields, should be largely inconsequential. But this is not true in limited information settings. Models of the effects of grades on wages typically focus on overall grade point averages as this is what traditionally appears on resumes (Piopiunk et al. (2020)). Indeed, that grade distributions are not fully known serves as an impetus for theoretical models of grade inflation resulting from competition between universities (Chan et al. (2007)), a finding that is consistent with Cornell's decision to release median course grades online but not to reveal them on the transcripts of their students (Bar et al. (2009)). Nominal grades may also matter to students for reasons beyond the traditional economic explanations, such as parental pressure or the psychological assessment of self-worth, a finding supported in the sociology and psychology literature (Rosenberg et al. (1995), Crocker et al. (2003), and Seymour & Hewitt (1997)).

Our results are consistent with prior work indicating the importance of grades to educational choices. Closest to our paper, Butcher et al. (2014) showed that a policy of restricting average grades to be no higher than a B+ in lower division classes resulted in substantial shifts towards science classes and science majors where the policy restriction was not as binding. More generally, the literature has focused on how students respond to grades either by taking the next course in the sequence or persisting in the major (Astorne-Figari & Speer (2018, 2019)). For example, McEwan et al. (2021) examines discontinuities in grades in introductory economics classes, demonstrating that those who just crossed the threshold were 18 percentage points more likely to major in economics. Much of this literature has also found that women respond more to grades (Rask & Bailey (2002), Rask & Tiefenthaler (2008), Ost (2010), Owen (2010)), though recent work suggests it may depend on the department in which the grades are received (Kaganovich et al. (2021), Kugler et al. (2021)). That women may value grades more is reflected in subjective expectation data from Zafar (2013), who shows that survey-elicited expected grades are more important in the determination of major choice for women, and Saltiel (2021), who shows that, for the same math ability, self-efficacy in math is lower for women. Our work complements the previous literature in that we model students' choices of courses, explicitly taking into account how one's abilities translate into grades differently across fields and across courses within fields. Hence our model explicitly incorporates the comparative advantage features found in Griffith (2010) who showed that performance in STEM classes relative to performance in all other classes was a key determinant in major choice.

Our paper also relates to a growing literature that empirically analyzes supply-side decision-making in higher education. For example, Epple et al. (2006) and Fu (2014) analyze how universities admit students and set tuition while Thomas (2021) examines the determinants of university course offerings. Our paper contributes to this literature by providing an empirical analysis of how grading policies are set in equilibrium. This builds upon descriptive evidence of grade inflation (Sabot & Wakeman-Linn (1991), Johnson (2003), and Babcock (2010)), policy experiments to curb grade inflation (Butcher et al. (2014) and Bar et al. (2009)), and theoretical work on grade inflation (Zubrickas (2015) and Chan et al. (2007)).

2 Data and Reduced-form Evidence

Our main dataset captures academic transcripts for all students at the University of Kentucky (UK) in the Fall of 2012. The data includes student demographics, pre-college academic measures, course enrollment, and grade information. We also have class evaluation surveys completed by students at the end of the semester that includes information on students' perceptions of the quality of the course, their expected grades, and number of hours spent per week studying for the class. Students are not identified in the evaluation forms outside of their cohort (freshman, sophomore, etc.) so we cannot link them to transcripts at the individual level.[4] We include economics and related fields as part of STEM because these majors exhibit similar patterns in earnings, course grades, and study times. We further limit the sample to courses with at least fifteen students in order to have enough observations per class for the analysis. See Appendix A.1 and Appendix Table A.1 for details on splitting departments into STEM/non-STEM and see Appendix A.3 for how courses are selected for the analysis.

Our sample yields approximately 58,000 student/class observations, with around 16,000 unique undergraduates. Table 1 provides summary statistics by gender. Women at UK have higher high school grades but lower ACT math scores. After entering college, women have substantially higher grades. The sharpest distinction between genders is in major selection. While women comprise a slight majority at UK, the ratio of men to women in STEM majors is 1.6.[5]

[4]Coverage in the evaluation data is incomplete as some departments opt out. For summary statistics and OLS analyses, we aggregate evaluation data to the class level and match to the transcript data. We match 76% of classes and restrict to classes with at least 50% response rate.

[5]The one exception is Biology, where women comprise a majority of enrollment.

Table 2 summarizes class-level characteristics by STEM-status of the course. STEM classes are almost double in size and give significantly lower grades compared to non-STEM courses. Students spend about an additional hour per week (40% more time) studying in a STEM course relative to a non-STEM course, yet earn lower grades. While women earn higher grades in STEM courses compared to men, the gap is smaller than the gap observed in non-STEM courses. This is somewhat surprising given that women study more than men (see Table 3 below) and, as we will show, increasing study time raises grades on average more in STEM courses than in non-STEM courses. These results suggest that part of the reason for the under-representation of women in STEM may be strategic sorting into non-STEM classes where women have a comparative advantage.

Table 3 presents OLS results showing the relationship between individual and class characteristics with grades and study hours after controlling for academic background measures. Selection into courses likely distorts these estimates, but these reduced form results help to describe the patterns in the data. The first column, which regresses individual course grades on characteristics, mirrors the patterns in Table 2. STEM classes are associated with lower grades and females earn higher grades even after conditioning on academic background measures. Classes with a higher fraction of female students also see higher grades. This result is consistent with the lack of a grade curve, otherwise the higher grades females receive would translate into lower grades for everyone else. It also suggests women are sorting into classes that give higher grades. Class size is negatively associated with grades, confounding two forces. On the one hand, students prefer higher grades which should lead to higher enrollments in classes that offer them. On the other hand, courses that have high intrinsic demand may give lower grades since these courses do not need to offer high grades to attract students, implying a negative relationship between grades and enrollment. The reduced-form results suggest that this latter effect dominates. The second column expands on the first by disaggregating the STEM/non-STEM classes into 14 department categories. Appendix Table A.1 shows which departments are included in which category. Results are qualitatively similar to the first set of results. The department category parameters in the first column of Appendix Table A.2 shows that while the general trend of STEM classes yielding lower grades remains intact, there is substantial variation across departments.

The third column on Table 3 shows a regression of class-level study hours on the average characteristics of the class. STEM classes are associated with about an extra half-hour of study, slightly less than what is seen in the descriptive statistics. This is consistent with STEM classes attracting students who are willing to study more and with the grading policies of STEM classes

Table 1: Descriptive Statistics by Gender

	Men	Women	p-value (H_0: Men = Women)
High school GPA	3.49 (0.472)	3.62 (0.401)	0.00
ACT Reading Score	26.1 (5.13)	26.0 (4.84)	0.37
ACT Math Score	25.7 (4.65)	23.9 (4.23)	0.00
Fall 2012 GPA	2.86 (0.938)	3.12 (0.848)	0.00
STEM Major	59.2%	37.6%	0.00
Minority	11.4%	13.6%	0.01
1st Gen	13.5%	15.0%	0.01

Note: Fall 2012 University of Kentucky undergraduate students, 7,904 men, 8,286 women. SAT scores are converted to equivalent ACT scores. 1st Gen is first-generation college students. Standard deviations in parentheses.

offering higher returns to studying. Classes with more women are associated with higher study times. Since this regression is done at the class level, we cannot separate out how much of this pattern is because women are sorting into classes where the returns to studying are higher or because women study more given the same incentives. However, both imply higher study times for women, a result consistent with survey evidence from Arcidiacono et al. (2012) and Stinebrickner & Stinebrickner (2012). Perhaps the most interesting coefficient is the one on average grades. Courses with higher grades are associated with *less* study time, suggesting grades should be interpreted as relative, not absolute, measures of accomplishment, as well as suggesting grade inflation may have negative consequences for learning. The fourth column expands STEM/non-STEM to the 14 department categories. Parameter estimates remain qualitatively similar, with the exception of Percent Female, which increases by about 75 percent. Department parameters are in the second column of Appendix Table A.2.

3 Model

Student i chooses n_i courses from the set $[1, \ldots, J]$. Let $d_{ij} = 1$ if j is one of the n_i courses chosen by student i and zero otherwise. Following Nevo et al. (2005), we assume the payoff associated with a bundle of courses is given by the sum of the payoffs for each of the individual courses where the

Table 2: Descriptive Statistics by Course Type

	STEM	Non-STEM
Class Size	80.2 (99.3)	41.4 (46.0)
Average Grade	2.94 (0.45)	3.27 (0.42)
Average Grade | Female	3.00 (0.56)	3.37 (0.43)
Study Hours	3.37 (1.42)	2.45 (0.81)
Percent Female	37.0%	58.3%
Percent Fem.Prof.	27.0%	46.4%
Percent Upper Level	40.6%	44.3%

Note: Fall 2012 University of Kentucky courses with enrollments of 15 or more students, 341 STEM courses, 743 non-STEM courses. For study hours, 327 STEM courses and 652 non-STEM courses. Standard deviations in parentheses. Testing for equality in means across STEM/non-STEM had p-values below 0.01 for all characteristics.

Table 3: Regressions of Grades and Study Time on Characteristics of the Individual and/or Class

Dependent Var.	(1) Grade		(2) Grade		(3) Study hours/week		(4) Study hours/week	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
STEM Class	-0.410	(0.010)	-		0.384	(0.097)	-	
Female	0.087	(0.009)	0.084	(0.009)	-		-	
Percent Female	0.360	(0.023)	0.416	(0.027)	0.571	(0.185)	0.731	(0.207)
Average Grade	-		-		-0.715	(0.077)	-0.689	(0.086)
ln(Class Size)	-0.032	(0.004)	-0.031	(0.005)	-0.180	(0.051)	-0.186	(0.053)
ACT Reading Score*	0.041	(0.005)	0.052	(0.005)	0.022	(0.028)	0.337	(0.161)
ACT Math Score*	0.139	(0.006)	0.132	(0.006)	0.116	(0.030)	0.300	(0.154)
High School GPA*	0.273	(0.005)	0.274	(0.005)	-0.570	(0.137)	-0.594	(0.159)
Department FE	No		Yes		No		Yes	

Note: * indicates variable is z-scored. Additional controls for grade regressions (1) and (2) include indicators for upper-level class (class number 300+), minority, freshman, first-generation, STEM major, Pell grant, in-state, % minority, % freshman, and % first-generation. There are 58,081 student/class observations. Additional controls in study hours regressions (3) and (4) include indicator for upper-level class, % minority, % freshmen, % first-generation, % STEM major, % Pell grant, and % in-state. There are 986 class observations from the evaluation data. Average Grade is calculated from the final grades for the course. Regressions (2) and (4) splits STEM/non-STEM into 14 department categories. See Appendix Table A.2 for By-Department parameters.

payoffs do not depend on the other courses in the bundle.[6] We specify the payoff for a particular course j as depending on student i's preference for the course, δ_{ij}, the amount of study effort the individual chooses to exert in the course, s_{ij}, and the course grade conditional on study effort, $g_{ij}(s_{ij})$. The individual's realized utility from choosing course j and exerting s_{ij} units of effort is given by:

$$U_{ij}(s_{ij}) = \phi_i g_{ij}(s_{ij}) - \psi_{ij} s_{ij} + \delta_{ij} \tag{1}$$

The grade student i receives in course j, g_{ij}, depends on the academic preparation of student i for course j, A_{ij}, the amount of study effort put forth by the student in the course, s_{ij}, the grading policies of the professor, and a mean-zero shock that is unknown to the individual at the time of course enrollment, η_{ij}:

$$g_{ij} = \beta_j + \gamma_j \left(A_{ij} + \ln(s_{ij})\right) + \eta_{ij} \tag{2}$$

Grading policies by the professors are then choices over an intercept, β_j, and a return to academic preparation and effort, γ_j. Gains from study effort enter as a log to capture the diminishing returns to studying. Along with the linear study effort cost defined in the utility function, this ensures an interior solution for the optimal amount of study time.

Students are assumed to know the grading policies of the professors.[7] They do not, however, know η_{ij}. Further, the cost of studying for a particular class is also partially unknown. Namely, study costs are given by:

$$\psi_{ij} = \psi_i \zeta_{ij} \tag{3}$$

where ψ_i is a student-specific (known) study cost and ζ_{ij} is the match-specific study cost revealed after the student's choice of classes. We assume that ζ_{ij} follows a log-normal distribution. Having the study costs evolve in this way facilitates the identification of the returns to studying; see Section 4.2.2.

After substituting the grading process (2) into the utility function (1), the optimal study effort

[6] For a model that includes complementarities in bundled choice, see Gentzkow (2007). The Gentzkow (2007) framework is not feasible in our setting because of the large number of potential course bundles. A natural concern with our setup is that students may balance hard courses with easier ones where now students would be choosing among all possible bundles of courses. In Appendix A.8.2 we show that our estimated model matches both the within-student distribution of high workload courses as well as the within-student distribution of STEM courses.

[7] Students have a number of formal and informal resources to learn about grading policies, be it through friends or course syllabi. Course evaluations, which show average expected grades for courses, are online and publicly available. See also Ferreyra et al. (2021) for a model of student effort as a function of college policies.

given a realization of ζ_{ij} can be found by differentiating $U_{ij}(s_{ij})$ with respect to s_{ij}:

$$s^*_{ij} = \frac{\phi_i \gamma_j}{\psi_i \zeta_{ij}} \tag{4}$$

Those who have lower study costs (ψ_i) and higher levels of academic preparation (A_{ij}) find courses with higher γ_j's relatively more attractive all else equal. Those who place a relatively high weight on expected grades (ϕ_i) study more conditional on choosing the same course.

Substituting the expression for optimal study time into the grade process equation yields:

$$g_{ij} = \beta_j + \gamma_j \left(A_{ij} + \ln(\phi_i) + \ln(\gamma_j) - \ln(\psi_i) - \ln(\zeta_{ij}) \right) + \eta_{ij} \tag{5}$$

Professors who set relatively higher values of γ_j see more study effort for two reasons. First, conditional on enrollment, higher γ_j's induce more effort. Second, all else equal, those with lower study costs are more likely to enroll in courses with higher γ_j's given the higher returns to studying.

Note that study time will be chosen optimally after the realization of the shock to study costs, ζ_{ij}, but ζ_{ij} is unknown at the time of course selection. The shock to grades, η_{ij}, is also unknown at the time when courses are chosen. Hence individuals maximize the expected utility of their course bundle taking into account their optimal response to the realizations of the ζ_{ij}'s. After substituting in the best study effort responses from (4) and the grade production process in (5) into (1) and taking expectations, the expected utility of course j can be written as:

$$\begin{aligned} \mathbb{E}(U_{ij}) &= \phi_i(E(g_{ij}) - \gamma_j) + \delta_{ij} & (6) \\ &= \phi_i \left[\beta_j + \gamma_j \left(A_{ij} + \ln[\phi_i] + \ln[\gamma_j] - \ln[\psi_i] - 1 \right) \right] + \delta_{ij} & (7) \end{aligned}$$

Students then solve the following maximization problem when choosing their optimal course bundle:

$$\begin{aligned} \max_{d_{i1},\dots,d_{iJ}} \quad & \sum_{j=1}^{J} d_{ij} \mathbb{E}(U_{ij}) & (8) \\ \text{subject to:} \quad & \sum_{j=1}^{J} d_{ij} = n_i, \;\; d_{ij} \in \{0,1\} \forall j \end{aligned}$$

where n_i is taken as given.

The key equations for estimation are then given by:

(i) the solution to the students maximization problem where (7) is substituted into (8),

(ii) the grade production process given in (5), and

(iii) the optimal study effort given in (4).

The next section describes the parameterizations used to estimate the model as well as the assumptions necessary to overcome the fact that our measures of study effort from the course evaluations are not linked to the individual's characteristics.

4 Estimation

We now describe the estimation of the model, beginning by showing how the model is parameterized. Next we show how to estimate the model when the unobserved preferences for particular courses are uncorrelated with the unobservables in the grade equation (no unobserved heterogeneity). We then extend our framework to the case where there are a finite number of unobserved types where the types affect both preferences and grades for different sets of courses.

4.1 Parameterizations

To estimate the model, we need to place some structure on course preferences, δ_{ij}, the value of grades, ϕ_i, and the cost of effort, ψ_i. Further, we must relate academic preparation, A_{ij}, to what we see in the data. We partition courses into K departments, $K < J$, where $k(j)$ gives the department for the jth course. Denote $w_i = 1$ if student i is female and zero otherwise. Denote X_i as the set of explanatory variables that affect academic preparation and cost of study effort. Denote Z_{1i} as the set of variables that affect department preferences and Z_{2ij} as the set of variables that affect the match between the student and the course. The list of variables used in X_i, Z_{1i}, and Z_{2ij} are given in Table 4. Preference shocks for courses are represented by ϵ_{ij}. We then parameterize the model as follows:

$$A_{ij} = w_i\alpha_{1k(j)} + X_i\alpha_{2k(j)} \tag{9}$$

$$\delta_{ij} = \delta_{0j} + w_i\delta_{1k(j)} + Z_{1i}\delta_{2k(j)} + Z_{2ij}\delta_3 + \epsilon_{ij} \tag{10}$$

$$\psi_i = \exp\left(\psi_0 + w_i\psi_1 + X_i\psi_2\right) \tag{11}$$

$$\phi_i = \phi_0 + w_i\phi_1 \tag{12}$$

There is no intercept in A_{ij} as it can not be identified separately from the β_j's. Note that the same variables enter into academic preparation and effort costs, only with different coefficients, and that the academic preparation variables are allowed to have different returns in each department. For example, ACT math scores may matter more to STEM classes than ACT reading scores. Preferences for courses, δ_{ij}, allow for both course fixed effects, δ_{0j}, as well as students with particular

Table 4: List of controls besides gender

	Covariates for academic preparation and cost of study effort
X_i	ACT reading, ACT math, high school grades, minority, first generation, unobserved type
	Covariates for preferences that vary by department
Z_{1i}	ACT reading, ACT math, high school grades, unobserved type
	Covariates for preferences that vary by class match
Z_{2ij}	female $\times$ female professor; freshmen and sophomore $\times$ STEM $\times$ female; (juniors and seniors) whether the course is required for the major, whether it is one of two or more courses that would fill a major requirement, whether the course is upper division; (sophomores) log number of courses opened up by taking the course, STAT212; (freshmen) log number of courses opened up by taking the course, CIS/WRD110;

Note: opened-up courses are ones where the particular course is a prerequisite; STAT212 is a statistics requirement typically taken as a sophomore; CIS/WRD110 is an English requirement typically taken as a freshman;

characteristics preferring courses in particular departments, $\delta_{1k(j)}$ and $\delta_{2k(j)}$. Note also that the effort costs are exponential in the explanatory variables, ensuring that effort costs are positive. Finally, preferences for grades are only allowed to vary by gender. In principle, we could allow them to vary with X_i as well, but this would substantially complicate both the model and identification.

