I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14036

Fear and Loathing in the Classroom: Why Does Teacher Quality Matter?

Michael Insler
Alexander F. McQuoid
Ahmed Rahman
Katherine Smith

JANUARY 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14036

Fear and Loathing in the Classroom: Why Does Teacher Quality Matter?

Michael Insler
United States Naval Academy

Alexander F. McQuoid

Ahmed Rahman
Lehigh University and IZA

Katherine Smith
United States Naval Academy

JANUARY 2021

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Fear and Loathing in the Classroom: Why Does Teacher Quality Matter?

This work disentangles aspects of teacher quality that impact student learning and performance. We exploit detailed data from post-secondary education that links students from randomly assigned instructors in introductory-level courses to the students' performances in follow-on courses for a wide variety of subjects. For a range of first-semester courses, we have both an objective score (based on common exams graded by committee) and a subjective grade provided by the instructor. We find that instructors who help boost the common final exam scores of their students also boost their performance in the follow-on course. Instructors who tend to give out easier subjective grades however dramatically hurt subsequent student performance. Exploring a variety of mechanisms, we suggest that instructors harm students not by "teaching to the test," but rather by producing misleading signals regarding the difficulty of the subject and the "soft skills" needed for college success. This effect is stronger in non-STEM fields, among female students, and among extroverted students. Faculty that are well-liked by students—and thus likely prized by university administrators—and considered to be easy have particularly pernicious effects on subsequent student performance.

JEL Classification: I20, I21, I23, J24

Keywords: soft standards, sequential learning, teacher value-added, rate my professor, higher education

Corresponding author:
Ahmed Rahman
Economics Department
College of Business
Lehigh University
621 Taylor Street
Bethlehem, PA 18015
USA

E-mail: asr418@lehigh.edu

1 Introduction

> When a man knows he is to be hanged in a fortnight,
> it concentrates his mind wonderfully.
>
> Samuel Johnson

Recent research has discovered large and persistent effects of teachers on student performance, but still fails to adequately uncover the specific channels (Rice (2003)). In a sequential learning framework we explore the channels through which instructor treatment in an initial period influences performance in the follow-on course in the sequence. Specifically, we decompose the "value-added" of each first-semester instructor into hard-skill and soft-skill components, each of which affects the student's subsequent performance. Observing grades from the second course in a two-course sequence, we find that while both channels impact longer-run learning, the soft-skill component appears to generate the larger impact.

Our identification strategy depends upon the random assignment of students to faculty and sections. Specifically, we use student panel data from the United States Naval Academy (USNA), where freshmen and sophomores must take a set of mandatory sequential courses, which includes courses in the humanities, social sciences, and STEM disciplines. Students cannot directly choose which courses to take nor when to take them. They cannot choose their instructors. They cannot switch instructors at any point. They must take the core sequence regardless of interest or ability. Our study is thus free of the selection concerns that plague most post-secondary educational environments. Due to unique institutional features, we observe students' administratively recorded grades at different points during the semester, including a cumulative course grade immediately prior to the final exam, a final exam grade, and an overall course grade, allowing us to separately estimate multiple aspects of faculty value-added.

Given that instructors determine the final grades of their students, there are both objective and subjective components of any academic performance measure. For a subset of courses in our sample, however, final exams are created, administered, and graded by faculty who do not directly influence the final course grade. This enables us to disentangle faculty impacts on objective measures of student learning within a course (grade on final exam) from faculty-specific subjective grading practices (final course grade).

Using the objectively determined final exam grade, we measure the direct impact of the instructor on the knowledge learned by the student. We will refer to this dimension of faculty quality as the "hard skills channel". Traditional measures of faculty quality mostly cannot directly isolate this knowledge transmission effect because grades are at least in part subjectively assigned by faculty. We further show that this hard skills channel does not suggest a "teaching to the test" effect, but rather shows genuine student learning of taught content.

Beyond this hard skills channel though, faculty can also shape student behaviors that are important for longer-run success. These may include how to allocate time to study, how to learn independently and without much hand-holding, and how to distinguish between easy and difficult subject areas. We will label the effect by which faculty may impair such skills as the "soft standards channel". When faculty set expectations that students do not need to put in significant effort to succeed in a discipline, and reward such behavior with easier grades, such low expectations may harm student performance in follow-on courses. To disentangle this effect from the hard skills channel, we use the subjective measure of professor quality stemming from the instructor-determined final course grade to separately identify the soft standards channel.

We find that the soft standards channel is particularly potent, and we further explore this mechanism to better understand how it impacts student learning. Disciplines that have arguably less sequential course content—such as English and social science courses—have a larger soft standards effect relative to STEM courses, which we attribute to the greater importance of soft skills in these disciplines. The soft standards channel carries through to the next course regardless of the amount of topical overlap.

To better understand this channel we merge data on student opinions of faculty from *ratemyprofessors.com* (RMP). A unique feature of this dataset is that it includes information on student opinions about faculty along two distinct dimensions. RMP includes not only the standard "overall rating" common to most teaching evaluations, it also includes students' rating on a dimension they care deeply about: instructor difficulty. Students seeking to minimize effort are particularly interested in which faculty are likely to have tough standards as these faculty will demand more effort for a given desired academic grade.

We find that instructors with higher overall ratings more severely harm sequential academic success, consistent with the soft standards channel. Overall ratings and difficulty ratings are highly negatively correlated, however, and upon accounting for difficulty ratings, overall ratings become positively related to sequential learning. Difficulty, however, has a much larger effect on sequential learning than overall ratings, consistent with our earlier finding that raising standards is a particularly important component of faculty quality.

Exploring the RMP data more deeply, we find that there are two types of particularly poor faculty. One type tends to be easy to identify: faculty with low difficulty and low overall ratings notably reduce student learning in sequential courses. Standard evaluations of teaching are likely to identify these low value-added faculty, so the policy implications of this finding are perhaps fairly straight-forward.

The second type of poorly-performing faculty are less likely to be identified as such. In fact, traditional measures of teaching quality are likely to wrongly identify them as high performers. These faculty tend to set low expectations and inflate grades. They are well-liked by students, who identify them as particularly easy (e.g., high overall rating, low difficulty rating in RMP data). Such faculty are likely well-known to colleagues who hear their praises from students and administrators, even as they generate negative externalities for colleagues who encounter their students in follow-on courses. Students encountering this type of faculty are notably worse off in sequential courses.

Given trends in higher education towards a more consumption-based model of learning—happier students are likely to pay higher prices, and higher grades keep students happy—this second type of faculty may be promoted and praised by administrators. A focus on student opinion forms in evaluating teacher performance coupled with grading discretion for faculty can alter instructor incentives in the classroom. This highlights a hard reality for teachers—aiding students cognitively is often unnoticed or unappreciated by the students themselves (Weinberg et al. (2009)) while facilitating their behaviors in preparing for subsequent coursework may often be outright punished.

To explore the persistence of faculty quality, we examine the impact of faculty in Calculus I and Calculus II on a third required semester of calculus. This three-semester sequence allows us to investigate the persistence of hard and soft skill transmission, as well as to see whether sequencing of faculty types matters. We find that both types of channels persist. While encountering a soft standards instructor at any point in one's education has persistent deleterious effects, the impact in the most recent semester is particularly pronounced. While the effect of the soft standards channel decays over time, the hard skills effect is similar for both Calculus I and Calculus II.

Finally, we explore the extent to which certain personality traits among students can help them overcome the longer-run detrimental impacts of soft standards professors. We find that students who are more introverted are less influenced by the soft standards channel, while extroverted individuals are more impacted. Previous literature suggests that extroverts tend to receive higher grades when instructors have greater autonomy over such grades (for example, giving more points for class participation as in Cain (2013)), which coupled with our findings suggest that the effects of soft standards should be expected to magnify for these cohorts. We also find that women tend to be more strongly negatively impacted by soft standards, consistent with previous studies that suggest that female students respond more strongly to initial grade signals than male students when it comes to choices such as major selection.[1]

The rest of the paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 discusses the data, while Section 4 develops the identification strategy. Results are presented in Section 5. Section 6 concludes.

[1] See for example the article "Why Are There So Few Women Economists?" in Chicago Booth Review, May 29, 2019.

2 Related Literature and Background

2.1 Teacher Value-added

There are essentially three big questions regarding the value-added of teachers for their pupils. One, how should we measure teacher value-added? Two, can these preferred measures of value-added be interpreted causally, or might they be driven by student sorting or some other omitted variable? Three, do our measures of value-added relate to true improvements in subsequent student outcomes, or are instructors merely spoon-feeding material or "teaching to the test?"

Most studies using student evaluations as a measure of teacher value-added focus on the students' self-professed amount of material learned, as well as students' overall rating of the instructor (Linask and Monks (2018)). But student evaluations in general will generate biased measures of value-added (Goos and Salomons (2017)).

Measuring instructor impacts on test scores is another approach. However teachers who are effective at improving test scores may not be effective at improving students' attitudes and behaviors towards learning. Teaching in other words can have multi-dimensional effects that need to be better understood (Blazar and Kraft (2017), Coenen et al. (2018)). One approach is to use alternative measures of teacher value-added, including non-cognitive skills proxied by absences, suspensions, and grade repetition (Jackson (2018)).

