Gender Stereotypes in Professor-Student Interactions*

Zachary Bleemer †

January 2019
Updated September 2019
Click Here for the Most Updated Version

Abstract

While third-party evaluators' gender biases have been shown to exacerbate labor market inequities, the role of gender stereotypes in subtly shaping interactions between students and their teachers and mentors remains largely unexplored outside of laboratories. In this study, I analyze a novel dataset of more than 1.2 million student evaluations written by UC Santa Cruz professors spanning 1965-1979 and 1999-2009, combined with detailed student transcript records, to identify professors' gender stereotypes and estimate their impact on students' educational decisions. I estimate each evaluation's genderedness by comparing the adjectives and adverbs used to describe different-gendered students who received the same letter grade in the same class, and characterize professors by the degree to which they tend to employ more male- and female-valence vocabulary in describing male and female students ($\hat{G}$). I then exploit plausibly-random professor assignments to students' first-quarter courses to quantify small but precisely-estimated effects of high-$\hat{G}$ professors on their students: students who take courses with high-$\hat{G}$ professors become more likely to take additional courses with that professor, take more courses in that field, and are more likely to earn a major in that field. These findings are highly robust to alternative specifications; persist in the presence of additional covariates measuring professors' gender, evaluative positivity, explicit gender bias, and attentiveness to students; and exhibit minimal heterogeneity by discipline, time, or other characteristics. The results suggest that both male and female students are encouraged by teachers whose presentation of constructive feedback adapts to the student's gender.

JEL Codes: C55, I23, J16, N32

*Thanks to Henry Brady, David Card, Len Carlson, Gregory Clark, Brad DeLong, Barry Eichengreen, Emily Eisner, Laura Giuliano, Anne Maclachlan, Rodolfo Mendoza-Denton, Jesse Rothstein, Martha Olney, Yotam Shem-Tov, Zoë Steier, Ellen and Eugene Switkes, Basit Zafar, and seminar participants at Georgia State, UC Davis, UC Berkeley, and UC Berkeley's Computational Text Analysis Working Group for valuable comments. The author was employed by the University of California in a research capacity throughout the period during which this study was conducted, and acknowledges financial support from UC Berkeley's Center for Studies in Higher Education and Institute for Research in Labor Economics. All errors remain my own.

†Department of Economics, UC Berkeley. E-mail: bleemer@berkeley.edu.

1 Introduction

Decades of scholarship have documented the prevalence of gender stereotypes and their role in shaping behavioral expectations (Broverman et al., 1972; Ellemers, 2018). In circumstances where third-party evaluators are judging applicants' expected performance, as in job candidate reviews, stereotypes have been shown to exacerbate inequities (Neumark, 1996; Riach and Rich, 2006; Quadlin, 2018; Sarsons, 2019), motivating a growing literature documenting differences in how men and women are described in employment-related evaluations like letters of recommendation (Schmader, Whitehead, and Wysocki, 2007; Madera, Hebl, and Martin, 2009; Dutt et al., 2016) and employee performance reviews (Biernat, Tocci, and Williams, 2012; Correll, 2019).[1] Less is known about how young workers and students respond to the stereotypes of the parents, teachers, and managers with whom they regularly interact, though the educational decisions often made on the basis of these adults' advice – like college persistence and major choice – are known to have high stakes (Card, 1999; Kirkeboen, Leuven, and Mogstad, 2016).[2] Whether teachers' and mentors' gender stereotypes facilitate or frustrate communication of constructive feedback to their younger mentees – and whether mentees respond positively or negatively to mentors whose feedback adapts to their mentees' gender – remains an open question.

The proliferation of digital text and text-analytical tools has substantially enhanced scholars' ability to observe gender stereotypes outside of laboratories, but previous studies scrutinizing text to identify gender stereotypes have faced two key challenges. The first challenge arises in disentangling gender stereotypes' specific contribution to the observed multidimensional differences between texts describing men and women. Unlike prior studies of how a subject's gender impacts the behavior of their evaluators and teachers (Bertrand and Mullainathan, 2004; Dee, 2005; Carlana, 2019), the subjects of evaluative text do not have randomly- or quasi-randomly-assigned genders, and descriptive differences could be confounded by selection bias or other factors. Second, as a result of data unavailability and limited research designs, it has proven challenging to causally link evaluators' gender stereotypes to differences in the actual outcomes of individuals – male or female – whom they evaluate.[3]

In this study, I present a massive new corpus of evaluative texts and a novel research design to study how the average "genderedness" of university professors' evaluations – that is, their systematic differential use of descriptive vocabulary in evaluations of male and female students – impacts their students' field of study choices. I study the University of California, Santa Cruz's "narrative evaluations," paragraph-length performance evaluations written by professors for each of their students alongside letter grades.[4] I estimate evaluations' genderedness by comparing the adjectives and adverbs used to describe students of different genders who received the same grade in the same course-term, and then define a characteristic of professors called $\hat{G}$, the degree to which they tend to employ more female-valence vocabulary in evaluating

[1] Sprague and Massoni (2005) and Schmidt (2015) document gendered language differences in teaching evaluations. Jakiela and Ozier (2018) find that even gendered pronoun use negatively correlates with female labor market participation in a cross-region setting.

[2] Carlana (2019) and Canning et al. (2019) show how specific beliefs held by teachers about students' relative ability and potential for growth contribute to student achievement gaps.

[3] Because this study's data only include male and female gender categories, I omit students who do not report a gender from the estimation sample and limit discussion to those two genders.

[4] Prior to 2000, in most cases students *only* received narrative evaluations in place of letter grades.

female students and more male-valence vocabulary in evaluating male students. I then consider the $\hat{G}$ of undergraduate students' first-quarter professors, estimating the impact of taking a course with a high-$\hat{G}$ professor on a student's likelihood of taking more courses in – or majoring in – the same field, compared to another student who earned the same grade in the same first-quarter course but with a lower-$\hat{G}$ professor.

The resulting analysis highlights the importance of distinguishing between sexist gender stereotypes as sometimes exhibited by third-party evaluators – which have been shown in many settings to exacerbate existing inequities – and value- and performance-neutral $\hat{G}$ measures estimated from evaluations primarily written to provide feedback to students. I estimate adjective gender valences that accord closely with historical norms: male students' work tends to be described as 'humorous', 'interesting', and 'philosophical', women's work as 'excellent', 'beautiful', and 'hard-working'. Female-valence words are more likely to be positive, and Humanities professors' evaluations exhibit more genderedness – using female-valence words to describe female students and vice-versa – than STEM professors'. Both male and female students who take courses with higher-$\hat{G}$ professors are more likely to take further courses with that professor, take more courses in that field, and are more likely to major in that field. This main finding is highly robust to alternative specifications; persists in the presence of additional covariates like professor gender and measures of evaluative positivity, explicit professor gender bias, and professor attentiveness; and exhibits minimal heterogeneity by field, course characteristics, or student characteristics. While there is some evidence that very high levels of $\hat{G}$ can be off-putting to students, these results suggest that professors with moderate $\hat{G}$ measures – who tend to use evaluative language subtly adapted to their students' genders – tend to be more encouraging to their students.

I begin in Section 2 by describing the specific setting of this study. The University of California, Santa Cruz (UCSC) was the largest of a slew of progressive colleges and universities founded in the 1960s that implemented a variety of contemporaneous educational innovations, including replacing letter grades with paragraph-length (and sometimes-longer) "narrative evaluations". Every student received an evaluation for every course, though evaluations for some large courses were written by graduate student assistants or using standardized rubrics. While narrative evaluations joined students' permanent records and may have been viewed by potential employers or graduate school admissions panels, their primary audience was the evaluated students themselves, who could view the evaluations following each term. Grades became mandatory in parallel with evaluations in 2000, and narrative evaluations became non-mandatory in 2010. Table 1 presents several anonymized examples of UCSC's student evaluations. For more details on UC Santa Cruz's student body, I have produced a companion interactive dashboard visualizing the longitudinal characteristics and long-run labor market outcomes of UC Santa Cruz's 1965-2010 students that is available online.[5]

I observe the approximately 1.2 million UCSC narrative evaluations written between 1965 and 1979 and between 1999 and 2009, written by more than 1,000 professors for about 75,000 students. I also observe each student's complete UCSC student transcript, including the grades they received in each course. As I discuss in Section 3, while the post-1999 records were obtained as a clean digital database, the earlier records were acquired as scanned student transcripts and transformed into a computer-readable database using the fOCR protocol (Bleemer, 2018), which combines multiple structured OCR transcriptions of each

[5] See https://www.universityofcalifornia.edu/infocenter/long-run-outcomes-uc-santa-cruz-alumni.

record into a high-quality composite for each student. While this study's main results are estimated using the 1999-2009 data, I duplicate the analysis in the historical records to test robustness and document surprising persistence over time in both measured word valences and the effect of $\hat{G}$ on students' educational choices.

Section 4 describes the study's novel empirical methodology for estimating professors' gendered language use. The main specification uses a fixed-effect linear regression model across narrative evaluations to predict students' gender by indicators for 1,600 frequently-used adjectives (with fixed effects for each letter grade in each course-term), while an alternative specification employs LASSO regularization to limit the set of gender-associated adjectives (Prollochs, Feuerriegel, and Neumann, 2018).[6] Both genders are associated with both positive and negative descriptive language – "original" but "uneven" for men, "lovely" but "tentative" for women– but female-valence adjectives tends to be more positive than male-valence adjectives, and are associated with higher average grades. Each evaluation is characterized by its measured female-genderedness $\hat{F}$ predicted from the model (excluding the fixed effects), and professors are assigned estimated $\hat{G}$ measures defined as the difference between the average normalized $\hat{F}$'s of their evaluations written for female and male students. The departments with the highest average $\hat{G}$ are literature and art, with the average professor providing evaluations with about 0.3 standard deviations more-female-valence language to female students relative to male students; the lowest-$\hat{G}$ departments were electrical engineering and applied math, in which the average professor's evaluations exhibited no measurable difference between male and female students. Departments explain 12 percent of variation in genderedness across professors, leaving substantial within-department variation across professors.

Having measured each UCSC professor's $\hat{G}$, in Section 5 I present the empirical methodology used to estimate the impact of having a high-$\hat{G}$ professor on student outcomes.[7] Assuming that the professors teaching students' first-quarter courses are quasi-randomly assigned (conditional on which courses the students enroll in), I estimate linear regression models of whether students persist in the course's field of study after their initial course, with fixed effects absorbing variation across course-letter-grade pairs and cohort years. Students who take the course with an professor with a 1 unit higher $\hat{G}$ – that is, professors who give their male students 1 s.d. more-male-gendered evaluations than their female students, and vice-versa – are more than 20 (s.e. 7) percentage points more likely to take another course from that professor, take about 1.5 (0.4) additional courses in that department, and are as much as 10 (3.2) p.p. more likely to earn a major in that field.

The estimated encouragement from high-$\hat{G}$ professors is similar for male and female students (though the effect appears slightly higher for female students), and other covariates that could explain field persistence – including professor gender, class size and gender composition, and measures of professors' evaluative attentiveness, evaluative positivity (measured using a standard sentiment analysis tool), and differential evaluative positivity by student gender – appear uncorrelated with the effect.[8] Students' encouragement by

[6]To avoid over-fitting concerns in the second-stage analysis below, first-quarter fall courses are held out of genderedness estimation.

[7]Importantly, these estimates describe the impact of high-$\hat{G}$ *professors* on student choice, not the effect of their own specific written evaluations, which could reflect other heterogeneity across students.

[8]Women are shown to become more likely to persist in a major when they have a female professor or more female students in their class (relative to impacts on male students), as has been shown previously in other settings (Bettinger and Long, 2005; Carrell, Page, and West, 2010; Zolitz and Feld, 2018).

high-$\hat{G}$ professors is also strikingly homogenous, with no observable differences in the effect over time, between STEM and non-STEM fields, among students with higher or lower grades, or many other student, professor, and class characteristics; however, high-$\hat{G}$ professors who also display observable gender bias – by generally providing less-positive evaluation to female students – are substantially less encouraging to female students.

I conduct a number of robustness checks to ensure that the estimated results are not sensitive to modeling choices or the particular setting of UCSC in the 2000s. In addition to estimating $\hat{G}$ using LASSO regularization to eliminate possibly-spurious correlations between word choice and student gender, I also estimate leave-one-out $\hat{G}$'s by professor to avoid over-fitting specific words used by few professors; neither meaningfully alters the reported estimates. The proposed causal research design is similar in spirit to a recent literature on 'judges designs', which exploit the random assignment of judges to criminal defendents (Aizer and Doyle, 2015; Dobbie, Goldin, and Yang, 2018); I show that first-quarter students' course characteristics (number of students; percent students female) cannot be explained by the course's quasi-randomly assigned professor's genderedness. Moreover, I show that professors with higher $\hat{G}$ values do in fact provide evaluations to first-quarter male students with more male-valence language (and vice-versa).

While UCSC required course evaluations for all courses until 2010, some professors had stopped taking them seriously in later years, providing evaluations like "The student received an A" or rubric evaluations in which words were chosen depending on the student's letter grade. I omit short evaluations (with fewer than 50 characters) – which excludes about 20% of evaluations in the 2000s – and the inclusion of course-grade fixed effects means that rubrics will not impact estimation of words' gender-valence. I also re-conduct all of the analysis described above using the 1965-1979 corpus of narrative evaluations (omitting grade fixed effects, since letter grades were not awarded at the time), a period of 'true believers' with very few short or rubric-generated evaluations. As described in Appendix A, I find highly-similar gendered language valences and cross-department patterns to those estimated in the 2000s. Only female students were encouraged by high-$\hat{G}$ professors at the time, though this may be an artifact of being unable to condition on course performance; the same is true in the 2000s absent grade-specific fixed effects. Female students' high-$\hat{G}$ encouragement in the 1970s exhibits similar robustness and homogeneity as in the 2000s.

This study contributes to methodological literatures about gender stereotypes and historical record digitization in addition to providing new evidence on gender stereotypes' role in pedagogy. First, it introduces a new measure of an important dimension of individual gender stereotypes: genderedness, or the degree to which people adapt their evaluative language to the gender of their subject. Exploiting an unusual university policy that resulted in millions of evaluations written by more than 1,000 professors for tens of thousands of students, I isolate the different descriptive language used in evaluations of highly-similar male and female students – students who enrolled in the same course at the same time and obtained the same grade.[9] This setting permits characterization of both the gender-valences of descriptive vocabulary – which will shortly be made available as an associated R package – and the characteristics of the professors who used them in

[9]While the gold standard in studies of gender stereotypes remains randomized control trials, it is likely impossible to obtain 'real-world' evaluative text in which the subject's gender is unknown (or randomly-assigned) to the evaluator. As a result, the highly-detailed UCSC information analyzed in this study may make it the best available setting to isolate differential evaluative language use by subject gender.

more- or less-gendered fashion.

Second, this study provides new evidence that while gender stereotypes are responsible for important labor market inequities, motivating policies like "blind auditions" (Goldin and Rouse, 2000) and the removal of gender-stereotypical decorations Cheryan et al. (2009), policies seeking to eliminate gender-specific differences how university teachers directly communicate with their students could be generally discouraging to both male and female students. First-quarter undergraduate students are shown to be relatively malleable in choosing their field of study, and professors' subtle gender-specific language appears to be an important manifestation of student-attentive pedagogy to which students positively respond. Students' encouragement by professors who employ gendered descriptive language also serves as an explanation for stereotypes' remarkable persistence since at least the 1970s, with the encouragement serving as positive feedback incentivizing continued use.

Finally, this is the first known study to analyze student transcript records digitized from PDF scans of the original documents, made possible by improvements in the computational identification of typos and other errors that typically frustrate the analysis of computer-recognized documents (Bleemer, 2018). The similarity in estimated results between the 1965-1979 digitized records and the 1999-2009 digital records described above also serves as a validation exercise for the quality of the historical records and the fOCR process that produced them, motivating additional research using digitized records to examine longitudinal changes in student behavior and university policies.

2 Background

The University of California, Santa Cruz was founded in 1965 as the eighth University of California campus, and one of three campuses founded in the 1960s.[10] Adopting a residential college model, with eight colleges by 1972 and ten by the mid-2000s, UCSC was intended as a university campus focused on undergraduate education and research; UCSC had no engineering program until 1997, and its professional schools and graduate programs remain small. UCSC was a low-selectivity public university generally accessible to high-performing California high school graduates: its 2000 freshman class of about 3,000 enrollees had an admissions rate of 83 percent and average SAT score of 1150. The student body was more white and less Asian than other California public universities – with the 2000 incoming class about 59 percent white, 20 percent Asian, and 14 percent Chicano/Latino – and tended to have a relatively larger proportion of female students (58 percent). 94 percent of new 2000 enrollees were California residents, typical of California's public universities at the time.[11]

UCSC's "Narrative Evaluation System" was one of a number of progressive educational innovations implemented by the university, and was one of the university's most popular institutions. While in the university's first years some courses did not provide evaluations, by the late 1960s students received paragraph-length evaluations written by their professors (or occasionally by their teaching assistants) in place of letter grades in most of their courses.[12] These evaluations typically included a short description of the course

[10] I am indebted to King (2018) for the historical material presented in this section.

[11] Statistics from https://www.universityofcalifornia.edu/infocenter/freshman-admissions-summary.

[12] Students intending to apply to graduate school were permitted to request letter grades in place of narrative evaluations, but

before detailing the student's performance.[13] UCSC student transcripts thus took the form of many-paged booklets, with the first page listing students' courses (and whether they passed the course) and each remaining page recording one or several evaluations. A set of sample evaluations can be seen in Table 1.

As a result, evaluations could be available to students' future employers and graduate school admissions panels, though their length implies that they were likely rarely used for this purpose.[14] Instead, narrative evaluations' primary role was to record professors' frank evaluations of their students' work for the benefit of the students themselves, who could observe their evaluations at the end of each term.[15] In their 2000 defense of narrative evaluations, the UCSC Alumni Association summarizes the evaluations as feedback that "give students careful and concise criticisms that help them understand the strength and weaknesses of their performance".[16]

By 2000, more than half of students were requesting letter grades in addition to narrative evaluations, and a faculty committee mandated letter grades in all courses starting in Fall 2000. That regime lasted until Fall 2010, when evaluations became non-mandatory.

UC Santa Cruz operates on a quarter system, with most students taking full course loads – usually three or four courses – during three quarters per year: Fall, Winter, and Spring. In the 1970s, many courses – especially courses taken in the first year – were listed in students' residential colleges instead of an academic department, and even in the 2000s most freshman-fall students enrolled in one college-specific course in their first quarter, outside of any academic department. In the 1970s, UCSC students enrolled in their first-quarter classes during a first-year orientation prior to the arrival of continuing students, but in recent years they have enrolled over the summer, often more than a month prior to moving to campus.