Having separate estimates by gender across all the relevant parameters will help uncover some of the driving forces behind the gender gap in STEM. For example, if female intrinsic demand for courses in STEM departments is relatively low ($\delta_{1k(j)}$ negative) while preferences for grades and cost of effort are relatively equal across males and females (ϕ_1 and ψ_1 close to zero), then changing grading policies will have little effect on the gender gap in STEM. On the other hand, if females have significantly different preferences over grades and study effort than males, then altering grading policies could affect the gender composition of classes and departments.

For ease of exposition, we first describe our estimation procedure absent unobserved heterogeneity. That is, we assume that the unobserved preferences for courses (the ϵ_{ij}'s) are uncorrelated with the unobserved grade shocks (the η_{ij}'s) and study-cost shocks (the ζ_{ij}'s). Under this assumption, there is no selection problem when estimating the grade process given our controls for ability and effort costs. This would be a selection-on-observables approach. We then extend the model to incorporate an additional form of selection by controlling for unobserved heterogeneity. We do so by incorporating a finite number of unobserved types. As illustrated in Table 4, a student's unobserved type affects their departmental preferences (through Z_{1i}), departmental abilities (through X_i), and study costs (again through X_i).

4.2 Estimation without Unobserved Heterogeneity

4.2.1 Grade parameters

Substituting the parameterizations for academic preparation, A_{ij}, the value of grades, ϕ_i, and study costs, ψ_i, into (5) yields the following reduced form grade equation:

$$g_{ij} = \theta_{0j} + \gamma_j \left(w_i \theta_{1k(j)} + X_i \theta_{2k(j)} \right) + \eta^*_{ij} \tag{13}$$

where:

$$\begin{aligned}
\theta_{0j} &= \beta_j + \gamma_j (\ln(\phi_0) + \ln(\gamma_j) - \psi_0) && (14)\\
\theta_{1k(j)} &= \alpha_{1k(j)} + \ln(\phi_0 + \phi_1) - \ln(\phi_0) - \psi_1 && (15)\\
\theta_{2k(j)} &= \alpha_{2k(j)} - \psi_2 && (16)\\
\eta^*_{ij} &= \eta_{ij} - \gamma_j \ln(\zeta_{ij}) && (17)
\end{aligned}$$

We estimate the reduced form parameters $\{\theta_{0j}, \theta_1, \theta_2\}$ as well as the structural slopes, the γ_j's—both relative to a normalization—using nonlinear least squares. A normalization must be made for every department as scaling up the θ's by some factor and scaling down the γ's by the same factor would be observationally equivalent. We set one γ_j equal to one for each department, a normalization that will be undone in Section 4.2.2. Denote C_k as the normalization for department k. We then estimate γ^N_j where $\gamma^N_j = \gamma_j / C_{k(j)}$. Similarly, we estimate $\theta^N_{1k(j)}$ and $\theta^N_{2k(j)}$ where $\theta^N_{1k(j)} = \theta_{1k(j)} C_{k(j)}$ and $\theta^N_{2k(j)} = \theta_{2k(j)} C_{k(j)}$.

The variation in the data used to identify $\{\theta^N_1, \theta^N_2\}$ comes from the relationship between student characteristics and grades in each department. The variation in the data used to identify the γ^N_j's is how these characteristics translate into grades relative to the normalized courses. Key to separately obtaining γ^N_j from $\{\theta^N_1, \theta^N_2\}$ is the parameterization of course-specific abilities as a weighted index of the characteristics of the students where the weights vary at the department level (see Equation (9)). Absent this restriction, we would not be able to separate out the parameters that are choices of the professors (the γ_j's) from course-specific abilities.

4.2.2 Study effort parameters

We next turn to recovering the study effort parameters. The course evaluation data give reported study hours for each individual in the classroom and we use these study hours as our measure of effort, s^*_{ij}. Taking logs of (4) and substituting in our parameterizations of ψ_i and ϕ_i yields:

$$\ln(s^*_{ij}) = \ln(\phi_i) + \ln(\gamma_j) - \ln(\psi_i) - \ln(\zeta_{ij}) \quad (18)$$

$$= \kappa_0 + w_i\kappa_1 - X_i\psi_2 + \ln(\gamma_j) - \ln(\zeta_{ij}) \quad (19)$$

where:

$$\kappa_0 = \ln(\phi_0) - \psi_0 \quad (20)$$

$$\kappa_1 = \ln(\phi_0 + \phi_1) - \ln(\phi_0) - \psi_1 \quad (21)$$

Recall that we had to normalize one γ_j for every department in the grade equation. Substituting in $\hat{\gamma}^N_j C_{k(j)}$ for γ_j in (19) and rearranging yields:

$$\ln(s^*_{ij}) - \ln(\hat{\gamma}^N_j) = \tilde{\kappa}_0 + w_i\kappa_1 - X_i\psi_2 + \kappa_{2k(j)} - \ln(\zeta_{ij}) \quad (22)$$

where $\kappa_{2k(j)} = \ln(C_{k(j)}/C_1)$ and $\tilde{\kappa}_0 = \kappa_0 + \ln(C_1)$. Here C_1 is the normalized course for the base department.

The course evaluation data cannot be linked to the individual data on grades and academic preparation. But the evaluation data does provide information about the year in school (cohort) of the evaluator (e.g., freshman, sophomore, junior, or senior). The observations we use in estimating the study parameters are then at the class-cohort level. Letting l_i indicate the cohort of student i and averaging (22) at the class-cohort level yields our estimating equation:

$$\frac{\sum_i (l_i = l) d_{ij} \ln(s^*_{ij})}{\sum_i (l_i = l) d_{ij}} - \ln(\hat{\gamma}^N_j) = \tilde{\kappa}_0 + w_{jl}\kappa_1 - X_{jl}\psi_2 + \kappa_{2k(j)} - \ln(\zeta_{jl}) \quad (23)$$

where w_{jl}, X_{jl}, and $\ln(\zeta_{jl})$ are the averages of w_i, X_i, and $\ln(\zeta_{ij})$ for students of cohort l enrolled in course j. $\ln(\zeta_{jl})$ is unobserved and assumed to be uncorrelated with w_{jl} and X_{jl}.

Estimates of (23) allow us to recover how observed characteristics (other than gender) affect study costs, $\hat{\psi}_2$. We can also partially undo the normalization on the γ's, solving for γ's that are normalized with respect to one course rather than one course in each department. Namely, let $\hat{\gamma}^P_j = \hat{\gamma}^N_j \exp(\hat{\kappa}_{2k(j)})$. $\hat{\gamma}^P_j$ then provides an estimate of γ_j/C_1.

The last normalization—the returns on preparation and study time in the only remaining normalized course—can be recovered by relating study hours back to grades. Namely, given the estimates of (23) we can back out an estimate of $\ln(\zeta_{jl})$. We then relate the average study shock for $\{j, l\}$ to the difference between average actual and average predicted grades for students in $\{j, l\}$ to recover C_1. Expected grades for student i in course j prior to the realizations of η_{ij} and ζ_{ij} are given by:

$$E(g_{ij}) = \theta_{0j} + \gamma_j \left(w_i \theta_{1k(j)} + X_i \theta_{2k(j)} \right) \tag{24}$$

where $g_{ij} - E(g_{ij}) = \eta_{ij} - \gamma_j \ln(\zeta_{ij})$. Evaluating expected grades using the estimates of the θ's and averaging across students in the same class and cohort, we obtain:

$$\frac{\sum_i (l_i = l) d_{ij} (g_{ij} - E(g_{ij}))}{\sum_i (l_i = l) d_{ij}} = C_1 \left[\hat{\gamma}_j^P \ln(\hat{\zeta}_{jl}) \right] + \eta_{jl} \tag{25}$$

where η_{jl} is the transitory shock to grades averaged at the $\{j, l\}$ level. Since we have assumed this shock is independent of $\ln(\zeta_{jl})$, we can estimate (25) by ordinary least squares.

However, two factors will lead the estimate of C_1 to be biased downward. First, grades are relative within a class. If there is a class-specific component to ζ_{ij} then this would factor into our estimate of β_j. Therefore, we replace $\ln(\zeta_{jl})$ with $\ln(\tilde{\zeta}_{jl})$, defined as the difference between $\ln(\zeta_{jl})$ and the mean (weighted by the number of respondents) of $\ln(\zeta_{jl})$ across cohorts in class j. The regression is then:

$$\frac{\sum_i (l_i = l) d_{ij} (g_{ij} - E(g_{ij}))}{\sum_i (l_i = l) d_{ij}} = C_1 \left[\hat{\gamma}_j^P \ln(\hat{\tilde{\zeta}}_{jl}) \right] + \tilde{\eta}_{jl} \tag{26}$$

Second, we only observe hours in discrete categories, introducing measurement error in our estimate of $\ln(\tilde{\zeta})$.[8] This measurement error will bias our estimate of C_1 downward. Appendix B.1 describes how we correct our estimate of C_1 for measurement error in $\ln(\tilde{\zeta})$. Given our estimate of C_1, we obtain estimates of γ_j and $\theta_{k(j)}$ using $\hat{\gamma}_j = \hat{C}_1 \hat{\gamma}_j^P$ and $\hat{\theta}_{k(j)} = \hat{\theta}_{k(j)}^N / (\exp(\hat{\kappa}_{2j}) \hat{C}_1)$.

4.2.3 Utility parameters

We now turn to the estimation of the parameters of the utility function. Substituting expected grades given in (24) into course utility given in (6), we can express expected utility for student i choosing course j as:

$$\mathbb{E}(U_{ij}) = \delta_{0j} + w_i \delta_{1k(j)} + Z_{1i} \delta_{2k(j)} + Z_{2ij} \delta_3 + \left(\widehat{E(g_{ij})} - \hat{\gamma}_j \right) (\phi_0 + w_i \phi_1) + \epsilon_{ij} \tag{27}$$

The variation in the data that identifies ϕ_0 and ϕ_1 comes from how individuals sort based on their comparative advantage in grades. Someone who is strong in mathematics will be more likely to

[8] Measurement error was not an issue when estimating the other study parameters because it was on the dependent variable. For this last parameter, however, measurement error is on one of the regressors.

sort into classes where the returns to ability in mathematics is high. The extent to which women are more or less likely to sort into classes based on where their abilities are rewarded identifies ϕ_1.

We assume that ϵ_{ij} is distributed Type 1 extreme value. If individuals were choosing one course, estimation of the parameters in (27) would follow a multinomial logit. Students, however, choose bundles of courses. Even though the structure of the model is such that there are no complementarities for choosing particular combinations of courses, the probability of choosing a particular bundle does not reduce to the probabilities of choosing each of the courses separately.

Choice set

Having defined student i's utility for each course j, we next describe the student's choice set. While our data set contains over 1,000 classes, students are precluded from registering for a substantial fraction of these courses. To account for the modifications that arise due to academic and administrative reasons, we use information on course pre-requisites, class enrollment capacity constraints, students' histories of courses from past semesters, and their AP exam results. Accounting for these factors results in students having on average 700 courses in their choice set.See Appendix A.4 for a description of the supplemental data we collected and how it was utilized to form the choice sets.

Simulated maximum likelihood

We use simulated maximum likelihood coupled with a nested fixed-point algorithm to estimate the choice parameters. To illustrate the approach, denote K_i as the set of courses chosen by i. Denote M_i as the highest payoff associated among the non-chosen courses:

$$M_i = \max_{j \notin K_i} \{\delta_{0j} + w_i \delta_{1k(j)} + Z_{1i}\delta_{2k(j)} + Z_{2ij}\delta_3 + \left(\widehat{E(g_{ij})} - \hat{\gamma}_j\right)(\phi_0 + w_i\phi_1) + \epsilon_{ij}\}$$

Suppose K_i consisted of courses $\{1,2,3\}$ and that the values for all the preference shocks, the ϵ_{ij}'s, were known with the exception of those for $\{1,2,3\}$. The probability of choosing $\{1,2,3\}$ could then be expressed as:

$$\begin{aligned} Pr(d_i = \{1,2,3\}) &= Pr(\overline{U}_{i1} > M_i, \overline{U}_{i2} > M_i, \overline{U}_{i3} > M_i) \\ &= Pr(\overline{U}_{i1} > M_i)Pr(\overline{U}_{i2} > M_i)Pr(\overline{U}_{i3} > M_i) \\ &= (1 - G(M_i - \overline{U}_{i1}))(1 - G(M_i - \overline{U}_{i2}))(1 - G(M_i - \overline{U}_{i3})) \end{aligned}$$

where $G(\cdot)$ is the extreme value cdf and $\overline{U}_{ij}$ is the flow payoff for j net of ϵ_{ij}.

Since the ϵ_{ij}'s for the non-chosen courses are not observed, we integrate them out of the likelihood function and approximate the integral by simulating their values from the Type I extreme value distribution. Denoting M_{ir} as the value of M_i at the rth draw of the non-chosen ϵ_{ij}s and R as the number of simulation draws, the full log-likelihood function is then given by:[9]

$$\ln \mathcal{L} = \sum_i \ln \left(\left[\sum_{r=1}^{R} \prod_{j=1}^{J} \left(1 - G(M_{ir} - \overline{U}_{ij})\right)^{d_{ij}} \right] / R \right) \tag{28}$$

While in theory, one could estimate all of the choice parameters $\{\delta_{0j}, \delta_{1k(j)}, \delta_{2k(j)}, \delta_3, \phi_0, \phi_1\}$ by numerically solving for the parameter values that maximize Equation (28), the large number of courses makes doing so computationally challenging. To circumvent this issue, we nest a fixed-point algorithm within the maximization routine that matches estimates of course-specific intercepts δ_{0j} directly to data on enrollment shares in the spirit of Berry et al. (1995). This approach imposes that the predicted enrollment shares exactly match the observed enrollment shares. The details of the fixed-point algorithm can be found in Appendix B.3.

Recovering the remaining structural parameters

The remaining structural parameters from the study effort estimation, Equation (23), are the study cost intercept, ψ_0, and the (relative) female study costs, ψ_1. These can be recovered using:

$$\begin{aligned} \hat{\psi}_0 &= \ln(\hat{C}_1) + \ln(\hat{\phi}_0) - \hat{\tilde{\kappa}}_0 \\ \hat{\psi}_1 &= \ln(\hat{\phi}_0 + \hat{\phi}_1) - \ln(\hat{\phi}_0) - \hat{\kappa}_1 \end{aligned}$$

The remaining structural parameters of the grade equation, Equation (13), are the course intercepts, β_j, and the female ability parameters $\alpha_{1k(j)}$. These can be recovered using:

$$\begin{aligned} \hat{\beta}_j &= \hat{\theta}_{0j} - \hat{\gamma}_j(\ln(\hat{\phi}_0) + \ln(\hat{\gamma}_j) - \hat{\psi}_0) \\ \hat{\alpha}_{1k(j)} &= \hat{\theta}_{1j} - \ln(\hat{\phi}_0 + \hat{\phi}_1) + \ln(\hat{\phi}_0) + \hat{\psi}_1 \end{aligned}$$

4.3 Estimation with Unobserved Heterogeneity

We now consider the case when one of the components of X_i and Z_{1i} is unknown to take into account correlation across outcomes for the same individual. We assume that this missing component takes on S values where π_s is the unconditional probability of the sth value. Types are identified through

[9] Our setup is similar in spirit to Nevo et al. (2005). The estimator in Nevo et al. (2005) randomly samples rankings of chosen options, computes likelihood contributions conditional on rankings, and averages across sampled rankings to simulate full likelihood. We simulate the stochastic utility of the best non-chosen course, compute likelihood contributions conditional on this stochastic utility, and average across simulation draws to simulate full likelihood.

the correlation of student grades in each of their chosen courses as well as the probabilities of choosing different course combinations after conditioning on the observables. In practice, we set S to three.

Including unobserved types helps account for selection into courses based in part on student abilities not captured by our observed measures. For example, there may be a type that has especially high grades in STEM courses and as a result will choose more STEM courses. But this type may also have a preference for STEM courses on top of any grade effect and this too is accounted for through the type component of Z_{1i}.

Integrating out over this missing component removes the additive separability of the log likelihood function suggesting that the estimation of the three sets of parameters (grades, course choices, and study time) can no longer be estimated in stages. However, using the insights of Arcidiacono & Jones (2003) and Arcidiacono & Miller (2011), it is possible to estimate some of the parameters in a first stage via a modified Expectations Maximization (EM) algorithm. Our full estimation procedure is described in Appendix B.2.

4.4 Discussion

Our model of course choices requires a number of assumptions. We discuss a few of these here as well as what patterns in the data we will miss if these assumptions are violated. In particular, we consider our assumptions regarding students' information about the grading processes, course bundling, and the way gender impacts course preferences.

We have assumed that students know the course-specific intercepts (β_j's) and slopes (γ_j's) for each class. The model can incorporate some limited forms of measurement error. For example, if a student is equally overoptimistic in all courses then the overoptimism will cancel out in the choice problem; it does not affect the utility ranking of the courses. If all students are overly optimistic about grades in particular departments then this too will not affect our estimates of the value of grades, even if this overoptimism varies between men and women as these will be captured in the course fixed effects (δ_{0j}) and female departmental effects ($\delta_{1k(j)}$). In the latter case, however, we will mistakenly attribute some of female departmental effects to differential optimism.

