

Maurer, Stephan; Schwerdt, Guido; Wiederhold, Simon

**Conference Paper**

## Understanding Gender Match Effects in Higher Education: The Role of Class Size

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2024: Upcoming Labor Market  
Challenges

**Provided in Cooperation with:**

Verein für Socialpolitik / German Economic Association

*Suggested Citation:* Maurer, Stephan; Schwerdt, Guido; Wiederhold, Simon (2024) : Understanding Gender Match Effects in Higher Education: The Role of Class Size, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2024: Upcoming Labor Market Challenges, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:

<https://hdl.handle.net/10419/302339>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# Understanding Gender Match Effects in Higher Education: The Role of Class Size

Stephan Maurer<sup>a</sup>, Guido Schwerdt<sup>b</sup>, Simon Wiederhold<sup>c</sup>

February 13, 2024

<sup>a</sup>*University of Edinburgh, UPF Barcelona School of Management, and CEP, [smaurer@ed.ac.uk](mailto:smaurer@ed.ac.uk)*

<sup>b</sup>*University of Konstanz, ifo Institute, CESifo, and ROA, [guido.schwerdt@uni-konstanz.de](mailto:guido.schwerdt@uni-konstanz.de)*

<sup>c</sup>*Halle Institute for Economic Research (IWH), University of Halle, ifo Institute, CESifo, IZA, and ROA; [simon.wiederhold@iwh-halle.de](mailto:simon.wiederhold@iwh-halle.de)*

---

## Abstract

It is widely believed that female students perform better when taught by female professors. However, little is known about the mechanisms explaining these gender match effects. Using administrative records from a German public university, which cover all programs and courses between 2006 and 2018, we show that gender match effects are sizable in smaller classes, but are absent in larger classes. These results suggest that direct and frequent interactions between students and professors are crucial for gender match effects to emerge. In contrast, the mere fact that one's professor is female is not sufficient to increase performance of female students.

*Keywords:* gender gap, role models, tertiary education, professors

*JEL:* I21, I23, I24, J16

---

---

\*We thank Sabrina Eisenbarth, Alexander Giessing, Brian Jacob, Stephanie Owen, Jörn-Steffen Pischke, Ludger Woessmann, Ulf Zölitz, seminar participants at Edinburgh, Hohenheim, Munich, and St. Gallen, and participants of the annual conference of the VfS Education Committee and the CESifo Area Conference on Economics of Education for helpful comments and discussions. We are particularly grateful to Eric Bettinger and Florian Hoffmann for sharing their class size information. We acknowledge that this research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2035/1 - 390681379. Christian König provided excellent research assistance.

## 1. Introduction

While the notion that female students experience substantial benefits when taught by female professors is widely acknowledged (see, for example, Bingham 2012, Warrell 2020), the bulk of evidence about gender match effects in higher education originates from small class contexts (e.g., Bettinger and Long (2005), Carrell et al. (2010)).<sup>1</sup> If such effects primarily arise because female professors serve as role models for female students, they can also be expected to occur in large classes. But if such effects additionally require more direct and frequent interactions between students and professors, findings from gender match effects in smaller classes might not extend to the larger class settings typical for many public universities worldwide.

In this paper, we study female gender match effects on student performance in a public university in Germany. Our analysis is based on administrative records for the universe of programs and courses in the period 2006 to 2018, providing considerable variation in class sizes. These data allow us to estimate female gender match effects by class size conditional on a rich set of student characteristics and program, course, semester, and lecturer fixed effects. In particular, we estimate the impact of female lecturers on the performance of female students, relative to how that same set of lecturers taught male students (i.e., lecturer set fixed effects), relative to the average performance in that course (i.e., course fixed effects), and relative to the average performance of female students with male lecturers (i.e., controlling for student gender). Grading in our setting is mostly anonymous, ensuring that the grade reflects learning, rather than any differential grading by gender match. Moreover, because our data include a large number of compulsory courses with different sizes, we can account for potential ability-based sorting of students

---

<sup>1</sup>Bettinger and Long (2005) study the effects of gender matching in higher education in classes with an average size of 31 students; in Carrell et al. (2010), the average class size is only 19 students.

to professors. We further address selection concerns by leveraging information on students' high school GPA, which is a powerful measure of students' academic ability in Germany.

Our results show that female students benefit more than their male peers from being taught by female professors. Overall, we find a female gender match effect of 7% of a standard deviation in grades. This estimate lies in between previously documented female gender match effects in tertiary education (Hoffmann and Oreopoulos (2009), Carrell et al. (2010)). Importantly, however, this average gender match effect masks a pronounced heterogeneity by class size. In small classes, gender match effects are substantial, implying performance gains for female students of 13% of a standard deviation and a reduction in the probability of failing an exam by 1.5 percentage points. These gender match effects are present along the whole grade distribution, pushing female students to obtain excellent rather than just good grades and reducing their probability of failure. In contrast, there are no gender match effects in large classes.<sup>2</sup>

In the German university system, students' leeway in course selection within programs is typically quite limited, in particular in early semesters, which provides an appealing context for examining gender match effects in the absence of randomized variation. We also conduct an extensive series of analyses to rule out that selective course choices by female students drive our results. Most importantly, results are very similar in strictly compulsory courses. We also show that female students do not systematically select into courses with female lecturers, nor do they systematically choose courses with younger or more lenient lecturers. Results are also robust to controlling for student fixed effects. To address potential sorting of female lecturers to courses of different sizes, we can also restrict the sample to lecturers who teach both small and large courses.

---

<sup>2</sup>We use the class size median to determine “small” (at or below median) and “large” (above median) classes, but verify the robustness to changes in the large-class cutoff. When estimating gender match effects by class size decile, we also find effects in medium-sized classes.

Our paper contributes to a growing literature that investigates female gender match effects in education. Several papers have shown that gender match effects matter for the educational production in schools (Dee (2005, 2007), Cho (2012), Parades (2014), Antecol et al. (2015), Muralidharan and Sheth (2016), Lim and Meer (2017, 2020), Oh (2023)). Moreover, exposure to role models in the form of advisors, mentors, or successful practitioners affects study choices and educational success in higher education, as well as occupational selection (Blau et al. (2010), Lyle and Smith (2014), Breda et al. (2018), Kofoed and McGovney (2019), Porter and Serra (2020), Canaan and Mouganie (2021, forthcoming), Agurto et al (2022)). Similarly, gender matches at the student-professor level have been shown to influence major and course choices at university (Dynan and Rouse (1997), Rask and Bailey (2002), Bettinger and Long (2005)) and positively affect student performance (Hoffmann and Oreopoulos (2009), Carrell et al. (2010)).

However, although a considerable body of research has underscored the significance of gender match effects in educational contexts, these findings do not hold universally. Drawing on a meta-analysis of 538 estimates and a newly conducted investigation encompassing 90 countries, a recent study by de Gendre et al. (2022) suggests that the effects of gender matching in education are often modest and vary substantially with context. We show that one factor determining the effectiveness of gender interactions in educational settings is the size of the class in which they occur.

The potential relevance of class size in the context of gender match effects becomes apparent upon recognizing that existing evidence, even that from higher education (Bettinger and Long (2005), Carrell et al. (2010)), is heavily skewed towards environments with *smaller* class sizes. This emphasis on smaller classes is largely due to the methodological reliance of the most rigorous studies on random student-to-teacher assignments, a strategy that is typically not feasible in larger classes. In a context without structured random assignment, Hoffmann and Oreopoulos (2009) document the existence of gender match effects in a setting involving larger class sizes. However, their estimated average

effects are only half as large as those observed in the smaller class context studied by Carrell et al. (2010). While our findings align with this overall pattern of heterogeneous effects suggested by the comparison of these two studies, inferring the role of class size for gender match effects from two cross-study comparisons is clearly limited due to differences in contexts and empirical approaches of these studies. We are the first to directly examine gender match effects across both small and large classes within a single university. Our findings indeed show that the conclusions drawn from earlier, small-class-centered research do not necessarily apply to the large-class settings found in many public universities.

The terms “gender match effects” and “role model effects” are sometimes used interchangeably and are often not precisely defined. Narrowly defined, role models effects may arise simply because just seeing that a female professor teaches a specific course may inspire female students in ways that lead to an increase in performance. However, our finding of a zero female-lecturer female-student interaction in larger classes casts doubt on role model effects operating in this narrow sense. If just seeing that the professor is female would be sufficient to trigger sizable role model effects, we should observe them in both smaller and larger classes. Instead, our main result, that gender match effects manifest exclusively in smaller classes, points towards the relevance of interactions between students and professors for producing economically significant gender match effects. This conclusion is further supported by an analysis of student evaluations at the university under study, which confirms that smaller classes facilitate more frequent and intense classroom interactions.