Instructor value-added measured by contemporaneous grades distributed by the same instructors may be inflated for various reasons (Palali et al. (2018)). Most studies look to student performance in follow-on courses or other longer-term outcomes. For example, studies such Figlio et al. (2015) and Xiaotao and Xu (2019) measure value-added as grade performance in self-selected follow-on courses in order to compare the teaching quality between tenured/tenure-track and non-tenured faculty.

Given that grades are typically chosen by faculty, measuring teaching quality in educational production is often done by examining the performance of students in subsequent classes. This sequential learning framework discussed in Shmanske (1988) has been used throughout the last several decades as a way to understand faculty value-added on student learning (see Weinberg et al. (2009) and Hoffmann and Oreopoulos (2009)). While the length of time one might expect the material to be carried into follow-on courses varies by discipline (Dills et al. (2016)), across disciplines subsequent course performance is a commonly used empirical strategy to identify teaching effectiveness.

With regards to causal interpretation, there are many potential challenges, such as student sorting towards preferred instructors. This is tackled effectively in Chetty et al. (2014a) for students in grades three through eight, but it remains a source of difficulty in most college environments. Value-added studies can also be complicated by the fact that in many institutions poor instructors teach fewer sections or less often, further biasing results (Feld et al. (2019)).

Finally an important question is whether or not ostensibly high value-added instructors, measured for example using test scores, actually help with longer-term student performance. The evidence is mixed. Again looking at elementary school instructors, Chetty et al. (2014b)) find positive instructor impacts for their students with respect to labor market and other socio-economic outcomes. On the other hand Carrell and West (2010) and Braga et al. (2014) show that students who have "popular" instructors tend to perform worse in follow-on college courses.

The impact of instructor standards on subsequent academic performance has been understudied, in part due to the difficulty of disentangling the various channels of faculty influence. For example, grading procedures designed to be objective are present in both Carrell and West (2010) and Braga et al. (2014), where common material is taught throughout each program with common exam questions for all students. While our setting has this feature as well, our additional measures of grade assignment with observable objective and subjective components at various points in the semester allow us to isolate these channels of faculty influence. Stroebe (2016) highlights the likelihood that grade leniency leads to worse academic outcomes but this is speculative, as it stems from indirect evidence from student evaluations.

2.2 Service Academy Research and External Validity

The three major United States service academies are the U.S. Naval Academy, the U.S. Military Academy (USMA), and the U.S. Air Force Academy (USAFA). These institutions have become increasingly prominent laboratories of applied economics in recent years, hosting research investigating a range of topics in labor economics, the economics of education, and behavioral economics.

USNA is a service academy with a liberal arts-style academic setting. Graduates earn a Bachelor of Science degree in one of approximately 25 majors. In this respect USNA is similar to USMA and USAFA; however USNA's academic setting is distinct from theirs in that the Naval Academy's faculty is at least fifty percent tenure-track civilian, career academics (this statistic tends to be as high as 60 percent in practice due to unfilled "billets" on the military side; see Keller et al. (2013)). In contrast, USMA's faculty model targets 25 percent civilian, while USAFA targets 29 percent (Keller et al. (2013)). The civilian tenure-track faculty are required to have earned a Ph.D and are evaluated for tenure according to guidelines that are comparable to those adopted by similar colleges and universities. The military faculty predominantly hold Masters Degrees (a small percentage have Ph.Ds). For these reasons, USNA is perhaps more academically comparable to other schools studied in the teacher value-added literature.

An additional benefit of our environment is the wide geographic and socioeconomic variety of students who attend USNA, which has a congressionally mandated mission to attract diverse students from across the nation in an attempt to provide an officer corps that is reflective of the wider naval force. As a result, the student body at USNA is more reflective of national college student populations than most universities (see Glaser and Insler (2020)).

2.3 Student Evaluations of Teaching (SET) via "Rate My Professor" (RMP)

Student Evaluations of Teaching (SETs) can be captured through faculty-rating sites such as *ratemyprofessors.com*. Given that participation is completely voluntary and that students must exert effort to go onto the site to find relevant professors, these evaluations are likely to be non-random. We can assume that students with strong opinions, either positive or negative, are more likely to take the time to voluntarily respond.

For RMP, the rating consists of five single-item questions concerning the easiness, clarity, helpfulness of the teacher, student's level of interest prior to attending class, and the use of the textbook during the course, which are aggregated into a single overall rating. Students are also asked to input information on their own course attendance and grade. In addition, they have the opportunity to add additional detailed comments about the course or the professor.

Previous studies have used data from RMP to answer three broad questions. First, to what degree do online evaluations and institution specific SETs substitute for one another? Brown et al. (2009) compare comments and ratings from the RMP website to institution specific SETs. They find ratings and comments on the website were very similar to those obtained through traditional evaluations, as they primarily concern teaching characteristics, personality, and quality. Ratings from the RMP website show statistically significant positive correlations (that exceed .60) with institutionally based SETs (Timmerman (2008)).

Second, do teaching evaluations capture quality or easiness of faculty? In general, more lenient instructors receive higher overall quality ratings (see Spooren et al. (2013)). Stuber et al. (2009) observe that, controlling for other predictors, instructor easiness predicted 50% of the variance in the scores on the overall quality measure. Timmerman (2008) found similar results and showed that this association can be partially explained by the fact that student learning is associated with student conceptions of an instructor's perceived easiness.

Third, do students choose courses and faculty members based on publicly available information (see Brown and Kosovich (2015))? While USNA students have no control over section placement in the first semester, they can provide input about class schedules in the second semester through a ranking system. However, the on-line system that solicits students' schedule rankings does not display faculty course assignments nor is such information publicly available through other channels. Our identification strategy thus limits concerns that selection of faculty members based on RMP information influences the results. However, the use of RMP ratings more commonly in the selection of courses and faculty confirms the view that student opinions do convey information that other students value.

3 Data Description

3.1 Institutional Data

We employ an administratively collected dataset of all USNA students enrolled during the academic years of 1997 through 2017. We observe every grade for every class taken by a student, along with course title and code, credit hours, and section number. Section number allows us to identify, for example, the subgroup of students in course X who sit in classroom Y on scheduled days A, B, and C at a specific time. We observe the instructor for each section, as well as instructor gender and type (civilian tenure-track faculty, civilian adjunct faculty, or officer). We observe grades at various points in the course: at six weeks, twelve weeks, end of classes, final exam, and overall course.

In addition to academic marks and course/section/instructor details, we observe a number of pre-USNA student characteristics: gender, math SAT score, verbal SAT score, high school standing, race/ethnicity, and feeder source (if any). All freshman are required to take the Myers-Briggs Type Indicators (MBTI) test, and we explore the role of student personality types in amplifying faculty standards in the analysis below. Summary statistics for students and faculty can be found in Table 1.

USNA provides an ideal setting to identify the effects of instructor standards on sequential learning, due to as-good-as-random assignment of students to initial and follow-on course-sections (and therefore instructors). Freshmen have little choice over the courses they take fall semester, and they have no choice over the section of a particular course nor the instructor. Students are not permitted to switch into nor out of sections to avoid certain instructors, before or during the fall semester, or to produce a more convenient schedule.[2] All freshmen must pass a set of

[2] The course schedules for first-semester freshmen are determined unilaterally by the registrar. USNA-based

11 core courses in a range of subject areas. It is possible for students to validate—or "test out of"—freshman year courses through USNA-administered placement exams or advanced placement (AP) scores (e.g., a student that validates Calculus I via AP scores will take Calculus II during the fall semester). The validation exams are the only form of indirect student input in the process.

The spring semester, however, is somewhat different. While most freshmen still have no choice over their spring semester courses (as most simply continue on within the core curriculum), they may attempt to select a daily schedule—and therefore course sections—based upon personal preferences. The online module for schedule selection does not include information on particular sections' instructors, and anecdotally students seem to rank their schedule options based on their preferences for the timing of free periods. The process employed by the registrar should create sections that are effectively random samples of the course population (with respect to both observable and unobservable characteristics). Brady et al. (2017) produces a battery of randomization tests for first-semester course assignments at USNA. However there may be some concern that follow-on course assignments are not quite so random, as students may somehow attempt to enroll in preferred time slots or with preferred instructors. To provide empirical evidence, we conduct balancing exercises to test that randomization holds for second-semester courses as well, at least with respect to observable characteristics.

For each section of each course in each semester we randomly draw 10,000 synthetic sections of equal size from the corresponding course's actual roster, without replacement. For each of these simulated sections, we compute the sum of each student-level observable characteristic: verbal SAT, math SAT, high school composite, and four separate personality scores. We next compute an empirical p-value for each section that is equal to the proportion of the 10,000 simulated sections that have summed values less than that of the actual section. Under random assignment, these p-value will be uniformly distributed. We conduct 882 distinct tests (six different second-semester courses across 21 academic years for each of the seven observable characteristics), using both a

sources for the information in this section are discussed in Brady et al. (2017).