3 Data

The data used in this study were provided by the UC Santa Cruz Office of the Registrar as part of its participation in the UC ClioMetric History Project, a massive data collection venture managed by UC Berkeley's Center for Studies in Higher Education and the UC Office of the President (Bleemer, 2018). The two subsections below describe the avaiable data in each of the two periods analyzed in this study.

3.1 1965-1979

In the early period, UCSC provided complete PDF transcript records for every enrolled student between 1965, when the university first enrolled students, and 1979, when the university began transitioning to a

this was hardly ever requested; an audit in the mid-1970s found that 0.003 percent of evaluations were provided as grades (King, 2018).

[13] A small number of large classes used evaluation rubrics with only small personalized differences between students, though the proportion of courses using rubrics grew over time.

[14] A reference from the UCSC Registrar notes that "In addition to the student, performance evaluations will be reviewed by college academic staff, by the student's department, and by anyone to whom a student opts to send the complete official transcript".

[15] When young alumni were asked in the mid-1970s which aspects of UCSC "were most important at the time they were students", the narrative evaluation system was the second-most-common response, ahead of faculty contact and major programs and behind only "student friendships" (Grant and Riesman, 1978).

[16] See https://senate.ucsc.edu/archives/Past%20Issues/narrative-evaluations/AlumniAssocNarrEvalNov2000.pdf.

digital record system. The records had been scanned from paper, and students' names were hand-transcribed. Each record contained three components:

1. Between one and four course record cards, which include permanent student characteristics as well as a table with one row for each course taken by the student (separated by quarter). Available permanent student characteristics include month of first enrollment, birth date, and home town. Course information includes name, department, number, pass/fail grade, and units received. Records are type-written.

2. Many pages of original course evaluations submitted by faculty. Course identifying information is type-written, but many evaluations are hand-written.

3. Several pages of aggregated course evaluations, typed by the Registrar's Office following reception of the original documents. Each evaluation is prefaced by course identifying information.

I process each of these PDF records into a high-quality computer-readable database using the fOCR protocol described in more detail in Bleemer (2018). First, each student record file is processed into XML documents by four separate OCR programs: OmniPage Ultimate, Adobe Acrobat DC 2018, ABBYY FineReader 12, and Tesseract 4.0. After identifying the template of each record page using 'fingerprint words' (like "Record Number" at the top of each course record), the four transcriptions of each course record card were concatenated by information type, eliminating most typos and otherwise-missing information. For example, the algorithm compares each transcription's text observed in the box where the student's year of enrollment was recorded; non-years are discarded, infeasible years are corrected into their most-likely feasible alternative (or discarded if there is no such alternative), and the most-frequently-transcribed year is recorded in the database. Tabular course records are similarly concatenated. Department codes are recognized from a complete dictionary, and infeasible course numbers are corrected or discarded. Finally, courses are matched across students to adjust remaining errors; any infrequently-occuring course that closely matches four of the five identifiable course features – course name, course number, course department, course year, and course quarter – is adjusted to match the more-frequently-occurring course.

Next, typewritten original and composite course evaluations are processed using regular-expression pattern recognition to identify the course's five identifiable features, which are matched to the courses recorded on the student's course record card. This results in a maximum of eight transcriptions of each course's evaluation, with four transcriptions of each of the original and composite records. Evaluations often also include additional course features like course section and professor, which are associated to the course using regular expressions.[17] Again, courses are concatenated across students to match professors: if a class has a single professor in a given quarter, then all students enrolled in that class are associated with the single professor.[18]

[17] When teaching assistants write the evaluation in the place of the professor, the TA's name is also captured. When the evaluation includes both a description of the course and an evaluation of the student's performance, the former is omitted by deleting language prior to "Evaluation:", which always proceeds the latter section.

[18] In particular, classes are defined as having a single professor if at least three students' records list a given professor's name and the second-most-prevalent "professor name" appears fewer times than the maximum of 3 and half the frequency of the most prevalent professor name. Professors are usually listed by their first initial and last name, though first initials are not always available, and sometimes the full first name is provided. Most professors' last names are unique within department, so I define an professor as a last name – department pair. Professors with last names that are not unique within department (that is, last names that

Student genders are identified by matching the human-transcribed first name provided in each student's record file to Social Security Administration records from their year of birth.[19] While UCSC was primarily an undergraduate institution in the 1970s, it also trained some graduate students, who are identified by their course enrollments and omitted from the sample.[20]

3.2 1999-2009

In the more recent period, all UCSC records were provided as digital extracts from the UCSC Registrar's internal database. I observe each student's initial year of enrollment and reported gender along with detailed course records, including each course's department, course number, grade earned, and the first and last names of assigned professors. I also observe students' declared majors and whether they ultimately earned a degree (as of mid-2019). A separate database contains students' narrative evaluations associated with each course, which can be linked to the course records by course-term. As in the historical records, professor genders are identified by matching first names to the SSA name database, with most unusual or androgynous names matched by hand using faculty web sites.

In both the 1970s and 2000s records, I aggregate departments into three disciplines: Humanities, Social Sciences, and STEM (which includes the natural and biological sciences, engineering, and mathematics and statistics). Arts fields are included with the Humanities, and UCSC's unique Community Studies major is included as a Social Science. The only available major which does not neatly fit into this categorization is Education, which I omit from all three discipline. Major categorization details are available from the author.

Full-sample descriptive statistics of both the 1970s and 2000s data are presented in Table 2. While the data quality of the 1970s records is lower than that of the 2000s records as a result of imperfect digitization, there is little reason to expect digitization errors to be correlated with the effects discussed below, suggesting that the noise primarily serves to attenuate the estimated results.[21] The table shows that the full sample includes records for about 27,000 1970s students and 49,000 2000s students; students' gender was evenly split in the 1970s, while 55 percent of 2000s students were female. The average 2000s UCSC student completed 31 courses and received 20 evaluations included in the estimation sample, which omits (a) evaluations with

---

are paired with multiple first initials; usually wife-husband pairs) are omitted from the sample, since it is not always clear which professor is teaching a given course. Once the first initial of each professor-department is determined, I search through all relevant evaluations to identify the professor's first name, if it appears on any evaluation (which occurs 62% of the time among professors who teach at least one course in which a first-year Fall student enrolls). Remaining professors first names are manually-identified by searching through oral histories of UCSC and other historical documents, with > 80 percent success (due to time constraints, I have not identified genders for professors who were not algorithmically matched to first names and who taught fewer than 150 course-students in the period). Finally, professors' genders are determined in the same way as student genders, matching to the SSA first name database, and those with androgynous first names are manually gendered using archival information (with 100 percent success).

[19] SSA records list the annual number of male and female American children born with each first name; I define students' gender when their names were more than 10 times more likely to be assigned to American newborns of one gender than the other, leaving 1-2 percent of names unmatched or androgynous. SSA records include more than 2,000 names for each gender in each year. I begin by matching students to SSA records from their birth year (or 1955, if birth year is unavailable), and then continue matching using subsequent and previous years if no match is identified. Data available at https://www.ssa.gov/oact/babynames/limits.html.

[20] In particular, any student who enrolls in a 200s or 300s level course in their first year of enrollment, or who ever enrolls in a "Graduate Internship" or "Teaching Supervision" course, is defined as a graduate student. The undergraduate record of undergraduates who continued enrollment as graduate students is maintained, but their graduate enrollment is omitted.

[21] Additional details and quality measures for the 1970s data are available in Appendix A.

fewer than 50 characters (like "The student received an A"), (b) evaluations that note that they were written by graduate student TAs instead of faculty, (c) evaluations written for courses without listed professors (including many independent studies), and (d) evaluations written for courses with multiple listed professors (since it's unclear which professor wrote the evaluation). 1970s students took 26 courses but are only associated with 7 evaluations each, both because evaluations were not provided for every course at the time and as a result of computational limitations in matching students' written evaluations to their courses. UC Santa Cruz students tend towards Social Science courses and majors on average, and male students were about 50 percent more likely to take classes or major in STEM fields than female students in both the 1970s and 2000s.

4 Gender Stereotype

In order to measure the degree to which professors employ gender stereotypes in their interactions with students, it is first necessary to precisely define a text-oriented measurement that closely corresponds to prevailing understandings of gender stereotypes. In her important review article on gender stereotypes, Naomi Ellemers notes that "both male and female evaluators tend to perceive and value the same performance differently depending on the gender of the individual who displayed this performance". Gender stereotypes reflect "how we *think* men and women differ from each other," and maintaining such general expectations in the case of specific interlocutors may itself "affect the way people attend to, interpret, and remember information about themselves and others" (Ellemers, 2018).

This definition of gender stereotype is challenging to operationalize even in natural-experimental settings, because differential treatment in response to the actions of differently-gendered individuals may reflect either preconceptions or real differences in average behavior by gender. Consider, for example, an professor who tends toward noting that her female students are "hard-working". This may be for at least three reasons:

1. The female students who enroll in the professor's course always tend to work harder than their male peers (or at least appear to work harder), and therefore work harder in this particular course;

2. The female students work harder in class than their male peers as a result of this the class's being taught by this particular professor;

3. The female students work only as hard as their male peers, but the professor nevertheless believes them to work harder as a result of a stereotype about female students.

The first of these explanations is not specific to this particular professor-student interaction, and the research design discussed below separately absorbs these possible average differences between male and female students' behavior.[22] But both the second and third explanations are important manifestations of

[22]Notice that this first explanation for descriptive differences between male and female subjects is a key confound for the measurement of stereotypes in other text corpora. Unlike unstructured corpora of text like those analyzed in the popular "Google N-grams" tool, UCSC students evaluations provide a fixed context in which evaluative language is used to describe different people.

gender stereotypes as described by Ellemers (2018), and could plausibly have positive or negative ramifications. The second explanation suggests that students' actual behavior changes when they're in the professor's class (presumably in response to some feature of the professor-student interaction), which could either reflect the student's heightened comfort ("being themselves") or an uncomfortable performative act ("playing the part"). Alternatively, the professor's evaluations might compliment (or criticize) characteristics that the students *do not* instantiate but which match their gender's stereotype, which could be self-verifying ("how I want to be seen"; see Swann (1983)) or offensively presumptuous ("that's not me at all").

Because I do not directly observe students' behavior outside of professor's evaluations, the estimates below conflate explanations (2) and (3) into a single dimension measuring the degree to which professors employ gender stereotypes in their interactions with students, which I refer to as $\hat{G}$. The following subsection presents an research strategy for estimating the male or female valence of each written evaluation, and the following subsection uses these genderedness measures to construct $\hat{G}$'s for every UCSC professor. While the tables and figures describe results for the 2000s sample of UCSC students, they are replicated in Appendix A for the 1970s sample of students with generally-similar results.

4.1 Language Measurement

This study defines the use of gender stereotypes as the increased likelihood of an professors' use of female-valence vocabulary to describe female students (or, equivalently, the professor's increased likelihood of using male-valence vocabulary to describe male students). I restrict the vocabulary to adjectives and adverbs, since these words usually describe the student's performance or output in evaluations – including nouns and verbs might heavily weight discipline-specific vocabulary that could confuse the analysis below – and characterize each evaluation by the presence or absence of every such word.[23]

I restrict the corpus of eligible evaluations to those written for students outside of their first-year Fall (to avoid re-using data in the first and second stages of the outcome models discussed below). Let F_i indicate whether student i is female (omitting the fewer than 0.5 percent of students who do not report male or female gender). Let W_{itce} be the large sparce matrix with a column for each of the 1,621 adjectives or adverbs that appear in at least 100 evaluations (to avoid over-fitting). I index student i's transcribed evaluation e from course c taken in quarter q in which she earned letter grade g.[24] In order to construct an index of the genderedness of a given evaluation, I estimate:

$$F_i = \alpha_{tcg} + \beta W_{itce} + \epsilon_{itceg} \tag{1}$$

where β is the parameter of interest. The inclusion of α_{cqg}, fixed effects for every letter grade by course-

When certain words are differentially used to describe female students' or their course performance, conditional on the course for which the students are being evaluated, the differential use reflects how female students are differentially evaluated for doing the same thing – taking the class – as their male peers. Analysis of unstructured corpuses could alternatively identify gendered correlations that result from women's being written about in different contexts than men, conflating genre differences with gender stereotypes.

[23] Adjectives are determined using the dictionary available at https://patternbasedwriting.com/elementary_writing_success/list-4800-adjectives/. Adverbs are defined as words created by adding 'ly' to other words, changing 'y' to 'i' accordingly.

[24] I use "course" to refer to a department-number offering and "class" to refer to a course as taught in a specific quarter.

quarter, means that β only captures differences in professors' language use between students who earned the same grade in the same class, avoiding cross-class variation which could arise from students' non-random sorting across professors, departments, or years.[25] The main specification estimates this model by OLS, though an alternative specification estimates it using a penalized LASSO regression (using 10-fold cross-validation to select λ) to select only gender-relevant vocabulary, forcing about half of the β coefficients to 0; see Appendix B.[26]

Figure 1 summarizes the male and female word valences estimated by Equation 1, characterizing the 40 words with the most-positive and most-negative robust t-statistics associated with their β coefficient. The size of a word's t-statistic is proportional to both the strength of its association with a particular gender as well as its frequency of use; high-t-statistic words are those that are both frequently-appearing and highly-gendered. The figure shows the estimated coefficient and 95-percent confidence intervals for each word, where confidence intervals in this context largely serve as a proxy for word use frequency; words with narrower confidence intervals are more-frequently-used in evaluations. The adjectives and adverbs most closely associated with male evaluations are 'late', 'humorous', and 'interesting'; with female evaluations, 'hard', 'excellent', and 'beautiful'. Less-common words strongly associated with male evaluations include 'satirical', 'wry', 'electric', and 'violent', while female students are associated with words like 'emotionally', 'gracefully', 'compassionate', and 'upbeat'. Nearly all of the most-female-gendered descritptive words are generally complimentary, while male-gendered words are more mixed between complimentary and critical. A full set of descriptive language gender valences will soon be released as an R package.

Let $\hat{F}^*_{itce} = W_{itce}\hat{\beta}$ be the partial predicted values of whether an evaluation is written for a female student, estimated using only the presence or absence of adjectives and adverbs in the evaluation (omitting the fixed effects). Then define

$$\hat{F}_{itce} = \frac{\hat{F}^*_{itce} - mean(\hat{F}^*_{itce})}{sd(\hat{F}^*_{itce})} \tag{2}$$

to be the normalized genderedness of the evaluation, which aids the values' interpretability.[27] An evaluation with $\hat{F}_{itce} = 1$, for example, includes adjectives and adverbs the combination of which are one standard deviation more likely to appear in the evaluation of a female student than a male student. Table 1 presents a set of sample evaluations ordered by estimated $\hat{F}$, providing examples of evaluations that are include more female- or male-valence language.

Table 3 presents OLS-estimated descriptive statistics of the estimated female-genderedness of students' evaluations, conditional on field of study and the year in which the course was taken. The first column shows that female students receive evaluations that are more female-gendered by 0.17 standard deviations on average compared to male evaluations, a large gender difference that nevertheless implies substantial overlap in the degree to which male and female students' evaluations are female-gendered. Evaluations for female students in STEM courses are less female-gendered by 0.09 s.d. than Social Science courses, which themselves provide female students with evaluations that are 0.13 s.d. less female-gendered than

[25] The 1970s version of this model omits all g indices, since grades were not awarded at the time.

[26] The R package $felm$ (version 2.8-2) is used to estimate all fixed-effect linear regressions in this study, while the $glmnet$ (2.0-16) package is used to estimate LASSO models.

[27] $\hat{F}^*_{itce}$ has a mean of 0.012 and a standard deviation of 0.082.

those of Humanities and Education courses (which provide female students evaluations that are 0.28 s.d. more female-gendered than evaluations for male students).

The third column of Table 3 shows weak evidence that female professors provide female students with evaluations that are more female-gendered by 0.06 standard deviations, though the difference is only statistically-significant at the 10 percent level. This suggests that the same descriptive language that evaluators tend to use in describing female students is also more frequently used by female professors in their evaluations, especially when describing female students. However, this result is explained by the fact that female professors are more likely to teach in Humanities fields; conditional on gender differences across fields, female professors provide somewhat more female-gendered evaluations to both male and female students, but the difference is statistically insignificant. The fourth column of Table 3 formalizes the claim that female-gendered language tends to be more evaluatively-positive: both male and female students who earn higher grades in the respective course (as measured by GPA normalized across all available grades) receive evaluations that are more female-gendered, with increases per standard deviation of grade by 0.2 s.d. for male students and 0.25 for female students.

These relationships highlight key features of evaluation genderedness that will provide important to modeling the impact of high-$\hat{G}$ professors – that is, professors who differentially use more female-valence vocabulary in their evaluations of female students and more male-valence vocabulary in their evaluations of male students – in the next section. Correlations between having high-$\hat{G}$ professors and student outcomes could be confounded by the correlation between evaluations' genderedness and students' performance, field, or other student-specific characteristics. It will be important to only compare outcomes for students who receive the same grade in the same course and to directly test whether the results are confounded by professors' evaluative positivity.

4.2 Professor Genderedness

Given this measure of the degree to which each evaluation employs female-gendered language, I define each professor p's $\hat{G}_p$ as the difference between the average $\hat{F}$ (estimated female-genderedness) of their evaluations written for female students and the average $\hat{F}$ of their evaluations of *male* students:

$$\hat{G}_p = \frac{\sum_e \hat{F}_{itce} F_i}{\sum_e F_i} - \frac{\sum_e \hat{F}_{itce}(1 - F_i)}{\sum_e (1 - F_i)} \tag{3}$$

Differencing removes any fixed component of professors' tendency to employ gendered language in evaluations, isolating the differential degree to which they target female-gendered vocabulary at female students (and vice-versa). I discuss the separate role of the two components of $\hat{G}_p$ in inflencing male and female students' outcomes in the Robustness section below. I describe professors with high $\hat{G}_p$ as employing gender stereotypes to a greater degree than low-$\hat{G}_p$ professors because their evaluations of female students differ from those of male students along the gender stereotype dimension defined by Equation 1. First-year Fall evaluations are omitted in order to characterize professors' $\hat{G}_p$ separately from their specific treatment of first-year students, and $\hat{G}_p$ is only calculated for professors who have written at least 25 male and 25 female evaluations in the database to minimize noise.

As discussed above in defining stereotypes in the context of this study, note that $\hat{G}_p$ *does not* characterize the degree to which professors differentially negatively evaluate or discourage male and female students, a more explicit measure of professor sexism. Instead, $\hat{G}_p$ reflects professors differential use of the female- and male-gendered vocabulary when evaluating male and female students, and may (among other things) reflect professors' attentiveness and personalized knowledge of their students. I define an explicit measure of professor sexism in the next section.