Our findings also have important implications for policies aiming at reducing gender gaps in higher education. Worldwide, gender gaps in science, technology, engineering, and mathematics (STEM) disciplines are particularly large, and the under-representation of women in STEM has frequently been discussed as being a driver of gender wage gaps (Beede et al. (2011), Blau and Kahn (2017)), a lack in innovation (Hong and Page (2004), Koning et al. (2021)), and the perpetuation of gender stereotypes (Diekmann et

al. (2010)). Intriguingly, exploiting the fact that we can observe professors and students across multiple academic fields, we find that gender match effects in smaller classes are *strongest* in STEM. Thus, a policy to attract more female professors in STEM programs may be effective in increasing the share of females who successfully complete STEM programs if applied in environments conducive to regular student-professor interactions. However, in larger education programs at public universities or in massive open online courses (MOOCs) with little interaction between students and professors, increasing the share of female professors in STEM might not yield comparably beneficial outcomes.

The remainder of the paper is structured as follows. Section 2 discusses the institutional background, describes our data, and lays out our empirical strategy. Section 3 presents our results and robustness checks. In Section 4, we discuss student-lecturer interactions as a mechanism explaining gender match effects in smaller class settings. Section 5 concludes.

## 2. Empirical Setup

### 2.1. Data and Institutional Background

We draw on the universe of bachelor-level exams taken at a medium-sized public university in Germany between 2006 and 2018. The university has 13 academic departments that offer different degree programs, which we henceforth call majors or programs.<sup>3</sup> For administrative purposes, the departments are further organized into three “sections”: (i) Political Science, Law & Economics, (ii) STEM (including Psychology), and (iii) Humanities (including several social sciences). Undergraduate majors are designed to be completed in three years, but it is quite common for students to take longer. Majors generally require a combination of compulsory courses, core elective courses, free elective

---

<sup>3</sup>We exclude programs that are taken as minors.

courses, and a final thesis for a total of 180 ECTS.<sup>4</sup> However, the share of compulsory vs. elective courses varies across majors. Students choose their major prior to enrollment. It is possible to change majors later on; however, this may prolong the duration of one’s studies, as not all previously completed courses are necessarily credited in the new major.

In our setting, an observation is the exam result in a given class taken by a given student in a given program and semester. We exclude law-related majors, as they have a very different grading scheme from other programs. This leaves us with 27 majors that cover Political Science & Economics, STEM, and Humanities.

Exams are graded on a scale from 1 to 5, with a total of 11 different possible grades.<sup>5</sup> Grades between the top grade of 1.0 and 4.0 are passing grades, the grade of 5.0 indicates a fail. To facilitate comparison, we standardize exam grades at the exam-semester level with mean 0 and standard deviation 1. We also reversed the usual German ordering so that higher values indicate better outcomes.

For every exam, there are usually at least two sittings, one immediately after the course and one several weeks later. Students who fail the first sitting can register for the second one, but students can also choose to take only the second sitting. In most courses, students can take at most two sittings. Failing a compulsory course twice typically means students have to leave their program and cannot enroll into the same program at any other public university in Germany. We exclude retries, second attempts, and any later attempts within the same course.<sup>6</sup> Courses can have up to two lecturers. We consider a course as female-taught if at least one of the lecturers is female, but we show that our results are robust to alternative codings (see Appendix Table A2).

---

<sup>4</sup>ECTS stands for European Credit Transfer and Accumulation System. One ECTS point corresponds to 25 to 30 hours of studying.

<sup>5</sup>Grades starting with 1, 2, and 3 can take three values each (e.g., 1.0, 1.3, and 1.7).

<sup>6</sup>In some cases, our data have several entries for a given exam-major-semester-student combination that are all coded as first attempt. If one of the grades is a fail, we consider the course as a fail. When there are several non-fails, we average the grades over all the attempts.



In addition to exam results, our data also contain rich student-level background characteristics that include gender, age, citizenship, and a student’s experience in their major (which we proxy by the academic year in which the first exam is taken). Based on the location where students finished high school, we can also construct a dummy for “local” students, which takes a value of 1 if students completed their high school education in the county where the university is located. Importantly, our data provide information on the GPA of the high school leaving exam, which we use as control for student ability.<sup>7</sup> Previous research has shown that high school grades in Germany are informative about student ability, as they correlate strongly with earnings (Schwerdt and Woessmann (2017)) and standardized test scores (Neumann et al. (2011)). In our data, we also observe a clear positive link between high school and university grades: For a one standard deviation increase in the high school grade, university grades on average improve by one third of a standard deviation (also see Appendix Figure A1).<sup>8</sup> We thus consider high school GPA to be a powerful measure of academic ability and a strong predictor of university exam performance.

We exclude observations where information on any of the student characteristics is missing. After applying these restrictions, 23,552 exam results were dropped, resulting in a final sample of 310,554 exam results from 18,592 distinct students.

Summary statistics are shown in Table 1. While our sample has 14% more female than male students, female lecturers only account for roughly a quarter of the courses taken. Female students take more courses taught by female lecturers than their male counterparts. Female students usually have better exam grades, and come to university

---

<sup>7</sup>Most of the students at the university we study come from federal states with centralized final high school exams, which facilitates the comparability of grades. Moreover, we also have information on the type of high school students attended and on the year in which they took the high school leaving exam. Our results are robust to allowing the association between high school grades and university exam grades to vary by the place of the high school, graduation year, and type of high school leaving exam (see Appendix Table A2).

<sup>8</sup>Luis Silva et al. (2022) even find that high school grades in Portugal are on average better at predicting study success at university than university admission tests.

with better high school GPAs and at a slightly younger age. The vast majority of bachelor students are German citizens, and about 13% of them attended high school in the county of their university.

## 2.2. Empirical Strategy

We are interested in female gender match effects, i.e., whether female students perform better when taught by female lecturers instead of male lecturers. Since students choose their program of study and, within a program, their elective courses, there are several potential confounders. However, our data allow us to follow the same lecturers and courses over time, exploiting changes in who teaches which courses. Specifically, for student  $i$  enrolled in program  $p$  (e.g., Economics) taking course  $c$  (e.g., Microeconomics I) in semester  $t$  (e.g., winter semester 2006/07), we estimate the following model:

$$\begin{aligned} grade_{ipct} &= \beta FemaleLecturer_{ct} \times FemaleStudent_i \\ &+ \gamma' StudentChars_{it} + \lambda' LecturerSet_{ct} \\ &+ \omega_p + \xi_c + \tau_t + \epsilon_{ipct}. \end{aligned} \tag{1}$$

The outcome of interest, *grade*, is standardized exam grades. *FemaleLecturer* is a dummy indicating whether the lecturer of course  $c$  in semester  $t$  is female (if there are two lecturers: if at least one of the lecturers is female). *FemaleStudent* is a dummy indicating whether student  $i$  is female, and the interaction between both dummy variables,  $FemaleLecturer \times FemaleStudent$ , is our key variable of interest with associated coefficient  $\beta$ . *StudentChars* is a vector of student characteristics: gender, final high school grade (standardized to mean 0 and standard deviation 1 in the overall sample), age, dummies for having a German citizenship and for having completed high school in the county where the attended university is located, respectively, and the starting year in the program (coded as the academic year in which we observe the first exam). With the

exception of age, student characteristics are time-invariant. *LecturerSet* are fixed effects for the combination of first and second lecturer. They are similar to lecturer fixed effects, but differentiate between situations where a lecturer teaches alone or with different other lecturers.<sup>9</sup>  $\omega$  denotes fixed effects for the program as part of which student  $i$  takes the course,  $\xi$  are course fixed effects, and  $\tau$  are semester fixed effects. Standard errors are twoway-clustered at the student and course level.<sup>10</sup>

Including this demanding set of fixed effects allows us to address multiple possible confounders in the estimation of gender match effects. In particular, we can account for different grading standards and gender shares across programs/courses and over time, systematic selection of students into courses that are perceived as easy or hard, and lecturers' teaching abilities. We identify effects from over-time changes in the gender of the lecturer(s) who teach a specific course, which could be due to, for example, sabbaticals, recruitment of new professors, or within-department reshuffling of teaching duties.

One remaining concern is that students systematically respond to changes in lecturer gender based on their own ability and gender. Such selection at the course-student level seems rather unlikely in the context of the German university system, where students commit to a specific program with a pre-defined set of courses upon entering the university. While there is some leeway to select courses within this pre-defined choice set, students generally have little insight or ability to time their courses strategically. Accordingly, Table 2 indeed shows that female students do not systematically sort into courses taught by female professors (Panel A). This holds both for higher- and lower-ability students. Female students neither differentially select into small or large courses taught by female lecturers (Panels B and C). In the remainder of Table 2, we provide additional evidence that there is no systematic selection of female students to lecturers of different age or

---

<sup>9</sup>That is, all courses taught only by lecturer A get a different lecturer set dummy from courses taught by lecturer A together with lecturer B.

<sup>10</sup>To estimate these models with many sets of fixed effects, we use the Stata command *reghdfe* (Correia (2014)).

perceived grading leniency.<sup>11</sup> Panels D and E indicate that female students are not more likely to attend classes with younger teachers (measured by the average age of both lecturers in two-instructor courses or by the actual age of the sole lecturer in single-instructor courses). Finally, in Panel F, we create a measure reflecting the average grades awarded by a lecturer in all the courses he or she previously (i.e., before the observed student took the course) taught. We here focus solely on single-instructor courses and weight all previous courses equally. Using this lecturer leniency measure, we find that female students do not systematically select into courses taught by lecturers who have previously given better grades.