Kolmogorov-Smirnov one sample equality of distribution test and a chi-squared goodness-of-fit test for each case. Results are summarized in Table 2.

The overwhelming number of tests fail to reject uniformity even at the 10% level. Overall mean and standard deviation across all p-value are 0.5 and 0.3, respectively, which are consistent with a Uniform(0,1) distribution. In summary, as there is little evidence to the contrary, the USNA procedure of allocating students to sections—at least for those courses used in our setting here—appears random with respect to students' observable characteristics.

3.2 RMP Data

Typically student opinion forms or teaching evaluations do not directly ask the student to assess the difficulty level of the instructor, especially relative to other instructors. Questions inquire about the student's overall impressions regarding the teacher, but the scope for response is often limited (e.g., positive, neutral, or negative). Further, while questions can cover different aspects of a course, responses often are highly correlated without much valuable variation. We use student opinions drawn from *ratemyprofessors.com*. Data from RMP is advantageous because survey responses capture multiple distinct dimensions of faculty characteristics as perceived by students.

More specifically, out of 686 separate instructors who teach within the core courses of interest, 321 have RMP profiles, with an average number of student opinions per instructor of 13.2.[3] Unsurprisingly, there is a high negative correlation between overall recommendation scores and level of difficulty scores (with correlation coefficient of -0.6). Summary Statistics based on RMP data are reported in Table 1.

[3] Note that for our baseline estimates we incorporate the average scores across all classes, including those not part of the core sequence. There is a high correlation between scores in core classes and upper level classes; results do not perceptibly change when excluding scores from upper level classes.

4 Econometric Approach

To identify aspects of faculty quality that affect college students' learning, we utilize the sequential structure of USNA's core curriculum. Our identification strategy depends on the random assignment of students in required courses for five sequences during the freshman and sophomore years. In the freshman year, all students are required to complete sequences in calculus (I and II), chemistry (I and II), English (I and II) and the social sciences. While the social science sequence is not strictly speaking a sequence—a history course and a political science course make up the content—the courses are framed as a social science sequence to freshmen during the registration period.[4] This sequence will in part help us identify the importance of the soft standards channel. In sophomore year, students take a sequence in physics (I and II), as well as a third semester of calculus. These five sequences cover both STEM and non-STEM courses and provide differing degrees of sequential content, ranging from strong sequential material in the natural sciences to minimal sequential material in the social sciences.

Our primary measures of faculty quality come from estimating faculty-specific effects on student outcomes (final exam grades and overall course grades) in the first semester of each sequence, after controlling for observable characteristics. We estimate the following equations:

$$Y^{1,f}_{ijks} = \Gamma^{f}_{k} + \theta^{1,f} X^{1}_{i} + \gamma_{year} + \gamma_{sem} + \gamma_{ks} + \epsilon_{ijks} \qquad (1)$$

$$Y^{1,c}_{ijks} = \Gamma^{c}_{k} + \theta^{1,c} X^{1}_{i} + \gamma_{year} + \gamma_{sem} + \gamma_{ks} + \epsilon_{ijks} \qquad (2)$$

where $Y^{1,c}_{ijks}$ represents the final grade student i received in section s of course j taught by professor k in the first course in the sequence, and $Y^{1,f}_{ijks}$ represents the final exam grade for the same $(ijks)$-quadruplet. X^{1}_{i} includes demographic information on students such as gender and ethnicity as well as pre-college academic information such as SAT scores and high school ranking. In addition, X^{1}_{i}

[4] Half of freshman are randomly assigned to take the history course in the fall followed by the political science course in the spring, while the other half start with the political science course in the fall and take the history course in the spring.

includes information on student personalities using Myers-Briggs personality types. Fixed effects for section, semester, and year are also included.

As discussed in Section 3, students receive grades at regular intervals throughout the semester as well as a final exam grade and a final course grade. Furthermore, the final exam for calculus, chemistry, and physics is an objective measure of student performance. The final exam is common for all students regardless of the instructor. The exam is written by a course coordinator and is graded by an external committee. As such, the instructor has no direct control over the final exam nor the final exam grade other than through teaching, which could include both genuine knowledge transmission as well as teaching to the test. The former should benefit achievement in the follow-on course, while the latter would impose a harmful effect or none at all.

Given the as-good-as randomization of students to course sections, from the perspective of the student, each faculty member is randomly assigned. Therefore the instructor effects, Γ_k^c and Γ_k^f, are not correlated with the error term and the instructor-specific coefficients capture the individual faculty members' value-added to the course grade and final exam grade in the first semester. As will become apparent below, Γ_k^c captures the effect stemming from the soft standards channel, and Γ_k^f captures the effect stemming from the hard skills channel. After estimating these instructor effects in the first semester, we standardize the estimates and include them in the grade determination model in the second semester to understand the impact of types of faculty on sequential learning.

In the analysis that follows, we explore the impact of these faculty specific effects on sequential learning using variants of the following estimation approach:

$$Y_{ijkls}^{2,c} = \alpha\Gamma_k^f + \beta\Gamma_k^c + \theta^2 X_i^2 + \delta_{sem} + \delta_{year} + \delta_{ls} + \eta_{ijkls} \tag{3}$$

where Y_{ijkls}^2 represents the final grade student i received in section s of follow-on course j currently taught by instructor l who had instructor k in the first course in the sequence. X_i^2 includes student-specific demographic information discussed above (or alternatively a student fixed effect) as well as previous course grade.

The two estimated instructor effects used in conjunction allow us to distinguish the hard skills and soft standards channels. While faculty have no direct control over grading of the common final exam, they retain some control over the final course grade. The common final exam has a common weight for all students regardless of professor (typically 30% of the overall grade), but since faculty have control over the assignments that make up the other portion of the grading scheme, the final course grade can reflect faculty tendencies in ways that the final exam grade does not. For the social science sequence, there is a final exam grade, but each professor has discretion over how to weight the assignment, what to include on the assignment, and what grade to assign.[5]

Our primary interest will be on the impact of the previous instructor in the sequence on follow-on course grades. The hard skills channel is captured through α while the soft standards channel is captured through β. Later in the analysis, we further adjust the specification to incorporate subjective student opinions from RMP data to more deeply explore the mechanisms through which faculty impact sequential student learning.

5 Results

In Table 3 column (1), we estimate the previous instructor final exam valued added on follow-on course grades, controlling for common fixed factors including year, semester, and current instructor-section fixed effects, which subsume course fixed effects. The estimated coefficient suggests that students in courses taught by faculty who have final exam scores that were better on average than expected based on pre-college student characteristics end up doing better in follow on courses. This finding is consistent with the view that these faculty are not simply teaching to the test, but are rather teaching important knowledge that carries through to sequential courses.

In column (2) we take the specification from column (1) and include our second measure of instructor effects based on course grades. To disentangle the soft standards channel from the hard skills channel, we include both measures of faculty quality described in Section 4: final course grade instructor effect (Γ_k^c from equation (2)) and final exam grade instructor effect (Γ_k^f from equa-

[5] The English sequence does not have a separate final exam grade, only a final course grade, and thus is not included in any model estimation that incorporates Γ_k^f.

tion (1)).[6] Interestingly, although average instructor effects on final exams continue to be positive and statistically significant, instructor effects based on course grades are negatively related to sequential learning. Soft standards and hard skills represent two distinct channels of faculty influence on sequential learning.

To test the robustness of this initial finding, we next exploit a unique feature of our academic setting. For STEM courses, the final exam is a common exam prepared by a single course coordinator and faculty do not grade their own students' exams. Column (3) again differentiates final exam and final course grade measures of faculty effects now using the subsample of courses that have common final exams, which offers the cleanest experiment available since no first-semester instructor grades her own students' final exams. Similar to the results in column (2), the impact of soft standards (β) has a negative effect on sequential learning, while the impact of hard skills (α) has a positive effect.

All else equal, faculty whose students do better on average on common final exams also do better in follow-on courses, consistent with the view that faculty quality measured in this way is capturing deeper learning (hard skills channel). In fact, this estimate may represents a lower bound, as it remains positive and significant despite the possibility that some instructors may be spoon-feeding material (i.e., teaching to the test), which cannot be directly observed in the data.

One possible limitation of columns (1)-(3) is that they do not control for student-specific aptitude, either in general or as it relates to specific course capabilities. Student capabilities may interact with faculty quality to influence sequential student learning. In column (4) we add general student characteristics in the form of student fixed effects (results are similar if we use specific characteristic controls) and a proxy for specific course capabilities in the form of the previous course grade. Results across columns (3) and (4) are comparable, and thus column (4) is our preferred specification due to its more robust set of controls.

The results thus far demonstrate that faculty whose students do better on the overall course grade tend to do worse in the follow-on course, *holding the hard skills channel constant*. This

[6] The correlation between the two measures is 0.41 for STEM courses and 0.85 for non-STEM courses that have final exams (history and political science).

finding is consistent with the view that faculty do more than just facilitate deeper learning, they also send signals about the difficulty of a discipline, generate enthusiasm for the discipline, and provide information about required effort levels. To the degree that faculty provide additional signals that are incompatible with the deeper knowledge taught to students, students end up performing worse in the follow-on course. This evidence suggests that faculty should consider the match between the demands of a discipline and student capabilities when signaling to students about how to approach sequential courses.