Figure 2 shows OLS estimates from a regression of $\hat{G}_p$ on academic department indicators, omitting departments with fewer than 5 covered professors and combining professors who teach residential college courses into a single 'department'. Professors in the hard sciences, engineering, and economics have the lowest measured genderedness, likely because evaluations in courses that teach specific skill sets are often restricted to limited functional language that leaves little room for substantive character description, while professors in writing, literature, and art courses have the highest measured genderedness levels. Average $\hat{G}_p$, which is measured in units of average standard deviations of evaluation genderedness between professors' evaluations of female and male students, ranges from approximately 0 in electrical engineering to 0.41 in English Literature. Field of study explains 12 percent of variation in $\hat{G}_p$ across 1,428 professors.[28]

5 Educational Outcomes

The key challenge in identifying professors' impacts on their students is students' non-random assignment to professors. If students choose professors based in part on characteristics correlated with professors $\hat{G}$, then apparent relationships between professor $\hat{G}$ and student outcomes could be the result of their selection. For example, if female students interested in pursuing a major try to avoid taking courses with gendered professors who they think might try to discourage them from their intended field of study, then $\hat{G}$ might appear to push female students out of fields of study (since more-committed female students take courses with less-gendered professors). In the opposing direction, if female students tend to major in Humanities disciplines (despite taking courses in other departments) and Humanities professors are more-gendered on average, then it would appear that $\hat{G}$ encourages female students into fields of study.

In order to avoid selection bias, the research design employed in this study restricts the analysis sample to courses taken by first-year students in their first quarter and estimates within-course-grade effects that compare the students' major choices with those of other students who enrolled in the same first-year course (and earned the same letter grade) in a different year (with an professor with a different $\hat{G}$). First-quarter students at UCSC made their course selection prior to arriving at the campus for the first time, and most of the departmental courses in which they enrolled were specifically targeted to first-quarter students, such that the students would have little choice over which professor with which to take the course even if they were choosing courses based on course professors. The students can therefore reasonably be treated as "professor takers", in the sense that they chose courses without choosing over professors. Similarly, professors are assigned to courses prior to knowing those course's enrollments, making them "student takers". I test this

[28] Female professors do not have higher measured levels of genderedness conditional on department. As a result, conditioning on professor gender has little impact on these departmental estimates (see Appendix Figure A-1).

assumption in the Robustness section below.

The only remaining dimension along which each student-professor match differs is time, with some matches happening earlier and others later in the 1965-1979 period. As a result, I estimate the following model of student outcomes Y_{ict}, including whether the student takes any more courses in that department or with that professor, the number of courses they take in the department, and whether the student earns a major in the department:

$$Y_{ictg} = \alpha_{cg} + \gamma_t + \beta_1 F_i + \beta_2 \hat{G}_{p_{ct}} \times (1 - F_i) + \beta_3 \hat{G}_{p_{ct}} \times F_i + \delta X_{ict} + \epsilon_{ict} \quad (4)$$

estimated over the sample of departmental first-year Fall courses c taken by UCSC students i in quarter t with professor p_{ct}. The parameters of interest are β_2 and β_3, which estimate the impact of professor $\hat{G}_{p_{ct}}$ on male and female students, respectively. Fixed effects α_{cg} and γ_t capture course-grade and time fixed effects. The main effects are estimated with an empty δX_{ict}, but additional controls will be added below to test the presence of alternative channels through which high-$\hat{G}_{p_{ct}}$ professors could encourage students to take more courses in their field other than their employment of gender stereotypes. Standard errors are two-way clustered by student and professor.[29]

The estimation sample for Equation 4 is substantially narrowed from the full set of students – by about 20 percent in the 2000s and 45 percent in the 1970s – for several reasons. First, I omit the small number of students whose first UCSC courses were not in the fall quarter, since by fall they may have obtained information about the faculty that would invalidate the quasi-random student-professor matching assumption discussed below. More importantly, I omit any student who did not take a first-quarter fall course satisfying the following criteria:

1. The course must be in an academic department, in order to test whether professors' characteristics influence students' persistence in that department. This eliminates both courses taught in students' residential colleges (a large share of 1970s first-quarter courses) and college writing courses (which were common in the 2000s).

2. The course must not be in mathematics. Mathematics courses are generally required by a large array of academic departments, and even a highly-'encouraging' math professor would likely students to take courses in any of an array of other departments, challenging identification of math professors' impact on student educational choices.[30]

3. The course must be taken for a grade, and must be taken from a professor for whom $\hat{G}$ can be calculated (that is, professors who have written evaluations for at least 25 male and female non-first-quarter students).

[29] While these standard errors could be downward-biased since they treat $\hat{G}$ as observed, the massive sample used to produce $\hat{G}$ leads to only minor changes in the estimates when bootstrapped; see Appendix Table **??**. I produce these estimates by bootstrapping Equations 1 and 2 800 times over the full evaluation sample and then using the estimates of $\hat{G}$ to bootstrapped estimates of 4 clustered by professor.

[30] Results including mathematics courses are nevertheless little-changed; see Appendix Table A-2.

The resulting "Estimation Sample" is described in Table 2. The students in the estimation sample are similar on observables to the full student sample, though they have slightly higher graduation rates (mostly in the social sciences). They took an average of 3.3 courses in their first year, receiving evaluations in 2.5 of them (though only 1.7 evaluations per student are eligible to be included in the analysis). Of those courses, about 45 percent were in the Social Sciences, 35 percent in STEM, and 20 percent in the Humanities.

Notice that the measured causal impact results from treatment with a high-$\hat{G}$ professor, not strictly treatment with stereotyped language. As a result, interventions altering the language used in written evaluations may not itself impact students' educational decisions according to this model. For example, high-$\hat{G}$ professors may elicit differential (more 'stereotypical') behavior from their male and female students, such that their gendered evaluations accurately reflected behavioral differences, and female students' preferences over those behavioral differences (not over evaluative language) could explain female students' tendancy to take more courses in high-$\hat{G}$ professors' departments. The terms "genderedness" and "gender stereotypes" capture this dualism: whether or not more-gendered professors inspire more stereotypical behavior among their students, the gendered language in their evaluations reflects a quality of the professors that encouraged female students into the professor's department. The relevant marginal adjustment would be from a high-$\hat{G}$ professor to a low-$\hat{G}$ professor, across all dimensions on which such professors differ on average.

5.1 Main Results

Table 4 presents estimates of β_1, β_2, and β_3 for a progressive series of students' enrollment and major choices. The first row shows that female students are more likely than male students to never take another course in departments in which they take first-quarter courses, a finding which persists on the intensive margin as well; Columns 3 and 6 show that women take fewer courses and are less likely to major in fields in which they took courses in their first-semester course departments, relative to their male peers. The first column also shows that female students who take an introductory course with a high-$\hat{G}$ professor become more likely to take another course in that field, while the positive effect on male students is statistically insignificant. However, both male and female students are substantially more likely to take additional courses with high-$\hat{G}$ professors. The standard deviation of professor genderedness is about 0.15, suggesting that male and female students with a one-s.d. more-$\hat{G}$ professor in a first-quarter course become about 3.5 (s.e. 1.03) percentage points more likely to take another course with that professor at some point in their academic career.

These short-run encouraging effects of high-$\hat{G}$ professors snowball into ramifications for students' entire university curriculum. A shift from the 25th percentile $\hat{G}$ to the 75th percentile $\hat{G}$ in a class's professor – from $\hat{G}$=0.04 to 0.21 – is expected to increase the number of courses in that field taken by each student by about 0.26 and increase their likelihood of earning a major in that field by about 1.4 percentage points, with effects slightly (but statistically-insignificantly) larger for women than for men.[31] Many students take a large number of courses in a field of study without ever declaring it their academic major; when students who took at least 9 courses in the department are included as "majors", my preferred definition, the increase

[31] Number of courses are winsorized at the 95th percentile to avoid results being driven by outliers; estimates are insensitive to alternative threshold choices.

in major likelihood increases to about 1.6 percentage points. While these effects are relatively small – out of a class of 100 students, the higher-$\hat{G}$ professor only encourages 2 students who would have otherwise chosen other majors to choose this field instead (or in addition to) – they nevertheless suggest that professors who adapt their evaluations to their students' genders *encourage* both male and female students to persist in that field.

One possible concern with these results is a mechanical correlation that could arise if professors who use unusual descriptive language in their evaluations tend to teach large numbers of female students who choose to major in that department. In this case, Equation 1 could over-fit those professors' descriptive word choices and associate them with female students, leading them to artificially-higher $\hat{G}$ and a correlation with major choice. Appendix Table **??** replicates Table 4 using leave-one-out (LOO) measures of $\hat{G}$, in which the underlying stereotype regression is run separately for each professor, omitting that professor from estimation.[32] The resulting leave-one-out predicted values are then used to estimate each professor's $\hat{G}$ level. The LOO estimates appear to strengthen slightly, but show a statistically-similar positive relationships across all findings, suggesting the absence of this confounding channel.

Of course, many other factors are also important to the major choice decision of first-quarter students, some of which are likely correlated across professors with their measured $\hat{G}$. Table 5 investigates whether high-$\hat{G}$ professors' encouragement into major choice can be instead explained by other characteristics of those professors, or the students who take courses from them, by adding covariates to X_{ict} in Equation 4. All covariates are added interacted with gender, estimating separate effects for male and female students. The first two columns show that the result magnitudes, but not their direction, are sensitive to the inclusion of course-grade and year fixed effects, while Column 3 replicates the final column of Table 4.

Column 4 adds an indicator for the instructor's gender, interacted with the student's gender. While the baseline results remain little-changed, female professors do seem to increase female students' likelihood of earning that major relative to male students (by about 2 percentage points, with the difference statistically-significant).[33]

Column 5 adds two characteristics of the courses in which first-quarter students enroll. While those students themselves are unlikely to have chosen the course on the basis of its professor, some of their more-senior peers might have done so, and the resulting course composition could thereby mediate high-$\hat{G}$ professors' impact on students. The new covariates measure the number of students in the course and the percent of students in the course who are female, both normalized across all courses. Again, the addition leaves the main results largely unchanged, though increases in class size and the proportion of female students appears to dissuade major choice by male students.[34]

Column 6 of Table 5 adds covariates directly measuring the evaluative positivity and negativity of the evaluations received by each student in the first-quarter course, testing whether high-$\hat{G}$ professors' impact on student major choice can be explained by high-$\hat{G}$ professors tending to provide more- or less-positive evaluations to their first-quarter students. Positivity and negativity are measured using a standard publicly-

[32] LOO estimates are currently available only for 1970s estimates.

[33] See Bettinger and Long (2005) and Carrell, Page, and West (2010).

[34] (Cohoon, 2001) and Zolitz and Feld (2018) show similar relative increases in female enrollment resulting from a higher proportion of female students in a class.

available sentiment analysis tool; each evaluation is assigned a measure of positivity and negativity, and I then normalize each measure across the full set of evaluations.[35] I find that positive and negative evaluations have large effects on students' persistence in the field, though the effects differ by gender–male students appear more sensitive to negative feedback (becoming 15 p.p. less likely to earn the major as a result of a 1 s.d. increase in negativity), while female students who receive 1 s.d. more-*positive* evaluations are 10 p.p. more likely to choose the major. While it is tempting to interpret these findings causally, with students responding to their professors' encouragement by choosing to persist in the field, they could alternatively reflect superior within-grade performance or particular student comparative advantages that would have led students to continue in the major irrespective of their professors' encouragement.[36] Nevertheless, Column 6 shows that professors' evaluate positivity in students' courses is not responsible for the relationship between high $\hat{G}$ and student persistence; adding measures of evaluative positivity hardly change the main coefficients.

Columns 7 and 8 test whether alternative characterizations of high-$\hat{G}$ professors absorb part of the main effect. Column 7 develops measures of professors' encouragement and sexism using the positivity and negativity of professors' evaluations written for other courses. I define professors' "average positivity" as the average difference between measured positivity and negativity in all non-first-quarter evaluations that they've written, and "average positivity by gender" as the difference between their average positivity for male students and their average positivity for female students. This latter definition can be understood as professors' explicit sexism, as opposed to their use of gender stereotypes when interacting with students; professors with high "sexism" tend to provide more-positive reviews to male students than to female students. Once again, adding these additional covariates hardly changes the main estimated results, though they are interesting to interpret in their own right; while 'sexist' professors appear encouraging to male students and discouraging to female students (though the coefficients are very noisily estimated), professors with high "average positivity" appear to *discourage* students; a 1 s.d. increase in average positivity causes a 3.5 p.p. decline in female students' likelihood of persisting in the major, with a smaller (and statistically insignificant) effect for male students. It appears that conditional on students' grades, having a more generally-positive professor actually leads students to leave the field, perhaps seeking more-critical feedback in other disciplines.

Finally, I develop a measure of professor attentiveness, on the supposition that higher-$\hat{G}$ professors encourage their students just because they write longer and more-attentive evaluations, which by their nature may be more gendered. In fact, even a tenth-order polynomial of evaluation length explains less than 1 percent of variation in evaluation's genderedness, but it could be that professors who better know their students could appear to have higher $\hat{G}$ but actually encourage their students for other reasons. I test this hypothesis by developing a measure of professors' evaluative attentiveness. For each course, I measure the degree of variation in the descriptive language used by the professor in the course WV_{ct}, relying on the fact that more-attentive professors are likely to provide more-personalized student evaluations that differ from

[35] I use the $QDAP$ sentiment dictionary to measure positivity and negativity, implemented using the $SentimentAnalysis$ R package, version 1.3-3.

[36] For analysis of how students of different genders differentially respond to professors' encouragement, including higher grades, see (Owen, 2010; Goldin, 2015; Kugler, Tinsley, and Ukhaneva, 2017).

each other, using the following metric:

$$WV_{ct} = \frac{1}{|E_{ct}|} \sum_{e \in E_{ct}} \Big(\frac{1}{W_e} \sum_{w \in W_e} \frac{\sigma_{wct} - 1}{|E_{ct}|} \Big) \tag{5}$$

namely, for every adjective or adverb $w \in W_e$ in evaluation e, the percent of other evaluations in class c in t that also used that word ($\sigma_{wct} - 1$), averaged across words within e and then averaged across evaluations E_{ct} written for that class. I also characterize the 'instructor word variation' of each professor by taking the average WV_{ct} for other classes taught by the same professor (excluding classes with first-year fall students), characterizing professors by their average level of variation in descriptive language.

Column 8 includes each of these attentiveness measures interacted with gender. I find that female students in classes that receive more-varying evaluations become somewhat more likely to persist in the major, though the result is only statistically significant at the 10 percent level. Otherwise, these measures of attentiveness do not appear to meaningfully contribute to students' major choice decision on top of the other factors influencing that choice, and do not meaningfully shift the main estimated coefficients, which remain approximately unchanged from their values estimated with a null X_{ict}.

5.2 Heterogeneity

Table 6 interacts $\hat{G}$ with class, professor, and student characteristics to measure how the relationship between $\hat{G}$ and subsequent major choice differs in different settings. In particular, I estimate:

$$\begin{aligned} Y_{ictg} = \alpha_{cg} + \gamma_t + \beta_1 F_i + \beta_2 \hat{G}_{p_{ct}} \times (1 - F_i) + \beta_3 \hat{G}_{p_{ct}} \times F_i + \beta_4 V_{ict} + \beta_5 V_{ict} * F_i + \\ \beta_6 \hat{G}_{p_{ct}} \times V_{ict} \times (1 - F_i) + \beta_7 \hat{G}_{p_{ct}} \times V_{ict} \times F_i + \epsilon_{ictg} \end{aligned} \tag{6}$$

where V_{ict} is a characteristic of the student, professor, or class. Table 6 estimates Equation 6 for many definitions of V_{ict}, including most of the covariates discussed in the previous subsection. Evidence of heterogeneity would appear as statistically-significant estimates of β_6 or β_7; for example, if the relationship between high-$\hat{G}$ professors and major choice weakens over time among male students, then I would estimate a negative β_6 when V_{ict} is defined as year.

In fact, Table 6 generally shows remarkably minimal evidence of heterogeneity. Despite the relatively-low $\hat{G}$ measures of STEM professors, the first column of Table 6 shows that high-$\hat{G}$ STEM professors are if-anything *more* encouraging to both male and female students than low-$\hat{G}$ STEM professors, though the difference is statistically insignificant. Interestingly, however, the relationship between $\hat{G}$ and encouragement appears substantially (though statistically-insignificantly) lower among female professors compared to male professors; while the relationship remains positive, it appears that the use of gender stereotypes by female professors hardly encourages male or female students, whereas high-$\hat{G}$ male professors appear much more encouraging.

I do not estimate any measurable heterogeneity in the main effect of professors' use of gender stereotypes by the number of students in the course, the percent female students in the course, the year in which the course was taught, or the grade that the student receives in the course; even students who earn very low

grades appear encouraged by professors with high $\hat{G}$. However, there are interesting interactions between professors' $\hat{G}$ measures and their use of positive and negative language. While the previous section shows that professors who tend to write more-positive evaluations tend to discourage major persistence, column 6 shows that that effect is wholly absorbed by heterogeneity by $\hat{G}$: professors who generally provide more-positive evaluations tend to be less encouraging even if they are high-$\hat{G}$ (for both male and female students), while having high $\hat{G}$ is more encouraging among professors who tend to give more-negative evaluations. While the main effect remains relatively large and positive, this suggests that the subtle gender-specific adaptations made by high-$\hat{G}$ professors are more impactful when used in providing more-critical feedback.

The seventh column shows that high-$\hat{G}$ 'sexist' professors – that is, professors who tend to provide more-negative feedback to female students relative to male students – are less-encouraging for female students than less-sexist professors. This unsurprising mediation suggests that female students are less receptive to stereotype-facilitated professor-student interactions when the professor also exhibits a tendency to provide more-critical feedback to female students.

Finally I find that the level of vocabulary variation employed by the professor in the course positively covaries with the main effect: students in courses in which their professors write more-personalized evaluations are more-encouraged by high-$\hat{G}$ professors than those in courses where the professor writes less-personalized reviews. This provides additional evidence that high-$\hat{G}$ professors are interacting differently with their students than low-$\hat{G}$ professors – apparently by adapting their interactions to their students' genders – and that students who are better-known by their professors are more encouraged to continue in the field of study by these interactions, no matter the student's gender and no matter the field.

5.3 Robustness

Table 7 presents a series of robustness checks testing some of the modeling assumptions discussed above. The first two columns test the conditional quasi-random assignment of male and students to high-$\hat{G}$ first-quarter instructors by attempting to predict the course's normalized number of students or percent female by the professors' $\hat{G}$ using Equation 4. As expected, there is no measurable correlation; courses taken by first-quarter students taught by high-$\hat{G}$ instructors have similar composition to those taught by low-$\hat{G}$ instructors, conditional on course-grade and year fixed effects.