In Section 3.2, we provide several additional pieces of evidence indicating that selection at the course-student level does not affect our results. Most importantly, we show that our results hold for compulsory courses and courses offered early in the study program, where course choice is even more limited or does not exist at all.

### 3. Results

#### 3.1. Main Results

Figure 1 provides a graphical illustration of our main result. It shows female gender match effects along deciles of class size. For class sizes in the lowest 5 deciles (corresponding to 74 or fewer students), we find positive, sizable, and statistically significant effects.<sup>12</sup> Pairing a female student with a female lecturer improves the student performance by 10–18% of a standard deviation in smaller classes. In terms of magnitude, the estimated gender match effects amount to 3–4 times the gender gap in exam performance.<sup>13</sup> However, for class sizes above the median, estimated female gender match effects decrease

---

<sup>11</sup>These are the only lecturer characteristics we can observe in the administrative data.

<sup>12</sup>Note that we proxy class size by the number of students taking the final exam. The actual number of students regularly attending the lectures is likely smaller than the number of exam-takers, as attendance is typically not compulsory at German universities.

<sup>13</sup>Conditional on other student characteristics and our set of fixed effects, female students perform 4% of a standard deviation worse than male students.

substantially in size. For the 6th, 7th and 8th decile in the class size distribution, we still find positive and sometimes marginally significant coefficients of around 4–8% of a standard deviation, whereas for the two highest deciles, coefficients are close to 0 and statistically insignificant. The heterogeneity by class size is also illustrated by the solid black lines, which depict a separate estimate of female gender match effects for class sizes below and above the median, respectively.

Table 3 shows our main result in regression table format. In Columns 1 and 2, we estimate female gender match effects in the whole sample, without or with controlling for a student’s high school GPA. In both cases, we find statistically significant average effects of around 7% of a standard deviation (corresponding to the dashed line in Figure 1). This effect size falls in between previous estimates of female gender match effects in tertiary education: Hoffmann and Oreopoulos (2009) find gains of up to 5% of a standard deviation for the University of Toronto, while Carrell et al. (2010) report effects of 10% of a standard deviation for the US Air Force Academy.<sup>14</sup>

In Columns 3 and 4 of Table 3, we directly compare gender match effects across small and large classes within the same university. The two columns correspond to the solid lines in Figure 1: They show that average positive female gender match effects are mostly driven by courses below the class size median, where we find an effect of 13.1% of a standard deviation. Above the median, the estimate is close to zero and statistically insignificant. We also report the  $p$ -value of a test for equality of the gender match effects for small and large courses, and can safely reject this hypothesis.<sup>15</sup>

---

<sup>14</sup>Hoffmann and Oreopoulos (2009) use a sample with an average class size of 81 students (contrasting with Carrell et al.’s (2010) 19 students) that excludes classes with fewer than 50 students. When we apply this minimum size criterion to our sample, the average class size increases to 188 students. In this restricted sample, we find an average teacher-student gender match effect of 4.3% of a standard deviation in grade performance ( $p = 0.115$ ), which is notably similar to the effect observed in Hoffmann and Oreopoulos (2009).

<sup>15</sup>Note that a significant portion of the variance in class size can be attributed to differences among academic study programs, accounting for approximately 41% of the total variance. Nevertheless, appreciable variation persists even when examining specific fields of study. For instance, the overall pattern of results is similar when we exclusively analyze the Bachelor’s program in Economics. In smaller classes, we

In Table 4, we examine from which part of the grade distribution the estimated female gender match effects come from. To do so, we replace the continuous grade outcome by a series of dummies that indicate whether students got an A, B, C, D, or failed. As can be seen in Panel A, female students that are paired with a female lecturer in a small class are 4.5 percentage points more likely to get an A, are 1.9 percentage points less likely to get a C, and are 1.5 percentage points less likely to fail a course. Female gender match effects thus seem to be present along the entire grade distribution: at the top, female students benefit from having a female lecturer by being more likely to receive excellent rather than just good grades; at the bottom, gender match effects materialize through a reduced risk of failing a course.<sup>16</sup> For large courses, we do not find gender match effects in any grade category.

### *3.2. Addressing Selection Concerns*

One fundamental challenge in analyzing student-teacher matches is the possibility that our findings simply reflect selection patterns. In the German university context, while instructors have no control over their course enrollees, students possess some degree of choice in which courses to enroll. Specifically, based on their chosen major, students are only required to complete a number of compulsory courses. Beyond these, they in principle have the freedom to choose elective courses from a specified list to accumulate the necessary credits, typically totaling 180 ECTS, for their degree completion. While our empirical model accounts for students' academic capabilities by controlling for high school GPA and incorporates fixed effects for both programs and courses, it could still

---

observe positive gender match effects of 11% of a standard deviation, while effects are insignificant, small, and even negative in larger classes. In the context of the Bachelor's program in Economics, fluctuations in class sizes can mainly be attributed to the fact that students are required to select one of six specialized areas (e.g., Applied Economics) following their first year of study. Each of these specializations entails a distinct set of compulsory courses, thus contributing to the variability in class sizes.

<sup>16</sup>Additionally, we also run unconditional quantile regressions in the sample of small courses (see Appendix Table A3). For implementation, we use the *rifhdreg* command for Stata (Rios-Avila (2020)). We find positive and significant gender match effects at the 10th, 25th, 50th, 75th, and 90th percentile of the grade distribution. The effects are most pronounced at the top and median of the distribution.

be that female students, possessing high ability not reflected in their high school GPA, systematically select courses specifically when those courses are taught by female lecturers. Below, we provide several pieces of evidence that such selection bias is unlikely to be present in our analysis.

First, we focus only on compulsory courses.<sup>17</sup> For this purpose, we reviewed module handbooks across all 27 majors from 2006 to 2018 to identify the compulsory courses specific to each major and entry cohort. This enabled us to supplement the exam data with identifiers for compulsory courses through string matching based on course names. Table 5 shows that the impact of having a female lecturer on female students' performance is consistent across small compulsory courses (Column 1) and small elective courses (Column 2).<sup>18</sup> In contrast, in larger courses, regardless of their compulsory or elective nature (Columns 3 and 4), we observe no discernible gender match effects. However, one caveat of the analysis of compulsory courses is that some programs (in particular, the Humanities) have only a limited number of such courses. Thus, we also examine courses taken during the first academic year. Even though not all of them are compulsory *de jure*, these first-year courses are often foundational courses, crucial for subsequent advanced studies, which are implicitly or explicitly recommended for early completion. The results from this analysis parallels findings from compulsory courses and our full sample, showing a significant female gender match effect in small early courses (Column 5) and no effect in larger ones (Column 6).<sup>19</sup>

---

<sup>17</sup>While the choice of course itself is non-negotiable, students can decide *when* to take a compulsory course in certain programs.

<sup>18</sup>This is in line with our result that female students do not systematically sort into courses taught by female, younger, or more lenient lecturers (Table 2).

<sup>19</sup>Note that we refrain from using only compulsory courses throughout for multiple reasons. First, the compulsory sample disproportionately represents certain fields with more compulsory courses; for instance, Economics exams constitute about 20% of all exams, but almost 30% of compulsory exams. Second, due to their very nature, compulsory courses are often not small (see number of observations in Table 5), which complicates our analysis of female gender match effects by class size. Third, the reliance on string matching to match compulsory courses to exams introduces potential measurement error in course classification. This is also another reason why we additionally show results for courses from the first two semesters.

Furthermore, even when considering non-compulsory courses, the likelihood of strategic course selection is constrained by the specific characteristics of the academic environment. First, there is almost no course duplication. Each course, whether elective or compulsory, is typically offered only once per semester by a particular lecturer. Second, the scheduling of courses is done on a semester-by-semester basis, with no advance public information on sabbaticals or potential faculty changes. Consequently, students are unable to foresee which instructor will teach a specific course in the upcoming year. These factors suggest that while high-ability female students could potentially try to avoid male instructors in their elective choices, doing so would mean sacrificing control over the elective content they pursue. This trade-off is unlikely to be appealing, as students generally prefer electives that align with their academic interests, and academic advisors typically recommend specializing through the choice of electives. Therefore, the range of electives students consider, especially those aiming to fulfill their degree requirements within a specific field (e.g., advanced macroeconomics or applied econometrics), is effectively quite limited, which substantially hampers the feasibility of systematically choosing courses based on the lecturer’s gender.

As a final check, we assess whether the female-male difference in various student characteristics differs between courses taught by female professors and courses taught by male professors (Appendix Table A1). To do so, we use five pre-determined student characteristics as outcome variables in the main estimation model outlined above (Pei et al. (2019)). We show the results of this balancing test across all classes as well as by class size and type of course (all vs. compulsory). We find little evidence for systematic differences: From the 30 coefficients of interest, only 4 are statistically significant at the 5%



level, and all coefficients are economically small.<sup>20</sup> Most importantly, we do not observe any sorting of students based on ability as measured by high school GPA.