But we might expect differential uncertainty over the grading process as some students will have had more time to learn about the processes or are better networked. If this were the case, we would expect upperclassmen to sort better into classes that reward their abilities than lowerclassmen. We might also expect women to be more informed about the grading processes of classes that are taken by more women. Given the model estimates, we can simulate course choices and test whether certain groups have higher or lower expected grades from their simulated choices than what would be expected from their actual choices. For example, if upperclassmen are more informed that lowerclassmen, then upperclassmen (lowerclassmen) will have higher (lower) expected grades for their actual choices than for their model-simulated choices. Appendix A.8.1 spells out how we test for differential sorting and provides evidence that our model is able to match the sorting patterns seen in the data.

Simulating course choices from the model can also be used to evaluate whether we are missing key features of the data regarding how students bundle courses. Namely, we can examine whether the model overpredicts the degree to which students take only classes that have high workloads or only classes that have low workloads. Similarly, given that STEM classes are associated with higher workloads, we can examine whether the model misses on how much students are balancing STEM classes with non-STEM classes. Appendix A.8.2 shows the results of this exercise, with the results suggesting our model is able to match the bundle of courses students take despite not allowing for interactions between the courses. While this may be surprising, Appendix A.8.2 also shows that study times are quite low at the University of Kentucky implying that balancing workloads may be of limited importance to these students.

Finally, although our model includes a variety of mechanisms to explain the gender gap in STEM, there may be features of key STEM courses that are deterrents for women that are not captured by the model. In Appendix A.8.3 we again use the model to see how well it predicts share female in various subsets of classes such as those required for a STEM major and with an enrollment of at least one hundred students. The simulated and actual share female are quite close in all cases. It may also be the case that women perform better in classes in taught by women, a feature not allowed for in our model. We show in Appendix A.8.3 that women do indeed perform slightly better in classes taught by women but the effect is very small at less than 0.006 grade points.

5 Demand-side Estimates

5.1 Preference estimates

Table 5 presents a subset of the preference parameters. Recall that the parameter on expected grades is identified from within-department variation in how abilities are rewarded in different classes. While both men and women value grades, women derive substantively higher utility from higher grades. The results show that women value grades 27% more than men. Female students prefer classes with female professors, with the estimate suggesting that women would have to be compensated about one-tenth of a grade point for the same class to be taught by a male professor.[10]

The first column of the second panel of Table 5 shows female preferences (relative to male preferences) for different departments, with the omitted category being Agriculture. The largest difference in preferences is between Engineering and Biology: 1.45, which translates into about 1.2 grade points. This helps to account for the severe under subscription to Engineering courses by women, where female share of enrollment is at 18 percent. No other category has female shares under 37 percent. Engineering is thus an outlier, with all the other category gaps at 0.898 or smaller ($\leq$ 0.77 grade points). Biology is also unique in that it is both in a STEM area as well as the category most preferred by females. Women are willing to accept 0.27 grade points lower to take a Biology course, compared to the next-preferred category, Psychology.

It is worth emphasizing that this represents preferences purged of grade considerations and sorting on other ability factors.[11] Education & Health, where women make up almost 70% of course enrollments, is shown to be similarly preferred to Chemistry & Physics, where women make up less than half of enrollments (See Table 8). The primary driver of women into Education & Health over Chemistry & Physics is the difference in grades and the matching of observed characteristics to characteristics of the department which we discuss next.

The second to fourth columns of Table 5 show how non-grade preferences for classes in particular departments vary by academic characteristics. The coefficients on ACT read, ACT math, and HS GPA show how these characteristics push or pull students into classes in particular departments relative to the omitted department, Agriculture. The most salient result is the strong positive correlation between ACT math scores and preferences for STEM departments.[12] Since men at Kentucky have higher ACT math scores (and lower HS grades) than their female counterparts, this too contributes to more men choosing STEM classes above and beyond the fact that higher ACT

[10]The coefficient may be biased upward due to aggregation of departments. To the extent female professors are more likely to be in departments females prefer and variation exists in aggregated groups, we may be capturing within-group department preferences (Carrell et al. (2010), Rask & Bailey (2002), Hoffmann & Oreopoulos (2009)).

[11]See Jacob et al. (2013), Jiang (2019), Kaganovich et al. (2021), Wiswall & Zafar (2015), and Zafar (2013) for other examples in the literature exploring non-grade preferences for departments or majors.

[12]The one exception is Mathematics, which may be due in part to students with lower mathematics skills being required to take additional remedial classes to satisfy general university requirements.

math scores are especially rewarded in the grading policies of STEM classes.

Table 5: Estimates of Preference Parameters

Preference for:		Coef.	Std. Error	
Expected grades (ϕ)		0.925	(0.007)	
Female x expected grade		0.248	(0.010)	
Female x female professor		0.106	(0.008)	
Departments	Female	ACT read*	ACT math*	HS GPA*
Biology	0.520	-0.231	0.057	-0.116
Psychology	0.209	-0.355	0.030	-0.193
English	0.129	0.033	-0.244	0.049
Education & Health	0.129	-0.335	-0.002	0.078
Chem & Physics	0.082	-0.138	0.152	-0.178
Mgmt. & Mkting.	-0.052	-0.222	0.138	-0.083
Regional Studies	-0.078	-0.218	-0.089	-0.018
Mathematics	-0.103	-0.175	-0.208	-0.194
Languages	-0.134	-0.101	0.041	-0.157
Social Sciences	-0.299	-0.179	-0.033	-0.181
Communications	-0.315	-0.252	0.040	-0.068
Econ., Fin., Acct.	-0.378	-0.286	0.128	-0.043
Engineering	-0.927	-0.298	0.494	0.074

Note: Specification also includes study costs, class-specific intercepts, parameters on university and major requirement courses, unobservable types, a measure of the "option value" of the course (log number of new courses that taking this course allows one to take), and coefficient on upper/lower-classmen cross level of the course and cross STEM cross gender. See Appendix Table A.5 for complete results. $*$ indicates variable is z-scored. *Female* are female non-grade preference for departments, compared to males. Department preferences are relative to Agriculture. STEM departments are bolded.

5.2 Study effort estimates

Estimates of the study cost parameters are presented in Table 6.[13] Women have 5.5% lower study costs than men, though the estimate is insignificant. Conditional on the same observed characteristics and taking the same class, women study almost 30% more than men (a result that is statistically significant), but our estimates of ϕ_0 and ϕ_1 imply that over 80% of their increased studying is due to preferences for grades.

The second set of columns shows how the returns to study effort vary across classes, taking the median γ class for each course grouping. The heterogeneity is quite large, with classes in STEM departments having the largest returns to studying. Doubling study effort would translate into one-third of a grade point increase in Engineering but would be less than half as effective in increasing grades in Education & Health.

5.3 Grade estimates

The estimated department-specific ability weights, the α's, are given in Table 7. These are calculated by taking the reduced-form θ's, undoing the normalization on the γ's, and subtracting off the part of the component that reflects study time (taken from ψ). The departments are sorted such that those with the highest female estimate are listed first.

The female coefficients shown in the first column suggest that women have a comparative advantage in non-STEM departments after accounting for differences in test scores and high school grades. This result makes sense in the context of the descriptive statistics presented in Table 2. Namely, women have higher grades than men in both STEM and non-STEM classes though the gap is smaller in STEM classes. Given that the returns to studying are higher in STEM classes and women study more than men, we would expect women to substantially outperform men in STEM classes absent women having a comparative advantage in non-STEM courses.

The second through fourth columns show the ability weights on the two components of ACT and high school grades. The returns to the different components of the ACT score are intuitive. The five departments classified as STEM have the five highest returns in ACT math score, with the highest return to ACT math found in math classes. Higher returns to ACT reading are found in (non-econ) Social Sciences, Psychology, English, and Languages.

[13] Due to incomplete coverage in the evaluation data, we make a number of cuts to the data for this analysis. For a detailed description of our sample selection rules, see Appendix A.3. For robustness to alternative sample selection rules, see Appendix A.5.

Table 6: Estimates of Study Costs and Departmental Returns to Studying

	Study Effort			Median γ
	Coef. (ψ)	Std. Error	Department	Coef.
Female	-0.055	(0.123)	**Engineering**	0.339
ACT Reading Score*	-0.022	(0.073)	**Mathematics**	0.333
ACT Math Score*	0.051	(0.086)	**Biology**	0.309
High school GPA*	-0.044	(0.085)	**Chem & Physics**	0.308
			Econ., Fin., Acct.	0.306
			Psychology	0.292
			Regional Studies	0.290
			English	0.270
			Communications	0.267
			Languages	0.247
			Social Sciences	0.220
			Agriculture	0.182
			Mgmt. & Mkting	0.166
			Education & Health	0.160

Note: * indicates variable is z-scored. Study costs also depend on minority, first generation, and unobserved type. STEM departments are bolded. Departments are sorted by their median value of γ.

These results stand in contrast to the reduced form results presented in Table 3 which suggested that test scores were relatively unimportant compared to high school grades in predicting college grades. The results in Table 7 suggest a more nuanced picture, with test scores—especially math test scores—being more important than high school grades in all STEM fields besides Biology, fields that are also characterized by lower average grades.

With the estimates of the grading equation, we can calculate expected grades for students in a representative class for each department. We create a representative class by taking an enrollment-weighted average of the β's and γ's for each department. We calculate expected grades in these representative classes separately by gender for two groups of students. The first group is unconditional where we look at all students and weight by the total number of classes each student takes. The second conditions on taking classes in the department where the weights are now given by the number of classes taken in the particular department.

Table 7: Estimates of Department-Specific Ability Weights (α)

	Female	ACT read*	ACT math*	HS GPA*
Education & Health	0.446	0.175	0.357	0.534
Regional Studies	0.245	0.080	0.532	0.686
Communications	0.188	0.137	0.126	0.568
Agriculture	0.153	0.238	0.555	0.893
Psychology	-0.077	0.386	0.445	0.683
English	-0.102	0.270	0.378	0.707
Languages	-0.167	0.265	0.416	0.605
Social Sciences	-0.204	0.432	0.346	0.789
Math	-0.243	-0.035	1.291	0.676
Mgmt. & Mkting	-0.305	0.173	0.382	0.731
Biology	-0.427	0.200	0.633	0.681
Engineering	-0.457	0.027	0.773	0.298
Econ., Fin., Acct.	-0.497	0.150	0.822	0.646
Chem. & Physics	-0.606	0.056	1.006	0.792

Note: $*$ indicates variable is z-scored. STEM departments are bolded. Departments sorted by female α. See Table A.6 for corresponding standard errors.

Results are presented in Table 8 and are sorted based on female unconditional grades that are shown in the first column. Four patterns stand out. First, there is positive selection into STEM courses: generally those who take STEM classes perform better than the average student. This is the not the case for all departments. Indeed, the second pattern is that negative selection is more likely to occur in departments with higher grades. Third, women are disproportionately represented in departments that give higher grades for the average student. Of the five departments that give the lowest grades—all of which are in the STEM umbrella—females are under-represented relative to the overall population in all but one, Biology. Finally, and consistent with Table 7, women have a comparative advantage in non-STEM courses. In all non-STEM courses, the unconditional expected grades for women are higher than those of men, in part because women study more. For STEM courses, the unconditional expected grades are higher for men, with the exception of Biology.

Table 8: Expected GPA for Average Classes By Department, Unconditional and Conditional on Taking Courses in that Department

	EGPA Females Unconditional	EGPA Females Conditional	EGPA Males Unconditional	EGPA Males Conditional	Share Female
Education & Health	3.52*	3.47	3.34	3.23	0.69
Communications	3.46*	3.38	3.25	3.13	0.56
Agriculture	3.37*	3.21	3.23	2.87	0.57
Mgmt. & Mkting	3.24*	3.36	3.16	3.28	0.50
Languages	3.23*	3.25	3.13	3.09	0.55
Regional Studies	3.21*	3.28	3.00	3.06	0.66
Social Sciences	3.12*	3.08	3.02	2.86	0.50
English	3.11*	3.09	2.98	2.96	0.65
Psychology	2.98*	2.95	2.85	2.71	0.67
Econ., Fin., Acct.	2.67	2.89	2.70	2.87	0.38
Biology	2.59	2.75	2.59	2.73	0.60
Math	2.56	2.58	2.58	2.65	0.47
Engineering	2.54*	2.83	2.60	2.93	0.18
Chem. & Physics	2.33*	2.55	2.43	2.69	0.47
Overall					0.54

Note: ∗ denotes statistically significant differences between female and male unconditional EGPA at the 5% level. Comparisons across unconditional and conditional EGPA for each gender separately were different for every department at the 5% level. 'Share Female' is % of enrollment in courses offered in the department that is female. 'Unconditional' represents the avg. grade outcome assuming that the entire student population enrolls in the course. STEM departments are bolded. Departments are sorted by female unconditional grades. See Table A.7 for corresponding standard errors.

5.4 Drivers of the STEM gap

Given the estimates of the student's choices over classes and effort and the estimates of the grading process, we now examine the sources of the male-female gap in choices of STEM classes. We focus our attention on freshmen and sophomore students because junior and seniors have already chosen their majors. Juniors and seniors can still change their choices in counterfactual simulations; however, changing characteristics of men and women would likely also lead to changes in majors, which is beyond the scope of this paper.[14] In all simulations, we change female parameters or characteristics to match male parameters or characteristics.[15]

Table 9 shows the share of classes taken in STEM for females and males who are freshmen and sophomores, as well as showing how these shares change for women as we change different characteristics.[16] We also report the difference between the male and female shares as a measure of the gender gap in STEM participation. The first two rows of Table 9 show that our model matches the data well. The model-predicted shares of STEM classes for men and women are 53.2% and 41.2%, respectively. The12.0 percentage point gap between the two shares is what we use as our baseline when comparing the drivers of the STEM gender gap.

The third row shows predictions when female preferences for grades are changed to be the same as male preferences for grades (ϕ_1 is set to zero). Equalizing grade preferences increases the share of classes women take in STEM to 45.1%. This reduces the gender gap in STEM by almost a third. This reduction arises both because STEM courses have lower grades and because women have a comparative advantage in non-STEM courses; lowering the value of grades weakens the importance of this comparative advantage.

The fourth and fifth rows change observed and unobserved abilities so the distribution is the same for men and women. The observed abilities affect both grades and the non-grade department-specific matches. Because men are relatively stronger on the math ACT and this makes STEM classes more attractive both through grades and through the department-specific match, equalizing observed abilities reduces the gender gap by 2.9 percentage points. The fifth row shows even stronger effects from equalizing unobserved ability, reducing the gender gap by 4.1 percentage points. We find that women have a comparative advantage in non-STEM courses, beyond what is associated with observable characteristics such as test scores. Because women value grades more, removing these relative advantages makes STEM coursework significantly more attractive to women and reduces the gender gap accordingly.

[14]Juniors and seniors only change their choices in these partial equilibrium counterfactual scenarios because counterfactual choices by freshmen and sophomores alter which courses are capacity constrained. However, these changes are generally very small because most juniors and seniors register before freshmen and sophomores and thus are not exposed to the effects of freshmen and sophomore choices on capacity constraints.

[15]Similar to Juniors and Seniors, male Freshmen and Sophomores only change their choices in these counterfactual scenarios because counterfactual choices by female Freshmen and Sophomores alter which course are capacity constrained. These effects are generally quite small so we omit them for the sake of brevity.

[16]Our counterfactual simulations hold the utilities of courses with $\gamma < 0.01$ fixed. See Appendix B.4 for how the counterfactual choice probabilities are calculated in the presence of capacity constraints.

Table 9: STEM Enrollment for Freshmen and Sophomores in Counterfactual Scenarios (Partial Eq)

	STEM Enrollment Share		
	Female	Male	STEM gap†
Data	40.9%	53.3%	
Baseline Model	41.2%	53.2%	12.0
Equalize grade preferences	45.1%		8.1
Shift obs. abil. incl. abil. tastes	44.1%		9.1
Shift unobs. abil. in grades	45.3%		7.9
Equalize unobs. pref. for depts.	41.4%		11.7
Female professor effect turned off	41.4%		11.8
Grade Around a B			
Same grades in all courses:◇ $\gamma_j = 0$	62.5%	65.3%	2.7
Same γ in all courses: $\gamma_j = 0.251$	60.9%	67.4%	6.5
γ_j's fixed at estimated values	56.2%	63.9%	7.7

Note: †: Female preference and ability parameters are adjusted to be identical to male preferences and abilities.
◇: 'Same grades in all courses' sets grades and γ's to be the same in all courses which is equivalent to turning off the effects of grades in course utility.

The next two counterfactuals (rows six and seven) equalize female unobserved preferences for departments and removes female preferences for female instructors respectively. Both changes reduce the gender gap, but only by a small amount (0.2 percentage points). Overall, we find non-grade preferences that are not already accounted for through other background measures are relatively unimportant to the STEM gap. However, the small effect of equalizing unobserved preferences for departments masks larger movements within STEM, lowering female participation in Biology and raising it in Engineering and Economics.

The final set of rows in Table 9 fix average grades in each course to a B under three partial equilibrium settings. First, we do this under the case where the γ_j's are set to zero. This is equivalent to removing grades from the utility function altogether; sorting on abilities still matters, but only through preferences. The results in the third-to-last row show substantial increases for both men and women in the share of their courses in STEM. That both groups see substantial increases is reflective of the much lower grades and higher effort demanded in STEM courses. But the results are especially large for women, increasing the share of courses in STEM by more than fifty percent and reducing the gender gap to 2.8 percentage points. This occurs because women value grades more than men and this effect is amplified further because of their comparative advantage in non-STEM courses.

The second scenario in the second-to-last row again fixes average grades in each course to a B but now assigns γ_j's to the median value across all courses (0.251). The β_j's are then adjusted to equalize average grades across courses. Relative to the case where the γ_j's are set to zero, the STEM share for females (males) falls (rises). This is again reflective of female comparative advantage in non-STEM subjects. The resulting gender gap is 6.5 percentage points. The third scenario (last row) fixes the γ_j's at their estimated values. Since STEM departments have higher γ's, this leads to lower STEM shares for both men and women, though the drop is larger for women. The estimated gender gap in this scenario is 7.7 percentage points, still substantially lower than the initial gap of 12 percentage points.