The second two columns of Table 7 document the relationship between first-quarter students' evaluations' $\hat{F}$ measures, their professors' average $\bar{F}_{Male}$ and $\bar{F}_{F}emale$ measures (the average non-first-quarter $\hat{F}$ values of professors' male and female evaluations, and the two components used to construct the single-dimensional $\hat{G}$), and students' persistence in the major. Column 3 shows that professors' degree of stereotyping is well-defined across courses; professors who tend to provide more male-valence evaluations to male students provide more male-valence evaluations to first-quarter male students, and those who tend to provide more female-valence evaluations to female students provide more female-valence evaluations to first-quarter female students. The fourth column interestingly shows that professors' gender stereotypes when interacting with male and female students similarly impact male and female students' likelihood of field persistence; male students, for example, are more likely to persist in a field if their professor employs male gender stereotypes in their evaluations of male students, but are also more likely to persist (with similar magnitude) if the

professor employs female gender stereotypes in their evaluations of female students. These results suggest an important symmetry in students' responses to professors' employment of stereotypes, justifying the main results' collapsing these gender-specific stereotype characteristics into the single $\hat{G}$ measure of the degree to which professors' evaluations adapt to their subjects' gender.

Finally the last column of Table 7 re-estimates Equation 4 allowing for a quadratic relationship between $\hat{G}$ and major persistence. The main linear effect remains positive and increases slightly, while the quadratic terms are negative (with the male quadratic term statistically-significant). The vertices of both quadratic relationships are positive and more than two standard deviations of $\hat{G}$ above mean, suggesting that only unusually-high values of $\hat{G}$ lead to declines in student persistence and justifying the simplification of the relationship to a single linear term in the main specification. Nevertheless, the negative quadratic terms signify an important limitation to the encouragement provided by professors' use of gender stereotypes; while small gender-specific adaptations in interactions with male and female students are encouraging, larger differences are likely to discourage students.

5.4 1970s Estimates

As discussed above, Appendix A entirely replicates the previous analysis using the 1965-1979 sample of student evaluations. The appendix shows that the words with the strongest gender valence in the 1970s are surprisingly similar to those identified in the 2000s stereotypes (Appendix Figure AA-1): the top male-valence words (by t-statistic) are 'original', 'entertaining', 'philosophical', and 'wry', while for female students they are 'sensitive', 'regularly', 'quiet', and 'nice'. Evaluations were somewhat more-gendered at the time – female students received evaluations that had higher $\hat{F}$ by 0.25 s.d., relative to 0.17 s.d. – but otherwise exhibited similar positivity and cross-departmental patterns. The main analytical difference is that letter grades were not assigned in the 1970s, so the fixed effects in Equation 4 are restricted to course and year effects. As a result, only female students appear to be encouraged by high-$\hat{G}$ professors (with similar coefficient magnitudes to the 2000s estimates), the same pattern that emerges in the 2000s estimates excluding grade controls.

Appendix Table AA-5 replicates Table 5, showing similar patterns to the 2000s results: female instructors and having more female students in the class increase female students' persistence relative to male students'; more-positive evaluations are correlated with higher persistence; there is no measurable relationship between measures of instructor attentiveness and course persistence. Interestingly, the relationship between $\hat{G}$ and female students' persistence weakens somewhat when the instructor attentiveness covariates are added, though the coefficient remains statistically significant at the 10 percent level.[37] As in the 2000s, it appears that high-$\hat{G}$ professors encourage persistence in their field of study among their first-quarter students, though some of the 1970s effect is absorbed by the fact that high-$\hat{G}$ instructors may differ in their evaluations' attentiveness to their students.

Appendix Table AA-6 shows similarly-minimal evidence of heterogeneity in the relationship between

[37] Note that the coefficient remains statistically significant at the 5% level if the measures of evaluative positivity are omitted, which may be endogenous to professors' $\hat{G}$, suggesting that the $\hat{G}$ measure remains an measurable contributor to female students' major choice.

professors' $\hat{G}$ and students' field persistence. Appendix Table AA-7 mirrors the robustness results reported in Table 7. Overall, this replication exercise provides some evidence supporting the external validity of the encouragement associated with professors' $\hat{G}$, and also shows surprising continuity over time in the gender-valences of descriptive vocabulary, the prevalence of gender stereotypes in written evaluations, and the impact of professors' adapting to their students' genders when interacting with them on the students' subsequent educational decisions.

6 Discussion and Conclusion

This study analyzes a massive rich database of 1.2 million evaluations of students written by their teachers to make three contributions to the existing literature on the impact of teachers' gender stereotypes on youths' educational decisions. First, it develops a new measure of gender stereotypes defined by language use, using the context of a controlled evaluative setting to identify descriptive words that teachers tend to use in describing the performance of male and female college students. Second, it employs this definition of stereotypes to characterize the degree to which a large group of university professors employ gender stereotypes in their interactions with students using a novel formula measuring the distance in gender-stereotype space between the evaluations written by each professor of male and female students. Finally, it exploits the quasi-random assignment of professors to first-semester students in two periods – the 1970s and the 2000s – to estimate how professors who adapt their student interactions to those students' genders impact those students' subsequent enrollment and major choices, finding that the employment of gender stereotypes increases male and female students' likelihood of taking more courses and earning majors in that field. This effect is separate from a number of other contributors to students' major choices and surprisingly homogeneous across a number of alternative settings, including different fields of study (STEM vs. non-STEM), different years, and different professor characteristics (like professors' attentiveness to their students or the degree of measurable sexism in their evaluations).

One plausible explanation for this relationship, which I can only observe in this setting among college-age students, is that male and female students may have been previously exposed to similar stereotyped vocabulary from their families and teachers, contributing to the vocabulary's familiarity (and possibly encouragement) when they arrived in their new university setting (Broverman et al., 1972).[38] A psychological literature pursues this possibility under the name of "self-verification theory", which "assumes that, out of a desire for social worlds that are coherent and predictable, people want others to see them as they see themselves" (Swann and Bosson, 2010); see Swann (1983). Another possibility is that professors' use of gender-tailored evaluative vocabulary could measure the degree to which the teachers attend or adapt to their students in some way orthogonal to the measures of attentiveness discussed in the analysis, with students encouraged by their professors' attentiveness (Corno and Snow, 1986; Bruhwiler and Blatchford, 2011); however, observed measures of professor attentiveness – including class size, variance in descriptive vocab-

[38] As a result of students' positive reaction (in terms of course enrollment and major choice) to professors' genderedness, a plausible explanation for the persistence of professors' gender stereotypes arises: professors' use of stereotyped language was encouraged by their students' enrolling in more courses in the same department, leading professors to continued use of the same gender stereotypes. The small literature on the persistence of gender stereotypes includes Ellemers and Barreto (2015).

ulary used in the course's evaluations, or variance in the descriptive vocabulary used by the professor in their other courses – do not absorb the estimated effect.

Crucially, this study does not claim to examine the role of "gender stereotypes" broadly, and it is possible that there are other ways not documented in this study which encourage or discourage students of either gender from certain fields of study. Instead, the study focuses on the relatively-subtle ways in which mentors adapt the language used in their written evaluations to their students genders, finding that such adaptations encourage students to take additional courses with that professor and elsewhere in the same field. There remains substantial room for future research in similar real-world quasi-experimental settings to better understand the full role of gender stereotypes in driving the important decisions made by young Americans, in the past and today.

References

Aizer, Anna and Joseph J. Doyle, Jr. 2015. "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges." Quarterly Journal of Economics 130 (2):759–803. URL Link.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." American Economic Review 94 (4):991–1013. URL Link.

Bettinger, Eric and Bridget Terry Long. 2005. "Do faculty serve as role models? The impact of instructor gender on female students." American Economic Review 95 (2):152–157. URL Link.

Biernat, Monica, M. J. Tocci, and Joan C. Williams. 2012. "The Language of Performance Evaluations: Gender-Based Shifts in Content and Consistency of Judgment." Social Psychological and Personality Science 3 (2):186–192. URL Link.

Bleemer, Zachary. 2018. "The UC ClioMetric History Project and Formatted Optical Character Recognition." Center for Studies in Higher Education Research Paper Series 18 (3).

Broverman, Inge K, Susan Raymond Vogel, Donald M Broverman, Frank E Clarkson, and Paul S Rosenkrantz. 1972. "Sex-role stereotypes: A current appraisal." Journal of Social Issues 28 (2):59–78.

Bruhwiler, Christian and Peter Blatchford. 2011. "Effects of class size and adaptive teaching competency on classroom processes and academic outcome." Learning and Instruction 21 (1):95–108. URL Link.

Canning, Elizabeth A., Katherine Muenks, Dorainne J. Green, and Mary C. Murphy. 2019. "STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes." Science Advances 5 (2). URL Link.

Card, David. 1999. "The Causal Effect of Education on Earnings." Handbook of Labor Economics 3:1801–1863. URL Link.

Carlana, Michela. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias." Quarterly Journal of Economics 134 (3):1163–1224. URL Link.

Carrell, Scott E., Marianne E. Page, and James E. West. 2010. "Sex and Science: How Professor Gender Perpetuates the Gender Gap." The Quarterly Journal of Economics 125 (3):1101–1144. URL Link.

Cheryan, Sapna, Victoria C. Plaut, Paul G. Davis, and Claude M. Steele. 2009. "Ambient Belonging: How Stereotypical Cues Impact Gender Participation in Computer Science." Journal of Personality and Social Psychology 97 (6):1045–1060. URL Link.

Cohoon, J. McGrath. 2001. "Toward improving female retention in the computer science major." Communications of the ACM 44 (5):108–114. URL Link.

Corno, Lyn and Richard E. Snow. 1986. "Adapting teaching to individual differences among learners." In Handbook of Research on Teaching 3, edited by M.C. Wittrock. New York: Macmillan, 605–629.

Correll, Shelley. 2019. "Inside the Black Box of Organizational Life: The Gendered Language of Performance Assessment." Manuscript .

Dee, Thomas. 2005. "A teacher like me: Does race, ethnicity, or gender matter?" American Economic Review 95 (2):158–165. URL Link.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." American Economic Review 108 (2):201–240. URL Link.

Dutt, Kuheli, Danielle L. Pfaff, Ariel F. Bernstein, Joseph S. Dillard, and Caryn J. Block. 2016. "Gender differences in recommendation letters for postdoctoral fellowships in geoscience." Nature Geoscience 9:805–809. URL Link.

Ellemers, Naomi. 2018. "Gender Stereotypes." Annual Review of Psychology 69:275–298. URL Link.

Ellemers, Naomi and Manuela Barreto. 2015. "Modern discrimination: how perpetrators and targets interactively perpetuate social disadvantage." Current Opinion in Behavioral Sciences 3:142–146. URL Link.

Goldin, Claudia. 2015. "Gender and the Undergraduate Economics Major." Manuscript URL Link.

Goldin, Claudia and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." American Economic Review 90 (4):715–741. URL

Link.

Grant, Gerald and Davis Riesman. 1978. The Perpetual Dream, chap. To Seem Small as it Grows Large: The Cluster Colleges at Santa Cruz. Chicago, IL: University of Chicago Press, 253–290.

Jakiela, Pamela and Owen Ozier. 2018. "Gendered Language." World Bank Policy Research Working Paper 8464. URL Link.

King, C. Judson. 2018. The University of California: Creating, Nurturing, and Maintaining Academic Quality in a Public University Setting. Berkeley, CA: Center for Studies in Higher Education Press.

Kirkeboen, Lars, Edwin Leuven, and Magne Mogstad. 2016. "Field of Study, Earnings, and Self-Selection." The Quarterly Journal of Economics 131 (3):1057–1111. URL Link.

Kugler, Adriana D., Catherine H. Tinsley, and Olga Ukhaneva. 2017. "Choice of Majors: Are Women Really Different from Men?" National Bureau of Economic Research Working Paper 23735. URL Link.

Madera, Juan M., Michele R. Hebl, and Randi C. Martin. 2009. "Gender and Letters of Recommendation for Academia: Agentic and Communal Differences." Journal of Applied Psychology 94 (6):1591–1599. URL Link.

Neumark, David. 1996. "Sex Discrimination in Restaurant Hiring: An Audit Study." Quarterly Journal of Economics 111 (3):915–941. URL Link.

Owen, Anne L. 2010. "Grades, Gender, and Encouragement: A Regression Discontinuity Analysis." The Journal of Economic Education 41 (3):217–234. URL Link.

Prollochs, Nicholas, Stefan Feuerriegel, and Dirk Neumann. 2018. "Statistical inferences for polarity identification in natural language." PLOS One 13 (12). URL Link.

Quadlin, Natasha. 2018. "The Mark of a Woman's Record: Gender and Academic Performance in Hiring." American Sociological Review 83 (2):331–360. URL Link.

Riach, Peter and Judith Rich. 2006. "An Experimental Investigation of Sexual Discrimination in Hiring in the English Labor Market." The B.E. Journal of Economic Analysis & Policy 6 (2):1–20. URL Link.

Sarsons, Heather. 2019. "Gender Differences in Recognition for Group Work." Journal of Political Economy URL Link.

Schmader, Toni, Jessica Whitehead, and Vicki H. Wysocki. 2007. "A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants." Sex Roles 57 (7):509–514. URL Link.

Schmidt, Ben. 2015. "Gendered Language in Teacher Reviews." Manuscript URL Link.

Sprague, Joey and Kelley Massoni. 2005. "Student Evaluations and Gendered Expectations: What We Can't Count Can Hurt Us." Sex Roles 53 (11-12):779–793. URL Link.

Swann, William B. 1983. "Self-verification: Bringing social reality into harmony with the self." Social psychological perspectives on the self 2:33–66. URL Link.

Swann, William B. and Jennifer K. Bosson. 2010. "Self and Identity." In Handbook of Social Psychology, 5th Ed., edited by Susan T. Fiske, Daniel T. Gilbert, , and Gardner Lindzey. New York, NY: McGraw-Hill, 589–628.

Zolitz, Ulf and Jan Feld. 2018. "The Effect of Peer Gender on Major Choice." University of Zurich, Department of Economics Working Paper 270. URL Link.

A Appendix A: Results from 1965=1979 UC Santa Cruz

This appendix complements the main estimates in this study by replicating the complete analysis for the 1965-1979 sample of UCSC students.

A.1 Student Descriptive Statistics

Appendix Table AA-1 presents expanded descriptive statistics of the 1965-1979 UCSC student records. Of the 27,000 undergraduate students in the sample, seven percent of students are missing years of enrollment, with the remaining years all falling between 1965 and 1984, when UCSC completed its conversion from paper to digital records. Major codes are unobserved for 14.1 percent of students, though some of those students may have had actually-blank "Major" fields as a result of dropping out prior to declaring a major.[39] Just over three-quarters of students' records list recognizable "home town" locations. While names are uniformly accurately observed, 1.4 percent of first names cannot be gendered.

The primary research design employed in this study focuses on the departmental courses in which students enrolled in their first-year Fall term. Table AA-1 shows that this includes 14,231 of the 26,934 UCSC students who enrolled between 1965 and 1979. Most students enrolled in three courses each quarter, and most first-year students took two of those courses in their residential college, not an academic department.[40] Only a small number of courses (especially in UCSC's early years) did not provide written evaluations (or they were not preserved by the Office of the Registrar). Many courses were taught by multiple professors in a given quarter, and others provided evaluations written by TAs instead of the professor; both were omitted.[41] Students whose first recorded quarter was not Fall, and who would therefore have little overlap with other new students in their initial courses, were also omitted. Finally, some narrative evaluations failed to match students' course records as a result of imperfect scanning and OCR, and were discarded.

Columns three and four of Table AA-1 provide descriptive statistics for the resulting sample of students with first-year Fall departmental evaluations, and the fifth column tests equality in each statistic between the first-year sample and the full sample. They tend to have started UCSC slightly later than the average student in the sample, as a result of lower prevalence of evaluation provision in UCSC's early years, but there are no statistically-significant differences in their enrollment age, gender, prior location of residence , or birth state. The first-year sample is less likely to have ended up an undeclared student and is more likely to have earned a major in the Humanities, Social Sciences and STEM; this is likely a result of the omission of non-degree students who did not begin in the Fall quarter and never declared a major, likely only enrolling for a short time. In general, the freshman sample appears smaller but otherwise-representative of typical first-year UCSC undergraduates in the 1960s and 1970s.

A.2 Gender Stereotypes

Figure AA-1 shows the 40 male and female words with the strongest association with each gender, as measured by t-statistic. The words are surprisingly-similar to those which appear in Figure1. Figure AA-2 shows the top ten positive- and negative-valence words associated with each gender.

Table AA-3 mirrors Table 3, replacing the grades with the measures of evaluative positivity and negativity defined in the text. It shows that female students and students with more-positive evaluations tend

[39] Some, but not all, of students who never declared a major had "Undeclared" listed as their major field.

[40] Residential college courses are omitted from the analysis below because the outcomes of interest – e.g. whether the student earned a major in that department – are not defined for residential colleges.

[41] When a course is taught by multiple professors I am typically unable to identify which professor taught each student. Evaluations written by TAs are substantially less informative about the professor's gender norms than those written by the professors themselves. Evaluations are marked as being written by a TA if they say so in the evaluation, as many do (e.g. noting "Written by FName LName, TA" at the bottom.)

to receive more female-gendered evaluations. Unlike in the 2000s case, female instructors provide more female-gendered evaluations to both male and female students, but especially to female students.

Figure AA-2 shows the average $\hat{G}$ measures of professors from each academic department, mirroring Figure 2. The highest-$\hat{G}$ departments are English Literature, Literature, and Theater; the lowest departments are Mathematics and (unlike in recent years) Spanish and French. The latter departments appear to have been more narrowly-tailored to teaching language proficiency in the 1970s, which generally inspired short and to-the-point evaluations, while in recent years their courses have overlapped to a greater degree with the Literature departments. While otherwise the department ordering looks fairly similar between the two periods, in general the 1970s coefficients are higher than those in the later period, reflecting larger average differences in the genderedness of evaluations that professors wrote for male and female students.

A.3 Results

Table AA-4 replicates Table 4, presenting the main results for the 1970s UCSC student cohort. Despite the 30+ year difference and a far smaller sample size, the results look very similar for female students, though in the 1970s there is no evidence that male students were encouraged by higher-$\hat{G}$ professors. An increase from the 25th to 75th percentile $\hat{G}$ professor increases the likelihood of a female student's earning my preferred measure of major choice (taking at least 9 courses in the field or declaring the major) by 2.1 p.p., compared to 1.8 p.p. in the 2000s. Unlike the 2000s, there is no evidence of increased course-taking from the same professor, though female students with high-$\hat{G}$ professors took more classes in the department.