### 3.3. *Heterogeneity*

In Table 6, we explore the heterogeneity of our results by broad academic field (see Section 2.1). As can be seen in the bottom of the table, the academic fields differ substantially in female lecturer share and class size. While most of the classes in the Economics & Political Science section are above the median in size, the opposite is true for the Humanities section.

When we pool all class sizes (Columns 1–3 of Table 6), we find similar positive and significant effects for STEM and Humanities majors, but no gender match effect in Economics & Political Science. This result can be explained by the fact that Economics & Political Science have a much higher share of large course observations (about 74%, as compared to 54% in STEM and 19% in Humanities). However, one particularly interesting result emerges when exploring effect heterogeneity by section and class size (see Columns 4–9). While, consistently across sections, gender match effects are statistically significant and sizable in smaller classes and insignificant in larger classes, female students in STEM benefit the most from being taught by female lecturers in smaller classes. However, the absence of significant gender match effects in larger classes indicates that even in settings in which female lecturers are most underrepresented, as is the case in STEM, merely witnessing another woman succeed in the same field is not sufficient for gender match effects to materialize.<sup>21</sup>

---

<sup>20</sup>Three of the four significant coefficients pertain to student age. In particular, we observe that the female-male difference in the age of students taught by a female lecturer is somewhat smaller than the female-male age difference of students taught by a male lecturer. However, the magnitude of the difference is small and we always control for student age in our regressions.

<sup>21</sup>Note that our estimate of the female gender match effect in large classes in STEM fields (Column 8 of Table 6) is positive and just shy of statistical significance at conventional levels ( $p=0.15$ ). Yet, the large class estimate amounts to only one-quarter of the estimate observed in smaller classes.

In Table 7, we investigate whether the benefits of being matched with a female lecturer accrue rather to higher-ability or to lower-ability female students, using students' high school GPA as a measure of academic ability. Columns 1 and 2 report results for students with a high school GPA above the median, Columns 3 and 4 restrict the sample to students with a below-median high school GPA. In both groups, we find that female students benefit from being paired with a female lecturer in a small class, while higher-ability students benefit even somewhat more (14.8% of a standard deviation, compared to 12.4% of a standard deviation for lower-ability students) (Columns 1 and 3). Higher-ability female students even somewhat benefit from having a female lecturer in large classes (Column 2), albeit to a smaller extent than in small classes.

### *3.4. Sensitivity Checks and Further Concerns*

A number of additional exercises, discussed in detail in the appendix and presented in Appendix Table A2, confirm the robustness of our results. These robustness checks include applying alternative definitions of “female-taught” or “large” courses, adding further high school controls, student fixed effects or program-by-semester fixed effects, allowing the effect of high school GPA to vary by high school type, location, and graduating year, and excluding students who drop out early. Moreover, while we have addressed the potential sorting of (high-ability) female students to courses taught by female instructors, another worry might be that (high-ability) female lecturers sort to small courses. However, we show that our results are robust to restricting the sample to lecturers who teach both large and small courses. We can also rule out that our results are driven by the fact that female lecturers are typically younger than male lecturers, potentially affecting their style of teaching, by allowing lecturer age to have a differential effect by student gender.

Another potential issue in any study of gender match effects using end-of-course grades as a measure of performance (e.g., Hoffmann and Oreopoulos (2009), Carrell et al. (2010)) is gender-biased grading, i.e., female graders giving better grades to female students.

For instance, Jansson and Tyrefors (2022) find evidence for same-sex bias in grading when exams are not anonymous. However, the institutional setting in our study renders gender-biased grading unlikely. Written exams are usually graded blind, with graders only knowing the student ID of the examinees, not their name or gender.<sup>22</sup> In addition, exams are often graded by teaching assistants and not by the lecturers themselves. Lecturers are therefore unlikely to know the gender of a student who wrote a given exam. The one major exception to this are so-called “seminars,” where students usually write and present a term paper. In these courses, a student’s identity is known to the grader. However, the class size of seminars is usually very small. In the Economics Department, for example, seminars are capped at 12 students. Given that gender match effects are also present in courses with 20, 30, and even 70 students (see Figure 1), gender-biased grading is unlikely to explain our findings.

#### 4. The Role of Student-Lecturer Interactions

Our evidence suggests that female gender match effects in higher education exist, but are strongly dependent on class size as they are not present in large courses. But why do these gender match effects exist, and why do they depend on class size? On the former question, we believe we have ruled out preferential grading and non-random assignment of students to lecturers. However, this still leaves at least two potential explanations: One explanation is gender-specific teaching skills, i.e., women might be better than men at teaching women. Another explanation is role model effects in a narrow sense, i.e., female lecturers motivating female students to do better. The difference between the two explanations is subtle, and we cannot distinguish between them empirically.

What can explain the class size gradient in female gender match effects? We believe that the intensity of student-teacher interactions is important. These interactions are

---

<sup>22</sup>Oral examinations are a possibility, but occur rarely. For instance, in the Economics Department there is no class where the grade is exclusively determined by an oral examination.

likely more frequent and of higher quality in smaller classes. Table 8 corroborates this claim based on data from course evaluations in the Economics Department. We observe that the larger the course, the less students feel that they can make comments, get useful feedback, or have the opportunity to ask questions.

The psychological literature also suggests the importance of student-teacher interactions. For instance, Buck et al. (2008) find that feeling a strong personal connection is necessary for being seen as a role model. Naturally, it seems easier to develop a personal connection with a lecturer in a small class than in an anonymous mass lecture. Additionally, Stout et al. (2010) show that female students are more likely to participate in class and seek help if their professor is female. It is likely that such behavior is more pronounced in small classes, where there is more opportunity to ask questions and interact with the professors.

It is worth considering potential alternative explanations for why female gender match effects mainly exist in smaller classes, although we cannot provide direct evidence for them. One possibility is that room size and climate may play a role, with larger classes potentially being louder and more difficult to concentrate in. Additionally, attendance may be lower in larger classes, reducing opportunities for students to interact with their professors. Another potential explanation is that the importance of teaching assistants (vis-à-vis professors) may be higher in larger classes, so the gender (and teaching style) of the professor matters less. Lastly, it is possible that students simply learn less in larger classes, which could also account for the absence of gender match effects.

These potential mechanisms reinforce the main message of the paper, namely that the female gender match effects observed in the literature are unlikely due to a simple role model effect of “seeing is believing”. If such role model effects were the only explanation, we would also expect to observe them in larger classrooms. However, the fact that gender match effects depend on classroom size suggests that the mechanisms underlying these effects may be more complex and context-dependent than previously thought.

## 5. Conclusion

We study whether female gender match effects in higher education depend on class size. To do so, we exploit rich administrative records from a German university, which cover all programs and courses in the period 2006 to 2018. We find that female gender match effects are substantial in smaller classes, implying performance gains of 13% of a standard deviation and a reduction in the probability of failing an exam by 1.5 percentage points if female students are taught by a female professor rather than a male professor. In contrast, there are no female gender match effects in large classes.

We are the first to show this quantitatively important interaction between female gender match effects and class size. Our findings add value to the expanding body of empirical research examining gender match effects in education, which has a predominant focus on smaller class settings. Notably, our results cast doubt on the idea that the sizable female gender match effects found in studies employing random student assignments in smaller classes simply transfer to the larger class settings common in public universities around the globe. In fact, our findings indicate that female gender match effects in such larger classes are likely substantially smaller.

Our findings also offer insights into the nature of female gender match effects. The mere knowledge that one's professor is female, which also students in large classes have, is apparently in itself not enough to increase the performance of female students. This suggests that the idea that gender match effects occur simply because female students are inspired by seeing another woman excel in a subject to the point of becoming a professor is too simplistic. Rather, our results suggest that gender match effects require direct and frequent interactions between students and professors, which is more typical in smaller classes.

Finally, our results also have important policy implications. Enrollment in tertiary education has increased in many countries in recent years, and the COVID-19 pandemic has

led to an increase in online education options in tertiary education, including MOOCs. These developments may result in more settings with larger class sizes and less direct and frequent interactions between students and professors. Our results suggest that this trend towards more online education may weaken the impact of policies designed to increase female graduation rates in traditionally male-dominated fields (such as STEM) by increasing gender diversity among professors.