6 Equilibrium Grading Policies

We now turn to the supply-side, examining how instructors set grading policies. In Section 5, we showed that grading policies differ significantly across departments. In particular, STEM courses generally have lower average grades but higher returns on effort, γ_j, compared to non-STEM courses. One principle goal of this paper is to analyze how these grading differences influence course choices and the implications for the gender gap in STEM.

However, this finding also prompts an additional question: Why do grading policies vary across courses? In particular, why do STEM courses have lower average grades but higher returns on effort than non-STEM courses? Understanding how professors choose grading policies is crucial to anticipating equilibrium responses to changes in the environment. For example, increasing STEM preparation in the hopes of increasing the number of STEM majors may be partially undone by how professors change their grading policies in response to the new environment.

The model we develop allows for grading policies to arise from differences in intrinsic demand of students. Heterogeneity in non-grade preferences δ_{ij} and abilities A_{ij} imply that some courses will be more popular than others even with homogeneous grading policies. These differences in intrinsic demand imply that the relationship between grading policies and the composition and outcomes of enrolled students differs across courses. A professor teaching an intrinsically popular course will need to grade especially harshly to achieve the same class size as a less popular course with average grading standards.

In addition to their grading policies, professors may be able to directly influence the demand for their course through exerting effort. Namely, we decompose the course fixed effect in the student's utility function, δ_{0j}, into intrinsic demand, δ^*_{0j}, and the effort of the professor, τ_j:

$$\delta_{0j} = \rho\tau_j + \delta^*_{0j} \tag{29}$$

where ρ measures how professor effort translates into course utility. We discuss our measure of effort in the next section. We also note that we do not have a way to recover ρ. As a result, we estimate the model under a variety of assumptions regarding its value.

Because grading and effort policies in all courses affect the choices of students, the composition of students in each course depends on the grading and effort policies of all professors. This general equilibrium feature means that each professor's optimal grading policy depends on the grading policies of all other professors. We assume professors do not collude when choosing grading policies implying policies are set in a non-cooperative game between professors.

We specify a professor's objection function and solve for parameter values that explain why the observed grading policies were optimal for professors. First, we estimate grading policy parameters and student preference parameters using the methods described in Section 4. Second, we derive the set of first order conditions which describe a pure-strategy equilibrium to the non-cooperative grade and effort policy setting game. This system of first order conditions describes how professor preference parameters, grading and effort policy parameters, and student parameters relate to one another when all other professors are setting grading and effort policies optimally. Finally, we solve for professor preference parameters that satisfy the set of first order conditions given estimates of the professor policy parameters and student preference parameters.

6.1 Measuring Professor Effort

To extract a measure of professor effort, we use three questions that were part of the class evaluations. Namely, did the instructor (1) present the material effectively, (2) stimulate interest in the subject, and (3) stimulate me to read further beyond the class? We average these three measures to create a student i's perception of professor effort in course j, $\tau^{(1)}_{ij}$.

There are at least two issues with using this average as a measure of effort. First, professors that give high grades may receive better evaluations because of the high grades rather than because of the effort exerted by the professor.[17] We are able to purge the effort measure of grade effects because the evaluation data contains the expected grade of each student filling out the evaluation. Using evaluation data across multiple semesters (Fall 2011 to Spring 2013), we regress $\tau_{ij}^{(1)}$ on a course fixed effect and dummy variables for each expected grade. The course fixed effect, $\tau_j^{(2)}$, gives us a measure of effort purged of the effect from offering high grades. The results from this regression are given in the top half of Appendix Table A.10 and show that higher expected grades are associated with better evaluations.

The second issue is that, conditional on the same amount of effort, some instructors may be better in the classroom than others. Since we are interested in discretionary effort rather than fixed instructor ability, we purge our effort measure of instructor effects, using multiple semesters of the evaluation data. To do so, we regress $\tau_j^{(2)}$ on an instructor fixed effect (taking advantage of the panel nature of the data) and log enrollment. The regression results in the bottom half of Appendix Table A.10 show that the coefficient on log enrollment is large and negative, implying that perceived quality of the class is lower when enrollment is high given the same instructor. We then subtract off the instructor fixed effect but leave in the effect of log enrollment: effort should be correlated with log enrollment if it is responding to characteristics of the class.[18] We then standardize this variable to have mean zero and standard deviation one. It is this standardized variable that we use for τ_j.

6.2 The Professor's Problem

We assume professors choose grading policy parameters β_j and γ_j as well as their effort, τ_j, to maximize an objective function that depends on (i) the number of students in their class, (ii) grades given in the course, (iii) the cost of assigning work (γ), and (iv) effort in the course that directly affects demand (τ). In particular, we specify the professor's objective function to penalize deviations from ideal log enrollment, e_{0j}, as well as the professor's ideal grade for the average student in the class, e_{1j}, ideal workload, e_{2j}, and ideal effort, e_{3j}, and where these ideals depend on observed and unobserved characteristics of the professor. This way of expressing the objective function is equivalent to a quadratic in each of the individual terms.

We specify the objective function this way in part because these are the measures we observe

[17] See Insler et al. (2021), Nelson & Lynch (1984), Zangenehzadeh (1988) who also find this positive relationship.

[18] There is potential for bias in this estimated effect of enrollment due to reverse causality. We explore this in detail in Appendix B.5. To summarize, we find that any potential bias would be economically insignificant.

in the data. Preferences over enrollments and workloads may relate to learning outcomes the professor values, but also impose time costs on the professor through increased student interaction as well as developing and grading (or supervising teaching assistants in administering and grading) assignments and exams. The professor may exert effort to affect course demand, but this too comes at a time cost. Having the professor directly value class grades (beyond their impact on enrollment) may reflect departmental norms that the professor would prefer not to violate. Further, giving too many low grades may result in student complaints and may increase demands on the professor's time.

Denote $\overline{G}_j(\beta, \gamma, \tau)$ as the expected average grade in class j given the vector of grading policies and effort choices for all courses β, γ, and τ. The dependence on β, γ, and τ comes through the composition of the students that take the course. Denote $P_{ij}(\beta, \gamma, \tau)$ as the probability i takes course j given the vector of grading policies. $\overline{G}_j(\beta, \gamma, \tau)$ and log enrollment in course j are given by:

$$\overline{G}_j(\beta, \gamma, \tau) = \beta_j + \gamma_j \left[\frac{\sum_i^N P_{ij}(\beta, \gamma, \tau) \left[A_{ij} + \ln(\phi_i) - \ln(\psi_i)\right]}{\sum_i^N P_{ij}(\beta, \gamma, \tau)} + \ln(\gamma_j) \right] \quad (30)$$

$$\ln\left[E_j(\beta, \gamma, \tau)\right] = \ln\left[\sum_i^N P_{ij}(\beta, \gamma, \tau) \right] \quad (31)$$

Then the objective function professor j maximizes is:

$$V_j(\beta, \gamma, \tau) = -\left(\ln\left[E_j(\beta, \gamma, \tau)\right] - e_{0j}\right)^2 - \lambda_1 \left(\overline{G}_j(\beta, \gamma, \tau) - e_{1j}\right)^2 - \lambda_2(\gamma_j - e_{2j})^2 - \lambda_3(\tau_j - e_{3j})^2 \quad (32)$$

where the coefficient on ideal log enrollment is normalized to one.[19]

Professors choose grading and effort policies given different innate demand for their courses. Absent the first term, professors would set γ_j to e_{2j} and set τ_j to e_{3j}. Given γ_j, they would then set β_j such that expected grades would equal e_{1j}. But with the first term, professors deviate from their ideal grades, workloads, and effort to mitigate the costs associated with having classes that are not the ideal size. If demand for a course would be above (below) e_{0j} when grades and effort were set to their ideal levels, professors adjust grades and effort downward (upward) to move enrollment closer to the ideal.

[19]The coefficient on one of the squared terms must be normalized to identify the model. As normalizing one of these coefficients to one is a monotonic transformation of the underlying utility function, the normalization has no implications for the counterfactual policy analysis.

6.3 Estimation

We use the first order conditions of the professor's objective function to form our estimating equations. After dividing by two, these are given by:[20]

$$0 = -\left(\ln\left[E_j(\beta,\gamma,\tau)\right] - e_{0j}\right)\frac{\partial \ln E_j}{\partial \beta_j} - \lambda_1\left(\overline{G}_j(\beta,\gamma,\tau) - e_{1j}\right)\frac{\partial \overline{G}_j}{\partial \beta_j} \tag{33}$$

$$0 = -\left(\ln\left[E_j(\beta,\gamma,\tau)\right] - e_{0j}\right)\frac{\partial \ln E_j}{\partial \gamma_j} - \lambda_1\left(\overline{G}_j(\beta,\gamma,\tau) - e_{1j}\right)\frac{\partial \overline{G}_j}{\partial \gamma_j} - \lambda_2(\gamma_j - e_{2j}) \tag{34}$$

$$0 = -\left(\ln\left[E_j(\beta,\gamma,\tau)\right] - e_{0j}\right)\frac{\partial \ln E_j}{\partial \tau_j} - \lambda_1\left(\overline{G}_j(\beta,\gamma,\tau) - e_{1j}\right)\frac{\partial \overline{G}_j}{\partial \tau_j} - \lambda_3(\tau_j - e_{3j}) \tag{35}$$

We allow for heterogeneity across professors in their preferences through the e_{lj}'s, specifying e_{lj} as:

$$e_{lj} = W_{lj}\Psi_l + \varepsilon_{lj} \tag{36}$$

where ε_{lj} are unobserved professor-specific tastes for the lth outcome. Since there are three first order conditions and four ε terms, we set ε_{0j} to zero. When l refers to enrollment, we specify W_{lj} to include a constant term and whether the courses is upper division, the latter allowing for instructors to prefer lower enrollments when the class is upper division. When l refers to grades or workload, we specify W_{lj} to include course category fixed effects, rank of the instructor, whether the instructor is female, and whether the course is upper division. Whether the instructor is female and whether the course is upper division are also interacted with STEM. This specification allows for differences across departments in the norms regarding grading policies as well as for heterogeneity in the costs of assigning and grading work due to the nature of the subject material. It also allows for incentives to differ depending on the rank of the instructor. Adjunct instructors may face incentives to offer higher grades than those who have permanent contracts. When l refers to effort, W_{lj} is a constant term given that our effort measure is defined based on deviations over time in course evaluations for the instructor.

The unobserved preferences ε_{1j}, ε_{2j}, and ε_{3j} in part determine the optimal choice of β_j, γ_j, and τ_j. The rest of this section shows how we obtain estimates of the parameters given the endogeneity of the grading and effort policies. The key identification assumption is that the unobserved professor preferences for grades, workload, and effort are uncorrelated with innate demand for the courses

[20] Note that when capacity constraints bind, the first terms in each of the expressions is zero. In this case, professors set their grades, workloads, and effort to their ideals. As a result, we do not use courses where capacity constraints bind in estimation. Given the parameter estimates we can, however, back out the corresponding unobserved preference terms using Equations (33)–(35) where the first terms of each are set to zero.

they are teaching after conditioning on W_{lj}. That is, if all professors were forced to set their policy parameters to β^0, γ^0 and τ^0, the corresponding enrollment would be uncorrelated with the unobserved preferences of the instructor. This assumption would be violated if, for example, professors who had unobserved preferences for high grades were assigned to courses based in part on whether the courses had high or low innate demand.

6.3.1 Recovering Ψ_0, Ψ_2, and λ_2

We estimate the professor parameters in three steps. In the first step, we estimate Ψ_0, Ψ_2, and λ_2 using the first two first order conditions for each course. After rearranging and differencing (33) and (34) to eliminate the λ_1 term and solving for γ_j we obtain:

$$\gamma_j = (1/\lambda_2)\ln[E_j]\,A_j - \Psi_0 A_j + W_{2j}\Psi_2 + \varepsilon_{2j} \tag{37}$$

where A_j is given by:

$$A_j = \left[\frac{\partial \ln[E_j]}{\partial \beta_j}\frac{\partial \overline{G}_j}{\partial \gamma_j} \Big/ \frac{\partial \overline{G}_j}{\partial \beta_j}\right] - \frac{\partial \ln[E_j]}{\partial \gamma_j} \tag{38}$$

Both the first and second terms of Equation (37) are correlated with ε_{2j} as ε_{2j} affects enrollment through γ_j and the corresponding derivatives of enrollment and grades with respect to γ_j. We create instruments for these two terms by evaluating A_j and $A_j \ln[E_j]$ at β^0, γ^0, and τ^0 for each of the courses using the estimated course choice model. Namely, we hold fixed the grading and effort policies in courses besides j and then evaluate model-predicted changes to log enrollment and grades at the counterfactual grading and effort policies for j. In practice, we evaluate the policies at the mean values for each parameter, though evaluating at alternative grading and effort policies produced similar results. The variation across classes is then driven by the innate demand for courses given fixed grading and effort policies.

Note that this strategy is contingent on knowing ρ (how professor effort translates into student utility). In particular, how changing τ affects innate demand depends on the value of ρ; choosing the wrong value will leave some correlation between the unobserved preference terms and our generated course demand. As we will show, whatever correlation is left is minor as our results are robust to choices of τ across a reasonable range of values for ρ.

6.3.2 Recovering Ψ_3 and λ_3

We follow a similar procedure for estimating λ_3 and Ψ_3, though now taking Ψ_0 as given from step 1. Now using Equations (33) and (35) to eliminate the λ_1 term and solving for τ_j we obtain:

$$\tau_j = (1/\lambda_3)B_j\left(\ln[E_j] - \Psi_0\right) + \Psi_3 + \varepsilon_{3j} \tag{39}$$

where B_j is given by:

$$B_j = \left[\frac{\partial \ln[E_j]}{\partial \beta_j}\frac{\partial \overline{G}_j}{\partial \tau_j} \bigg/ \frac{\partial \overline{G}_j}{\partial \beta_j}\right] - \frac{\partial \ln[E_j]}{\partial \tau_j} \tag{40}$$

We then instrument for $B_j\left(\ln[E_j] - \Psi_0\right)$ by evaluating B_j and $\ln[E_j]$ at the common grading and effort policies used in step 1, β^0, γ^0, and τ^0.

6.3.3 Recovering Ψ_1 and λ_1

In the final step, we recover estimates of Ψ_1 and λ_1 using Equation (33) and our estimates from step 1 for Ψ_0. Equation (33), the first order condition with respect to β_j, can be rewritten as:

$$\overline{G}_j(\beta, \gamma, \tau) = -(1/\lambda_1)C_j\left(\ln[E_j] - \Psi_0\right) + W_{1j}\Psi_1 + \varepsilon_{1j} \tag{41}$$

where C_j is given by:

$$C_j = \left[\frac{\partial \ln[E_j]}{\partial \beta_j} / \frac{\partial \overline{G}_j}{\partial \beta_j}\right]$$

We then instrument for $C_j\left(\ln[E_j] - \Psi_0\right)$ following the procedure in step 1, evaluating C_j and $\ln[E_j]$ with the policy parameters set to β^0, γ^0 and τ^0 in each course.

6.4 Supply-side Results

We estimate professor preference parameters when the return to professor effort, ρ, is set to 0.05. Sample selection for the professor estimation is discussed in A.3. Appendix Table A.11 shows estimates when ρ is set to 0 and 0.2. We believe 0 to 0.2 span the possible values for ρ. Zero is clearly a lower bound—effort expended preparing to effectively present class material or stimulate current and future interest in the subject should not lower enrollment. For the upper bound of 0.2, note that our effort measure has a standard deviation of 1. The structure of the model is such that a one standard deviation increase in effort results in slightly less than a ρ standard deviation increase in log enrollment.[21] Higher values of ρ would imply that student demand for a course would be unrealistically controllable by the professor's endogenous effort choices.

[21] The extent to which effects on log enrollment are smaller than a ρ standard deviation increase depends on the number of students eligible to take the class.

Table 10 shows the estimates of the professor preference parameters. Professors (STEM and non-STEM) who teach upper-level classes prefer higher grades, lower workloads, and smaller class sizes relative to those teaching lower-level classes, perhaps reflecting the more specialized nature of these courses. Tenured and tenure-track faculty prefer lower grades compared to lecturers and this is especially so for assistant professors. Instructors who are not on the tenure track may have an incentive to offer higher grades as their contracts may depend on teaching evaluations, which in turn rise with expected grades (see Table A.10). Female professors teaching in non-STEM departments have higher ideal grades than their male counterparts. However, in STEM departments, there is little difference in grading across professor gender.

Department-specific parameters are listed in order of decreasing size of the ideal grade of professors in the department. Professors in Management and Marketing and Education and Health departments have the highest ideal grades while professors in English and Mathematics departments have the lowest ideal grades. Higher ideal grades are also generally associated with lower ideal workloads. It is worth emphasizing that these ideal grades and workloads are not driven by direct student demand for courses; they may be set by norms in the department, perhaps following the lead of senior faculty or influenced by instructors' own experiences as undergraduates.

While Table 10 reveals how professors would prefer to assign grades and workloads, student demand for courses induces deviations from these ideals to achieve enrollments that are closer to the e_{0j}. One can see these demand adjustments directly in the rearranged first order conditions given in Equations (37), (39), and (41). The first terms of each equation show how deviations from ideal enrollments (demand adjustments) impact professor choices.

Table 11 reports averages of these demand adjustment terms by department relative to the average across all courses. Demand adjustments are sorted from lowest to highest based on the adjustment to grades. In response to higher student demand, STEM departments (along with Psychology and Management and Marketing) give out lower grades and assign higher workloads, compared to non-STEM departments. The difference in grades between Biology and English (the two extremes) due to demand factors is about 0.5 grade points. English has the lowest ideal grades with the exception of Math (Table 10), yet offers grades around the median in equilibrium (Table 8) due to the relatively low demand for English courses. In contrast, Biology is close to the median on ideal grades, yet gives substantially lower grades due to the high demand for Biology courses.