To explore this finding more carefully, we next consider faculty impacts in STEM and non-STEM courses. Column (5) includes an interaction term between an indicator variable for STEM course sequences (calculus, physics, and chemistry) and the previous instructor effects. We find no differential effect for the hard skills channel, but we do find that the soft standards channel operates differently between STEM and non-STEM courses. The sequential impact of the soft standards effect for a non-STEM course is -0.14 compared to -0.08 for a STEM course. The results suggest that the impact faculty have on student performance in sequential courses goes beyond just learned material as the soft standards effect is *more* pronounced for courses that are *less* sequential in terms of topical material.

The influences of teachers beyond just information transmission may matter more for certain types of students. For example, Rask and Tiefenthaler (2008) argue that women are more responsive to grade signals than men. In column (6), we estimate the impact of the previous faculty effects across genders. We find two dimensions of differentiation. In terms of the hard skills channel, female students who have faculty that are higher quality teachers do even better in follow-on courses than men who have the same high quality faculty. When looking at the soft standards channel, however, the impact of lower standards faculty is significantly larger for women when compared to men. This soft standards effect is roughly 20% larger for women than for men. Female students may respond more acutely to the signals provided by faculty that go beyond learned material.

Clear patterns emerge from Table 3. Teacher quality matters for two distinct reasons. On the one hand, high quality teachers are those who facilitate learning and knowledge accumulation, and these faculty improve student outcomes in sequential courses. On the other hand, teachers set standards, and those that send signals about the relative ease of a course damage student success in follow-on courses. We now turn to exploring these signaling effects further.

5.1 Student Ratings and Sequential Learning

To better understand the signals being sent by faculty, we next turn to student perceptions of faculty quality. Using data from *ratemyprofessors.com*, we include student perceptions of faculty overall and in terms of faculty difficulty. Our interest is in better understanding how students receive and process signals sent via faculty about the rigor and value of a course, and what impact these signals have on sequential learning. We further use the RMP data to build profiles of perceived instructor traits that are most harmful and helpful to students' sequential learning.

In Table 4, column (1), we start with an alternative measure of faculty quality using the overall RMP rating of the faculty member who taught the initial course in the sequence. Given our findings above from empirically-derived measures of faculty value-added stemming from grade information, it is perhaps not too surprising that faculty who are perceived to be higher quality by students are associated with lower learning in follow-on courses. While perhaps not surprising, it is nonetheless disheartening that student evaluations of faculty as high quality are negatively related to subsequent learning.

However, student opinion is likely to be influenced in large part by required effort demanded by faculty, which is captured in the RMP data through a measure of instructor difficulty. A stark visualization of the negative relationship between these two measures is shown in Figure 1. While some faculty who are truly poor teachers structure courses in confusing ways that raise the cost of learning to students, the more likely explanation of the observed negative correlation here is that effort is costly, and faculty who are more challenging demand more effort from their students, leading to low overall ratings. In column (2), we include both RMP measures of faculty quality for initial course instructors, and find that conditional on the level of difficulty, overall ratings are now positively related to subsequent learning. The impact of a professor who is deemed difficult is better for subsequent learning in follow-on courses, and the magnitude of the effect is about four times larger than the impact of a professor with a higher overall rating.

To the degree that student opinions play a role in the institution's evaluation of its faculty—and assuming that universities place greater weight on student learning as opposed to student happiness—these results suggest that extreme care should be given in how student opinions are utilized. A naive approach to evaluating student opinions of the quality of the faculty will tend to identify as excellent those who in reality harm rather than help subsequent learning.

In column (3), we include both measures of RMP faculty quality along with our grade-based measures of faculty quality (Γ_k^c and Γ_k^f). Student perceptions of faculty quality may not perfectly align with actual faculty quality, nor will the signals emitted by faculty be perfectly correlated with the reception and extraction of information by students. Figure 2 shows the relationship between instructor effects based on overall course grades (Γ_k^c) and RMP student ratings of faculty. In column (3), all four measures of faculty qualities are statistically significant, suggesting each is capturing a unique aspect of faculty influence. The inclusion of grade-based measures of faculty quality reduces the impact of the difficulty ranking, confirming the relationship between Γ_k^c and student perceptions of difficulty.

Column (4) shows the robustness of the results when restricting the sample to courses that have a common final exam. The results highlight the myriad of ways that faculty impact students in the classroom, from knowledge transmission to standards setting. Students are positively impacted by a mixture of "subjective standard setting" (via the RMP difficulty measure), "objective standard setting" (via the soft standards channel), and knowledge transmission (via the hard skills channel). The remaining RMP overall rating effect may represent an aspirational channel. After accounting for knowledge transmission and standards setting, faculty that students hold in high esteem tend to improve sequential learning outcomes. Figure 3 provides an illustrated representation of the impact of all four aspects of faculty quality on sequential GPA.

The information captured in our RMP data differs from what is commonly found on university administered student opinion forms. RMP captures something that administrators are perhaps uncomfortable collecting, but which students themselves care deeply about: how difficult is the professor? Effort minimizers will want to avoid faculty who demand significant effort. Learning is hard work requiring sustained effort, which is costly. Students are revealing in RMP opinions not just information regarding the difficulty of the instructor, but more importantly the amount of effort required by the instructor. The finding that higher effort results in sustained learning is not perhaps surprising, although the self-revelation mechanism may be.

5.2 Non-linear Impacts of Faculty Types

In Table 5, we expand on the results by allowing for non-linear impacts of different types of faculty based on categorizations related to ratings. Table 4 suggests that difficulty and overall ratings are providing useful information about what faculty are doing in the classroom, but the revealed information may not map monotonically into the ratings. We thus explore alternative signal extraction mechanisms based on various classifications of the underlying data.

In column (1), we consider dummy variables for the top 25% and bottom 25% of faculty in terms of either overall ratings or difficulty ratings, which allows for the possibility that ratings may not have a linear effect on sequential learning. Our conjecture is that faculty who are rated highly overall may be too lax on students, and thus not induce sufficient effort for deep learning, which would negatively impact future learning. However, faculty who are ranked low overall may just be poor teachers, who similarly fail to produce deep learning. Focusing on the upper and lower quartiles allows us to potential identify such non-linear impacts.

Consistent with that hypothesis, we find that relative to the middle 50% of the distribution, faculty in the top 25% and the bottom 25% in the overall ratings each diminish future learning. While the effects are larger for those in the bottom 25%, the finding of a negative effect for the top 25% is notable. Difficulty ratings on the other hand appear to follow a more monotonic relationship, with the easiest faculty (bottom 25%) harming sequential learning relative to a faculty member in the middle 50% of the distribution, while the hardest faculty generate improved sequential learning at a rate that mirrors that at the bottom. These findings give pause to the interpretation that faculty should be well-regarded by their students, as popular instructors appear to on average harm their students' achievement in follow-on courses.

In column (2), we consider the robustness of this finding by grouping the faculty into alternative categories. We consider four categories of overall ratings: those faculty at the very top and very bottom (top 10% and bottom 10%) as well as those near the top and bottom (between 10th and 25th percentiles, and between 75th and 90th percentiles). Those at the absolute top of the overall ratings harm sequential learning more than those simply near the top, while those at the very bottom appear to harm students in ways that are similar to those near the bottom of the overall ratings. Faculty at the extremes of overall ratings—those that are well liked and those that are reviled—are both detrimental for student learning in sequential courses when compared to faculty in the middle of the ratings distribution. In column (3), we repeat this exercise, but instead focus on splitting up the distribution more finely for difficulty ratings. These follow a more expected monotonic relationship. As difficulty ratings increase, sequential learning increases.

Column (4) considers the relative impact of all measures of faculty quality by now including soft standards and hard skills channels (Γ_k^c and Γ_k^f, respectively). Compared to column (1), the impact of the top 25% in both difficulty and overall ratings is reduced. The significance of the bottom 25% is maintained, however, which suggests that students who identify faculty as particularly poor overall or low difficulty rating do even worse even after accounting for the empirically-derived channels of faculty quality.

5.3 High Quality and Low Quality Faculty: Beware the Schmopes

Taken together, the results in columns (1) through (4) suggest that there are likely important bundles of characteristics for faculty who embody the extremes of both ratings. We thus group faculty into four different categories in column (5) of Table 5. Faculty that are both well-liked and considered very difficult (top 25% of each RMP rating distribution) are likely what most faculty aspire to be: challenging and demanding, but generating devotion and enthusiasm based on superior teaching. Such unicorns are extremely rare in the data, making up just under 2% of the overall faculty, and just over 1% of the total number of observations. We find no evidence that these faculty impact sequential learning, although this may be related to the small sample size. The other three groupings ("High Difficulty, Low Overall"; "Low Difficulty, Low Overall"; "Low Difficulty, High Overall"), however, are all statistically different from the excluded group of faculty who are in the middle of the distribution on at least one of these dimensions.