Table AA-5 replicates Table 5, showing how the impact of $\hat{G}$ interacts with other factors influencing students' major choice. As in the 2000s, female students are more likely to earn majors (relative to male students) in departments where their first-quarter class was taught by a female teacher or had more female students. Female students were discouraged by larger first-quarter classes, and female students who received more-positive evaluations were more likely to earn the major. While professor attentiveness is not measurably related to female students' likelihood of earning the major, adding it as a covariate absorbs about 1.5 p.p. of the main effect, rendering it statistically significant at only the 10 percent level.

Interestingly, not only does professors' measured $\hat{G}$ not effect male students' major choice, but their major choice appears statistically unrelated to any observed factor in any model. It appears that 1970s male students' majors were not very sensitive to their first-quarter courses, and may have been more-strongly predetermined than either female students' majors at the time or than contemporary male or female students'.

Table AA-6 replicates Table 6. While the relationship between high-$\hat{G}$ professors and student encouragement exhibited minimal heterogeneity among the 2000s students, there is no evidence of the heterogeneity in the estimated effect on female students along *any* observed margin in the 1970s students, perhaps in part as a result of the smaller (and noisier) sample.

Finally, Table AA-7 replicates the robustness estimates reported in Table 7. As in the 2000s data, professors' $\hat{G}$ measures are conditionally statistically uncorrelated the proportion of courses that are female or the number of students in their courses, while the average genderedness or professors' male and female non-first-quarter evaluations are strong predictors of their first-quarter male and female evaluations' genderedness, respectively. Most of the encouragement effect of high-$\hat{G}$ professors for female students loads onto professors' average female evaluations' genderedness, whereas the effect was more equally-distributed in the 2000s data. The final column shows no evidence of a quadratic relationship between $\hat{G}$ and student major choice in the 1970s, justifying the linear specification.

B Appendix B: Replication Using LASSO to Estimate Equation 1

Estimating Equation 1 could over-fit the true gender valences of the included adjectives and adverbs, many of which are likely to actually have no gender valence at all. Text analysis scholarship frequently replaces OLS regression estimation with LASSO estimation in order to avoid such over-fitting. While regularization is unnecessary in this context – there are about 1,600 adjectives in the sample of more than 1,000,000 evaluations, which makes over-fitting unlikely, and even the small contributions that some words might add to estimated gender valences are of great interest in this study – I re-estimate Equation 1 using LASSO (choosing λ optimally using 10-fold cross-validation) and replicate Tables 4 to 6.

Table BB-1 shows that the main results slightly attenuate when using the LASSO measure of professor $\hat{G}$, likely as a result of the biased predictions of the LASSO estimator (which assumes that words with weak gender-valences actually have no such valence), but remain highly statistically significant.

Table BB-2 shows highly-similar patters to Table 5. The slight coefficient attenuation leaves the estimated effect of professor $\hat{G}$ on female students statistically significant at only the 10 percent level.

Table BB-3 similarly shows highly-similar patters to the main results, though the inclusion of an interaction with professor word variation, while itself noisily estimated, drives the main effect for female students to 0. The other models remain largely maintain statistical significance, despite the coefficients' slight attenuation.

In short, while LASSO regularization is unnecessary in this context as a result of the massive size of UCSC's narrative evaluation corpus (and the relatively-small number of adjectives and adverbs used to describe students), replacing OLS with LASSO little changes the resulting estimates of interest.

Table 1: Sample Anonymized 1999-2009 UCSC Narrative Evaluations

Department	Anonymized Evaluation	Gender	Grade	$\hat{F}$	$\hat{Pos}$
Biology	XXX was a **wonderful** student in this course: **bright**, engaged, and articulate. She brought **tremendous** enthusiasm as well as care to all that she did. XXX earned **high** As on both her midterms (99% and 98% **respectively**). Her attendance to lecture was **nearly perfect**, and to section **absolutely perfect**, and in both settings she was a **regular** and **intelligent** participant. She was always **willing** to ask questions too, which is a boon to **any** professor who hopes her students are **following** the **material** (A+). As part of a group project on immigrant health care, XXX took the **lead** on discussing risk factors and risk conditions that make **good** health among immigrants so **challenging**. Her analysis of **legal** frameworks was also **effective**. XXX was a **comfortable**, **informative** presenter with a **strong** well argued thesis (A/A+). XXX **final** paper on the **same** topic was **similarly effective** (A). **Lovely** work!	F	A	2.67	0.08
Environmental Science	XXX was a **quiet** yet **diligent** member of this class, who developed a **close understanding** of the course **material**. Her term paper examined the XXX, providing both **detailed** analysis of its activities and situating these in the **wider** context of XXX. The paper was **clearly written** and well-researched and demonstrated XXX analytical ability.	F	P	1.76	0.43
Japanese	**Overall**, XXX did a **good** job during this course. She **conscientiously** and **diligently** participated in class, submitting most assignments on **time**. Her **final** grade was 74 out of 100. However her **poor** performance on the **final written** examination resulted in lowering her **final** grade. What she needs to work on is **increasing** her confidence in regards to speaking the language. She was **good** to have in class as her participation benefited all.	F	C	0.43	0.70
Astronomy	XXX never attended discussion section and **only** handed in five of the nine homework assignments. She scored below **average** on both the midterm and **final**.	F	P	-0.29	-1.47
History	XXX work in this **lower**-division survey of **Early** Medieval Europe was **passing**, although somewhat **problematic**, **overall**. He attended section **sporadically**, often arrived **late**, and did not participate in discussion. His first paper, on XXX did not display a unified style and also needed **better** grounding in **historical** context and in its source analysis. His second paper was puzzling, since the topic fell almost **completely outside** the range of the **material** covered in this class; this was thus a failing essay. XXX midterm was **good** and his **final passing**.	M	C	-1.45	-2.32
Psychology	XXX gave a well-presented and **insightful** **oral** seminar, produced a well-**written** and **original** paper, and did well on the chapter quizzes. Class participation was **active**, and was **usually** well-**informed** and **productive**.	M	A-	-2.55	0.53
Residential College	XXX is one of the most **entertaining** writers I have encountered. **Even** his **weekly short** responses to the texts made me laugh out loud. His prose is **creative** and **clever**, and quite **witty**. His first paper, though **very** well **written**, was **shy** on analysis. He **quickly** eradicated this flaw, and generated **smart** essays – **even** when a thesis didn't appear **promising** – for example, one on the immaturity of the characters in XXX – he could create a **smart**, **introspective** and well-integrated argument. **Unfortunately**, XXX missed many classes and did not always contribute **positively** to the class environment when he was **present**. Despite his disruptive and sometimes confrontational behavior, he did listen and react to the conversation at hand and would reintegrate himself into discussion. For his student-led presentation, XXX presented a well-researched **historical** **background** to XXX. He did not work with the **other** students presenting, but did provide **valuable** input to the discussion.	M	B-	-3.07	-0.81

Note: Sample anonymized narrative evaluations provided to UC Santa Cruz students between 1999 and 2009, with the department of the course in which the evaluation was provided, the student's gender, the accompanying letter grade, the evaluation's predicted genderedness $\hat{F}$ (see Equation 2; more-positive implies more female-valence vocabulary), and the evaluations measured positive or negative sentiment $\hat{Pos}$ (see text for details; normalized across evaluations, with more-positive values implying a more-positive evaluation). **Adjective and adverbs included in estimating Equation 2 are in bold**; those with coefficients above 0.01 (male) are in cyan, and those below -0.01 (female) are in orange.

Source: UC-CHP UCSC Student Database

Table 2: Descriptive Statistics of UCSC Students

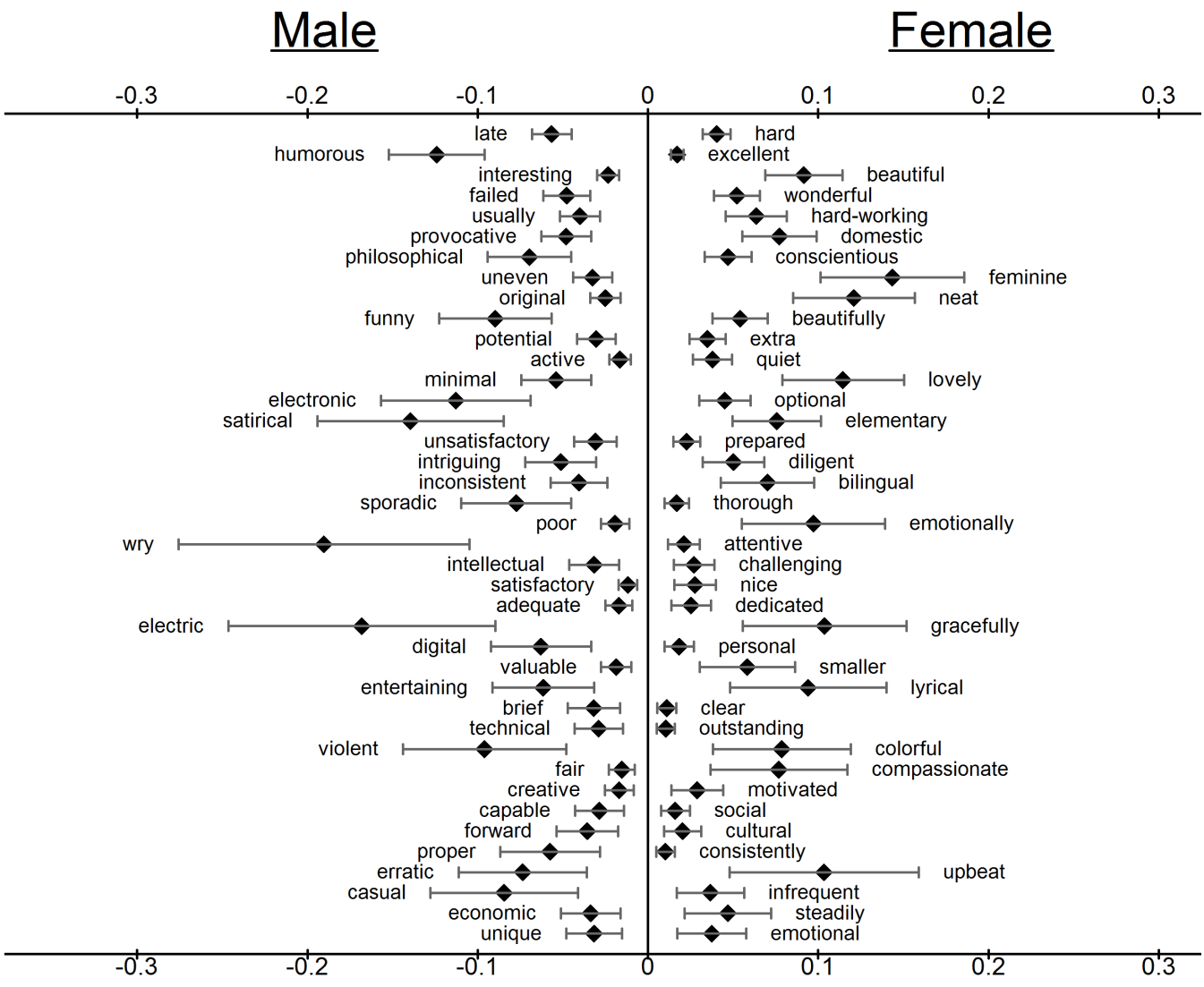
	UCSC 1965-1979 Cohorts				UCSC 1999-2009 Cohorts			
	Full Sample		Est. Sample		Full Sample		Est. Sample	
	F	M	F	M	F	M	F	M
% Graduate	-	-	-	-	77.1	75.8	78.8	77.8
% Major by Discipline								
Humanities	22.6	20.3	24.2	21.1	20.3	20.2	21.8	21.3
Social Sciences	35.9	37.3	41.9	42.3	57.8	47.9	59.2	50.0
STEM	14.6	21.8	17.3	25.3	19.7	29.9	18.0	27.7
None	27.7	21.4	17.2	12.0	6.5	6.2	5.4	5.3
# Courses[1]	25.8	26.2	2.8	2.7	30.6	30.7	3.3	3.3
# Eval.[1]	6.8	7.0	1.6	1.6	20.7	20.0	2.5	2.5
% Courses by Discipline[1]								
Humanities	25.3	20.9	32.0	25.4	19.4	20.7	22.2	21.4
Social Sciences	35.9	35.7	43.6	40.4	45.2	37.6	48.1	38.2
STEM	20.8	29.1	22.0	32.8	24.0	32.7	29.7	40.3
# Students	13,283	13,265	6,976	7,069	26,911	21,960	21,338	17,299
% by Gender	50.0	50.0	49.7	50.3	55.1	44.9	55.2	44.8

Note: Count and proportion statistics describing undergraduate UC Santa Cruz students by cohort year (first year of enrollment) and gender (Female or Male). Separate statistics for the full sample of students and for students in the Estimation Sample, which includes all students who enrolled in at least one eligible course in their first-quarter Fall; see text for details. Graduation defined as degree attainment by Spring 2019; unavailable for earlier cohort. Majors are recorded as 'none' if the student leaves the university prior to declaring a major or earns a degree with an individual or otherwise-uncategorized major; double-major students can be double-counted across disciplines. Number of courses and number of evaluations are sums across students' undergraduate tenure, or in their first quarter for the estimation sample. Courses taught in residential colleges and college writing courses are in no discipline. Students not recorded as male or female are omitted; earlier cohort student gender determined by SSA matching procedure described in the text.

[1] Course descriptive statistics for the estimation sample are restricted to the courses taken in students' first quarter.

Source: UC-CHP UCSC Student Database

Figure 1: 1999-2009 Within-Course Gendered Adjective Associations by Gender



Note: The coefficient estimates and 95 percent confidence intervals of the 40 words with the highest and lowest estimated t-statistics from Equation 1, which estimates a fixed-effect OLS regression model of evaluated students' gender by indicators for the presence of each adjective and adverb that appears in at least 100 evaluations, estimated across 1999-2009 UCSC student evaluations along with course-term-grade fixed effects. Standard errors are robust. Terms are presented in order of t-statistic magnitude (testing the null hypothesis that β is 0). Source: UC-CHP UCSC Student Database

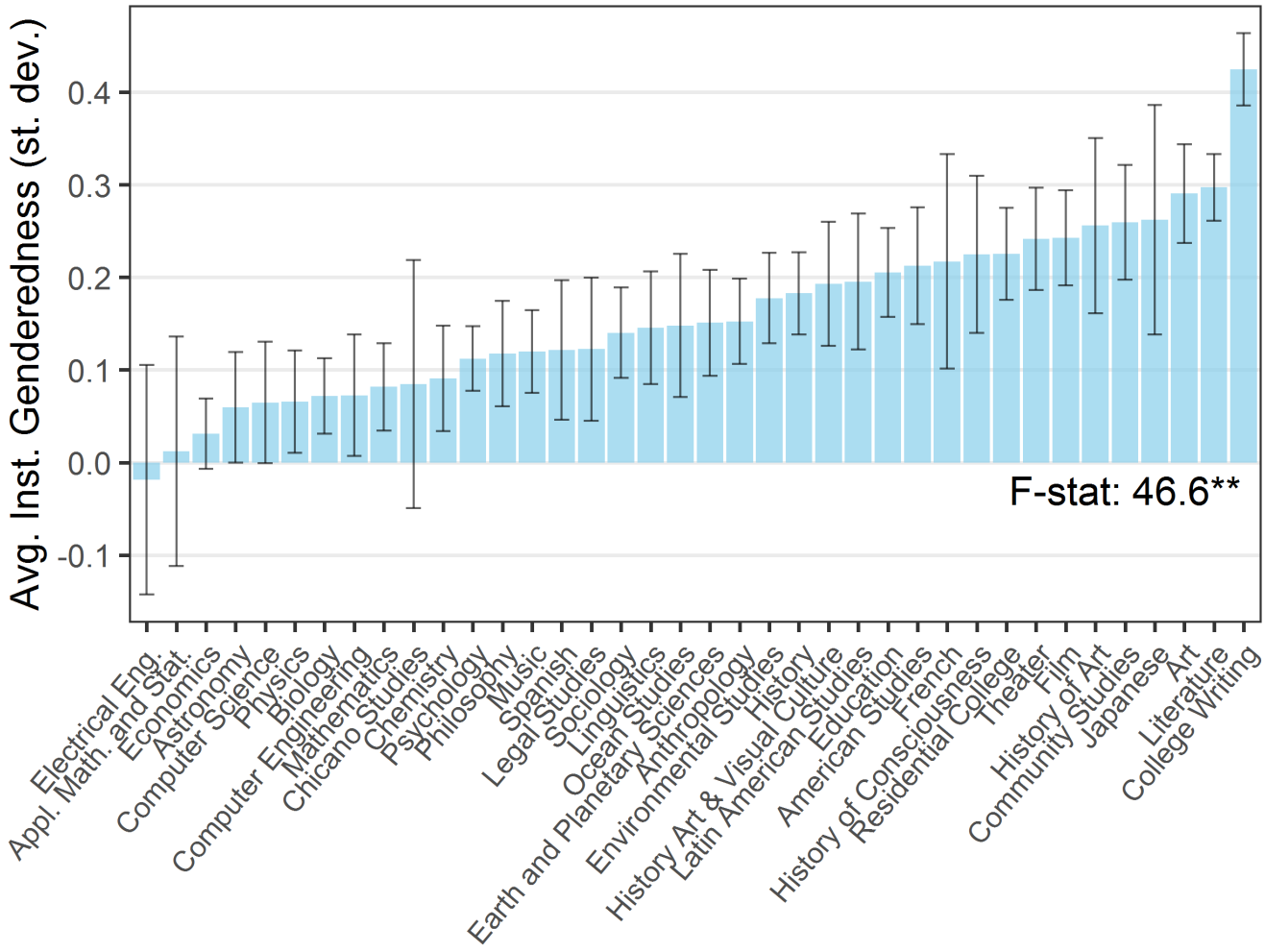
Table 3: Characteristics of 1999-2009 UCSC Evaluations' Level of Female-Stereotyped Language

Dep. Var:	Pred. Female Evaluation ($\hat{F}$)				
Female	0.165**	0.278**	0.144**	0.129**	0.214**
	(0.008)	(0.013)	(0.012)	(0.008)	(0.018)
Female × Soc. Sci		-0.125**			-0.104**
		(0.018)			(0.020)
Female × STEM		-0.218**			-0.150**
		(0.019)			(0.021)
Male × Female Professor			0.016		0.046
			(0.032)		(0.032)
Female * Female Professor			0.060†		0.047
			(0.035)		(0.034)
Male × Grade				0.205**	0.208**
				(0.008)	(0.009)
Female × Grade				0.253**	0.246**
				(0.009)	(0.010)
Department FEs	X	X	X	X	X
Year FEs	X	X	X	X	X
Observations	1,105,793	1,105,793	1,000,574	878,972	798,047