Using the results in Tables 10 and 11, we can decompose the gaps in grades and workloads between STEM and non-STEM courses into contributions from student demand and professor

Table 10: Estimates of Professor Preferences

	Ideal grade		Ideal workload		Ideal effort		Ideal log enrl	
λ (Preference Weight)	2.773	(0.436)	34.009	(3.203)	0.316	(0.096)	1.000	——
Constant	2.581	(0.128)	0.322	(0.046)	-0.132	(0.054)	5.189	(0.615)
Upper-level Class	0.530	(0.080)	-0.108	(0.046)			-1.584	(0.512)
Upper-level X STEM	-0.098	(0.071)	0.076	(0.023)				
Grad. Student	-0.008	(0.051)	0.006	(0.016)				
Lecturer	0.119	(0.053)	-0.022	(0.015)				
Asst. Prof.	-0.106	(0.055)	0.040	(0.016)				
Tenured Prof.	-0.059	(0.048)	0.011	(0.014)				
Female Prof.	0.107	(0.033)	0.002	(0.010)				
Female Prof. X STEM	-0.074	(0.065)	0.000	(0.020)				
Mgmt & Mkting	0.324	(0.089)	-0.043	(0.027)				
Education & Health	0.229	(0.069)	-0.007	(0.021)				
Communications	0.093	(0.070)	0.051	(0.022)				
Biology	0.055	(0.132)	-0.002	(0.035)				
Engineering	0.019	(0.084)	0.088	(0.026)				
Psychology	-0.006	(0.103)	0.066	(0.030)				
Regional Studies	-0.015	(0.074)	0.086	(0.024)				
Econ., Fin., Acct..	-0.026	(0.104)	0.026	(0.029)				
Language	-0.053	(0.066)	0.076	(0.021)				
Social Science	-0.129	(0.064)	0.035	(0.020)				
Chem & Physics	-0.130	(0.108)	0.057	(0.030)				
English	-0.231	(0.082)	0.126	(0.026)				
Math	-0.322	(0.083)	0.133	(0.023)				

Note: The weight on ideal log enrollment, λ_0, is normalized to 1. The return to professor effort, ρ, is set to 0.05. The base professor rank category is adjunct instructors contracted by the course/semester. Lecturers are offered longer-term contracts and are salaried. See Appendix Table A.11 for parameter estimates at alternative values of ρ at 0 and 0.2. STEM departments are bolded. Baseline department is Agriculture.

Table 11: Demand Adjustments Relative to the Mean ($\rho = 0.05$)

	Grades	Workload	Effort
Biology	-0.3102	0.0654	-0.1303
Psychology	-0.2992	0.0725	-0.1257
Mgmt & Mkting	-0.2664	0.0621	-0.112
Econ., Fin., Acct.	-0.2484	0.0567	-0.1033
Chem & Phys.	-0.0799	0.0177	-0.0332
Engineering	-0.0694	0.025	-0.0214
Education & Health	-0.0419	0.0073	-0.0209
Social Sciences	-0.0107	-0.0124	-0.0031
Math	-0.0097	-0.0011	-0.0004
Agriculture	0.0394	-0.0118	0.0152
Communications	0.0939	-0.0114	0.0394
Languages	0.1549	-0.0343	0.0636
Regional Studies	0.1867	-0.0203	0.0727
English	0.1896	-0.0762	0.0746

Note: Demand adjustments are calculated from the first terms of Equations (37), (39), and (41).

preferences. In particular, we examine the share of the gap in grades and workloads that is due to differences in i) demand (Table 11), ii) level of course offerings (Table 10 rows 3-4), iii) rank of the instructor (Table 10 rows 5-8), iv) female professor representation (Table 10 rows 9-10), and v) departmental effects (Table 10 rows 11-23). Results are presented in Table 12.

The first column of Table 12 shows how demand factors vary across STEM and non-STEM courses. These are calculated by weighting the department demand adjustments in Table 11 by the number of courses in each STEM and non-STEM department respectively. The first panel shows that differences in demand result in STEM grades being 0.15 grade points lower than non-STEM grades. This represents 35% of the average difference between STEM and non-STEM course grades. The second panel shows that differences in demand account for a similar share of the differences in workloads across STEM and non-STEM departments.

The next set of columns come from calculating how the components of ideal grades and workloads given in Table 10 vary by department. More upper-division classes are offered in non-STEM departments. This coupled with upper-level non-STEM classes having higher ideal grades accounts for about 18% of the difference between STEM and non-STEM grades. Despite the substantial heterogeneity in ideal grades across instructor rank, this accounts for very little of the differences between STEM and non-STEM courses. Differences in female representation coupled with women in non-STEM fields having higher ideal grades accounts for a little under 10% of the non-STEM difference. The remaining 35% of grade differences across STEM and non-STEM courses is accounted for by the department-specific intercepts, which reflect factors such as departmental norms.

Table 12: Decomposing STEM/Non-STEM Differences in Grades and Workload ($\rho = 0.05$)

	Demand Adjust	Upper-level Class	Faculty Rank	Female Faculty	Dept. Prefs.	Total Effect
STEM grade	-0.1087	-0.055	-0.0093	-0.0291	-0.1054	-0.3074
Non-STEM grade	0.0413	0.0215	0.0036	0.0114	0.0412	0.1190
Diff.	0.1500	0.0765	0.0129	0.0405	0.1466	0.4264
Shares	35.18%	17.94%	3.03%	9.50%	34.38%	100%
STEM workload	0.0255	0.0248	0.0012	-0.0378	0.0254	0.039
Non-STEM workload	-0.0097	-0.0097	-0.0005	-0.0375	-0.0099	-0.0673
Diff.	-0.0352	-0.0345	-0.0017	0.0003	-0.0353	-0.1063
Shares	33.11%	32.46%	1.60%	-0.28%	33.21%	100%

Note: Decomposition of STEM/non-STEM differences in grades and workloads (γ) are calcuated by using Ψ estimates of department-category intercepts, instructor rank, upper-level class, and female professor from Table 10 and demand-side adjustments calculated in Table 11.

6.5 General equilibrium counterfactuals

Given the estimates of professor preferences, we can now examine how equilibrium grading practices would change in counterfactual scenarios. We focus on three sets of counterfactuals, the first two of which examine the sources of the grading differences across fields and their implications for STEM enrollment. First, we examine the effects of equating intrinsic demand (i.e. non-grade preferences) for STEM and non-STEM courses. Second, we examine the effects of removing observed differences in preferences between STEM and non-STEM professors. Finally, we consider a counterfactual that could actually be implemented, examining the effects of mandating an average grade of a B in all courses. This last counterfactual mirrors the partial equilibrium case considered in the last row of Table 9 but now allows professors to adjust workloads and effort. Note that all counterfactuals hold choice of major by juniors and seniors fixed. They should therefore be interpreted as short-run results, with larger long-run impacts likely occurring as students adjust their majors.

Counterfactual results are presented in Table 13. The first row shows the data. The second row equates intrinsic demand. There are a number of demand-side mechanisms, including department preferences and previously passed courses, that may lead to heterogeneous demand for STEM and non-STEM courses. To equalize intrinsic demand, we turn off preferences for grades by setting ϕ_0 and ϕ_1 to zero and then solve for a value Δ such that decreasing the utility of all STEM classes by Δ equates average class sizes in STEM and non-STEM courses. Δ thus quantifies the extent to which demand for STEM courses is greater than demand for non-STEM courses in a scenario where grades are irrelevant. Then, to simulate a counterfactual scenario where intrinsic demand is equal in STEM and non-STEM courses, we subtract Δ from the utility of all STEM classes, allow students to value grades according to the estimated values of ϕ_0 and ϕ_1, and then solve for the equilibrium grading policies and the resulting course enrollments.[22] Given the same intrinsic demand, STEM courses would have lower enrollments because of professor preferences for lower grades and higher workloads. The gap in grades across STEM and non-STEM courses shrinks from 0.43 grade points in the baseline to 0.16 grade points when intrinsic demand is equalized. Hence much of the differences in grades across STEM and non-STEM courses is due to substantially higher demand for STEM courses.

The next three rows of Table 13 equate observed professor preferences by averaging out differences due to the rank and gender composition of the instructors as well as the departmental effects. In particular, we modify instructor preferences for ideal grades and/or ideal workloads given in Equation (36) as follows:

$$e^*_{lj} = \frac{\sum_{j'} W_{lj'}\Psi_l}{J} + \varepsilon_{lj} \tag{42}$$

This removes all systematic differences in ideal grades and/or workloads due to instructor characteristics or departments; however, it retains idiosyncratic preference terms ε_{lj}. Given the new professor preferences, we then solve for the equilibrium grading and effort policies. Equalizing professor preferences across both ideal grades and workloads (row 3) reduces the average differences in

[22] We solve for the optimal grading and effort policies in this counterfactual scenario via a fixed-point algorithm. Given guesses of the grading and effort choices, we solve for student choice probabilities, taking into account capacity constraints as in B.4. We then update the grading and effort policies using the first order conditions of the professor's maximization problem.

grades across STEM and non-STEM to 0.23 points, showing the large role demand factors play in grading policies. Decreasing grading differences leads to increases in STEM enrollments, especially for women. Women increase the share of classes they take in STEM by 6.3 percentage points; the corresponding increase for men is 5 percentage points. Rows 4 and 5 show that while equalizing ideal grades and workloads both increase STEM participation, equalizing ideal grades has a larger effect, especially for women. Indeed, equalizing ideal workloads actually increases the STEM gap.

The last three rows of Table 13 show the partial and general equilibrium results from imposing that each class' average grade be a 3.0. Row 6 shows the results equivalent to that of the last row of Table 9 but now including juniors and seniors. Here the γ_j's and τ_j's are fixed at their estimated values. The share of STEM classes increases for men and women by 8.6 and 12 percentage points, respectively. Row 7 then shows what happens when professors are able to partially undo the effects of the policy by changing their workloads (γ_j's) and effort (τ_j's). Professor responses to the policy lower the increases for men and women to 7.4 and 10.1 percentage points. But the effects on STEM enrollment—especially for women—remain large. Finally, row 8 shows the equivalent of row 7 but where the curve only applies to lower-division courses. Since over 80% (65%) of STEM (non-STEM) enrollment is in lower-division courses, the effects remain large, only dropping STEM enrollment by 1.1 percentage points relative to having the curve affect all courses.

7 Conclusion

The number of STEM graduates—especially for under-represented groups—has been an ongoing concern. At the same time, STEM courses are on average associated with lower grades and higher study times, both factors that may deter enrollment. Using administrative data form the University of Kentucky, we estimate a model of course choices to understand what influences STEM enrollment and how those influences differentially affect men and women. While we show that a variety of factors influence how students choose courses, we find that differences in grading policies play an important role in suppressing STEM demand and this is particularly true for female students.

One issue with policies aimed at reducing grading differences is that instructors may respond to these policies by changing other aspects of their courses. To capture these responses—and to understand the source of grading differences more generally—our analysis treats grading policies as equilibrium objects chosen by instructors in competition with one another. Taking into account these equilibrium responses, we show that a policy of curving all courses around a B would increase

Table 13: Counterfactuals with Endogenous Professor Responses

	Class size		STEM Enrollment Share			Weighted Avg. Grade		
	STEM	Non-STEM	Overall	Female	Male	Overall	STEM	Non-STEM
Baseline	82.6	45.0	41.8%	34.6%	49.5%	3.013	2.763	3.197
Equate Demand	45.8	59.3	23.2%	18.3%	28.4%	3.102	2.978	3.141
Equate Prof. Pref	93.7	40.6	47.5%	40.9%	54.5%	3.130	3.010	3.241
Equate Prof. Grade Pref	89.9	42.1	45.5%	38.9%	52.6%	3.124	3.018	3.215
Equate Prof. γ Pref	86.4	43.4	43.8%	36.4%	51.6%	3.010	2.766	3.205
Grade Around 3$^{\diamond}$								
Partial Eq.*	103.1	36.9	52.2%	46.6%	58.2%	3.000	3.000	3.000
General Eq.	99.9	38.2	50.6%	44.7%	56.9%	3.000	3.000	3.000
General Eq. Lower Div	97.7	39.0	49.5%	43.4%	56.0%	3.056	2.997	3.116

Note: ρ is fixed at 0.05. Weighted refers to weighting by class enrollment. $\diamond$: "Grade Around 3" adjusts mean grade in all courses to a B, affecting both males and females. Professors change grading strategies based on student responses to changes in preferences and abilities for general equilibrium analysis. $*$: Partial equilibrium results correspond to the last row in Table 9. While Table 9 shows STEM shares for only freshmen and sophomores, the results presented here are for all students. See Appendix Table A.12 for changes to γ and counterfactuals evaluated at alternative ρ values.

overall STEM participation by 8.8 percentage points (a 21% increase) and female STEM participation by 10.1 percentage points (a 29% increase). Changing grading practices to mitigate large departmental-differences in average grades then results in substantial increases in STEM enrollment and a shrinking of the gender gap.

There are at least two reasons why our estimates likely understate the long run effects of equalizing average grades across classes. First, our counterfactual holds choice of major fixed for juniors and seniors; later cohorts will be able to also respond to the policy by shifting into STEM majors. Second, the shifting composition of STEM classes towards more women may have a positive feedback effect through changing the climate of the classes. Weighed against these positive effects, increases in the supply of STEM majors may result in lower wage premiums for STEM majors, partially undoing the effects of the policy.

References

Altonji, J. G., Arcidiacono, P. & Maurel, A. (2016), *Handbook of Labor Economics*, Vol. 5, Elsevier, chapter 7.

Altonji, J. G., Blom, E. & Meghir, C. (2012), 'Heterogeneity in human capital investments: High school curriculum, college major, and careers', **4**(1), 185–223.

Altonji, J. G., Kahn, L. B. & Speer, J. D. (2016), 'Cashier or consultant? entry labor market conditions, field of study, and career success', *Journal of Labor Economics* **34**(S1), 361–401.

Arcidiacono, P., Aucejo, E. M. & Spenner, K. (2012), 'What happens after enrollment? an analysis of the time path of racial differences in GPA and major choice', *IZA Journal of Labor Economics* **1**(1), 1–24.

Arcidiacono, P. & Jones, J. B. (2003), 'Finite mixture distributions, sequential likelihood and the EM algorithm', *Econometrica* **71**(3), 933–946.

Arcidiacono, P. & Miller, R. A. (2011), 'Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity', *Econometrica* **79**(6), 1823–1868.

Astorne-Figari, C. & Speer, J. D. (2018), 'Drop out, switch majors, or persist? The contrasting gender gaps', *Economics Letters* pp. 82–85.

Astorne-Figari, C. & Speer, J. D. (2019), 'Are changes of major, major changes? The roles of grades, gender, and preferences in college major switching', *Economics of Education Review* **70**, 75–93.

Babcock, P. (2010), 'Real costs of nominal grade inflation? new evidence from student course evaluations', *Economic Inquiry* **48**(4), 983–996.

Bagues, M., Labini, M. S. & Zinovyeva, N. (2008), 'Differential grading standards and university funding: Evidence from Italy', *CESifo Economics Studies* **52**(2), 149–176.

Bar, T., Kadiyali, V. & Zussman, A. (2009), 'Grade information and grade inflation: The Cornell experiment', *Journal of Economic Perspectives* **23**(3), 93–108.

Berry, S., Levinsohn, J. & Pakes, A. (1995), 'Automobile prices in market equilibrium', *Econometrica* **63**(4), 841–890.

Brint, S., Cantwell, A. M. & Saxena, P. (2012), 'Disciplinary categories, majors, and undergraduate academic experiences: Rethinking Bok's "underachieving colleges" thesis', *Research in Higher Education* **53**(1), 1–25.

Butcher, K. F., McEwan, P. J. & Weerapana, A. (2014), 'The effects of an anti-grade-inflation policy at Wellesley College', *Journal of Economic Perspectives* **8**(3), 189–204.

Carrell, S. E., Page, M. E. & West, J. E. (2010), 'Sex and science: How professor gender perpetuates the gender gap', *Quarterly Journal of Economics* **125**(3), 1101–1144.

Chan, W., Hao, L. & Suen, W. (2007), 'A signaling theory of grade inflation', *International Economic Review* **48**(3), 1065–1090.

Chapman, B. (2014), 'Top-performing high school seniors can get free ride to state colleges for science studies', *New York Daily News* **May**.

Crocker, J., Luhtanen, R. K., Cooper, M. L. & Bouvrette, A. (2003), 'Contingencies of self-worth in college students: Theory and measurement', *Journal of Personality and Social Psychology* **85**(6), 894–908.

Epple, D., Romano, R. & Sieg, H. (2006), 'Admission, tuition, and financial aid policies in the market for higher education', *Econometrica* **74**(4), 885–928.

Ferreyra, M., Garriga, C. & Diaz, A. S. (2021), 'raising college access and completion: How much can free college help?', Working paper.

Freeman, D. G. (1999), 'Grade divergence as a market outcome', *Journal of Economic Education* **30**(4), 344–351.

Fu, C. (2014), 'Equilibrium tuition, applications, admissions, and enrollment in the college market', *Journal of Political Economy* **122**(2), 225–281.

Gemici, A. & Wiswall, M. (2014), 'Evolution of gender differences in post-secondary human capital investments: College majors', *International Economic Review* **1**(55), 23–56.

Gentzkow, M. (2007), 'Valuing new goods in a model with complementarity: Online newspapers', *American Economic Review* **97**(3), 713–744.

Graeml, Y. (2019), 'Senators introduce bill seeking to help international stem students work in us', *The Daily Californian* .
URL: *www.dailycal.org/2019/06/12/senators-introduce-bill-seeking-to-help-international-stem-students-to-work-in-us/*

Griffith, A. L. (2010), 'Persistence of women and minorities in STEM field majors: Is it the school that matters?', *Economics of Education Review* **29**(6), 911–922.