Which faculty are most associated with sequential learning? Those faculty who bundle together characteristics of high difficulty and low likability. One interpretation of this finding is that poor teachers are considered to be difficult by students because of lack of clarity in lecture and course structure, and this experience pushes students to invest in studying on their own to learn the material, resulting in deeper learning which is carried through to the next semester. However, we suspect that the issue is more likely to be explained by faculty who demand a lot of their students, forcing students to exert costly effort. This learning by effort leads to deeper learning and sequential success, but also engenders animosity towards the professor, resulting in high difficulty and low overall ratings.

Still examining column (5), faculty who are considered very easy and poor overall do notable damage to sequential learning. These faculty are in fact likely to be poor teachers, who perhaps minimize effort themselves, resulting in an easy and poor course. While students may not enjoy exerting effort, they are not totally unaware of their opportunity cost of time, and may feel cheated by these low engagement faculty, resulting in a signal (via RMP) to peers that the course is easy while also signaling disapproval of the behavior of the professor.

However, the grouping that appears to be most problematic are those with high overall ratings and low difficulty ratings. These faculty severely harm sequential learning, and more perniciously, are likely to be faculty who are praised by administrators for achieving high engagement from students (expressed through high overall opinions by students). These faculty—which we dub "a Seemingly Conscientious and Hardworking Mentor, an Obtuse and Perfunctory Educator" or *Schmopes*[7]—are deeply problematic because they damage student learning and are elevated within

[7] Inspired by discussions with our students who justified the value of this type of professor by commenting that they give hope to students, one rather perspicacious student retorted with "hope, schmope."

the university system as role model faculty.

The perniciousness of this type of faculty is confirmed in column (6). Even after including grade-based measures of faculty quality (Γ_k^c and Γ_k^f), the effect of having a *Schmope* as an initial course instructor remains detrimental to sequential learning. These faculty so heavily anchor and influence student beliefs about a discipline and about themselves that the effects persist even after controlling for all other measures of faculty quality.

The results suggest that *Schmopes* are damaging for sequential student learning but raise current excitement and regard for a class (and for the teacher). Assuming a university cares more deeply about sequential learning rather than student happiness, our results suggest that *Schmopes* are a serious problem in a university setting. While the role of *Schmopes* should be minimized, in light of the positive regard students hold for these faculty, there is perhaps some role for them in academia. Terminal courses or courses unrelated to a major could perhaps be a productive spot for these faculty. Furthermore, if student opinion is weighted too highly in evaluation of faculty, the incentives are such that non-*Schmopes* may act as if they are *Schmopes* to minimize unfair comparisons. At a minimum, cultural norms regarding *Schmopes* should be decidedly different. These faculty should be regarded as dubious educators and limited role models, who may provide some cheerleading value to students in controlled environments.

5.4 Persistence in the Knowledge Channel and Expectations Channel

We next exploit an additional unique feature of our setting: a third semester of a calculus sequence. As with our two semester sequences analyzed above, the third semester of calculus is required and students are as-good-as randomly assigned to classes, as demonstrated by our balancing tests. An additional caveat to the analysis in this case is that while students cannot choose faculty directly, there are two different options for the third semester of calculus. Students can choose to meet the requirement using either a course focused on vector fields (SM221) or a course focused on optimization (SM223). Approximately 60% of students take the vector fields option. Selection of the course is usually influenced by major choice, which occurs after freshman year. For example, economics majors tend to take the optimization sequence. While students have more choice with this third math sequence course, they are still limited in their ability to select specific faculty.

Following the methodology above, we estimate a hard skills effect and a soft standards effect for faculty in both Calculus I and Calculus II. We then explore how these faculty effects evolve over time as they impact the third semester of calculus. To show the consistency of our results with the previous analysis, column (1) of Table 6 re-estimates our preferred specification restricting the sample to only Calculus II observations. Consistent with the previous results for all courses, the hard skills channel has a positive and significant effect on sequential learning. Faculty soft standards continues to have a detrimental effect on sequential learning.

Column (2) shows the impact of faculty from Calculus I and Calculus II on Calculus III performance. Consistent with previous results, the hard skills channel is still positive and significant for both Calculus I and Calculus II. Additionally, faculty soft standards effects are negative and statistically significant. These results provide additional support for the mechanisms identified previously, and provide additional insights into why faculty quality matters. The magnitudes of the hard skills channels for Calculus I and Calculus II faculty are similar, suggesting that knowledge base building in sequential courses has long run impacts on future learning. It also suggests that having a professor who is better at stimulating knowledge accumulation is valuable regardless of when in the process one encounters her.

Soft Standards effects seem to diminish over time. While both Calculus I and Calculus II soft standards effects are negative and significant, and continue to be notably larger than the hard skills channel (impacts are 2 to 3 times larger), they appear to diminish over time. The impact of a particularly lenient professor (a one standard deviation increase in Γ_k^c) in Calculus I is -0.081, while a similarly lenient professor in Calculus II reduces sequential learning by nearly twice as much, -0.153. This suggests that the standards channel identified here continues to influence student behavior, but the influence is reduced over time as students adjust to new surroundings. It is worth emphasizing that the soft standards effect in the first semester is not zero: standards set by faculty persist well past the first course.

To explore the soft standards channel in more detail, we consider non-linear effects of different types of faculty standards. For faculty who are very lenient, the effects may be very different than for faculty who have tougher standards. Column (3) groups faculty into three categories (for both Calculus I and Calculus II): soft standards faculty (top 25% of the standards distribution), tough standards faculty (bottom 25% of the standards distribution), and neither (middle 50% of the distribution). This classification allows us to explore non-linear effects. For Calculus I, the magnitude of the impact of soft standards is about half the impact of tough standards. For Calculus II, the magnitude of soft standards is similar to that of tough standards. Patterns in column (3) reinforce that tough standards have similar effects regardless of the timing, while the impact of soft standards diminishes over time. Soft standards in Calculus I have only one-half of the impact of soft standards in Calculus II.

Column (4) uses a more refined categorization: the top and bottom 25% of the standards distribution is broken up into the top 10% and the next 15%, for both Calculus I and II. Once again, similar patterns emerge. The impact of tough standards for each portion of the distribution is similar for Calculus I and II, except for the toughest faculty in Calculus II who have a significantly larger impact on Calculus III grades. The impact of soft standards, while nonlinear, is monotonic: the more lenient the professor, the worse students do in follow-on courses. Soft standards reduce academic success, but the effect diminishes over time. In fact, some leniency in Calculus I (top 10-25% of faculty standards) has no significant effect on Calculus III grades, while similar leniency in Calculus II reduces academic performance in Calculus III. However, particularly soft standards faculty in Calculus I continue to hinder students two semesters later. Although the persistence of the effect of a soft standards instructor diminishes over time, it is easy to imagine that soft standards may influence student choices in other dimensions, such as choice of major. Encountering a lenient instructor early on in a college career may influence a student to choose a major that is not well suited to the student's comparative advantage (see Insler et al. (2020)).

The results in columns (2)-(4) raise the issue of the sequencing of faculty standards. If the effect of soft standards diminishes over time, and the effect of tough standards persists, how detrimental is a soft standards professor if a student is likely to encounter a tough standards professor later on in the sequence? Column (5) introduces dummies for sequential draws for students. If a student draws a lenient professor in both Calculus I and in Calculus II, how much does this affect her outcome in Calculus III? Drawing two consecutive soft standards faculty in Calculus I and II lowers a student's overall grade in Calculus III by -0.25 (relative to drawing two consecutive middling faculty), which is equivalent to about a quarter of a standard deviation of Calculus III grades. Drawing two consecutive tough standards faculty has a mirrored effect, raising Calculus III grades by 0.32. Interestingly, regardless of the timing of the sequence, drawing one tough standard and one soft standard professor cancels each other out, with no net effect on Calculus III grades. The impact of tough standards professors persists regardless of when you draw them, as can be seen in the sequences where you draw one middle standards professor and one tough standards professor. Drawing the tough standards professor first or second in this case has a similar positive effect on your Calculus III grade. On the other hand, the impact of a soft standards professor followed by a middle standards professor has only one-half of the impact of drawing a middle standards professor followed by a soft standards professor.

Sequencing is further explored using the more disaggregated classifications (as found in column (4)). Figure 4 shows the estimated coefficients based on alternative sequences of faculty standards in Calculus I (columns) and Calculus II (rows). The first panel shows the impact of different types of faculty in Calculus II assuming one drew an especially soft standards professor in Calculus I. The impact of that initial soft standards draw is blunted only if one draws a sufficiently tough standards professor in Calculus II. The diminishing effect of a soft standards professor in Calculus I can also be seen in column (2), where a middle standards professor is able to offset the damage done from the low expectations set in Calculus I. Conversely, a very tough standards professor in Calculus I (column (5)) is offset by a soft standards professor in Calculus II because of the impact of a more recent soft standards professor.