## References

- [1] Antecol, Heather, Ozkan Eren, and Serkan Ozbeklik, “The Effect of Teacher Gender on Student Achievement in Primary School”, *Journal of Labor Economics* 33 (2015): 63-89.
- [2] Agurto, Marcos, Muchin Bazan, Siddarth Hari, and Sudipta Sarangi, “To Inspire and to Inform: The Role of Role Models”, Working Paper, 2022.
- [3] Beede, David, Tiffany Julian, David Langdon, George McKittrick, Beethika Khan, and Mark Doms, “Women in STEM: A Gender Gap to Innovation”, U.S. Department of Commerce Economics and Statistics Administration Issue Brief, 04-11, 2011.
- [4] Bettinger, Eric P., and Bridget Terry Long, “Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students”, *AEA Papers and Proceedings* 95 (2005): 152-157.
- [5] Bingham, Liz, “Role models are essential to help women reach the top”, *The Guardian* 2018, Sep 27 <https://www.theguardian.com/careers/role-models-gender-barrier>.
- [6] Blau, Francine D., Janet M. Currie, Rachel T.A. Croson, and Donna K. Ginter, “Can Mentoring Help Female Assistant Professors? Interim Results from a Randomized Trial”, *AEA Papers and Proceedings* 100 (2010): 348-352.
- [7] Blau, Francine D., and Lawrence M. Kahn, “The gender wage gap: Extent, trends, and explanations”, *Journal of Economic Literature*, 55 (2017): 789-865.
- [8] Breda, Thomas, Julie Grenet, Marion Monnet, Clementine Van Effenterre, “Can Female Role Models Reduce the Gender Gap in Science? Evidence from Classroom Interventions in French High Schools”, PSE Working Papers Nr. 2018-06 (2018).

- [9] Buck, Gayle A., Vicki L. Plano Clark, Diandra Leslie-Pelecky, Yun Lu, and Patricia Cerda-Lizarraga, “Examining the Cognitive Processes Used by Adolescent Girls and Women Scientists in Identifying Science Role Models: A Feminist Approach”, *Science Education* 92 (2008): 688-707.
- [10] Canaan, Serena, and Pierre Mouganie, “Does Advisor Gender Affect Women’s Persistence in Economics?”, *AEA Papers and Proceedings* 111 (2021): 112-116.
- [11] Canaan, Serena, and Pierre Mouganie, “The Impact of Advisor Gender on Female Students’ STEM Enrollment and Persistence”, *Journal of Human Resources* forthcoming.
- [12] Cho, Insook, “The Effect of Teacher-Student Gender Matching: Evidence from OECD Countries”, *Economics of Education Review* 31 (2012): 54-67.
- [13] Carrell, Scott E., Marianne E. Page, and James E. West, “Sex and Science: How Professor Gender Perpetuates the Gender Gap”, *Quarterly Journal of Economics* 125 (2010): 1101-1144.
- [14] Correia, Sergio. “REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects”, *Statistical Software Components* S457874, Boston College Department of Economics (2014, revised 2019).
- [15] Dee, Thomas, S., “A Teacher Like Me: Does Race, Ethnicity or Gender Matter?”, *AEA Papers and Proceedings* 95 (2005): 158-165.
- [16] Dee, Thomas S., “Teachers and the Gender Gaps in Student Achievement”, *Journal of Human Resources* 42 (2007): 529-554.



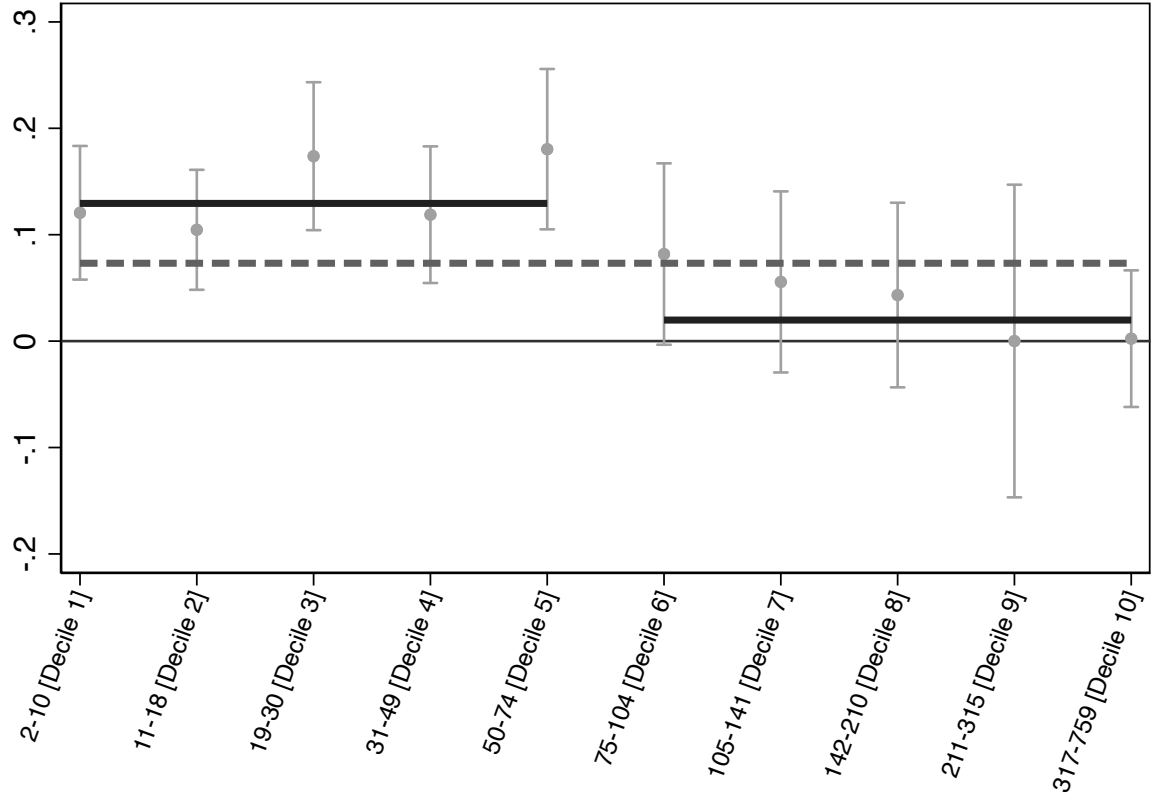
- [17] de Gendre, Alexandra, Jan Feld, Nicolas Salamanca, and Ulf Zölitz, “Do Same-Sex Teachers Affect Test Scores and Job Preferences? A Super-Study and a Meta-Analysis on Role Model Effects in Education”, Working Paper, 2022
- [18] Diekmann, Amanda B., Elizabeth R. Brown, Amanda M. Johnston, and Emily K. Clark, “Seeking Congruity Between Goals and Roles: A New Look at Why Women Opt Out of Science, Technology, Engineering, and Mathematics Careers”, *Psychological Science*, 21 (2010): 1051-1057.
- [19] Dynan, Karen E., and Cecilia Elena Rouse, “The Underrepresentation of Women in Economics: A Study of Undergraduate Economics Students”, *Journal of Economic Education* 28 (1997): 350-368.
- [20] Hoffmann, Florian, and Philip Oreopoulos, “A Professor Like Me. The Influence of Instructor Gender on College Achievement”, *Journal of Human Resources* 44 (2009): 480-494.
- [21] Hong, Lu, and Scott E. Page, “Groups of diverse problem solvers can outperform groups of high-ability problem solvers”, *Proceedings of the National Academy of Sciences* 101 (2004): 16385-16389.
- [22] Jansson, Joakim, and Björn Tyrefors, “Grading Bias and the Leaky Pipeline in Economics: Evidence from Stockholm University”, *Labour Economics* 78 (2022).
- [23] Kofoed, Michael S., and Elizabeth McGovney, “The Effect of Same-Gender or Same-Race Role Models on Occupation Choice”, *Journal of Human Resources* 54 (2019): 430-467.
- [24] Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson, “Who do we invent for? Patents by women focus more on women’s health, but few women get to invent”, *Science* 372 (2021): 1345-1348.

- [25] Lim, Jaegeum, and Jonathan Meer, “The Impact of Teacher-Student Gender Matches. Random Assignment Evidence from South Korea”, *Journal of Human Resources* 52 (2017): 979-997.
- [26] Lim, Jaegeum, and Jonathan Meer, “Persistent Effects of Teacher-Student Gender Matches”, *Journal of Human Resources* 55 (2020): 809-835.
- [27] Lyle, David S., and John Z. Smith, “The Effect of High-Performing Mentors on Junior Officer Promotion in the US Army”, *Journal of Labor Economics* 32 (2014): 229-258.
- [28] Muralidharan, Karthik, and Ketki Sheth, “Bridging Education Gender Gaps in Developing Countries. The Role of Female Teachers”, *Journal of Human Resources* 51 (2016): 269-297.
- [29] Neumann, Marko, Ulrich Trautwein, and Gabriel Nagy, “Do Central Examinations Lead to Greater Grading Comparability? A Study of Frame-of-Reference Effects on the University Entrance Qualification in Germany”, *Studies in Educational Evaluation* 37 (2011): 206-217.
- [30] Oh, Byeung-Kuk, “The Impact of Teacher Gender on Students’ Class Rank”, *Applied Economics Letters* 2023: DOI: 10.1080/13504851.2023.2299244
- [31] Paredes, Valentina, “A Teacher Like Me or a Student Like Me? Role Model Versus Teacher Bias Effect”, *Economics of Education Review* 34 (2014): 38-49.
- [32] Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt. “Poorly Measured Confounders Are More Useful on the Left Than on the Right”, *Journal of Business and Economic Statistics* 37 (2019): 205-216.