Gross, N. (2019), 'Your gi bill benefits: Everything you need to know', *Military Times* .
URL: *www.militarytimes.com/education-transition/2019/07/20/gi-bill-benefits-guide/*

Hastings, J. S., Neilson, C. A. & Zimmerman, S. D. (2013), Are some degrees worth more than others? evidence from college admission cutoffs in chile, Technical Report 19241, National Bureau of Economic Research.

Hinz, G. (2019), 'STEM students would get free tuition under Il. bill', *Crain's Chicago Business* . **URL:** *www.chicagobusiness.com/greg-hinz-politics/stem-students-would-get-free-tuition-under-illinois-bill*

Hoffmann, F. & Oreopoulos, P. (2009), 'A professor like me the influence of instructor gender on college achievement', *Journal of Human Resources* **44**(2), 479–494.

Insler, M., McQuoid, A. F., Rahman, A. & Smith, K. (2021), Fear and loathing in the classroom: Why does teacher quality matter?

Jacob, B., McCall, B. & Stange, K. M. (2013), College as country club: Do colleges cater to students' preferences for consumption?, Technical report, National Bureau of Economic Research.

Jiang, X. (2019), Women in stem: Ability, preference, and value.

Johnson, V. E. (2003), *Grade inflation: A crisis in college education*, Springer.

Kaganovich, M., Taylor, M. & Xiao, R. (2021), Gender differences in persistence in a field of study.

Kirkeboen, L. J., Leuven, E. & Mogstad, M. (2016), 'Field of study, earnings, and self-selection', *Quarterly Journal of Economics* **131**(3), 1057–1111.

Kugler, A., Tinsley, C. & Ukhaneva, O. (2021), 'Choice of majors: Are women really different from men?', *Economics of Education Review* **81**, 1–19.

Latek, T. (2019), 'Lawmakers drawing up proposals for 2020 general assembly', *Kentucky Today* . **URL:** *kentuckytoday.com/stories/lawmakers-drawing-up-proposals-for-2020-general-assembly*

McEwan, P. J., Rogers, S. & Weerapana, A. (2021), 'Grade sensitivity and the economics major at a women's college', *AEA Papers and Proceedings* **11**, 102–106.

Nelson, J. P. & Lynch, K. A. (1984), 'Grade inflation, real income, simultaneity, and teaching evaluations', *Journal of Economic Education* pp. 21–37.

Nevo, A., Rubinfeld, D. L. & McCabe, M. (2005), 'Academic journal pricing and the demand of libraries', *American Economic Review* **95**(2), 447–452.

Ost, B. (2010), 'The role of peers and grades in determining major persistence in the sciences', *Economics of Education Review* **29**(6), 923–934.

Owen, A. (2010), 'Grades, gender, and encouragement: A regression discontinuity analysis', *The Journal of Economic Education* **81**, 217–234.

Piopiunk, M., Schwerdt, G., Simon, L. & Woessmann, L. (2020), 'Skills, signals, and employability: An experimental investigation', *European Economic Review* **123**, 1–25.

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.

Rask, K. N. & Bailey, E. M. (2002), 'Are faculty role models? Evidence from major choice in an undergraduate institution', *Journal of Economic Education* **33**(2), 99–124.

Rask, K. & Tiefenthaler, J. (2008), 'The role of grade sensitivity in explaining the gender imbalance in undergraduate economics', *Economics of Education Review* **27**(6), 676–687.

Rosenberg, M., Schooler, C., Schoenbach, C. & Rosenberg, F. (1995), 'Global self-esteem and specific self-esteem: Different concepts, different outcomes', *American Sociological Review* **60**(1), 141–156.

Sabot, R. & Wakeman-Linn, J. (1991), 'Grade inflation and course choice', *Journal of Economic Perspectives* pp. 159–170.

Saltiel, F. (2021), 'What's math got to do with it? multidimensional ability and the gender gap in STEM', Working paper.

Seymour, E. & Hewitt, N. M. (1997), *Talking about leaving*, Westview Press, Boulder, CO.

Stevens, R. H. (2021), 'How congress can help get more women and girls involved in stem education', *The Hill* .

URL: *https://thehill.com/blogs/congress-blog/education/554019-how-congress-can-help-get-more-women-and-girls-involved-in-stem*

Stinebrickner, R. & Stinebrickner, T. (2014), 'A major in science? Initial beliefs and final outcomes for college major and dropout', *Review of Economic Studies* **81**(1), 426–472.

Stinebrickner, R. & Stinebrickner, T. R. (2012), 'Learning about academic ability and the college dropout decision', *Journal of Labor Economics* **30**(4), 707–748.

Thomas, J. (2021), What do course offerings imply about university preferences?

U.S. Bureau of Labor Statistics (2019), 'Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity - labor force statistics from the current population survey, table 11'.

Wiswall, M. & Zafar, B. (2015), 'Determinants of college major choice: Identification using an information experiment', *Review of Economic Studies* **82**(2), 791–824.

Zafar, B. (2013), 'College major choice and the gender gap', *Journal of Human Resources* **48**(3), 545–595.

Zangenehzadeh, H. (1988), 'Grade inflation: A way out', *Journal of Economic Education* pp. 217–226.

Zubrickas, R. (2015), 'Optimal grading', *International Economic Review* **3**(56), 751–776.

A Data Appendix

In this section we describe how we processed the data as well as showing additional results for the various estimation stages. The appendix covers the following:

1. how departments are aggregated into our fourteen categories,
2. additional parameters from the motivating regressions on grades and hours studied,
3. our sample selection procedures for the various structural estimation stages,
4. how the choice sets of the students were constructed,
5. robustness of the study effort parameters to different sample selection rules,
6. additional student preference parameters,
7. additional tables with standard errors,
8. testing model assumptions,
9. the inputs to our measure of professor effort,
10. professor preference estimates and equilibrium counterfactuals under different values of the returns to professor effort (ρ).

A.1 Aggregation of departments

Table A.1 shows the aggregation of departments into our fourteen categories. We partitioned departments into these categories by first grouping departments by their School organization. UK has Colleges of: Agriculture, Arts and Sciences, Business and Economics, Communication and Information, Design, Education, Engineering, and Fine Arts. Within the Colleges, departments were further grouped based partly on shared core requirements and cross-listed coursework. Finally, some departments were carved out (ex. Psychology as its own category) or inserted into a category (ex. all Fine Arts departments into Communications) manually, mostly due to department size.

A.2 Additional parameters from the motivating regressions

Table A.2 shows estimates of the department indicator variables from specifications (2) and (4) of Table 3. Results are sorted by the coefficient on grades. The first column shows that the coefficients

Table A.1: Aggregation of Departments

Categories	Departments
Agriculture	Agricultural Biotechnology, Agricultural Economics, Agricultural Ed, Agriculture General, Animal & Food Sciences, Biosystems & Agr Engineering, Environmental Studies, Forestry, Landscape Architecture, Plant Pathology, Plant & Soil Sciences, Sustainable Agriculture
Regional Studies	Appalachian Studies, Family Sciences, Gender & Women's Studies, Hispanic Studies, Latin American Studies
Communications	Arts Admin, Communication, Communication & Info Studies, Fine Arts - Music, Fine Arts - Theatre Arts, Schl Of Journalism & Telecomm, Schl of Art & Visual Studies, Schl of Interior Design
Education & Health	Allied Health Ed & Research, Comm Disorders, Community & Leader Dev, Dept of Gerontology, Dietetics & Nutrition, Early Child, Spec Ed, Rehab, Ed, Ed Curriculum & Instr, Ed Policy Studies & Eval, Ed, Schl & Counsel Psych, Health Sci Ed, Kinesiology- Health Promotion, Lib & Info Sci, Nursing, Public Health, STEM Ed, Social Work
Engineering	Chemical & Materials Engineering, Civil Engineering, Computer Science, Electrical & Computer Engineering, Engineering, Mechanical Engineering, Mining Engineering, Schl of Architecture
Languages	Linguistics, Modern & Classical Languages, Philosophy
English	English
Biology	Biology, Entomology
Mathematics	Mathematics, Statistics
Chem & Physics	Chemistry, Earth & Environmental Sciences, Physics & Astronomy
Psychology	Psychology
Social Sciences	Anthropology, Geography, History, Political Science, Schl of Human Environmental Sciences, Sociology
Mgmt. & Mkting.	Aerospace Studies, Department of Mgmt, Dept of Mkt & Supply Chain, Merchand, Apparel & Textiles, Mil Sci & Leadership
Econ., Fin., Acct.	Accountancy, Economics, Dept of Finance & Quantitative Methods

Note: STEM departments are bolded.

Table A.2: Reduced-form Grade Regression (2) and Study Hours (4) Department Parameters

	(2) Grade		(4) Hours	
Department	Coef.	Std. Err.	Coef.	Std. Err.
Education & Health	0.275	(0.026)	0.005	(0.185)
Communications	0.170	(0.024)	0.310	(0.181)
Mgmt. & Mkting	0.154	(0.030)	-0.012	(0.226)
Languages	0.034	(0.027)	-0.022	(0.182)
Regional Studies	-0.046	(0.031)	-0.018	(0.202)
Social Sciences	-0.085	(0.025)	-0.058	(0.181)
Engineering	-0.148	(0.027)	0.929	(0.193)
English	-0.164	(0.034)	0.282	(0.223)
Econ., Fin., Acct.	-0.246	(0.027)	0.564	(0.223)
Psychology	-0.276	(0.030)	0.233	(0.266)
Math	-0.389	(0.025)	0.591	(0.197)
Biology	-0.438	(0.028)	0.142	(0.250)
Chem. & Physics	-0.507	(0.026)	0.427	(0.223)

Note: Agriculture is the excluded department-category. STEM departments are bolded.

on grades are lowest for the STEM classes plus Psychology. There is over a 0.75 grade point gap between the highest grading department (Education & Health) and the lowest grading department (Chemistry & Physics). The second column shows that STEM departments also have the highest coefficients for hours of study, with Biology being the one exception.

A.3 Sample selection

We now describe our sample selection rules for the various stages of the structural estimation. We restrict courses to those that have enrollment of at least 15 undergraduates. This cuts the 2,026 classes observed in the population to 1,084. The total number of individual-course observations resulting from this cut is 58,081 with 16,190 unique students. We then remove specialized classes that would result from taking a second course in a sequence as the decision process is very different for these courses. The restriction we impose is that at least 99 students had the course in their choice set and that less than 50% of those who had the course in their choice set took the course.

Imposing this restriction results in 1,003 courses chosen by 16,079 unique students. This represents our baseline data for the choices of courses and grades.

There is an additional restriction imposed in the grade estimation. Namely, there are 18 courses where all students received the exact same grade, accounting for 518 individual-course observations, or less than 1% of enrollments in the baseline data. The courses are still part of our course choice problem but are not used in the estimation of grades. Instead, the expected grades for students in these courses was set to what it was in the data, 4.0, with γ for these courses set to 0.

Estimates of the grade parameters show an additional 7 courses where the estimate of γ was less than 0.01 (estimates of γ are constrained to be greater than zero). For these courses, what yields high grades is fundamentally different from those of other courses in the same department. These courses account for 224 individual-course observations, or less than 0.5% of enrollments in the baseline data. For the purposes of estimating study times (where one of the inputs is $\ln(\gamma)$) and the professor estimation (where γ is a choice), we do not use these courses. For our counterfactuals, we fix the grading policies of these courses at what we observe in the data.

For the study effort analysis, observations are at the course-cohort level with an initial sample of 2,449 course-cohort evaluations. In principle there could have been 4,012 observations if there was a student from each cohort in the class who also filled out a course evaluation. The 2,449 is then the result of some courses either not having students in a particular cohort or having students in a particular cohort where none filled out the course evaluation. Our sample is further cut to 2,411 once courses with γ less than 0.01 are removed.

We implement a number of additional restrictions on the sample for the study effort analysis. The cohort of the student in the evaluation data is based on the students' self reports while in the administrative data it is based on our calculations given the academic records of the student. We define the response rate for the course-cohort as the number of course-cohort observations in the evaluation data divided by course-cohort enrollment in the registrar data. Because we want the average characteristics for a particular course-cohort from the registrar data to match the characteristics of those who filled out the evaluations, we restrict our analysis to course-cohorts where the response rate on the evaluations was between 70% and 101%. Imposing this restriction reduced our number of observations to around 850. Out of concerns for selection issues for under-classmen taking upper-level classes, we focus on lower-level classes, reducing the number of observations to 533. Finally, we focus on classes with at least 25 students due to concerns about measurement error in γ, reducing the number of observations to 390. We examine the robustness of our study effort results to alternative sample cuts in Section A.5.

For the professor estimation, we do not use courses where γ is less than 0.01. We also do not use courses that hit their capacity constraint as the professor maximization problem is different when the capacity constraint binds. This reduces our number of courses to 943. We also have to impose sample restrictions given the evaluation data. Here we need professors to have at least two measures of effort across Fall 2011 to Spring 2013, in addition to having one of those measures be for our semester of analysis, Fall 2012. This reduces our sample to 744 courses.

For those for whom we do not see their actual effort, we can still recover their unobserved preferences for ideal grades and workloads using the first two first-order conditions of the professor's problem evaluated at their actual grading policies and using the estimates of the professor's parameters. Then, for those for whom we can recover ideal grades, workloads, and effort because we see their evaluation data, we recover the marginal distribution of ideal effort given ideal grades and workload. Under the assumption that each of our measures is normally distributed, we can simulate an effort draw for those for whom we do not have evaluation data that leads to the same correlation patterns for ideal grades, workload and effort as those for whom we do have evaluation data.

A.4 Construction of students' choices sets

To construct an accurate choice set of classes for students, we account for several administrative and academic rules, as well as each student's academic history. Each student's choice set is constructed as follows:

1. Student academic history: We track each student's transcript history for seven previous semesters (from Fall 2008 to Spring 2011), and eliminate from the choice set classes that the student has taken and passed. When students are observed in classes that they have already taken in previous semesters, we put these classes back into their choice sets.

2. Class prerequisites: Taking advantage of each student's academic history, we link this information to each class' prerequisites as defined in the 2012-2013 UK Undergraduate Bulletin. Students are not allowed to register for classes unless they have satisfied all prerequisites. When students are seen taking courses for which they have not fulfilled prerequisites, we assume the professor has made a request for an exemption on behalf of the student, and put these classes back into their choice sets.

3. AP Exams: We have access to students' Advanced Placement exam score history. We use the information in the UK Undergraduate Bulletin to allow students the option to bypass introductory courses for relevant subjects upon scoring at or above the minimum threshold (usually a score of 3 or above on the AP exam).

4. Room capacity constraint: We have access to timestamps for every class that students register for during the semester. For each student, we capture the first observed time-stamp, which corresponds to the first class the student registered for during Fall 2012. We also have access to the room capacity for every class. For every class in our sample, we isolate the time-stamp of the 'last-in-line' student who registered for the course. For each class, this last timestamp is compared against every student's first time-stamp. If the student's first time-stamp is later than the class's last timestamp, and if the class reached room capacity, we preclude this class from the student's choice set. Building room capacity this way creates an enrollment order for every class. We maintain this framework in our counterfactual analysis to repopulate classrooms when utility parameters or grading policies change.

Table A.3 shows the average share of STEM and non-STEM classes available by cohort after implementing each of the steps above. By construction, removing already-taken courses by freshmen has no effect on the choice set (Restriction 1). For seniors, however, almost 10% and 5% of courses are ruled out in STEM and non-STEM respectively. That the share ruled out is higher in STEM is consistent with the higher demand for STEM courses. Removing courses where the pre-requisites have not been met (Restriction 2) substantially changes choice sets, and this is especially true in STEM courses. Over 40% of STEM courses are ruled out of the choice set because of either not meeting prerequisites or already having taken the course. The comparable number of non-STEM courses is around 20%. The additional changes to the choice set from AP exams (Restriction 3) or capacity constraints (Restriction 4) are small.

A.5 Robustness of study effort parameters to different selection rules

Table A.4 shows how our estimates of equation (23) change as we remove some of sample restrictions described in A.3. The first column shows our preferred model. The second column adds in courses with less than 25 students. The final column additionally adds upper-level courses. The patterns are similar across the three columns, indicating that our study effort results are robust to alternative sample restrictions.

Table A.3: Share of Courses Available by Cohort and STEM Classification after Choice Set Modifications

Modification	Freshmen	Sophomores	Juniors	Senior	Overall
STEM departments					
(1)	1.000	0.953	0.921	0.906	0.946
(2)	0.550	0.541	0.552	0.555	0.550
(3)	0.562	0.547	0.556	0.558	0.556
(4)	0.553	0.546	0.555	0.557	0.553
non-STEM departments					
(1)	1.000	0.974	0.962	0.952	0.972
(2)	0.802	0.785	0.790	0.789	0.792
(3)	0.804	0.787	0.791	0.790	0.793
(4)	0.799	0.786	0.790	0.789	0.791

Modification (1) removes courses that were already taken.

Modification (2) removes courses where the prerequisites are not met based on transcripts.

Modification (3) adds in courses where prerequisites were met by AP exams.

Modificiation (4) removes courses where capacity constraints are met and adds in courses where the student enrolled in the course despite not meeting the prerequisites.

Table A.4: Study Effort Robustness

	Preferred Model	Remove Class Size Restriction	Also Add Upper-level Classes
Constant	1.920	1.973	1.840
Female	0.293	0.397	0.307
ACT Reading	0.022	-0.029	0.050
ACT Math	-0.051	-0.073	-0.058
HS GPA	0.044	-0.048	-0.101
Minority	-0.311	-0.542	-0.117
1st Gen	-0.200	-0.022	-0.096
Type 2	-1.620	-1.705	-1.509
Type 3	-1.353	-1.148	-1.042
Regional Studies	0.384	0.287	0.284
Communications	-0.369	-0.502	-0.601
Education & Health	-0.331	-0.437	-0.352
Engineering	0.147	0.603	0.505
Languages	0.537	0.462	0.408
English	0.229	0.209	0.165
Biology	0.249	0.113	0.329
Math	-0.152	-0.231	-0.243
Chem. & Physics	0.438	0.361	0.325
Psychology	0.241	0.117	0.156
Social Science	0.126	0.055	0.067
Mgmt.. & Marketing	-0.844	-1.039	-0.910
Econ., Fin., Acct.	0.324	0.277	0.238
Upper class			0.097

A.6 Additional student preference parameters

Table A.5 shows the full set of student preference parameters (see Table 5 for a subset of the parameters). The parameters not discussed in the body of text also follow the expected patterns. The more courses opened up by a class (ln Open Class), the more appealing the class is for sophomores and even more so for freshmen. For junior and seniors, courses that fill their declared majors are associated with higher utilities as are upper-level classes in general.