The results from our exploration of the three-semester calculus sequence suggest that the impact of soft standards diminishes over time, while the impact of tough standards is persistent and of a similar magnitude regardless of when a student encounters such a professor. While the impact of soft standards diminishes, the effects are particularly acute in the semester immediately following. The impact of a soft standards professor in the previous semester has a larger impact than any other type of professor in the analysis. Thus the timing of encountering a soft standards professor matters more for sequential learning than the timing of encountering a tough standards professor.

5.5 Mechanism Robustness

To conclude the analysis, we consider additional characteristics of faculty and students that may help to explain why the soft standards channel is so impactful. In the process, we show the robustness of the channel and its impact on sequential learning. We consider the role of student personality, faculty type, and gender in turn.

Student Personality

To start, we consider how student characteristics may influence the response to standards set by faculty. Student personality characteristics—as measured by the Myers-Briggs test—may help to explain the magnitude of the soft standards channel.[8] In column (1) of Table 7, we interact our measure of faculty soft standards (Γ_k^c) with the MBTI extroversion-introversion scale. The more extroverted a student, the more detrimental the impact of having a lenient professor. To consider potential nonlinear impacts between extroverts and introverts, in column (2) we include indicator variables for students in the top 25% of the introversion scale ("High Introversion") and in the top 25% of the extroversion scale. We find that it is the high introverts who are least affected by lenient faculty. This is consistent with the view that more introverted students are more focused on their own sense of performance rather than the external signals given by faculty.

In column (3), we consider a second measure of personality from the MBTI: thinking/feeling. The thinking/feeling dimension focuses on how an individual tends to make decisions. Those who

[8] All students must take this personality test early in their freshman year.

score high on the thinking scale put more weight on objective principles and facts, while those who score higher on the feeling scale put more weight on the people involved and personal concerns. Consistent with this characterization, students who are more "thinking" are less influenced by the faculty standard-setting, while those who make decisions based on the people involved ("feeling") are more influenced. The results are consistent with the view that the soft standards channel works through influencing students about the amount of effort and dedication required in subsequent courses.

Faculty Types

We also consider characteristics of faculty within this context. USNA has both civilian professors as well as military instructors (each making up about 50% of the faculty). We consider whether students respond differently to soft standards and hard skills channels when taught by a military officer. Column (4) finds that soft standards are more damaging to subsequent learning when they come from a military officer although the effect is still present for civilian faculty. We interpret this finding as consistent with the view that soft standards are important in any context, but are amplified when students feel a greater sense of homophily towards instructors. There is no differential effect from the hard skills channel, which is consistent with the interpretation that this channel captures deeper learning that improves sequential academic success.

Gender and Grade Signals

Finally, in column (5) of Table 7 we consider the gender of the student alongside the gender of the professor. As in Table 3, we find that female students respond more to standards-setting, resulting in lower performance than their male classmates in sequential courses. We do not find any evidence that female faculty are treated differently by students—either male or female—in terms of how students respond to standards. But female students perform worse in follow-on courses due to soft standards, irrespective of the gender of the instructor.

There are two broad interpretations we can make here, one related to how instructors may grade female students differently, the other to how female students may respond differently to faculty standards. Regarding the first of these, studies suggest that instructors tend to praise female students for “good” behavior, regardless of its relevance to content or to the lesson at hand, and tend to criticize male students for “bad” behavior (Golombok et al. (1994)). One potential consequence is to make females appear to be better than they may really be, and to make this “goodness” appear more important than their academic competence. Gender biases towards appropriate behavior of female students may contribute to misleading signals regarding academic performance.

On the other hand, many studies also suggest that females take grade signals as information regarding their own ability more seriously than their male counterparts (Rask and Tiefenthaler (2008), Goldin and Guerrieri (2019)). A recent study for example highlights that male students are more likely to request a change in grade from the instructor compared with female students (Li and Zafar (2020)). These studies often point out that women receiving poor grades become more discouraged about their innate ability to do well. Our study complements these works by highlighting a corollary finding, that higher standards can also elicit greater effort, producing better academic achievement later on.

6 Conclusion

Student performance in sequential courses is impacted by the quality and characteristics of faculty in previous courses. We explore why teaching quality matters for the accumulation of human capital in post-secondary education. Effective instructors are able to transfer knowledge and elicit labor effort from students. Utilizing as-good-as random assignment of faculty and students in a variety of sequential courses and observed grades at different phases during the course, we identify two distinct channels through which teacher quality matters, a hard skills channel and a soft standards channel.

Both channels are important for understanding the ways in which faculty have persistent influence on students’ academic performance. We find that the soft standards channel is particularly impactful on subsequent student performance. To explore this channel more deeply, we merge our institutional data with student opinions of faculty from *ratemyprofessors.com*. We find that overall ratings are negatively associated with lower sequential learning if faculty difficulty is not accounted for. If faculty difficulty ratings are included, overall ratings are positively associated with sequential learning, but the impact of difficulty rating is more important. Overall and difficulty ratings are negatively correlated, suggesting that while induced effort is key to longer term learning, students' dislike of upfront costly effort negatively influences opinions of faculty quality.

Using a three-semester sequence in calculus, we consider the persistence and sequential timing of faculty on student performance. We find that the magnitude of the soft standards channel diminishes over time, while the impact of tough standards persists regardless of the timing. The hard skills channel remains persistent over time.

In addition, we explore student and faculty traits that amplify the soft standards channel. Those who rank more highly in terms of introversion and thinking on the MBTI tend to be less affected by standard-setting, while female students tend to respond more strongly to instructor leniency. The impact of faculty standards is stronger in non-STEM courses compared to STEM courses.

Finally, we identify two types of faculty that are particularly damaging. One is easy to identify using traditional measures of student opinions: faculty who are considered easy and poor overall. The second type, however, is more problematic as they tend to be well-liked by students, in part because they are considered easy. *Schmopes* may provide some edification, but they are fundamentally dubious educators and limited role models.

What do we want higher education to achieve? Undergraduate education involves the inculcation of a variety of soft skills, such as good study habits, an appreciation for the time necessary to do the work, an ability to work with others, among other behavioral traits. Colleges are also often anxious to please their students. Complaints regarding dry lectures, excessive difficulty of material or perceived lack of face-to-face time with faculty can worry administrators who rely on money from tuition-paying students, alumni, and donors. Colleges are after all businesses with their own underlying revenue and cost structures. The increasing financial pressures to placate students may ultimately be incompatible with faculty objectives to facilitate the behaviors that lead to long-term success in college (and beyond).

References

Blazar, D. and Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational evaluation and policy analysis*, 39(1):146–170.

Brady, R. R., Insler, M. A., and Rahman, A. S. (2017). Bad company: Understanding negative peer effects in college achievement. *European Economic Review*, 98:144–168.

Braga, M., Paccagnella, M., and Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, 41:71–88.

Brown, C. L. and Kosovich, S. M. (2015). The impact of professor reputation and section attributes on student course selection. *Research in Higher Education*, 56(5):496–509.

Brown, M. J., Baillie, M., and Fraser, S. (2009). Rating ratemyprofessors. com: A comparison of online and official student evaluations of teaching. *College Teaching*, 57(2):89–92.

Cain, S. (2013). *Quiet: The power of introverts in a world that can't stop talking*. Broadway Books.

Carrell, S. E. and West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9):2633–79.

Coenen, J., Cornelisz, I., Groot, W., Maassen van den Brink, H., and Van Klaveren, C. (2018). Teacher characteristics and their effects on student test scores: A systematic review. *Journal of economic surveys*, 32(3):848–877.

Dills, A., Hernández-Julián, R., and Rotthoff, K. W. (2016). Knowledge decay between semesters. *Economics of Education Review*, 50:63–74.

Feld, J., Salamanca, N., and Zölitz, U. (2019). Students are almost as effective as professors in university teaching. *Economics of Education Review*, 73:101912.

Figlio, D. N., Schapiro, M. O., and Soter, K. B. (2015). Are tenure track professors better teachers? *Review of Economics and Statistics*, 97(4):715–724.

Glaser, D. J. and Insler, M. (2020). The deleterious effects of fatigue on exam timing. *USNA Working Paper Series*.

Goldin, C. and Guerrieri, V. (2019). Why are there so few women economists? *Chicago Booth Review*.

Golombok, S., Fivush, R., and Fivush, G. (1994). *Gender development*. Cambridge University Press.

Goos, M. and Salomons, A. (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education*, 58(4):341–364.

Hoffmann, F. and Oreopoulos, P. (2009). Professor qualities and student achievement. *The Review of Economics and Statistics*, 91(1):83–92.

Insler, M., Rahman, A., and Smith, K. (2020). Herding - a theory and some evidence from college major selections. Technical report, USNA Working Paper Series.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Keller, K. M., Lim, N., Harrington, L. M., O'Neill, K., and Haddad, A. (2013). The mix of military and civilian faculty at the united states air force academy.

Li, C. H. and Zafar, B. (2020). Ask and you shall receive? gender differences in regrades in college. Technical report, National Bureau of Economic Research.