Note: Estimated coefficients and standard errors from OLS regression models of $\hat{F}_{itce}$, the normalized estimated degree of female-gender stereotype in a given 1999-2009 UCSC written evaluation, on the student's gender and gender's interactions with the course's discipline, the professor's gender, and the normalized grade earned in the course, with fixed effects by department and quarter. Standard errors are two-way clustered by student and professor. Professor gender is determined by SSA matching procedure described in the text. Grade is measured by grade points (from 0 to 4) and normalized. Course categorization into Social Science and STEM fields follows UCSC's disciplinary organization. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

Figure 2: UCSC 1999-2009 Academic Departments' Average Professor $\hat{G}$



Note: Fixed effect estimates from an OLS regression model of estimated 1999-2009 UCSC faculty $\hat{G}$ – their tendency to use male-gendered language in evaluations of male students (and vice-versa) – on department indicators, with 95 percent confidence intervals using robust standard errors. A one unit change in professor $\hat{G}$ corresponds to an professor who writes evaluations for female students that are relatively one standard deviation more female-gendered than their evaluations for male students. The F-statistic tests the coefficients' joint difference from 0. Source: UC-CHP UCSC Student Database

Table 4: Within-Course Effect of 1999-2009 UCSC First-Quarter Professor $\hat{G}$ on Enrollment and Major Choice

Dep. Var. (%):	One More Course		Number of	> Eight	Earned	> Eight Courses
	Department	Professor	Courses (#)	Courses	Major	or Earned Major
Female	-4.25**	-1.4**	-0.43**	-2.84**	-2.01**	-2.32**
	(0.79)	(0.50)	(0.08)	(0.63)	(0.61)	(0.66)
Male × Prof. $\hat{G}$	3.43	22.9**	1.42**	9.31**	7.82*	8.28*
	(3.04)	(6.81)	(0.44)	(3.22)	(3.40)	(3.44)
Female × Prof. $\hat{G}$	9.84**	24.1**	1.62**	10.95**	8.77**	10.27**
	(2.72)	(6.87)	(0.40)	(3.08)	(3.18)	(3.24)
Course-Grade FEs	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X
# Observations	65,450	65,450	65,450	65,450	65,450	65,450
# Students	37,827	37,827	37,827	37,827	37,827	37,827
# Professors	918	918	918	918	918	918
Mean of Y	69.2	23.0	6.35	33.8	35.7	40.5

Note: Estimated coefficients and standard errors from OLS regression models of 1999-2009 UCSC students' enrollment and major choices on their gender and their first-quarter (Fall) freshman professors' $\hat{G}$, with fixed effects by course-grade and start year. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

Table 5: Within-Course Effect of 1999-2009 UCSC First-Quarter Professor $\hat{G}$ on Major Choice, with Covariates

Dep. Var:	> Eight Courses or Earned Major in Same Field							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-2.89*	-2.57**	-2.32**	-2.85**	-2.01**	-2.20**	-2.19**	-2.29**
	(1.38)	(0.63)	(0.66)	(0.80)	(0.71)	(0.74)	(0.73)	(0.74)
Male ×								
Prof. $\hat{G}$	20.74*	4.12	8.28*	9.27**	10.06**	9.98**	10.26**	9.68*
	(9.14)	(3.07)	(3.44)	(3.55)	(3.51)	(3.53)	(3.85)	(4.14)
Female Prof.				-1.37	-0.32	-0.27	-0.53	-0.45
				(1.17)	(1.17)	(1.15)	(1.16)	(1.13)
# Stud. In Class[1]					-3.95*	-4.25*	-4.12*	-3.98*
					(1.91)	(1.91)	(1.89)	(1.94)
% Fem. In Class[1]					-4.47**	-4.42**	-4.59**	-4.61**
					(1.14)	(1.15)	(1.16)	(1.16)
Pos. Sent.[2]						6.40	7.36	7.30
						(4.79)	(4.88)	(5.00)
Neg. Sent.[2]						-14.69*	-15.88*	-15.91*
						(7.37)	(7.43)	(7.53)
Prof. Avg. Pos.[2]							-2.21	-2.44†
							(1.35)	(1.35)
Prof. Pos. by Gender[2]							32.62	30.33
							(41.96)	(42.03)
Class Word Var.[3]								0.47
								(0.88)
Prof. Word Var.[3]								-0.68
								(1.20)
Female ×								
Prof. $\hat{G}$	13.67	9.63**	10.27**	10.55**	9.28**	8.53*	10.81**	10.28**
	(12.92)	(3.04)	(3.24)	(3.38)	(3.42)	(3.52)	(3.67)	(3.99)
Female Prof.				0.77	0.50	0.60	0.39	0.51
				(1.03)	(1.04)	(1.04)	(1.05)	(1.00)
# Stud. In Class					-2.08	-2.27	-2.31	-2.20
					(1.96)	(1.96)	(1.94)	(1.95)
% Fem. In Class					-0.49	-0.46	-0.54	-0.56
					(1.11)	(1.12)	(1.13)	(1.14)
Pos. Sent.						9.60*	11.54**	12.52**
						(4.17)	(4.25)	(4.46)
Neg. Sent.						-5.02	-5.89	-5.03
						(8.11)	(8.17)	(8.21)
Prof. Avg. Pos.							-3.45**	-3.63**
							(1.22)	(1.21)
Prof. Pos. by Gender							-4.99	-7.16
							(40.73)	(40.38)
Class Word Var.								1.37†
								(0.74)
Prof. Word Var.								-1.44
								(1.05)
Course FEs		X						
Course-Grade FEs			X	X	X	X	X	X
Year FEs		X	X	X	X	X	X	X
# Observations	75,170	75,170	65,450	62,875	62,171	62,171	62,171	62,111
# Students	41,996	41,996	41,996	41,486	41,287	41,287	41,287	41,262
# Professors	938	938	938	891	874	874	874	872
Mean of Y	39.9	39.9	39.9	39.7	39.8	39.8	39.8	39.8

Note: Estimated coefficients and standard errors from OLS regression models over first-quarter courses of an indicator for 1999-2009 UCSC students' earning a major (or taking more than eight classes) in the same field as that course on their gender interacted with the course's professor's $\hat{G}$ and additional covariates, with fixed effects by course-grade and start year. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%. [1] Measures are normalized. [2] Positive and Negative Sentiment of the student's own evaluation, as measured using sentiment analysis with the QDAP dictionary; see the text. Professor average positivity measured as the average of their evaluations across all non-first-quarter students; professor positivity by gender is the differnece between their average positivity in evaluations of non-first-quarter male students and that of female students. [3] Measures of the average variation in evaluations across students within the class and of the professor's average variation in evaluations across students in their other, non-first-quarter classes; see the text for details.

Source: UC-CHP UCSC Student Database

Table 6: Heterogeneity in Within-Course Effect of 1999-2009 UCSC First-Quarter Professor $\hat{G}$ on Major Choice

Dep. Var:	> Eight Courses or Earned Major in Same Field									
Interaction Variable (Var.):	STEM Course	Female Professor	# Stud. in Class[1]	% Fem. in Class[1]	GPA	Prof. Avg. Pos.[2]	Prof. Pos. by Gender[2]	Year	Class Word Variation[3]	Prof. Word Variation[3]
Female	-1.95*	-2.98**	-2.53**	-1.69**	-2.38**	-2.48**	-2.21**	-2.23**	-2.42**	-2.34**
	(0.96)	(0.83)	(0.65)	(0.61)	(0.67)	(0.69)	(0.76)	(0.69)	(0.70)	(0.71)
Male × Prof. $\hat{G}$	7.96†	11.61**	7.93*	10.31**	8.06*	9.08*	11.21**	8.48*	10.25**	8.94*
	(4.13)	(4.15)	(3.49)	(3.37)	(3.46)	(3.73)	(3.87)	(3.47)	(3.95)	(4.27)
Female × Prof. $\hat{G}$	9.19*	13.91**	10.80**	8.47*	10.61**	12.02**	14.23**	9.98**	13.02**	10.72**
	(3.61)	(3.88)	(3.33)	(3.33)	(3.25)	(3.55)	(3.62)	(3.22)	(3.75)	(4.07)
Var.		-0.44	-1.45†	-3.97**		0.74	-0.78		-0.19	-0.52
		(1.54)	(0.84)	(1.17)		(0.49)	(0.80)		(0.63)	(0.78)
Female × Var.	-0.88	2.54†	0.53	3.72**	-0.29	-0.29	0.76	-0.41	0.14	0.06
	(1.31)	(1.30)	(0.43)	(0.63)	(0.45)	(0.48)	(0.70)	(0.64)	(0.56)	(0.66)
Male × Prof. $\hat{G}$ × Var.	3.07	-7.90	-2.44	0.99	0.14	-5.69*	-2.46	-0.12	4.74†	3.33
	(7.13)	(7.15)	(2.86)	(3.25)	(2.40)	(2.54)	(2.96)	(3.71)	(2.74)	(3.59)
Female × Prof. $\hat{G}$ × Var.	3.08	-10.33†	-0.49	-0.09	-1.18	-5.39*	-6.89**	2.95	5.40*	2.66
	(7.20)	(5.81)	(2.82)	(3.77)	(2.41)	(2.38)	(2.51)	(3.30)	(2.48)	(3.03)
Course-Grade FEs	X	X	X	X	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X	X	X	X	X
# Observations	75,170	72,127	74,325	74,325	65,450	75,170	75,170	75,170	74,832	74,832
# Students	41,996	41,486	41,806	41,806	37,827	41,996	41,996	41,996	41,920	41,920
# Professors	938	891	919	919	918	938	938	938	919	919
Mean of Y	40	39.7	40	40	40.5	40	40	40	39.9	39.9

Note: Estimated coefficients and standard errors from OLS regression models over first-quarter courses of an indicator for 1999-2009 UCSC students' earning a major (or taking more than eight classes) in the same field as that course on their gender interacted with the course's professor's $\hat{G}$ interacted with additional covariates, with fixed effects by course-grade and start year. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%. [1] Measures are normalized. [2] Positive and Negative Sentiment of the student's own evaluation, as measured using sentiment analysis with the QDAP dictionary; see the text. Professor average positivity measured as the average of their evaluations across all non-first-quarter students; professor positivity by gender is the differnece between their average positivity in evaluations of non-first-quarter male students and that of female students. [3] Measures of the average variation in evaluations across students within the class and of the professor's average variation in evaluations across students in their other, non-first-quarter classes; see the text for details.

Source: UC-CHP UCSC Student Database

Table 7: Robustness of 1999-2009 UCSC First-Quarter Professor $\hat{G}$ and Effect on Major Choice

	% of Class Female	# Students in Course	$\hat{F}$	>8 Courses or Earned Major	>8 Courses or Earned Major
Female	0.001	-0.00	0.01	-2.27**	-2.18**
	(0.004)	(0.00)	(0.01)	(0.68)	(0.67)
Male × Prof. $\hat{G}$	-0.025	-0.04			15.87**
	(0.085)	(0.07)			(5.29)
Female × Prof. $\hat{G}$	-0.084	-0.05			14.16**
	(0.089)	(0.07)			(5.37)
Male × $\bar{F}_{Male}$			0.74**	-8.80*	
			(0.13)	(3.53)	
Male × $\bar{F}_{Female}$			-0.07	7.62*	
			(0.16)	(3.66)	
Female × $\bar{F}_{Male}$			0.27*	-10.18**	
			(0.13)	(3.28)	
Female × $\bar{F}_{Female}$			0.41*	10.14**	
			(0.17)	(3.47)	
Male × (Prof. $\hat{G}$)^2					-17.27*
					(7.73)
Female × (Prof. $\hat{G}$)^2					-8.42
					(7.37)
Course-Grade FEs	X	X	X	X	X
Year FEs	X	X	X	X	X
# Observations	74,325	74,325	75,170	75,170	75,170
Mean of Y	-0.04	0.01	-0.02	39.95	39.95

Note: Estimated coefficients and standard errors from OLS regression models over 1999-2009 UCSC students' first-quarter courses, with fixed effects by course-grade and start year. First two columns model course characteristics (normalized percent of female students in the course and number of students in the course) by gender interacted with the professor's $\hat{G}$ as placebos. The next two columns model $\hat{F}$ (the estimated female-genderedness of students' evaluations) and whether students earned a major (or took more than eight classes) in the same field as the course by gender interacted with the average genderedness of evaluations written by the professor for non-first-quarter male ($\bar{F}_{Male}$) and ($\bar{F}_{Female}$) students. The last column models students' major choice by gender interacted with a quadratic term in professor's $\hat{G}$. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

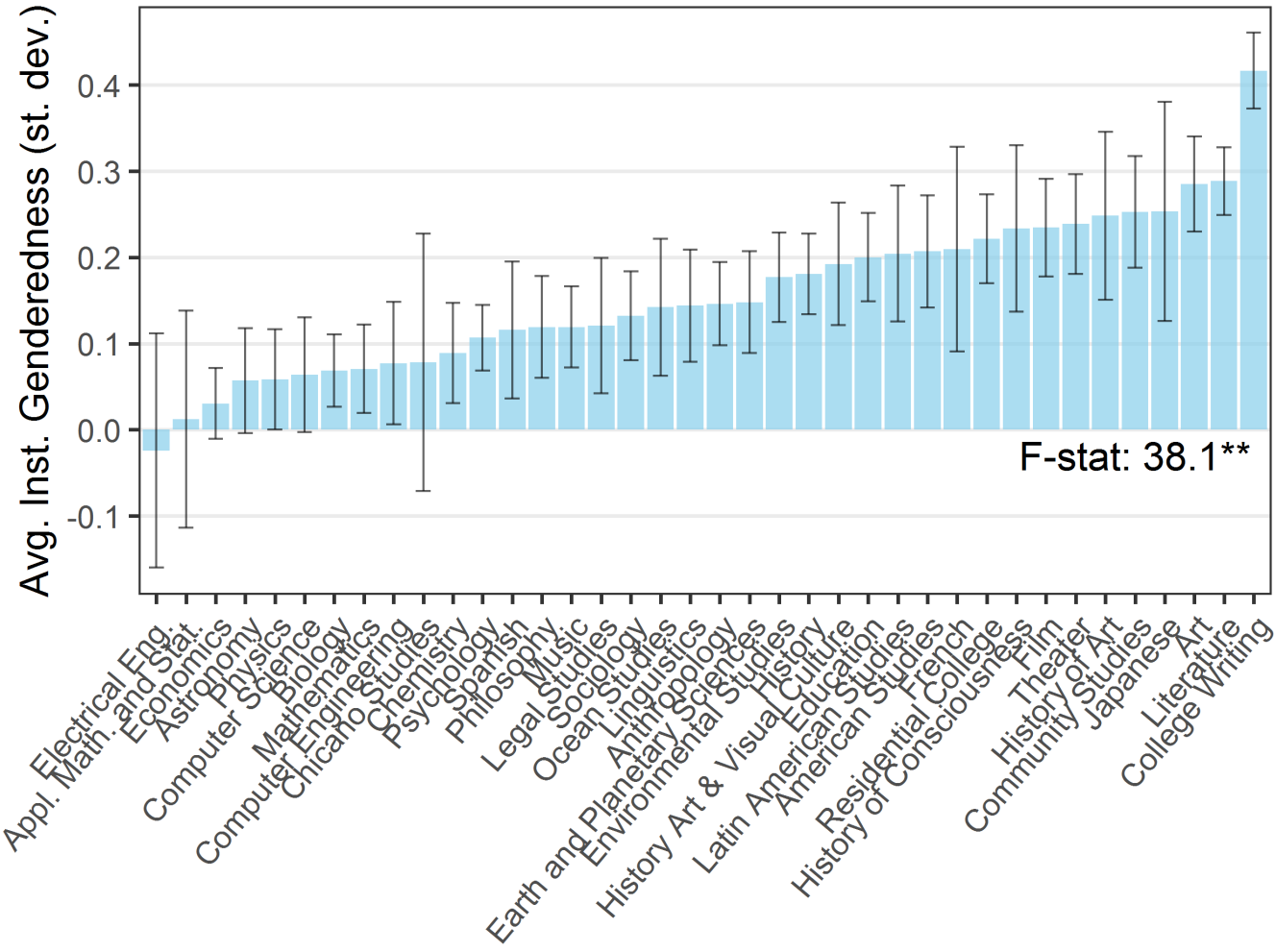
Table A-1: Within-Course Effect of 1965-1979 UCSC First-Quarter Professor Leave-One-Out (LOO) $\hat{G}$ on Enrollment and Major Choice

Dep. Var. (%):	One More Course		Number of	> Eight	Earned	> Eight Courses
	Department	Professor	Courses (#)	Courses	Major	or Earned Major
Female	-4.49**	-2.14**	-0.53**	-3.79**	-5.32**	-5.06**
	(1.07)	(0.77)	(0.09)	(0.86)	(1.15)	(1.06)
Professor $\hat{G}$	-5.57	3.69	-0.22	0.27	-6.61	-4.21
	(4.36)	(3.99)	(0.31)	(2.96)	(4.02)	(3.71)
Female $\times$ Prof. $\hat{G}$	10.76*	2.08	1.12**	8.09*	14.65**	14.51**
	(4.81)	(3.46)	(0.37)	(3.57)	(4.81)	(4.39)
Course FEs	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X
# Observations	14999	15086	14999	14999	13146	14999
# Students	11056	11111	11056	11056	9676	11056
# Professors	541	541	541	541	539	541
Mean of Y	68.3	13.1	4.46	19.8	32.8	33.5

Note: Estimated coefficients and standard errors from OLS regression models of 1999-2009 UCSC students' enrollment and major choices on their gender and their first-quarter (Fall) freshman professors' leave-one-out (LOO) $\hat{G}$, with fixed effects by course-grade and start year. LOO $\hat{G}$ is calculated by separately estimating Equation 1 for each professor, leaving that professor out of the estimation; predicted values are then estimated for that professor and normalized across professors prior to estimating $\hat{G}$. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

Figure A-1: UCSC 1999-2009 Academic Departments' Average Professor $\hat{G}$, Conditional on Professor Gender



Fixed effect estimates from an OLS regression model of estimated 1999-2009 UCSC faculty $\hat{G}$ – their tendency to use male-gendered language in evaluations of male students (and vice-versa) – on department indicators and a female indicator, with 95 percent confidence intervals using robust standard errors. A one unit change in professor $\hat{G}$ corresponds to an professor who writes evaluations for female students that are relatively one standard deviation more female-gendered than their evaluations for male students. Professor gender is determined by SSA matching procedure described in the text. The F-statistic tests the coefficients' joint difference from 0. Source: UC-CHP UCSC Student Database