- [33] Porter, Catherine, and Danila Serra, “Gender Differences in the Choice of Major: The Importance of Female Role Models”, *American Economic Journal: Applied Economics* 12 (2020): 226-254.
- [34] Rask, Kevin N., and Elizabeth M. Bailey, “Are Faculty Role Models? Evidence from Major Choice in an Undergraduate Institution”, *Journal of Economic Education* 33 (2002): 99-124.
- [35] Rios-Avila, Fernando. “Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition”, *Stata Journal* 20 (2020): 51-94.
- [36] Schwerdt, Guido, and Ludger Woessmann, “The Information Value of Central School Exams”, *Economics of Education Review* 56 (2017): 65-79.
- [37] Silva, Pedro Luis, Carla Sá, and Rciardo Biscaia, “High School and Exam Scores: Does Their Predictive Validity for Academic Performance Vary with Programme Selectivity?”, IZA DP No. 15350 (2022).
- [38] Stout, Jane G., Nilanjana Dasgupta, Matthew Hunsinger, and Melissa A. McManus, “STEMing the Ide: Using Ingroup Experts to Inoculate Women’s Self-Concept in Science, Technology, Engineering, and Mathematics (STEM)”, *Journal of Personality and Social Psychology* 100 (2010): 255-270.
- [39] Warrell, Margie, “”Seeing Is Believing: Female Role Models Inspire Girls To Think Bigger”, *Forbes* 2020, Oct 9, <https://www.forbes.com/sites/margiewarrell/2020/10/09/seeing-is-believing-female-role-models-inspire-girls-to-rise/>.

## Figures and Tables

Figure 1: Class size heterogeneity of female gender match effects



*Notes:* Figure shows estimated female gender match effects and their 95% confidence intervals by class size decile. Dependent variable: Exam grades, standardized to mean 0 and standard deviation 1 at the exam-semester level. Estimations control for student characteristics (gender, high school GPA, age, German citizenship, being a local student, and first year in major) and for class size decile, program, course, semester, and lecturer set fixed effects. Dashed line indicates the average female gender match effect in the whole sample (see Column 2 of Table 3). Black lines depict gender match effects for class sizes below and above the median, respectively (see Columns 3 and 4 of Table 3). *Data source:* Administrative student records.

Table 1: Summary statistics

	Overall			Females			Males		
	Mean	Med	SD	Mean	Med	SD	Mean	Med	SD
Panel A: Exam-level variables									
Female lecturer	0.272	0	0.445	0.321	0	0.467	0.213	0	0.410
Exam grade	0	0.149	0.976	0.032	0.185	0.946	-0.038	0.099	1.010
Failed exam	0.129	0	0.335	0.103	0	0.303	0.160	0	0.367
Class size	121.1	74	131.1	111.0	70	122.3	133.1	80	140.0
Observations		310,554			168,788			141,766	
Panel B: Student-level variables									
Female student	0.532	1	0.499						
HS GPA	-0.149	-0.073	1.017	-0.049	-0.073	1.002	-0.262	-0.236	1.022
Age student	21.377	20.933	2.879	21.264	20.780	2.975	21.505	21	2.759
Native student	0.967	1	0.180	0.964	1	0.185	0.969	1	0.173
Local student	0.127	0	0.334	0.111	0	0.314	0.147	0	0.354
First year in major	2011.7	2012	3.9	2011.7	2012	4.0	2011.8	2012	3.8
Observations		18,592			9,890			8,702	

*Notes:* Table presents summary statistics for exam-level variables (Panel A) and student-level variables (Panel B). HS GPA refers to the final high school GPA; standardized to mean 0 and standard deviation 1 in the overall sample. Students' migration background is based on citizenship. Local students completed their high school in the county where the university is located. First year refers to the first academic year in which a student appears in our data in a given major. *Data source:* Administrative student records.

Table 2: Patterns of lecturer-student matching

	(1)	(2)	(3)
	Panel A: Dep Var: Lecturer female		
Female student	-0.000 (0.001)	0.000 (0.001)	-0.000 (0.001)
Observations	310,554	138,233	169,940
	Panel B: Dep Var: Course small & lecturer female		
Female student	0.000 (0.001)	-0.000 (0.001)	0.001 (0.001)
Observations	310,554	138,233	169,940
	Panel C: Dep Var: Course large & lecturer female		
Female student	-0.000 (0.001)	0.000 (0.001)	-0.001 (0.001)
Observations	310,554	138,233	169,940
	Panel D: Dep Var: Lecturer age		
Female student	0.016 (0.027)	0.015 (0.038)	0.017 (0.034)
Observations	266,039	120,621	143,610
	Panel E: Dep Var: Lecturer age (only courses with one lecturer)		
Female student	0.022 (0.028)	0.032 (0.041)	0.017 (0.035)
Observations	235,631	104,637	129,285
	Panel F: Dep Var: Lecturer prior course grades (only courses with one lecturer )		
Female student	-0.000 (0.001)	-0.001 (0.002)	0.001 (0.002)
Observations	243,064	104,927	136,385
Students	All	Above median HS GPA	Below median HS GPA

*Notes:* Dependent variables: Panel A: Binary variable indicating whether the course is taught by at least one female lecturer (“female-taught”); Panel B: Binary variable indicating whether the course is female-taught and the course size is at most 74 students; Panel C: Binary variable indicating whether the course is female-taught and the course size is above 74 students; Panel D: Lecturer age, averaged over the up to two lecturers teaching a given course (estimations without missing observations for both lecturer ages; if only one lecturer age is missing, the age of the other lecturer is used); Panel E: like Panel D, but only for courses with one lecturer; Panel F: average prior grades given by a lecturer in all previous courses, only for courses with one lecturer (see text for details). Column 1 has all students, while we consider only students with an above-median (below-median) final high school GPA in Column 2 (Column 3). All regressions control for student characteristics (high school GPA, age, German citizenship, being a local student, and first year in major) and for program, course, and semester fixed effects. Standard errors, twoway-clustered at the student and course level, in parentheses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.

Table 3: Female gender match effects by class size

	(1)	(2)	(3)	(4)
Female lecturer	0.071***	0.073***	0.131***	0.018
× female student	(0.019)	(0.018)	(0.018)	(0.029)
Student characteristics				
High school GPA		0.435***	0.379***	0.494***
		(0.007)	(0.007)	(0.009)
Female student	0.035**	-0.060***	-0.069***	-0.060***
	(0.014)	(0.013)	(0.014)	(0.017)
Student age	-0.035***	-0.007***	-0.004	-0.012***
	(0.003)	(0.002)	(0.002)	(0.003)
Native student	0.376***	0.227***	0.205***	0.234***
	(0.031)	(0.027)	(0.032)	(0.036)
Local student	-0.138***	-0.128***	-0.112***	-0.143***
	(0.017)	(0.013)	(0.015)	(0.018)
First year in major	0.016**	0.014**	0.005	0.028***
	(0.007)	(0.006)	(0.006)	(0.010)
Class size	All	All	Small	Large
$p$ -value large = small				<.001
Observations	310,554	310,554	155,591	154,960

*Notes:* Dependent variable: Exam grades, standardized to mean 0 and standard deviation 1 at the exam-semester level. High school GPA standardized to mean 0 and standard deviation 1 in the overall sample. All regressions control for program, course, semester, and lecturer set fixed effects. Small courses have 74 or fewer students, large courses have 75 or more students. Students' migration background is based on citizenship. Standard errors, twoway-clustered at the student and course level, in parentheses. In the bottom of the table, we report the  $p$ -value of a test for the equality of the gender match effects in small and large classes, based on a model that includes interactions between all variables and a dummy for small courses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.

Table 4: Female gender match effects for different grade categories in small classes

Dep Var:	(1) Grade A	(2) Grade B	(3) Grade C	(4) Grade D	(5) Fail
Panel A: Small Classes					
Female lecturer	0.045***	-0.006	-0.019***	-0.005	-0.015***
× female student	(0.007)	(0.007)	(0.005)	(0.003)	(0.004)
Mean dependent variable	0.286	0.380	0.179	0.066	0.089
Observations	155,591	155,591	155,591	155,591	155,591
Panel B: Large Classes					
Female lecturer	0.002	0.005	-0.006	-0.002	0.001
× female student	(0.006)	(0.010)	(0.006)	(0.006)	(0.007)
Mean dependent variable	0.134	0.293	0.267	0.136	0.169
Observations	154,960	154,960	154,960	154,960	154,960

*Notes:* Dependent variable: Binary variables indicating the four major grade categories (Columns 1–4) and binary variable taking a value of 1 if the student failed the exam, zero otherwise (Column 5). All regressions control for student characteristics (gender, high school GPA, age, German citizenship, being a local student, and first year in major) and for program, course, semester, and lecturer set fixed effects. In Panel A, sample is restricted to classes with 74 or fewer students; in Panel B, sample is restricted to classes with 75 or more students. Standard errors, twoway-clustered at the student and course level, in parentheses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.