Linask, M. and Monks, J. (2018). Measuring faculty teaching effectiveness using conditional fixed effects. *The Journal of Economic Education*, 49(4):324–339.

Palali, A., Van Elk, R., Bolhaar, J., and Rud, I. (2018). Are good researchers also good teachers? the relationship between research quality and teaching quality. *Economics of Education Review*, 64:40–49.

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

Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes.* ERIC.

Shmanske, S. (1988). On the measurement of teacher effectiveness. *The Journal of Economic Education*, 19(4):307–314.

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4):598–642.

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science*, 11(6):800–816.

Stuber, J. M., Watson, A., Carle, A., and Staggs, K. (2009). Gender expectations and on-line evaluations of teaching: Evidence from ratemyprofessors. com. *Teaching in Higher Education*, 14(4):387–399.

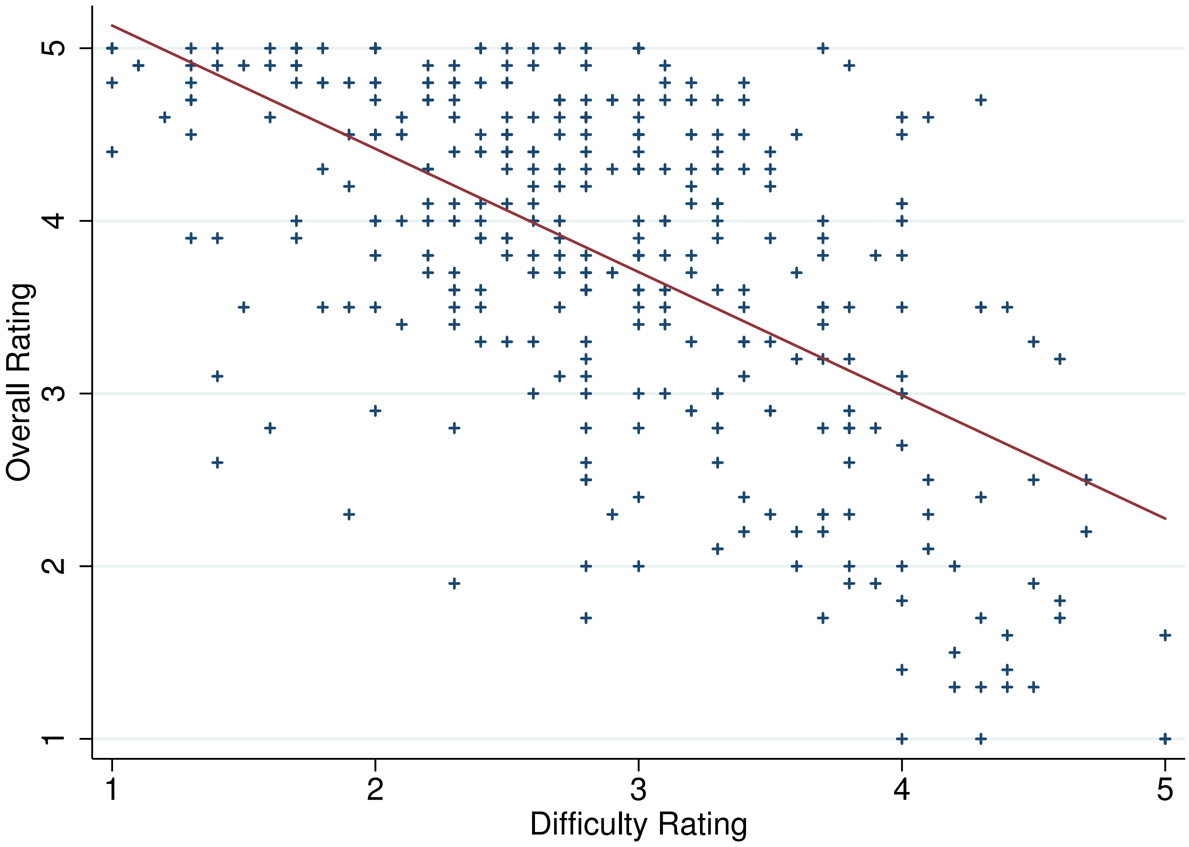
Timmerman, T. (2008). On the validity of ratemyprofessors. com. *Journal of Education for Business*, 84(1):55–61.

Weinberg, B. A., Hashimoto, M., and Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education*, 40(3):227–261.

Xiaotao, F. and Xu, D. (2019). Does contractual form matter? the impact of different types of non-tenure-track faculty on college students' academic outcomes. *Journal of Human Resources*, 54(4):1081–1120.