A.7 Tables with standard errors

Tables 7 and 8 in the text are presented without standard errors to enhance readability. Tables A.6 and A.7 below present the complete tables. For expected GPA results in Table A.7, parameter estimates are drawn 1,000 times to bootstrap errors.

A.8 Testing model assumptions

The demand-side estimation entails a number of strong assumptions. We discuss some of these here as well as provide auxiliary tests of their validity. These tests involve comparing the outcomes for simulated course choices versus actual course choices. To simulate course choices, we use the following procedure:

1. draw unobserved preferences for each student;

2. beginning with the student with the earliest time stamp, take the n courses with the highest utility (the sum of expected utility plus the unobserved preference) where n is the number of courses the student chose in the data;

3. repeat step 2 for each additional student based on the ordering of the time stamps and removing courses from subsequent students' choice sets once the capacity constraint is reached.

A.8.1 Knowledge of the grading process

We are assuming that students know the returns to their abilities in different classes (the γ_j's) as well as the grading intercepts (the β_j's). Students, however, may have different levels of uncertainty about the true parameters. If this were the case, we might expect that upperclassmen were able to sort better than lowerclassmen. We might also expect women to be more informed about classes

Table A.5: Complete Table - Estimates of Preference Parameters

	E(grades)		Fem x E(grade)		Fem x Fem Prof							
All Cohorts x	0.925	(0.007)	0.248	(0.010)	0.106	(0.008)						
	Req. Class		ln Open Classes		STEM Class		Fem x STEM Class					
Freshmen x	2.948	(0.024)	0.483	(0.005)	0.472	(0.020)	-0.021	(0.028)				
Sophomores x	1.428	(0.023)	0.400	(0.004)	0.421	(0.017)	-0.135	(0.025)				
	Major req. (A)		Major req. (B)		Upper-lev. Class							
Upperclassmen x	3.538	(0.009)	2.080	(0.019)	1.841	(0.008)						
	Female		ACT read		ACT Math		HS GPA		Type 2		Type 3	
Regional Studies	-0.078	(0.030)	-0.218	(0.024)	-0.089	(0.026)	-0.018	(0.024)	-1.640	(0.077)	-1.589	(0.037)
Communications	-0.315	(0.025)	-0.252	(0.018)	0.040	(0.020)	-0.068	(0.018)	-1.023	(0.045)	-1.548	(0.024)
Education & Health	0.129	(0.025)	-0.335	(0.020)	-0.002	(0.022)	0.078	(0.020)	-1.244	(0.051)	-2.092	(0.031)
Engineering	-0.927	(0.036)	-0.298	(0.021)	0.494	(0.023)	0.074	(0.022)	-0.834	(0.066)	-1.299	(0.032)
Languages	-0.134	(0.027)	-0.101	(0.020)	0.041	(0.022)	-0.157	(0.020)	-1.399	(0.059)	-1.528	(0.029)
English	0.129	(0.033)	0.033	(0.025)	-0.244	(0.028)	0.049	(0.025)	-0.455	(0.077)	-1.638	(0.042)
Biology	0.520	(0.029)	-0.231	(0.021)	0.057	(0.023)	-0.116	(0.022)	-0.668	(0.063)	-1.207	(0.032)
Mathematics	-0.103	(0.029)	-0.175	(0.019)	-0.208	(0.021)	-0.194	(0.019)	-1.309	(0.051)	-1.121	(0.027)
Chem. & Physics	0.082	(0.030)	-0.138	(0.019)	0.152	(0.021)	-0.178	(0.020)	-1.185	(0.053)	-1.210	(0.027)
Psychology	0.209	(0.028)	-0.355	(0.022)	0.030	(0.025)	-0.193	(0.022)	-0.919	(0.060)	-1.137	(0.033)
Social Sciences	-0.299	(0.025)	-0.179	(0.018)	-0.033	(0.020)	-0.181	(0.018)	-1.099	(0.045)	-1.351	(0.024)
Mgmt. & Mkting	-0.052	(0.030)	-0.222	(0.023)	0.138	(0.025)	-0.083	(0.023)	-1.500	(0.078)	-1.594	(0.038)
Econ., Fin., Acct.	-0.378	(0.029)	-0.286	(0.019)	0.128	(0.022)	-0.043	(0.020)	-1.891	(0.067)	-1.525	(0.029)

Note: Major req. (A) refers to whether the course was required for the major; Major req. (B) refers to whether the course was one of two or more required for the major.

Table A.6: Estimates of Department-Specific Ability Weights (α) with Standard Errors

	Female		ACT read*		ACT math*		HS GPA*	
Education & Health	0.446	(0.152)	0.175	(0.055)	0.357	(0.072)	0.534	(0.103)
Regional Studies	0.245	(0.123)	0.080	(0.054)	0.532	(0.089)	0.686	(0.114)
Communications	0.188	(0.121)	0.137	(0.049)	0.126	(0.037)	0.568	(0.144)
Agriculture	0.153	(0.179)	0.238	(0.097)	0.555	(0.116)	0.893	(0.157)
Psychology	-0.077	(0.090)	0.386	(0.054)	0.445	(0.059)	0.683	(0.066)
English	-0.102	(0.140)	0.270	(0.081)	0.378	(0.093)	0.707	(0.112)
Languages	-0.167	(0.076)	0.265	(0.066)	0.416	(0.080)	0.605	(0.121)
Social Sciences	-0.204	(0.074)	0.432	(0.070)	0.346	(0.060)	0.789	(0.118)
Math	-0.243	(0.054)	-0.035	(0.028)	1.291	(0.109)	0.676	(0.067)
Mgmt. & Mkting	-0.305	(0.112)	0.173	(0.082)	0.382	(0.112)	0.731	(0.212)
Biology	-0.427	(0.074)	0.200	(0.048)	0.633	(0.071)	0.681	(0.082)
Engineering	-0.457	(0.086)	0.027	(0.037)	0.773	(0.084)	0.298	(0.051)
Econ., Fin., Acct.	-0.497	(0.075)	0.150	(0.046)	0.822	(0.092)	0.646	(0.085)
Chem. & Physics	-0.606	(0.062)	0.056	(0.034)	1.006	(0.059)	0.792	(0.053)

Note: * indicates variable is z-scored.

where there are more women. Both of these have implications for what the model predicts versus what is seen in the data.

To the extent that students do not know the grading processes in different classes, we would expect upperclassmen to have better information than lowerclassmen. In this case, the model would overpredict how well lowerclassmen sort into classes that match their abilities, with the opposite occurring for upperclassmen.

To test whether upperclassmen are sorting better than lowerclassmen, we simulate course choices using the model estimates and calculate expected grades using both the simulated and actual choices. If expected grades in the simulated courses were on average higher than in the actual course choices for lowerclassmen, then this would support upperclassmen being better informed. However, average expected grades for freshmen, sophomores, and upperclassmen in simulated versus actual course are all within 0.01 grade points. For freshmen, simulated expected grades are 2.984 versus 2.986 for actual expected grades; for sophomores, simulated expected grades are 3.015 versus 3.018 for actual expected grades; for upperclassmen, simulated expected grades are 3.078 versus

Table A.7: Expected GPA for Average Classes By Department, Unconditional and Conditional on Taking Courses in that Department with Standard Errors

	EGPA Females Unconditional		EGPA Females Conditional		EGPA Males Unconditional		EGPA Males Conditional	
Education & Health	3.52	(0.04)	3.47	(0.04)	3.34	(0.04)	3.23	(0.04)
Communications	3.46	(0.11)	3.38	(0.10)	3.25	(0.10)	3.13	(0.09)
Agriculture	3.37	(0.04)	3.21	(0.04)	3.23	(0.04)	2.87	(0.03)
Mgmt. & Mkting	3.24	(0.12)	3.36	(0.12)	3.16	(0.11)	3.28	(0.12)
Languages	3.23	(0.04)	3.25	(0.04)	3.13	(0.04)	3.09	(0.04)
Regional Studies	3.21	(0.03)	3.28	(0.03)	3.00	(0.03)	3.06	(0.03)
Social Sciences	3.12	(0.03)	3.08	(0.03)	3.02	(0.03)	2.86	(0.03)
English	3.11	(0.04)	3.09	(0.04)	2.98	(0.04)	2.96	(0.04)
Psychology	2.98	(0.02)	2.95	(0.02)	2.85	(0.02)	2.71	(0.02)
Econ., Fin., Acct.	2.67	(0.03)	2.89	(0.03)	2.70	(0.02)	2.87	(0.02)
Biology	2.59	(0.02)	2.75	(0.02)	2.59	(0.02)	2.73	(0.02)
Math	2.56	(0.02)	2.58	(0.02)	2.58	(0.02)	2.65	(0.02)
Engineering	2.54	(0.04)	2.83	(0.03)	2.60	(0.02)	2.93	(0.02)
Chem. & Physics	2.33	(0.02)	2.55	(0.02)	2.43	(0.02)	2.69	(0.02)

Note: Standard errors generated from bootstrapping with 1,000 draws.

3.069 for actual expected grades.

With regard to information varying for men and women about the β_j's and γ_j's for particular courses, an implication could be that women have better information in courses where women are a greater share of the enrollment. We would then expect that women would do a better job matching their abilities to the grading policies for these classes than what the model would predict. We calculate average expected grades for females in each class based on their actual choices as well as simulated choices from the model. The difference between female expected grades at the class level for actual choices and simulated choices should be positively related to the share female in the class if women have better information about grading policies in classes where there are more women. This turns out not to be the case. Indeed, regressing the difference between actual expected grades and simulated expected grades on share female shows a small, negative, and insignificant relationship.

A.8.2 Balancing effort across classes

Key to keeping the model tractable is the assumption that the utility the student receives from a course does not depend on the other chosen courses. This rules out students balancing workloads by taking a mix of easy and hard classes. To evaluate this assumption, we can see how the model does at predicting the within-student distribution of class workloads. We simulate the model to predict the share of classes each student takes that are above the median γ (workload) class and then calculate the standard deviation across students. If the model substantially over-predicts this standard deviation, then this would be evidence that students were balancing their workloads: the model would be over-predicting the number of students who were taking all hard classes as well as over-predicting the number of students taking all easy classes. This test suggests that bundling is not a major concern, with the model-predicted standard deviations actually being slightly lower than what is seen in the data. For the share of classes above the median workload, the model-predicted standard deviation is 0.3103 versus 0.3237 in the data.

We also predict the share of classes each student takes in STEM. As STEM classes typically require more work, if bundling is a serious issue we would expect students to balance STEM classes with non-STEM classes. The model-generated standard deviation of the share of classes taken in STEM across students would then be too large. Like the previous model test, there is no evidence that the model is missing students balancing STEM and non-STEM classes. For share STEM, the model-predicted standard deviation is 0.3264 versus 0.3431 in the data.

While this may seem surprising, the evaluation data shows that average course study time is 2.6 hours per week with 99% classes associated with less than 9 hours of study time per week. This implies that students have copious time for leisure and other non-academic activities; as such, balancing workloads across courses may be of limited importance to students.

A.8.3 Flexibility of gender preferences

Although our model includes a variety of mechanisms to explain the gender gap in STEM, there may be features of key STEM courses that are deterrents for women that are not captured by the model. To investigate this issue, we use the model to simulate course choices and then examine how share female varies across different subsets of simulated and actual courses. The first two columns of Table A.8 show share female in simulated and actual lower-division courses while the third and fourth columns also restricts to courses with 100 or more students. Rows 1 and 2 look at all

Table A.8: Share Female in Lower-division STEM classes, Predicted versus Actual

	Lower-division		Plus Enroll≥100	
	Actual	Simulated	Actual	Simulated
STEM, All	0.4584	0.453	0.488	0.4821
STEM minus Biology, All	0.4303	0.4253	0.4523	0.4502
STEM, Lowerclassmen	0.4775	0.4715	0.5048	0.4977
STEM minus Biology, Lowerclassmen	0.4483	0.4458	0.4685	0.468
STEM, required, All	0.4499	0.4471	0.4839	0.4805
STEM minus Biology, required,All	0.4162	0.4163	0.4448	0.4461
STEM, required, Lowerclassmen	0.4658	0.4675	0.5004	0.4974
STEM minus Biology, required Lowerclassmen	0.4296	0.4389	0.4597	0.4654

STEM lower-division courses and all STEM lower-division courses taking out Biology. Rows 3 and 4 examine the same courses as rows 1 and 2 but only for lowerclassmen. Rows 5 through 8 repeats the analysis of rows 1 through 4 but further restricts the classes to those that are required for a STEM major. Regardless of the set of classes, or whether we are only looking at lowerclassmen, the simulated share female is within one percentage point of the actual share female in all cases, suggesting that our model is capturing well how women and men are distributed across courses in different departments.

Our first stage estimation was simplified computationally by estimating the reduced form grading processes separately by department. But this limited the ways in which gender could interact with the characteristics of the course. One possibility would be that women receive higher grades in courses taught by women. To see whether this is an issue, we look at the residuals of the grade equation for women and see what the mean value is in courses taught by women. The average difference between the student's actual versus predicted grade for women in classes where the instructor is female was less than 0.006. So while the number is positive, it is quite small.

A.9 Inputs to our measure of professor effort

Outside of a time-use survey or rigidly prescribed schedules (for example, unionized manufacturing jobs) it is often difficult to gather data on worker effort. For professors, where their time could have multiple uses (for example, data analysis or writing an article/book could yield benefits for

Table A.9: Correlation among Class Evaluation and Grades Students Expect to Receive

	Expected Grade	Q09	Q13	Q19	(Q09+Q13+Q19)/3
Expected Grade	1.0000				
Q09	0.1590	1.0000			
Q13	0.1997	0.7374	1.0000		
Q19	0.1805	0.5983	0.7317	1.0000	
(Q09+Q13+Q19)/3	0.2025	0.8633	0.9256	0.8815	1.0000

Note: Expected Grades are grades students expect to receive (as indicated on class evaluations). Questions are asked on the evaluation in a 5-point Likert scale, and asks: Did the instructor (1) present the material effectively - Q09, (2) stimulate interest in the subject - Q13, and (3) stimulate me to read further beyond the class - Q19?

both research and teaching), even direct measures of inputs become problematic. Instead, we use information about students' receptivity to the professor's teaching to capture a measure of the professor's effort, τ. Of the twenty questions in the evaluations, we focus on three with students answering on a five-point Likert scale:

- Q09: Did the professor present class materials effectively?
- Q13: Did the professor stimulate your interest in the subject?
- Q19: Did the professor stimulate you to read further in the subject beyond the class?

We average across the questions to create a raw measure of effort. This raw measure suffers from correlation with expected grade. Table A.9 shows professors who are generous with grades may receive higher evaluations regardless of instruction effort. We purge this expected grade effect in the top half of Table A.10.

In addition, professors may have differing amounts of natural ability to teach in an engaging manner. Since our interest is discretionary effort in response to demand-side factors, we remove permanent characteristics of the professor. We use five semesters of evaluation data (Fall 2011 to Spring 2013) to have enough professor-level observations, and regress the purged effort measure on instructor fixed effect and enrollment. The results are in the bottom half of Table A.10. We subtract the instructor fixed effect but leave in the effect of enrollment. This final measure is our measure of professor effort, τ.

Table A.10: Professor Effort Residualization & Regression of Effort Measure on Log Enrollment

	Coef.	Std. Err.
Expected Grade		
A	1.1773	(0.0559)
B	0.9736	(0.0559)
C	0.7606	(0.0567)
D	0.4305	(0.0485)
log(class size)	-0.0901	(0.008)

Note: Dependent variable in top half of the table is the average of three questions in a 5-point Likert scale asked in class evaluations: Did the instructor (1) present the material effectively, (2) stimulate interest in the subject, and (3) stimulate me to read further beyond the class? Regressors include class times semester fixed effects. Dependent variable in the bottom half of the table is the average of the three evaluation question minus the grade effects estimated in the first half of the table. Regressors include professor and semester fixed effects.

A.10 Professor preference estimates and equilibrium counterfactuals under different values ρ

In the paper, we assume that ρ, the measure of how professor effort translates to course utility (and thus higher enrollment), is set to 0.05. We also estimate the professor preference parameters under the assumption that $\rho = 0$ and $\rho = 0.2$. We feel that this is a reasonable range, representing scenarios where professor effort does not impact course utility to where a professor can largely control enrollment with didactic preparation. Table A.11 shows estimates of professor preferences (shown in Table 10) at alternate values of ρ. Table A.12 shows the general equilibrium counterfactual results (shown in Table 13) at alternative values of ρ.