Table A-2: Within-Course Effect of 1999-2009 UCSC First-Quarter Professor $\hat{G}$ on Enrollment and Major Choice, Including Mathematics

Dep. Var. (%):	One More Course		Number of	$>$ Eight	Earned	$>$ Eight Courses
	Department	Professor	Courses (#)	Courses	Major	or Earned Major
Female	-4.39**	-1.39**	-0.42**	-2.36**	-1.53**	-1.92**
	(0.73)	(0.47)	(0.07)	(0.53)	(0.53)	(0.56)
Male $\times$ Prof. $\hat{G}$	3.78	20.70**	1.32**	8.86**	7.47**	7.68**
	(2.78)	(6.74)	(0.38)	(2.78)	(2.87)	(2.94)
Female $\times$ Prof. $\hat{G}$	10.57**	22.97**	1.52**	9.26**	7.30*	8.84**
	(2.40)	(6.48)	(0.36)	(2.81)	(2.86)	(2.92)
Course-Grade FEs	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X
# Observations	73,203	73,203	73,203	73,203	73,203	73,203
# Students	40,299	40,299	40,299	40,299	40,299	40,299
# Instructors	957	957	957	957	957	957
Mean of Y	70.8	22.7	6.08	30.9	32.7	37.1

Note: Estimated coefficients and standard errors from OLS regression models of 1999-2009 UCSC students' enrollment and major choices on their gender and their first-quarter (Fall) freshman professors' $\hat{G}$, with fixed effects by course-grade and start year. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges and college writing are omitted, but mathematics courses are included. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

Table A-3: Within-Course Effect of 1965-1979 UCSC First-Quarter Professor $\hat{G}$ on Enrollment and Major Choice, with Bootstrapped Standard Errors

Dep. Var. (%):	One More Course		Number of	> Eight	Earned	> Eight Courses
	Department	Professor	Courses (#)	Courses	Major	or Earned Major
Female	-4.71**	-2.23*	-0.58**	-4.14**	-5.78**	-5.67**
	(1.49)	(1.03)	(0.11)	(1.10)	(1.43)	(1.37)
Professor $\hat{G}$	-2.64	5.30	0.03	2.19	-0.43	1.13
	(5.38)	(5.15)	(0.41)	(4.13)	(4.98)	(4.63)
Female $\times$ Prof. $\hat{G}$	6.04	1.17	0.71*	5.16†	8.77*	9.26*
	(4.06)	(2.82)	(0.31)	(3.00)	(3.89)	(3.65)
Course FEs	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X
# Observations	15059	15149	15059	15059	13200	15059
# Students	11092	11149	11092	11092	9707	11092
# Professors	542	543	542	542	541	542
Mean of Y	68.2	13.1	4.45	19.8	32.7	33.4

Note: Estimated coefficients and standard errors from OLS regression models of 1999-2009 UCSC students' enrollment and major choices on their gender and their first-quarter (Fall) freshman professors' $\hat{G}$, with fixed effects by course-grade and start year. Standard errors are bootstrapped to account for variation in $\hat{G}$; see text for details. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

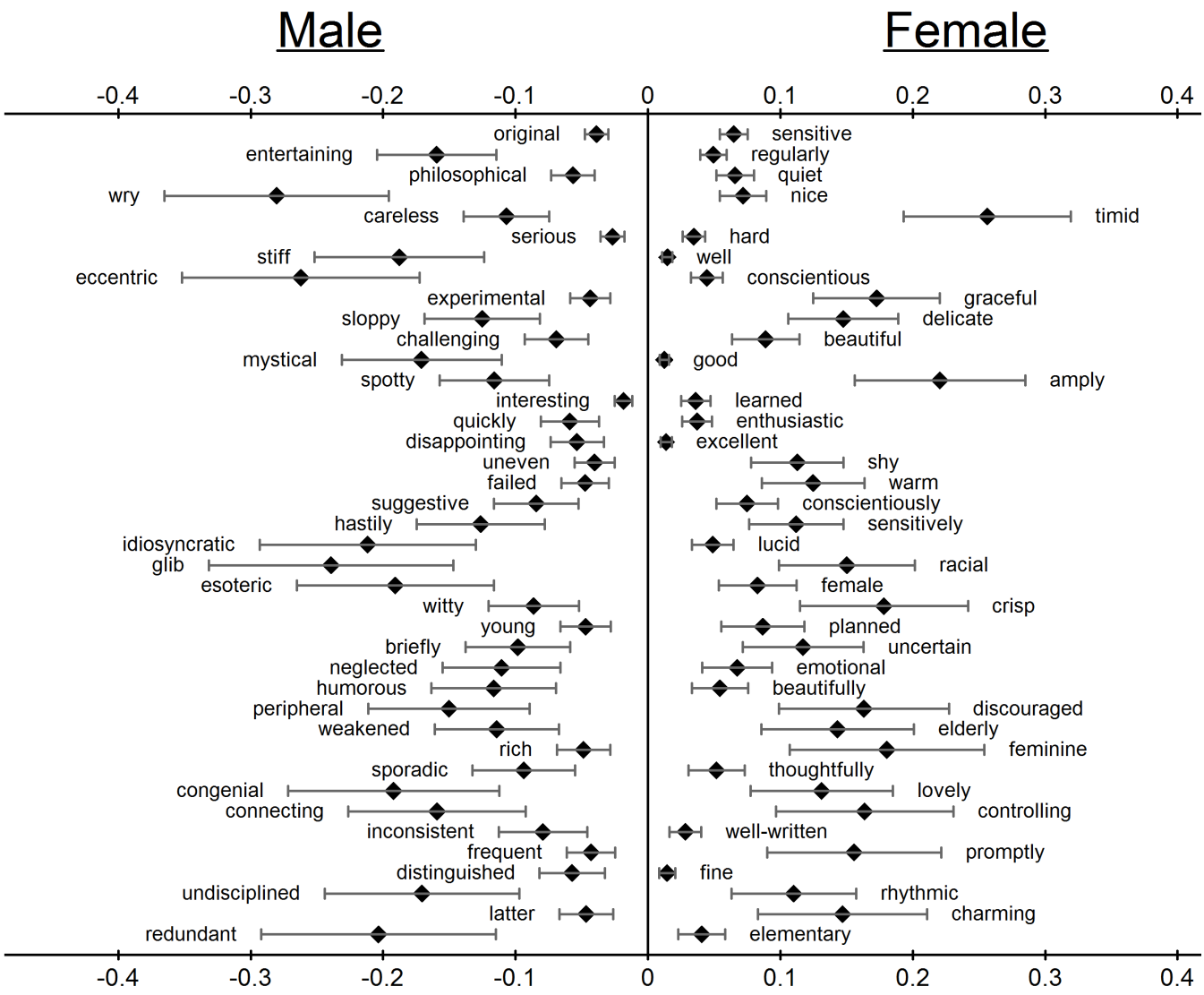
Table AA-1: Data Quality and Descriptive Statistics of 1965-1979 UCSC Students

	All Undergraduates		First-Year Fall Sample		
	% Missing	Average	% Missing	Average	Equality p-value
Year of Enrollment	7.0	1973.3	0.0	1973.8	0.000
Initial Age	27.8	19.8	14.2	19.8	0.899
% Female	1.4	50.0	1.3	50.3	0.484
% CA Resident	22.9	80.6	9.4	80.7	0.782
% Born in CA	12.7	79.1	0.6	78.6	0.332
% Within 30 mi. of UCSC	25.9	23.4	13.0	22.8	0.236
Female					
% Humanities Major	13.2	23.3	12.5	25.2	0.006
% Soc. Sci. Major	13.2	35.0	12.5	40.8	0.000
% STEM Major	13.2	14.4	12.5	17.0	0.000
% Undeclared[a]	13.2	27.7	12.5	17.2	0.000
Male					
% Humanities Major	14.7	20.9	14.5	21.8	0.158
% Soc. Sci. Major	14.7	36.7	14.5	41.5	0.000
% STEM Major	14.7	21.7	14.5	25.2	0.000
% Undeclared[a]	14.7	21.4	14.5	12.0	0.000
Observations	26,934		14,231		

Note: Sample averages and missing data proportions for the full sample of 1965-1979 UC Santa Cruz undergraduates and for those with evaluated first-year Fall courses in academic departments. P-values from t-tests between the two sample means. Age is calculated as the difference between year of birth and year of enrollment. Student gender is determined by matching first names with the contemporaneous SSA name-gender database 20 years prior; "missing" reflects androgynous names. California residency is measured as the student's listed home town being in California, and distance to campus is measured from the town's centroid. All fields except gender (which is determined from manually-keyed student names) and birth state (determined from first three digits of manually-keyed SSNs) are derived from fOCR-processed UCSC transcripts (Bleemer, 2018). [a] Student majors are "undeclared" when they leave university prior to declaring a major and when they graduate either with an individual major or without declaring a major.

Source: UC-CHP UCSC Student Database

Figure AA-1: 1965-1979 Within-Course Gendered Adjective Associations by Gender



Note: The coefficient estimates and 95 percent confidence intervals of the 40 words with the highest and lowest estimated t-statistics from Equation 1, which estimates a fixed-effect OLS regression model of evaluated students' gender by indicators for the presence of each adjective and adverb that appears in at least 100 evaluations, estimated across 1999-2009 UCSC student evaluations along with course-term fixed effects. Standard errors are robust. Terms are presented in order of t-statistic magnitude (testing the null hypothesis that β is 0). Source: UC-CHP UCSC Student Database

Table AA-2: 1965-1979 Ten Most-Gendered Positive And Negative Adjectives and Adverbs by Gender

	Female		Male	
Positive	sensitive	graceful	entertaining	congenial
	quiet	delicate	interesting	distinguished
	nice	beautiful	witty	clever
	well	good	humorous	decisive
	conscientious	amply	rich	constructive
Negative	timid	hesitant	careless	spotty
	hard	unequal	stiff	disappointing
	uncertain	suffering	eccentric	uneven
	afraid	tentative	sloppy	failed
	racist	unsure	challenging	hastily

Note: The ten positive and negative words with the most negative or positive t-statistics from an OLS regression across 1965-1979 non-first-quarter UCSC student evaluations of an indicator for the student's being female on indicators for the presence of each adjective or adverb, including course-term fixed effects. Word positivity measured using the QDAP dictionary. Source: UC-CHP UCSC Student Database

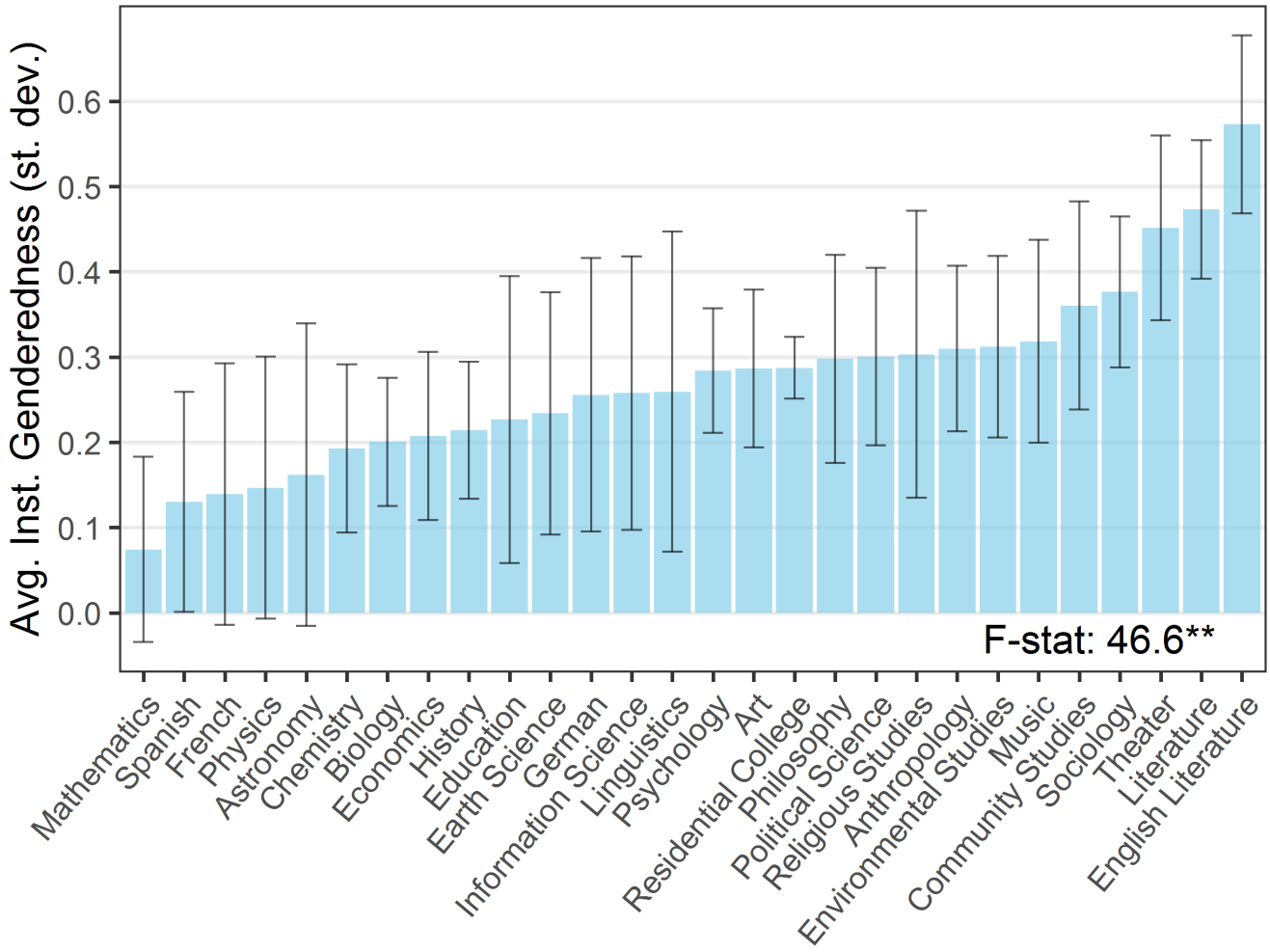
Table AA-3: Characteristics of 1965-1979 UCSC Evaluations' Level of Female-Stereotyped Language

Dep. Var:	Pred. Female Evaluation				
Female	0.253**	0.296**	0.239**	0.250**	0.277**
	(0.008)	(0.012)	(0.011)	(0.008)	(0.018)
Female × Soc. Sci		-0.037*			-0.026
		(0.017)			(0.021)
Female × STEM		-0.120**			-0.095**
		(0.018)			(0.022)
Male × Female Professor			0.055†		0.066*
			(0.032)		(0.031)
Female × Female Professor			0.102**		0.097**
			(0.027)		(0.028)
Male × Pos. Sent.				0.077**	0.076**
				(0.007)	(0.008)
Female × Pos. Sent.				-0.122**	-0.124**
				(0.006)	(0.007)
Male × Neg. Sent.				0.078**	0.075**
				(0.007)	(0.008)
Female × Neg. Sent.				-0.075**	-0.080**
				(0.006)	(0.007)
Department FEs	X	X	X	X	X
Year FEs	X	X	X	X	X
Observations	153,743	153,743	109,458	153,743	109,458

Note: Estimated coefficients and standard errors from OLS regression models of $\hat{F}_{itce}$, the normalized estimated degree of female-gender stereotype in a given 1965-1979 UCSC written evaluation, on the student's gender and gender's interactions with the course's discipline, the professor's gender, and the measured positivity and negativity of the student's evaluation (measured using sentiment analysis with the QDAP dictionary), with fixed effects by department and quarter. Standard errors are two-way clustered by student and professor. Professor gender is determined by SSA matching procedure described in the text. Grade is measured by grade points (from 0 to 4) and normalized. Course categorization into Social Science and STEM fields follows UCSC's disciplinary organization. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

Figure AA-2: UCSC 1965-1979 Academic Departments' Average Professor $\hat{G}$



Note: Fixed effect estimates from an OLS regression model of estimated 1965-1979 UCSC faculty $\hat{G}$ – their tendency to use male-gendered language in evaluations of male students (and vice-versa) – on department indicators, with 95 percent confidence intervals using robust standard errors. A one unit change in professor $\hat{G}$ corresponds to an professor who writes evaluations for female students that are relatively one standard deviation more female-gendered than their evaluations for male students. The F-statistic tests the coefficients' joint difference from 0. Source: UC-CHP UCSC Student Database

Table AA-4: Within-Course Effect of 1965-1979 UCSC First-Quarter Professor $\hat{G}$ on Enrollment and Major Choice

Dep. Var. (%):	One More Course		Number of	$>$ Eight	Earned	$>$ Eight Courses
	Department	Professor	Courses (#)	Courses	Major	or Earned Major
Female	-4.72**	-2.18*	-0.58**	-4.14**	-5.78**	-5.67**
	(1.42)	(1.04)	(0.11)	(1.06)	(1.48)	(1.36)
Professor $\hat{G}$	-2.64	5.07	0.03	2.18	-0.44	1.12
	(4.01)	(3.79)	(0.28)	(2.84)	(3.60)	(3.34)
Female $\times$ Prof. $\hat{G}$	6.05	1.15	0.71*	5.16†	8.77*	9.25**
	(4.03)	(2.98)	(0.30)	(2.86)	(3.89)	(3.60)
Course FEs	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X
# Observations	15059	15148	15059	15059	13200	15059
# Students	11092	11148	11092	11092	9707	11092
# Professors	542	543	542	542	541	542
Mean of Y	68.2	13.1	4.45	19.8	32.7	33.4

Note: Estimated coefficients and standard errors from OLS regression models of 1999-2009 UCSC students' enrollment and major choices on their gender and their first-quarter (Fall) freshman professors' $\hat{G}$, with fixed effects by course and start year. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

Table AA-5: Within-Course Effect of 1965-1979 UCSC First-Quarter Professor $\hat{G}$ on Major Choice, with Covariates