Table 5: Female gender match effects by class size: Compulsory vs. elective courses

	(1)	(2)	(3)	(4)	(5)	(6)
Female lecturer	0.128***	0.129***	0.022	0.008	0.139***	0.010
× female student	(0.041)	(0.019)	(0.028)	(0.046)	(0.038)	(0.033)
Course type	Comp.	Elect.	Comp.	Elect.	First 2 semesters	
Class size	Small	Small	Large	Large	Small	Large
Observations	23,515	132,043	104,935	50,002	36,474	90,964

*Notes:* Dependent variable: Exam grades, standardized to mean 0 and standard deviation 1 at the exam-semester level. All regressions control for student characteristics (gender, high school GPA, age, German citizenship, being a local student, and first year in major) and for program, course, semester, and lecturer set fixed effects. Small courses have 74 or fewer students, large courses have 75 or more students. Standard errors, twoway-clustered at the student and course level, in parentheses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.



Table 6: Female gender match effects by class size and broad field

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female lecturer × female student	0.025 (0.031)	0.107*** (0.028)	0.085*** (0.026)	0.078** (0.030)	0.181*** (0.036)	0.118*** (0.026)	-0.005 (0.041)	0.046 (0.032)	-0.017 (0.052)
Class size	All	All	All	Small	Small	Small	Large	Large	Large
Broad field	Econ & PolSci	STEM	Humanities	Econ & PolSci	STEM	Humanities	Econ & PolSci	STEM	Humanities
Fem. lecturer share	0.258	0.195	0.381	0.333	0.220	0.404	0.232	0.174	0.280
Observations	104,507	120,364	91,507	27,050	55,765	74,243	77,446	64,589	17,248

*Notes:* Dependent variable: Exam grades, standardized to mean 0 and standard deviation 1 at the exam-semester level. All regressions control for student characteristics (gender, high school GPA, age, German citizenship, being a local student, and first year in major) and for program, course, semester, and lecturer set fixed effects. Small courses have 74 or fewer students, large courses have 75 or more students. The allocation of programs to broad fields follows the administrative division of the university. Econ & PolSci includes the programs Economics and Political & Administration Sciences. STEM includes the programs Biological Sciences, Chemistry, Computer Science, Information Engineering, Life Science, Mathematics, Molecular Materials Science, Nanoscience, Physics, Psychology. Humanities includes the programs British and American Studies, German Literature, French Studies, History, Italian Studies, Cultural Studies of Antiquity, Literature-Art-Media, Philosophy, Slavistik/Literature, Sociology, Spanish Studies, Linguistics, Sports science. Financial Mathematics is offered jointly by the Department of Mathematics and the Department of Economics and is allocated to both Econ & PolSci and STEM. Standard errors, twoway-clustered at the student and course level, in parentheses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.

Table 7: Female gender match effects by class size and high school GPA

	(1)	(2)	(3)	(4)
Female lecturer × female student	0.148*** (0.027)	0.058* (0.031)	0.124*** (0.023)	-0.006 (0.032)
Class size	Small	Large	Small	Large
Student high school GPA	Above median		Below median	
Observations	65,198	72,974	87,921	81,977

*Notes:* Dependent variable: Exam grades, standardized to mean 0 and standard deviation 1 at the exam-semester level. All regressions control for student characteristics (gender, high school GPA, age, German citizenship, being a local student, and first year in major) and for program, course, semester, and lecturer set fixed effects. Small courses have 74 or fewer students, large courses have 75 or more students. Standard errors, twoway-clustered at the student and course level, in parentheses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.

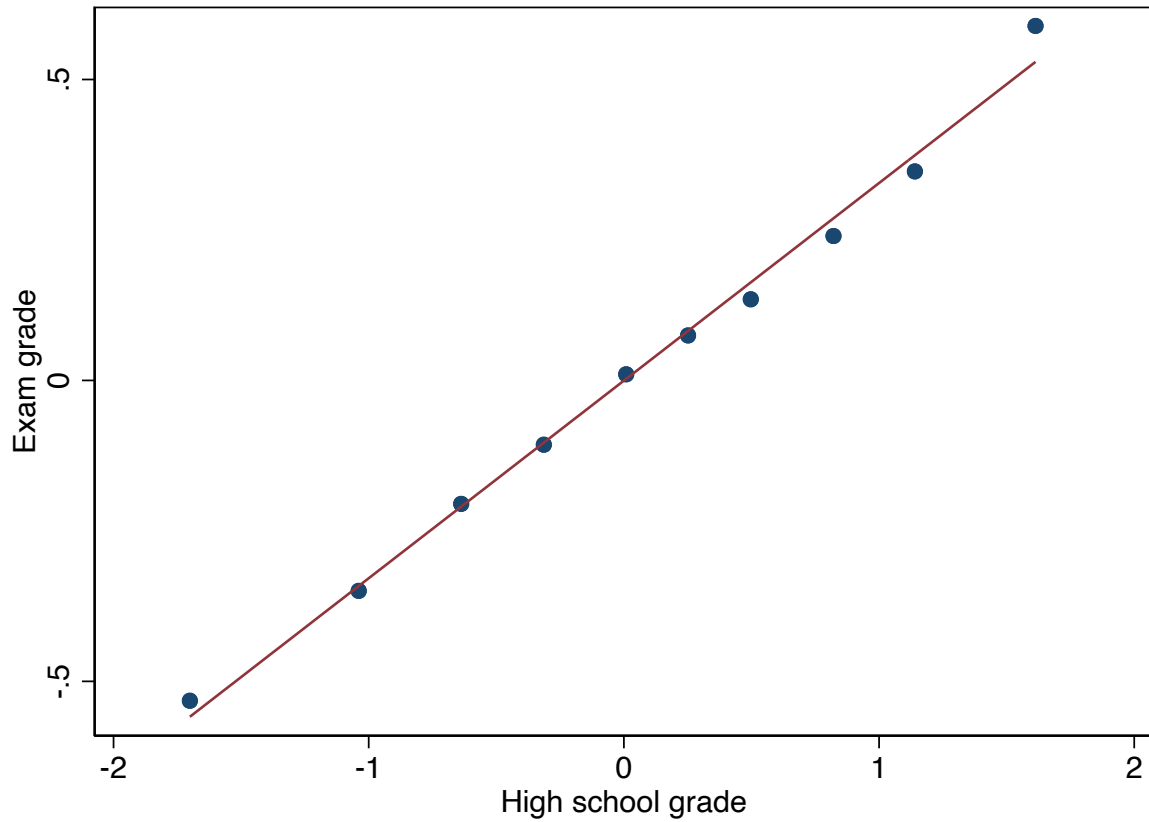
Table 8: Differences in lecturer-student interactions by class size from student evaluations

Dep Var:	Can make questions and comments		Get useful feedback		Opportunities to ask questions	
	(1)	(2)	(3)	(4)	(5)	(6)
Class size	-0.001*** (0.000)		-0.002*** (0.001)		-0.002** (0.001)	
Large class		-0.232*** (0.080)		-0.404*** (0.131)		-0.327** (0.122)
Mean Dep Var	4.574		4.258		4.469	
Observations	71		71		46	

*Notes:* Dependent variable: Course average of student replies to the questions indicated in the column header. The full questions read: “I feel I can ask questions and make comments at any time” (Columns 1 and 2); “I get useful feedback and advice from the lecturer when I ask” (Columns 3 and 4); “I have enough opportunities to ask questions” (Columns 5 and 6). Responses were given on a 5-point Likert scale, ranging from “strongly agree” (=5) to “strongly disagree” (=1). *Large class* is a binary variable, taking a value of 1 if the course had more than 74 filled-out evaluations, zero otherwise. Robust standard errors in parentheses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Student evaluations from all Economics classes for the winter semester 2018/19.

## Appendix

Figure A1: Relationship between exam grades and high school GPA



*Notes:* Binned scatterplot of the bivariate relationship between exam grades and final high school GPA. Exam grades are standardized to mean 0 and standard deviation 1 at the exam-semester level. High school GPA is standardized to mean 0 and standard deviation 1 in the overall sample. For both grade variables, the usual German ordering is reversed so that higher values indicate better outcomes. *Data source:* Administrative student records.

Table A1: Balancing tests

	Female gender match effect coefficient in					
	All classes (1)	Small classes (2)	Large classes (3)	Comp. (4)	Small comp. (5)	Large comp. (6)
<i>Outcome variable:</i>						
HS GPA	-0.006 (0.012)	-0.010 (0.019)	0.006 (0.011)	-0.001 (0.016)	-0.020 (0.047)	0.005 (0.016)
Age student	-0.104** (0.050)	-0.176** (0.072)	0.007 (0.043)	-0.045 (0.059)	-0.372** (0.167)	0.006 (0.053)
Native student	-0.001 (0.002)	-0.001 (0.004)	-0.002 (0.003)	-0.001 (0.004)	0.016** (0.007)	-0.003 (0.004)
Local student	0.004 (0.005)	0.008 (0.007)	0.000 (0.004)	0.001 (0.005)	0.004 (0.016)	0.000 (0.005)
First year in major	0.002 (0.013)	0.021 (0.016)	-0.017 (0.017)	0.009 (0.012)	0.021 (0.035)	0.008 (0.012)
Observations	310,554	155,591	154,960	128,455	23,515	104,935

*Notes:* Table shows results from regressing a number of predetermined student characteristics on the interaction of female student and female lecturer. All regressions control for the student being female and for program, course, semester, and lecturer set fixed effects, as well as for the four student characteristics that are not used as outcome variable in the respective regression. Columns 1–3 report results for compulsory and elective courses, Columns 4–6 report results for compulsory courses only. Small courses have 74 or fewer students, large courses have 75 or more students. HS GPA refers to the final high school GPA; standardized to mean 0 and standard deviation 1 in the overall sample. Students' migration background is based on citizenship. Local students completed their high school in the county where the university is located. First year refers to the first academic year in which a student appears in our data in a given major. Standard errors, twoway-clustered at the student and course level, in parentheses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.