B Methods Appendix

This appendix provides additional details regarding our empirical methods:

1. how we account for measurement error in the study effort shocks,
2. our modified EM algorithm to recover the parameters of the grade process and conditional probabilities of being each unobserved type,
3. how we solve for counterfactual choice probabilities in the presence of capacity constraints,

Table A.11: Estimates of Professor Preferences at Alternative ρ Values

		$\rho = 0.2$		$\rho = 0$		$\rho = 0.2$		$\rho = 0$	
		Ideal grade				Ideal workload			
	λ	2.775	(0.451)	2.771	(0.431)	34.906	(3.433)	33.703	(3.147)
	Constant	2.582	(0.131)	2.581	(0.127)	5.211	(0.637)	5.181	(0.632)
Gender	Female Prof.	0.107	(0.033)	0.107	(0.033)	0.002	(0.010)	0.002	(0.010)
	Female X STEM	-0.074	(0.065)	-0.074	(0.065)	0.000	(0.020)	0.001	(0.020)
Rank	Grad. Student	-0.008	(0.051)	-0.008	(0.051)	0.006	(0.016)	0.006	(0.016)
	Lecturer	0.119	(0.053)	0.119	(0.054)	-0.021	(0.015)	-0.023	(0.015)
	Asst. Prof.	-0.106	(0.055)	-0.106	(0.055)	0.040	(0.016)	0.040	(0.016)
	Tenured Prof.	-0.059	(0.048)	-0.059	(0.048)	0.011	(0.014)	0.010	(0.014)
Class	Upper-level Class	0.530	(0.082)	0.530	(0.079)	-0.107	(0.046)	-0.109	(0.046)
	Upper X STEM	-0.098	(0.071)	-0.098	(0.071)	0.075	(0.023)	0.076	(0.023)
Dept.	Regional Studies	-0.015	(0.074)	-0.015	(0.074)	0.086	(0.024)	0.086	(0.024)
	Communications	0.092	(0.070)	0.093	(0.070)	0.052	(0.022)	0.051	(0.022)
	Education & Health	0.229	(0.069)	0.229	(0.069)	-0.006	(0.021)	-0.007	(0.021)
	Engineering	0.018	(0.084)	0.019	(0.084)	0.089	(0.026)	0.088	(0.026)
	Languages	-0.053	(0.066)	-0.053	(0.066)	0.076	(0.021)	0.076	(0.021)
	English	-0.231	(0.082)	-0.231	(0.082)	0.126	(0.026)	0.127	(0.026)
	Biology	0.054	(0.134)	0.055	(0.132)	0.001	(0.035)	-0.003	(0.035)
	Math	-0.323	(0.084)	-0.322	(0.083)	0.135	(0.023)	0.133	(0.023)
	Chem & Physics	-0.131	(0.109)	-0.130	(0.108)	0.059	(0.030)	0.056	(0.030)
	Psychology	-0.006	(0.103)	-0.006	(0.102)	0.068	(0.030)	0.066	(0.030)
	Social Sciences	-0.129	(0.064)	-0.129	(0.064)	0.036	(0.020)	0.035	(0.020)
	Mgmt & Mkting	0.323	(0.089)	0.324	(0.089)	-0.042	(0.027)	-0.043	(0.027)
	Econ., Fin., Acct.	-0.027	(0.105)	-0.026	(0.104)	0.028	(0.029)	0.026	(0.029)
		Ideal log enrl				Ideal prof. effort			
	λ	1.000	-	1.000	-	0.637	(0.098)	0.000	(0.000)
	Constant	5.211	(0.637)	5.181	(0.632)	-0.267	(0.055)	0.000	(0.000)
Class	Upper-level Class	-1.603	(0.528)	-1.578	(0.525)				

Note: Ideal enrollment λ is normalized to equal 1. When $\rho = 0$, all professors are assumed to exert zero effort, so weights and ideal effort are not estimated. The base for Rank is "Instructor," who are adjunct instructors contracted by the course/semester. "Lecturers" are offered longer-term contracts and are salaried.

Table A.12: Counterfactual Scenarios in General Equilibrium at Alternative ρ Values

		Class size		STEM Enrollment Share		
	ρ	STEM	Non-STEM	Overall	Female	Male
Baseline		82.6	45.0	41.8%	34.6%	49.5%
Equal Demand	0	45.7	59.4	23.1%	18.3%	28.3%
	0.05	45.8	59.3	23.2%	18.3%	28.4%
	0.2	46.6	59.0	23.6%	18.7%	28.9%
Equal Prof. Pref	0	93.7	40.6	47.5%	40.9%	54.5%
	0.05	93.7	40.6	47.5%	40.9%	54.5%
	0.2	93.6	40.6	47.4%	40.9%	54.4%
Equal Prof. γ Pref	0	86.4	43.4	43.8%	36.4%	51.6%
	0.05	86.4	43.4	43.8%	36.4%	51.6%
	0.2	86.5	43.4	43.8%	36.4%	51.7%
Equal Prof. Grade Pref	0	90.0	42.1	45.5%	38.9%	52.6%
	0.05	89.9	42.1	45.5%	38.9%	52.6%
	0.2	89.8	42.1	45.5%	38.9%	52.6%
Grade Around 3$^\diamond$	0	100.0	38.1	50.6%	44.8%	57.0%
	0.05	99.9	38.2	50.6%	44.7%	56.9%
	0.2	99.2	38.4	50.2%	44.3%	56.6%

Note: $\diamond$: "Grade around 3" adjusts mean grade in all courses to a B, affecting both males and females. Professors change grading strategies based on student responses to changes in preferences and abilities for general equilibrium analysis.

and

4. showing that bias in our professor effort measure from reverse causality is small.

B.1 Measurement error in the study effort shocks

This section describes how we deal with measurement error in the study effort shocks, $\ln(\tilde{\zeta})$. We begin by isolating cohort-class combinations where there is one student. For each of these combinations, we calculate our estimate of $\ln(\tilde{\zeta})$. The standard deviation of these estimates gives us an estimate of the standard deviation of the individual-specific residuals from the data where study times are grouped into the discrete categories from the course evaluations. Label this standard deviation σ_ζ.

Our next step generates simulated data from the continuous study time process (so the hours are not lumped into bins), using the cohort-class data employed and the corresponding parameter estimates from Equation (23) and where the the $\ln(\tilde{\zeta})$'s are drawn from a normal distribution with mean zero and standard deviation σ_ζ. We then calculate what the residuals would be if this simulated data was now discretized in the manner used in the course evaluation data. Label these new residuals $\ln(\tilde{\zeta}^*)$.

We then estimate the following regression:

$$\ln(\tilde{\zeta}) = \beta_0 + \beta_1 \ln(\tilde{\zeta}^*) + \epsilon_1 \tag{43}$$

where β_1 gives the signal to noise ratio. We then inflate our estimate of C_1 from Equation (25) by dividing it by β_1.

B.2 Modified EM algorithm

We now describe our estimation procedure in the presence of unobserved heterogeneity. First consider the parameters of the grade process and the course choices. With unobserved heterogeneity, we now need to make an assumption on the distribution of η_{ij}, the residual in the grade equation. We assume the error is distributed $N(0, \sigma_\eta)$. In theory, one could use the structural choice likelihood in Equation (28) to capture the likelihood of making observed course choices; however, maximizing Equation (28) at every iteration of the EM algorithm is computationally infeasible. Instead, we construct an alternative course choice likelihood function based on a flexible analog of the structural model. For the reduced-form choice problem, we abstract away from the bundling of courses, treating each course choice as its own decision problem. In order to facilitate computation, at points we break down the problem into the probability of taking a course from department k and then the probability choosing the specific course j:

$$p_{ijk} = p_{ik}p_{ij|k}$$

We specify the reduced-form payoff of taking class j as:

$$v_{ij} = (\phi_1^* + w_i\phi_2^*)g_{ij}(\gamma_j^N, \theta_{j(k)}^N, X_i) + \delta_{0j}^* + w_i\delta_{1j}^* + Z_{1i}\delta_{2k(j)}^* + Z_{2ij}\delta_3^* + \epsilon_{ij}^* \tag{44}$$

where $g_{ij}(\cdot)$ represents the expected grade of student i in course j and ϵ_{ij}^* is assumed to follow a nested logit structure with nesting at the department level characterized by ν. The full set of choice parameters is then $\varphi = \{\phi^*, \delta^*, \nu\}$. Note that although we will not be interpreting estimates of φ, the structure of utility in Equation (44) is very similar to the structure in Equation (27).[23] This ensures that conditional type probabilities from this specification will be appropriate for classifying students for the estimation of Equation (27).

Let φ represent the parameters of this flexible choice process. The integrated log likelihood is then:

$$\sum_i \ln\left(\sum_{s=1}^{S} \pi_s \mathcal{L}_{igs}(\theta, \gamma) \mathcal{L}_{ics}(\varphi)\right) \tag{45}$$

where $\mathcal{L}_{igs}(\theta, \gamma)$ and $\mathcal{L}_{ics}(\varphi)$ are the grade and choice (of courses) likelihoods respectively conditional on i being of type s.

We iterate on the following steps until convergence, where the mth step follows:

1. Given the parameters of the grade equation and choice process at step $m-1$, $\{\theta^{(m-1)}, \gamma^{(m-1)}\}$ and $\{\varphi^{(m-1)}\}$ and the estimate of $\pi^{(m-1)}$, calculate the conditional probability of i being of type s using Bayes rule:

$$q_{is}^{(m)} = \frac{\pi_s^{(m)} \mathcal{L}_{igs}\left(\theta^{(m-1)}, \gamma^{(m-1)}\right) \mathcal{L}_{ics}\left(\varphi^{(m-1)}\right)}{\sum_{s'} \pi_{s'}^{(m)} \mathcal{L}_{igs'}\left(\theta^{(m-1)}, \gamma^{(m-1)}\right) \mathcal{L}_{ics'}\left(\varphi^{(m-1)}\right)} \tag{46}$$

2. Update $\pi_s^{(m)}$ using $\left(\sum_{i=1}^{N} q_{is}^{(m)}\right)/N$.

[23] The structure of utility in Equation (44) differs from the structure in Equation (27) in three ways: First, Equation (44) does not subtract γ_j from expected grades. Second, Equation (44) assumes nested logit preference shocks while Equation (27) assumes independent Type 1 extreme value errors. Finally, Equation (44) assumes contemporaneous choices are independent while Equation (27) models students choosing bundles of courses simultaneously.

3. Using the $q_{is}^{(m)}$s as weights, obtain $\{\theta^{(m)}, \gamma^{(m)}, \varphi^{(m)}\}$ by maximizing:

$$\sum_i \sum_s q_{is}^{(m)} \left(\ln\left[\mathcal{L}_{igs}\left(\theta,\gamma\right)\right] + \ln\left[\mathcal{L}_{ics}\left(\varphi\right)\right]\right) \tag{47}$$

To facilitate computation, the maximization step (step 3) is conducted in stages. Denote $f(g_{ij}, \gamma_j^N, \theta_{k(j)}^N)$ as the likelihood of observing g_{ij} given the parameters γ_j^N and $\theta_{k(j)}^N$. Denote $\varphi(!A)$ as φ absent the Ath component. Finally, denote d_{ij} as an indicator for whether i chose course j, d_{ijk} as an indicator for whether i chose course j in department k, and d_{ik} as an indicator for whether i's choice was in department k. Maximization then proceeds as follows:

1. For each department $k \in K$, taking φ as given, choose γ_j^N and θ_k^N to maximize:

$$\sum_i \sum_{j\in k} d_{ijk} \left(\ln[f(g_{ij}, \gamma_j^N, \theta_k^N)] + \ln[p_{ij|k}(\gamma_j^N, \theta_k^N, \varphi)]\right) \tag{48}$$

2. Taking γ_j^N, θ_k^N, and $\varphi(!\phi^*, !\delta_3^*)$ as given, choose ϕ^* and δ_3^* to maximize:

$$\sum_i \sum_j d_{ij} \ln\left[p_{ijk}\left(\gamma_j^N, \theta_k^N, \varphi(!\phi^*, !\delta_3^*), \phi^*, \delta_3^*\right)\right] \tag{49}$$

3. For each department $k \in K$, taking γ_j^N, θ_k^N, and $\varphi(!\delta_0^*)$ as given, choose δ_{0j}^* (relative to one course in each department) to maximize:

$$\sum_i \sum_{j\in k} d_{ijk} \ln\left[p_{ij|k}\left(\gamma_j^N, \theta_k^N, \varphi(!\delta_{0j}^*), \delta_{0j}^*\right)\right] \tag{50}$$

4. Taking γ_j^N, θ_k^N, and $\varphi(!\delta_1^*, !\delta_2^*, !\nu)$ as given, choose $\delta_{1j(k)}^*$, $\delta_{2j(k)}^*$, and ν to maximize:[24]

$$\sum_i \sum_k d_{ik} \ln\left[p_{ik}\left(\gamma_j^N, \theta_k^N, \varphi(!\delta_1^*, !\delta_2^*, !\nu), \delta_1^*, \delta_2^*, \nu\right)\right] \tag{51}$$

The advantage of this sequential strategy is it limits the number of parameters being estimated at each stage and limits the number of times the 1003 choice probabilities are calculated for each individual. Further, when the 1003 choice probabilities are calculated within the maximization routine at step 2, the number of parameters we are maximizing over is limited.

Once the algorithm has converged, we have consistent estimates of $\{\theta, \gamma, \varphi\}$ as well as the conditional probabilities of being in each type. We can use the estimates of q_{is} as weights to form the average type probabilities of students of year in school l in class j to then estimate the parameters of the study process in (23). Finally, we use the estimates of q_{is} as weights in estimating the structural choice parameters using (28).

[24] At this step we also recover the δ_{0j}^*'s for the normalized courses in each department from step 3.

B.3 Fixed-point algorithm

We now describe our fixed-point algorithm that is used in each calculation of the student choice likelihood. Let $\tilde{\Theta} = \{\delta_{1k(j)}, \delta_{2k(j)}, \delta_3, \phi_0, \phi_1\}$ represent choice parameters other than δ_{0j}, let S_j^d represent the share of students choosing course j in the data, and let $S_j\left(\delta_{0j}, \tilde{\Theta}\right)$ represent the predicted share of students choosing course j as a function of δ_{0j} and other choice parameters. Given a new guess of $\tilde{\Theta}$, we use the δ_{0j}'s from the previous guess $\delta_{0j}^0(\tilde{\Theta})$ and calculate $S_j\left(\delta_{0j}^0, \tilde{\Theta}\right)$. The mth iteration of the fixed-point problem updates δ_{0j}^m using:

$$\delta_{0j}^m = \delta_{0j}^{m-1} + \ln\left[S_j^d\right] - \ln\left[S_j\left(\delta^{m-1}, \tilde{\Theta}\right)\right] \tag{52}$$

Given the δ_{0j}^m, we update $S_j\left(\delta_{0j}^m, \tilde{\Theta}\right)$. These steps are repeated until the predicted and actual enrollment shares are arbitrarily close.

B.4 Counterfactuals in the presence of capacity constraints

Embedded within each counterfactual are each student's conditional choice probabilities. In order to ensure that capacity constraints are not exceeded, we work backward based on the registration ordering given by the time stamps (see Section A.4). We proceed in the following manner for each student n where n refers to the the ordering based on the student's time stamp:

1. Calculate the choice probabilities for student n over the courses where the student has met the prerequisites and where the course is not already filled.

2. If adding the choice probabilities to the probabilities of the previous $n-1$ students does not cause any of the classes to exceed the course capacity, proceed to the next student.

3. If one or more of the courses exceeds capacity in step 2, identify the class where adding n's probability causes the capacity constraint to be exceeded by the greatest amount. Label this excess capacity c and the choice probability p. Note that $c < p$ as the course previously had open space.

4. Assign the probability that the course identified in step 3 is in n's choice set as $1 - \frac{c}{p}$. Take the choice probabilities when this course is in the choice set, multiply them by $1 - \frac{c}{p}$, and add them to the number of enrollees in each course. This ensures that the identified course will be exactly filled. Repeat step 1 for student n, taking into account that all probabilities from the new choice set will be multiplied by $\frac{c}{p}$ and where the identified course is no longer in n's choice set.

The algorithm ensures that any capacity constrained courses are exactly filled, with filled courses no longer available to individuals with later time stamps.

B.5 Size of bias in our professor effort measure

Our measure of professor effort is biased due to reverse causality: professor effort affects enrollment. In this section we illustrate why this is bias is likely small. We begin by approximating log enrollment for instructor i in course j at time t as:

$$\ln(E_{ijt}) = \delta_{1i} + \delta_{2j} + \delta_{3t} + \rho\tau_{ijt} + \epsilon_{1ijt} \tag{53}$$

In this case ρ gives the return to effort, τ_{ijt}.

Now express effort, τ_{ijt} as depending on innate demand—everything in the equation above besides the τ_{ijt} term—plus some instructor and time fixed effects and an error:

$$\tau_{ijt} = \alpha\left(\delta_{1i} + \delta_{2j} + \delta_{3t} + \epsilon_{1ijt}\right) + \delta_{4j} + \delta_{5t} + \epsilon_{2ijt} \tag{54}$$

α then gives how innate demand affects effort. Note that our assumption is that $\alpha < 0$ as higher innate demand implies less effort.

We can then express equation (54) with respect to $\ln(E_{ijt})$:

$$\tau_{ijt} = \alpha\ln(E_{ijt}) - \alpha\rho\tau_{ijt} + \delta_{4j} + \delta_{5t} + \epsilon_{2ijt} \tag{55}$$

Rearranging the terms shows the bias we would get in our estimate of α because log enrollment depended on effort.

$$\tau_{ijt} = \frac{\alpha\ln(E_{ijt})}{1+\alpha\rho} + \delta^*_{4j} + \delta^*_{5t} + \epsilon^*_{2ijt} \tag{56}$$

where the starred variables are just scaled up by $\frac{1}{1+\alpha\rho}$.

Equation (56) is what we use to estimate professor effort. The coefficient on log enrollment is less than zero. This tells us two things under the assumption that ρ is greater than zero, that is, professor effort positively affects enrollment. First, α is indeed negative. Second, $1 > (1+\alpha\rho) > 0$. So the magnitude of our coefficient on effort is biased upward by the factor $\frac{1}{1+\alpha\rho}$. The closer this number is to one, the smaller the bias.

Our estimated coefficient on log enrollment is -0.09 (see Table A.10). Note that -0.09 is actually biased upward in magnitude. So the actual bias factor is between 1 and $\frac{1}{1-0.09\rho}$. If ρ is 0.05, the bias factor is then less than 1.0045 (where 1 would be no bias). When ρ is 0.2, the bias factor is less than 1.018. In this latter case, the true α would be between -0.088 and -0.09, suggesting any bias in our effort measure is small.