Dep. Var:	> Eight Courses or Earned Major						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-6.72**	-5.67**	-7.04**	-6.45**	-6.23*	-6.58**	-5.77*
	(1.69)	(1.36)	(1.49)	(1.52)	(2.42)	(2.40)	(2.44)
Male ×							
Prof. $\hat{G}$	-9.24	1.12	-1.99	-1.83	-2.10	-2.54	-3.33
	(6.57)	(3.34)	(3.60)	(3.60)	(3.61)	(3.69)	(3.89)
Female Prof.			-2.36	-1.31	-1.25	-2.00	-2.06
			(1.80)	(1.77)	(1.76)	(1.84)	(1.86)
# Stud. In Class[1]				-0.92	-0.92	-0.93	-0.76
				(0.96)	(0.96)	(0.96)	(0.95)
% Fem. In Class[1]				-1.63	-1.59	-1.58	-1.55
				(1.01)	(1.01)	(1.01)	(1.02)
Pos. Sent.[2]					9.25	8.66	9.49
					(6.29)	(6.30)	(6.33)
Neg. Sent.[2]					1.80	2.16	4.10
					(9.19)	(9.20)	(9.26)
Prof. Avg. Pos.[2]						4.05†	3.94†
						(2.27)	(2.28)
Prof. Pos. by Gender[2]						-25.16	-24.92
						(41.25)	(40.55)
Class Word Var.[3]							-0.07
							(0.86)
Prof. Word Var.[3]							-0.72
							(1.65)
Female ×							
Prof. $\hat{G}$	-0.62	10.38**	9.95**	8.66**	8.69*	8.68*	7.16†
	(6.38)	(3.16)	(3.38)	(3.36)	(3.38)	(3.45)	(3.74)
Female Prof.			2.09	1.62	1.68	1.31	1.34
			(1.59)	(1.66)	(1.63)	(1.70)	(1.74)
# Stud. In Class				-2.28*	-2.26*	-2.27*	-1.88*
				(1.01)	(0.99)	(0.98)	(0.94)
% Fem. In Class				1.60†	1.57	1.54	1.62†
				(0.95)	(0.96)	(0.96)	(0.96)
Pos. Sent.					10.55†	10.92†	8.95
					(5.94)	(5.89)	(5.67)
Neg. Sent.					-9.06	-9.26	-12.19
					(10.89)	(10.92)	(10.74)
Prof. Avg. Pos.						1.50	1.25
						(2.30)	(2.31)
Prof. Pos. by Gender						-27.05	-33.56
						(33.02)	(31.71)
Class Word Var.							-1.04
							(0.99)
Prof. Word Var.							-1.67
							(1.71)
Course FEs		X	X	X	X	X	X
Year FEs		X	X	X	X	X	X
# Observations	15,059	15,059	14,114	13,658	13,658	13,658	13,571
# Students	11,092	11,092	10,555	10,302	10,302	10,302	10,251
# Professors	542	542	451	441	441	441	432
Mean of Y	33.4	33.4	34.0	33.9	33.9	33.9	33.9

Note: Estimated coefficients and standard errors from OLS regression models over first-quarter courses of an indicator for 1965-1979 UCSC students' earning a major (or taking more than eight classes) in the same field as that course on their gender interacted with the course's professor's $\hat{G}$ and additional covariates, with fixed effects by course and start year. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%. [1] Measures are normalized. [2] Positive and Negative Sentiment of the student's own evaluation, as measured using sentiment analysis with the QDAP dictionary; see the text. Professor average positivity measured as the average of their evaluations across all non-first-quarter students; professor positivity by gender is the differnece between their average positivity in evaluations of non-first-quarter male students and that of female students. [3] Measures of the average variation in evaluations across students within the class and of the professor's average variation in evaluations across students in their other, non-first-quarter classes; see the text for details.

Source: UC-CHP UCSC Student Database

Table AA-6: Heterogeneity in Within-Course Effect of 1965-1979 UCSC First-Quarter Professor $\hat{G}$ on Major Choice

Dep. Var:	$>$ Eight Courses or Earned Major in Same Field								
Interaction Variable (Var.):	STEM Course	Female Professor	# Stud. in Class	% Fem. in Class	Prof. Avg. Pos.	Prof. Pos. by Gender	Year	Class Word Variation	Prof. Word Variation
Female	-5.21**	-7.45**	-6.12**	-5.74**	-5.71**	-5.95**	-5.60**	-5.16**	-5.34**
	(1.66)	(1.58)	(1.42)	(1.33)	(1.37)	(1.34)	(1.35)	(1.41)	(1.51)
Male $\times$ Prof. $\hat{G}$	0.82	-3.73	-0.41	1.52	1.89	-0.14	1.42	2.47	-0.87
	(3.76)	(4.16)	(3.65)	(3.34)	(3.38)	(3.30)	(3.45)	(3.56)	(3.43)
Female $\times$ Prof. $\hat{G}$	8.88**	9.67*	10.96**	10.52**	11.26**	9.93**	10.43**	9.38**	6.70*
	(3.37)	(3.91)	(3.27)	(3.25)	(3.23)	(3.13)	(3.26)	(3.14)	(3.20)
Var.		-5.70†	0.49	-3.35**	-0.69	2.13†		-1.30	-1.10
		(3.30)	(1.11)	(1.11)	(0.84)	(1.09)		(1.24)	(1.23)
Female $\times$ Var.	-2.71	7.79*	-3.15**	4.46**	-0.38	-1.65	-1.77	-1.14	-0.51
	(3.58)	(3.24)	(1.21)	(1.31)	(1.11)	(1.28)	(1.12)	(1.14)	(1.26)
Male $\times$ Prof. $\hat{G}$ $\times$ Var.	0.59	10.98	-7.15	6.46*	2.51	-2.37	0.54	3.15	2.90
	(9.07)	(8.59)	(4.43)	(3.03)	(2.31)	(3.13)	(2.82)	(3.50)	(3.51)
Female $\times$ Prof. $\hat{G}$ $\times$ Var.	12.48	-0.22	3.10	0.96	2.98	0.27	2.16	2.89	0.56
	(11.39)	(8.33)	(3.89)	(3.15)	(2.21)	(3.07)	(2.67)	(3.54)	(3.47)
Course FEs	X	X	X	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X	X	X	X
# Observations	15059	14114	14587	14578	15059	15059	15059	14585	14585
# Students	11092	10555	10839	10836	11092	11092	11092	10839	10839
# Professors	542	451	530	530	542	542	542	529	529
Mean of Y	33.4	34	33.3	33.3	33.4	33.4	33.4	33.3	33.3

Note: Estimated coefficients and standard errors from OLS regression models over first-quarter courses of an indicator for 1965-1979 UCSC students' earning a major (or taking more than eight classes) in the same field as that course on their gender interacted with the course's professor's $\hat{G}$ interacted with additional covariates, with fixed effects by course and start year. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%. [1] Measures are normalized. [2] Positive and Negative Sentiment of the student's own evaluation, as measured using sentiment analysis with the QDAP dictionary; see the text. Professor average positivity measured as the average of their evaluations across all non-first-quarter students; professor positivity by gender is the differnece between their average positivity in evaluations of non-first-quarter male students and that of female students. [3] Measures of the average variation in evaluations across students within the class and of the professor's average variation in evaluations across students in their other, non-first-quarter classes; see the text for details.

Source: UC-CHP UCSC Student Database

Table AA-7: Robustness of 1965-1979 UCSC First-Quarter Professor $\hat{G}$ and Effect on Major Choice

	% of Class Female	# Students in Course	$\hat{F}_{itce}$	>8 Courses or Earned Major	>8 Courses or Earned Major
Female	-0.070**	-0.01	0.09**	-5.86**	-5.31**
	(0.018)	(0.01)	(0.03)	(1.36)	(1.54)
Male × Prof. $\hat{G}$	-0.150	-0.17			3.77
	(0.121)	(0.13)			(6.33)
Female × Prof. $\hat{G}$	-0.201†	-0.19			9.83†
	(0.119)	(0.14)			(5.42)
Male × $\bar{F}_{Male}$			0.57**	0.37	
			(0.13)	(3.50)	
Male × $\bar{F}_{Female}$			0.23	3.67	
			(0.14)	(3.98)	
Female × $\bar{F}_{Male}$			0.43**	-7.31*	
			(0.12)	(3.29)	
Female × $\bar{F}_{Female}$			0.32*	16.46**	
			(0.14)	(3.89)	
Male × (Prof. $\hat{G}$)^2					-4.28
					(9.88)
Female × (Prof. $\hat{G}$)^2					0.53
					(9.25)
Course-Grade FEs	X	X	X	X	X
Year FEs	X	X	X	X	X
# Observations	14,664	14,673	15,149	15,063	15,063
Mean of Y	0.06	0.14	-0.01	33.45	33.45

Note: Note: Estimated coefficients and standard errors from OLS regression models over 1965-1979 UCSC students' first-quarter courses, with fixed effects by course-grade and start year. First two columns model course characteristics (normalized percent of female students in the course and number of students in the course) by gender interacted with the professor's $\hat{G}$ as placebos. The next two columns model $\hat{F}$ (the estimated female-genderedness of students' evaluations) and whether students earned a major (or took more than eight classes) in the same field as the course by gender interacted with the average genderedness of evaluations written by the professor for non-first-quarter male ($\bar{F}_{Male}$) and ($\bar{F}_{Female}$) students. The last column models students' major choice by gender interacted with a quadratic term in professor's $\hat{G}$. Standard errors are two-way clustered by student and professor. Professors' $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

Table BB-1: Within-Course Effect of 1999-2009 UCSC First-Quarter Professor LASSO $\hat{G}$ on Enrollment and Major Choice

Dep. Var. (%):	One More Course		Number of	$>$ Eight	Earned	$>$ Eight Courses
	Department	Professor	Courses (#)	Courses	Major	or Earned Major
Female	-4.18**	-1.48**	-0.41**	-2.67**	-1.71**	-2.01**
	(0.76)	(0.50)	(0.09)	(0.61)	(0.65)	(0.67)
Male $\times$ Prof. $\hat{G}$	1.58	23.91**	1.13*	7.02*	7.58*	7.98*
	(2.90)	(7.92)	(0.47)	(3.28)	(3.30)	(3.44)
Female $\times$ Prof. $\hat{G}$	8.86**	26.05**	1.21**	7.49*	5.90†	7.58*
	(2.81)	(7.97)	(0.41)	(3.11)	(3.17)	(3.34)
Course-Grade FEs	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X
# Observations	65450	65450	65450	65450	65450	65450
# Students	37827	37827	37827	37827	37827	37827
# Professors	918	918	918	918	918	918
Mean of Y	69.2	23	6.35	33.8	35.7	40.5

Note: Estimated coefficients and standard errors from OLS regression models of 1999-2009 UCSC students' enrollment and major choices on their gender and their first-quarter (Fall) freshman professors' LASSO $\hat{G}$ (measured by estimating Equation 1 by LASSO, with optimal 10-fold-cross-validation λ, instead of OLS), with fixed effects by course-grade and start year. Standard errors are two-way clustered by student and professor. Professors' LASSO $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%.

Source: UC-CHP UCSC Student Database

Table BB-2: Within-Course Effect of 1999-2009 UCSC First-Quarter Professor LASSO $\hat{G}$ on Major Choice, with Covariates

Dep. Var:	> Eight Courses or Earned Major							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-3.13*	-2.20**	-2.01**	-2.58**	-1.83**	-2.83*	-3.19*	-3.56**
	(1.27)	(0.64)	(0.67)	(0.83)	(0.70)	(1.15)	(1.32)	(1.34)
Male ×								
Prof. $\hat{G}$	20.40*	3.43	7.98*	8.97*	9.45**	9.55**	9.23*	8.89*
	(10.40)	(3.02)	(3.44)	(3.52)	(3.53)	(3.51)	(3.80)	(3.97)
Female Prof.				-1.42	-0.35	-0.29	-0.56	-0.45
				(1.17)	(1.16)	(1.15)	(1.16)	(1.12)
# Stud. In Class					-3.84*	-4.17*	-4.00*	-3.88*
					(1.92)	(1.92)	(1.89)	(1.94)
% Fem. In Class					-4.54**	-4.48**	-4.64**	-4.65**
					(1.14)	(1.14)	(1.16)	(1.16)
Pos. Sent.						6.88	7.81	7.78
						(4.81)	(4.89)	(5.02)
Neg. Sent.						-14.38*	-15.43*	-15.38*
						(7.26)	(7.32)	(7.48)
Prof. Avg. Pos.							42.78	39.17
							(41.49)	(41.50)
Prof. Pos. by Gender							-1.99	-2.24†
							(1.34)	(1.33)
Class Word Var.								0.55
								(0.89)
Prof. Word Var.								-0.80
								(1.17)
Female ×								
Prof. $\hat{G}$	14.89	7.03*	7.58*	7.57*	6.63†	5.99†	7.30*	7.04†
	(14.27)	(2.97)	(3.34)	(3.44)	(3.49)	(3.52)	(3.72)	(3.93)
Female Prof.				0.86	0.56	0.66	0.45	0.60
				(1.02)	(1.03)	(1.02)	(1.03)	(0.99)
# Stud. In Class					-2.06	-2.26	-2.25	-2.16
					(1.96)	(1.96)	(1.95)	(1.96)
% Fem. In Class					-0.54	-0.51	-0.58	-0.58
					(1.11)	(1.12)	(1.13)	(1.14)
Pos. Sent.						10.43*	12.33**	13.18**
						(4.15)	(4.23)	(4.46)
Neg. Sent.						-4.00	-4.62	-3.97
						(8.07)	(8.11)	(8.15)
Prof. Avg. Pos.							12.61	8.15
							(39.84)	(39.17)
Prof. Pos. by Gender							-3.04*	-3.33**
							(1.23)	(1.21)
Class Word Var.								1.47*
								(0.74)
Prof. Word Var.								-1.73†
								(1.03)
Course FEs		X						
Course-Grade FEs			X	X	X	X	X	X
Year FEs		X	X	X	X	X	X	X
# Observations	75,170	75,170	65,450	62,875	62,171	62,171	62,171	62,111
# Students	41,996	41,996	41,996	41,486	41,287	41,287	41,287	41,262
# Professors	938	938	938	891	874	874	874	872
Mean of Y	39.9	39.9	39.9	39.7	39.8	39.8	39.8	39.8

Note: Estimated coefficients and standard errors from OLS regression models over first-quarter courses of an indicator for 1999-2009 UCSC students' earning a major (or taking more than eight classes) in the same field as that course on their gender interacted with the course's professor's LASSO $\hat{G}$ (measured by estimating Equation 1 by LASSO instead of OLS) and additional covariates, with fixed effects by course-grade and start year. Standard errors are two-way clustered by student and professor. Professors' LASSO $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%. [1] Measures are normalized. [2] Positive and Negative Sentiment of the student's own evaluation, as measured using sentiment analysis with the QDAP dictionary; see the text. Professor average positivity measured as the average of their evaluations across all non-first-quarter students; professor positivity by gender is the differnece between their average positivity in evaluations of non-first-quarter male students and that of female students. [3] Measures of the average variation in evaluations across students within the class and of the professor's average variation in evaluations across students in their other, non-first-quarter classes; see the text for details.

Source: UC-CHP UCSC Student Database

Table BB-3: Heterogeneity in Within-Course Effect of 1999-2009 UCSC First-Quarter Professor LASSO $\hat{G}$ on Major Choice

Dep. Var:	> Eight Courses or Earned Major in Same Field									
Interaction Variable (Var.):	STEM Course	Female Professor	# Stud. in Class	% Fem. in Class	GPA	Prof. Avg. Pos.	Prof. Pos. by Gender	Year	Class Word Variation	Prof. Word Variation
Female	-1.66†	-2.49**	-2.21**	-1.63**	-2.06**	-2.13**	-1.97**	-1.88**	-2.12**	-2.00**
	(0.97)	(0.87)	(0.63)	(0.62)	(0.67)	(0.69)	(0.75)	(0.69)	(0.68)	(0.70)
Male × Prof. $\hat{G}$	9.52*	10.78*	7.80*	9.32**	7.58*	9.75**	9.85*	8.15*	9.35*	9.04*
	(4.27)	(4.37)	(3.47)	(3.38)	(3.46)	(3.74)	(4.42)	(3.45)	(4.23)	(4.32)
Female × Prof. $\hat{G}$	8.83*	8.41†	7.86*	6.16†	7.82*	9.83**	10.42*	6.98*	8.74*	6.13
	(3.89)	(4.31)	(3.41)	(3.47)	(3.32)	(3.68)	(4.16)	(3.19)	(3.90)	(3.93)
Var.		-0.82	-1.47†	-3.74**		0.91*	-0.83		0.15	-0.51
		(1.51)	(0.86)	(1.11)		(0.45)	(0.81)		(0.65)	(0.77)
Female × Var.	-0.84	1.91	0.75	3.50**	-0.14	-0.25	0.47	-0.52	0.32	0.23
	(1.26)	(1.26)	(0.55)	(0.67)	(0.43)	(0.48)	(0.67)	(0.59)	(0.60)	(0.63)
Male × Prof. $\hat{G}$ × Var.	-3.20	-5.48	-1.66	-1.55	2.83	-7.30**	-1.36	0.46	2.51	3.28
	(6.69)	(6.83)	(3.22)	(3.02)	(2.38)	(2.56)	(3.72)	(3.47)	(3.55)	(4.30)
Female × Prof. $\hat{G}$ × Var.	-6.30	-2.30	-2.41	-0.66	0.16	-5.98*	-4.41	4.90	0.45	-1.30
	(6.77)	(6.07)	(2.37)	(4.25)	(2.69)	(2.33)	(3.10)	(3.12)	(2.88)	(3.49)
Course-Grade FEs	X	X	X	X	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X	X	X	X	X
# Observations	75170	72127	74325	74325	65450	75170	75170	75170	74832	74832
# Students	41996	41486	41806	41806	37827	41996	41996	41996	41920	41920
# Professors	938	891	919	919	918	938	938	938	919	919
Mean of Y	40	39.7	40	40	40.5	40	40	40	39.9	39.9

Note: Estimated coefficients and standard errors from OLS regression models over first-quarter courses of an indicator for 1999-2009 UCSC students' earning a major (or taking more than eight classes) in the same field as that course on their gender interacted with the course's professor's LASSO $\hat{G}$ (measured by estimating Equation 1 by LASSO, with optimal 10-fold-cross-validation λ, instead of OLS) interacted with additional covariates, with fixed effects by course-grade and start year. Standard errors are two-way clustered by student and professor. Professors' LASSO $\hat{G}$ is defined as the estimated difference between their average use of gendered adjectives and adverbs for female and male students, with a higher value corresponding to a greater difference; see the text for details. "Number of Courses" and "> Eight Courses" refers to courses in the same department as the freshman course. Courses taken in residential colleges, college writing, and mathematics are omitted. Statistical significance: † 10%, * 5%, ** 1%. [1] Measures are normalized. [2] Positive and Negative Sentiment of the student's own evaluation, as measured using sentiment analysis with the QDAP dictionary; see the text. Professor average positivity measured as the average of their evaluations across all non-first-quarter students; professor positivity by gender is the differnece between their average positivity in evaluations of non-first-quarter male students and that of female students. [3] Measures of the average variation in evaluations across students within the class and of the professor's average variation in evaluations across students in their other, non-first-quarter classes; see the text for details.

Source: UC-CHP UCSC Student Database