Appendix Table A2 shows the robustness of our results to different specifications and parameterizations. As discussed in Section 2.2, classes can have up to two lecturers. In our main analysis, we keep all classes and define a course as female-taught if at least one lecturer is female. We check the robustness of our results to this definition in Panels A and B. In Panel A, we restrict the sample to classes with only one lecturer, where the definition of a female-taught course is unambiguous. In Panel B, we keep all classes, but define a course as female-taught if both lecturers are female. Our results are robust to both alternative definitions of female-taught courses.

One worry regarding the marked difference in the gender match effects for small and large courses is that the two types of courses may be taught by systematically different lecturers. To alleviate this concern, we restrict the sample to courses where at least one of the lecturers taught at least one large and one small course over our period of observation. The size of the interaction term, shown in panel C of Appendix Table A2, is somewhat smaller than in the baseline, but the results show the same pattern: Positive and significant female gender match effects for small courses, none for larger ones. Panel D further explores the role of teacher characteristics by considering teacher age. Given that female lecturers are on average younger than male ones, our estimated gender match effects could also be due to female students reacting differentially to younger lecturers. We therefore interact teacher age with student gender and include this as an additional control.<sup>23</sup> Our results are not affected by adding this control, indicating that our gender match effects are not simply age effects.

In Panel E of Appendix Table A2, we account for variation in school quality in Germany over time and across space by interacting the student high school GPA with indi-

---

<sup>23</sup>When there are two lecturers, we take the average of their ages. We lose observations since our dataset does not provide birth year information for every lecturer. When the age for only one lecturer is missing, we take the value for the other lecturer to calculate the age.

cators of the location of the high school, graduation year, and broad types of high school leaving exam.<sup>24</sup> Again, our results remain essentially unchanged.

In a similar vein, Panel F of Appendix Table A2 makes use of the fact that we observe several exams per student, allowing us to account for student fixed effects. Coefficients decrease by about half in this specification, however without altering our basic pattern: A sizable female gender match effect in small classes and no effect in large classes. Another potential confounder could be that specific departments hired more female lecturers over time and also changed exam standards, entry requirements, or other aspects of teaching. In Panel G, we therefore include major-by-semester fixed effects. Results are virtually identical to our baseline findings.

One possibility for students to differentially select into courses taught by men or women might be by postponing courses. In Panel H of Appendix Table A2, we thus define the usual study semester in which a given course is taken by calculating the modal study semester in which students in a given major take this exam. Based on this, we create a dummy variable indicating whether students take the course in the usual study semester (or earlier). Including this indicator as an additional control variable does not affect our results. The same holds for Panel I, where we exclude students who studied less than three semesters in their major. The latter check shows that our results are not driven by students who drop out early.

In our main analysis, we have defined large and small classes based on the median of the overall class size distribution. Based on the idea that the intensity of student-teacher interaction depends on class size, we consider this to be the most sensible approach. This is also in line with the pattern observed in Figure 1, showing that female gender match

---

<sup>24</sup>Location is measured by the county of high school graduation for students who graduated from high school in Germany. For students who completed high school abroad, we use the country of graduation. The most common type of high school leaving exam is the regular “Abitur” taken at standard upper secondary high schools. Other common types include Abitur at more specialized high schools, diplomas that allow university attendance only in some specific programs (“fachgebundene Hochschulreife”) or various types of vocational or second-chance education programs that award a university entrance qualification.

effects strongly decrease above the median of the class size distribution. However, one disadvantage of this approach is that some majors are very small and might thus not have many large classes, whereas for other majors, most classes might be large. In Panel J of Appendix Table A2, we therefore use major-specific medians to define large and small classes. This change in the definition of the class size cutoff leaves our results for small classes unchanged, as we continue to find a large female gender match effect. However, we now also observe a statistically significant, albeit much smaller, effect in large classes. This is likely due to the fact that in some majors, “large” classes by our definition are in fact small. In the programs “Slavistic” and “Cultural Studies of Antiquity”, for example, the median number of exam takers is 7. In “French Studies”, the median is 8.

Finally, in Panel K of Appendix Table A2, we check whether our results depend on the standardization of exam grades at the exam-semester level. We use raw exam grades that follow the German system from 1 (very good) to 5 (fail). In line with this new ordering, we now obtain a negative point estimates on the female lecturer female student interaction, but otherwise the same qualitative result: In small classes, female students paired with female lecturers receive significantly better (i.e., lower) grades, which is not the case in large classes.

Table A2: Robustness

Class size	All (1)	Small (2)	Large (3)
Panel A: Classes with only one lecturer			
Female lecturer	0.081***	0.132***	0.029
× female student	(0.019)	(0.018)	(0.034)
Observations	279,930	142,550	137,377
Panel B: Alternative treatment definition: Both lecturers female			
Female lecturer	0.077***	0.129***	0.024
× female student	(0.019)	(0.018)	(0.033)
Observations	310,554	155,591	154,960
Panel C: Only lecturers that teach both large and small courses			
Female lecturer	0.054**	0.112***	0.033
× female student	(0.026)	(0.029)	(0.031)
Observations	207,603	64,339	143,261
Panel D: Controlling for lecturer age × student gender			
Female lecturer	0.087***	0.135***	0.035
× female student	(0.019)	(0.019)	(0.028)
Observations	266,039	129,483	136,556
Panel E: Additional high school controls			
Female lecturer	0.065***	0.117***	0.014
× female student	(0.017)	(0.017)	(0.027)
Observations	310,465	155,359	154,737
Panel F: Controlling for student fixed effects			
Female lecturer	0.038**	0.071***	0.002
× female student	(0.016)	(0.016)	(0.025)
Observations	310,095	154,393	153,700
Panel G: Controlling for major × semester fixed effects			
Female lecturer	0.073***	0.132***	0.018
× female student	(0.018)	(0.018)	(0.028)
Observations	310,550	155,587	154,922
Panel H: Controlling for taking a course at the usual time			
Female lecturer	0.074***	0.133***	0.019
× female student	(0.019)	(0.018)	(0.029)
Observations	299,616	145,638	153,975
Panel I: Dropping students who study less than 3 semesters in major			
Female lecturer	0.079***	0.129***	0.025
× female student	(0.019)	(0.018)	(0.030)
Observations	286,683	150,429	136,249
Panel J: Using major-specific medians to define large courses			
Female lecturer	0.073***	0.141***	0.063***
× female student	(0.018)	(0.027)	(0.021)
Observations	310,554	53,511	256,746
Panel K: Raw exam grades			
Female lecturer	-0.059***	-0.114***	-0.012
× female student	(0.019)	(0.018)	(0.030)
Observations	310,554	155,591	154,960

*Notes:* Dependent variable: Exam grades, standardized to mean 0 and standard deviation 1 at the exam-semester level (raw exam grades in Panel K). All regressions control for student characteristics (gender, high school GPA, age, German citizenship, being a local student, and first year in major — the only exception being Panel F, where these get captured by the student fixed effects) and for program, course, semester, and lecturer set fixed effects. With the exception of Panel J, small courses have 74 or fewer students, large courses have 75 or more students. In Panel E, we allow the effect of the high school GPA to vary by the place of the high school, graduation year, and type of high school leaving exam. Standard errors, twoway-clustered at the student and course level, in parentheses. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.



Table A3: Quantile regression results for small classes

	(1)	(2)	(3)	(4)	(5)
Female lecturer × female student	0.128*** (0.026)	0.172*** (0.021)	0.168*** (0.019)	0.084*** (0.011)	0.047*** (0.015)
Percentile	10th	25th	50th	75th	90th

*Notes:* Sample restricted to small courses with 74 or fewer students (155,591 observations). Dependent variable: Exam grades, standardized to mean 0 and standard deviation 1 at the exam-semester level. Quantiles are computed over the within-course standardized grade. All regressions control for student characteristics (gender, high school GPA, age, German citizenship, being a local student, and first year in major) and for program, course, semester, and lecturer set fixed effects. Standard errors based on a bootstrap with 100 repetitions. Significance levels: \*  $p < .10$ , \*\*  $p < .05$ , \*\*\*  $p < .01$ . *Data source:* Administrative student records.