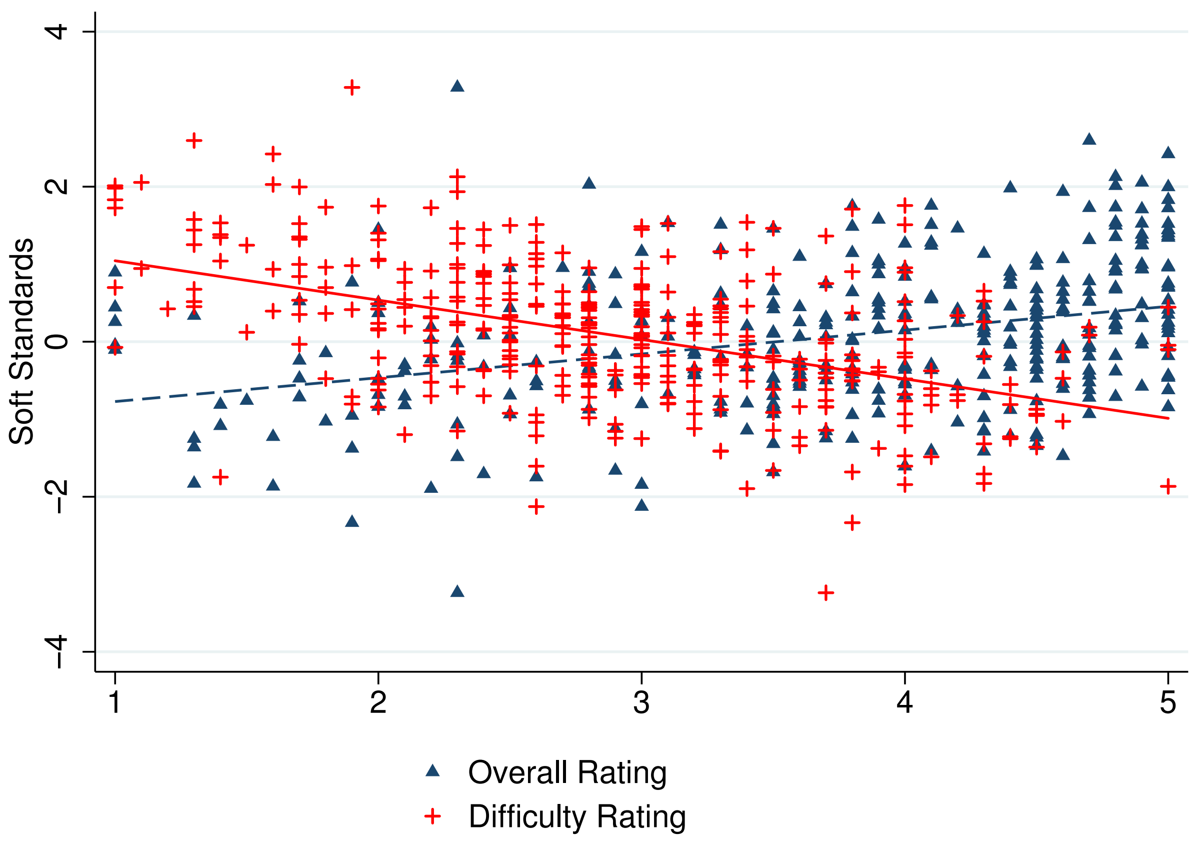
Figures

Figure 1: *Ratemyprofessors.com* Difficulty Rating vs. Overall Rating



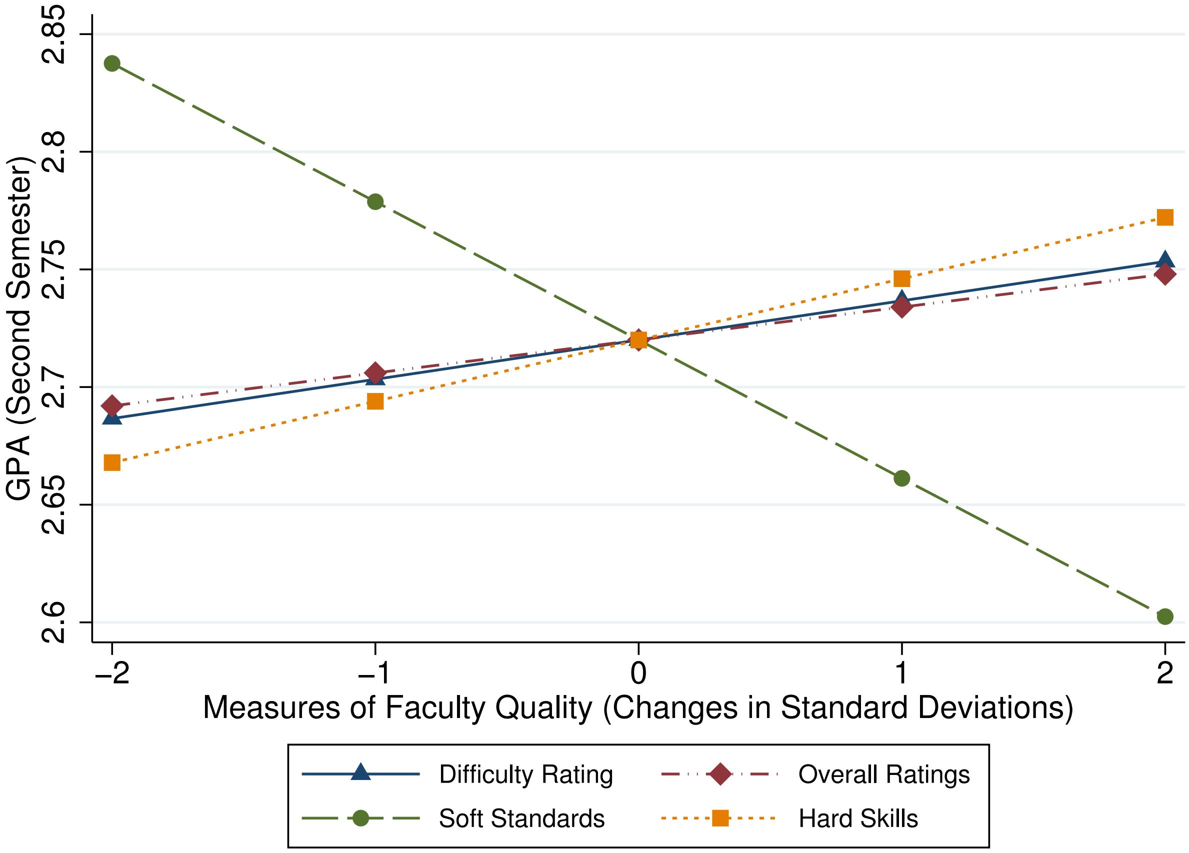
Note: Figure shows the relationship between overall rating and difficulty rating for instructors who teach the first course in a core sequence. Data was drawn from *ratemyprofessors.com*. Ratings for each category are recorded on a 1 (poor) to 5 (excellent) scale.

Figure 2: Soft Standards Instructor Effect, Difficulty Rating, and Overall Rating



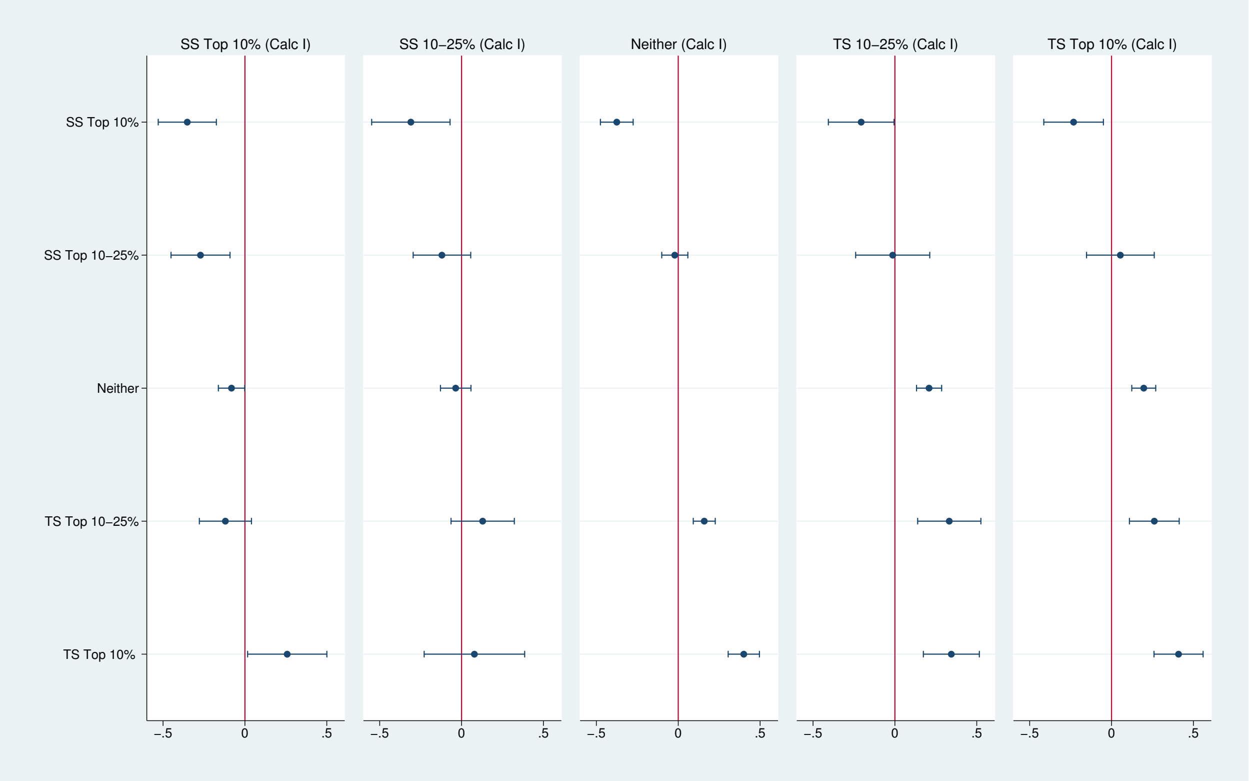
Note: Figure shows the relationship between soft standard effect (Γ_k^c) and *ratemyprofessors.com* ratings (overall rating and difficulty rating).

Figure 3: Counterfactual Effects of Aspects of Teacher Quality on Sequential GPA



Note: Figure shows the predicted effect on second semester GPA from encountering different types of faculty in the first semester. Measured in standard deviation changes in faculty quality based on estimates from Column (4) in Table 4.

Figure 4: Impact of Sequencing of Faculty Standards in Calculus I and II on Calculus III Grades



Note: Figure shows indicator coefficients for different faculty standards sequencing. Columns refer to faculty types in Calculus I, while rows refer to faculty types in Calculus II. The impact of a particular sequence of faculty types is thus given by the column and the row. The categories used are based on the distribution of faculty standards in Calculus I and II. The distribution was split into 5 categories: top 10% of standards distribution (SS), next 15% of standards distribution (SS Top 10-25%), middle 50% of the distirbution, next 15% of the standards distribution (TS top 10-25%), and the bottom 10% of the standards distribution (TS top 10%).

Table 1: Student and Faculty Characteristics - Summary Statistics

	Mean	Standard Deviation	Min	Max
Student Characteristics (N=23,023)				
Pre-College Characteristics:				
SAT Verbal	666.2	71.6	410	800
SAT Math	643.4	76.9	230	800
High School Standing Proxy	546.4	143.2	137	800
Female	0.196	0.398	0	1
Minority	0.266	0.442	0	1
Athlete	0.270	0.444	0	1
Myers-Briggs Personality:				
Extroversion / Introversion	2.95	23.4	-57	51
Sensing / Intuition	7.8	23.8	-51	67
Thinking / Feeling	16.0	22	-43	65
Judging / Perceiving	4.8	25	-61	55
Feeder Source:				
Foundation	0.053	0.225	0	1
NAPS	0.176	0.381	0	1
Boost	0.0005	0.02	0	1
Nuclear	0.015	0.12	0	1
Sequential Course Grades				
Overall Grade (N=77,705)				
First Course	2.83	0.87	0	4
Second Course	2.73	0.95	0	4
Final Exam Grade (N=61,411)				
First Course	2.38	1.15	0	4
Second Course	2.26	1.20	0	4
Faculty Characteristics (first course in sequence)				
Full Sample (N=686)				
Course Grade	2.91	0.84	1	4
Final Exam Grade	2.52	1.14	0	4
Female	0.265	0.441	0	1
Rate My Professor Sample (N=321)				
Number of Ratings	13.2	12.2	1	91
Overall Rating	3.8	1.01	1	5
Difficulty Rating	2.9	0.85	1	5
Course Grade	2.87	0.85	1	4
Final Exam Grade	2.49	1.14	0	4

Notes: Table contains sample statistics for students, course, and faculty-level variables. Data covers all core courses from 1997-2017 at the United States Naval Academy. Faculty-level variables are drawn from USNA administrative data as well as from *ratemyprofessors.com*. SAT scores include converted ACT scores. High school standing proxy is an administratively developed score for academic performance in high school. Myers-Briggs data is collected for each student in their freshman year. The variable measures the intensity of personality along four distinct dimensions. Positive numbers refer to the first category listed and negative numbers refer to the second category listed. For example, -12 for the Sensing / Intuition dimension would refer to an mildly intuitive personality. Feeder source refers to admissions pipelines distinct from the traditional "direct from high school" method. The five sequential course sequences are in English, chemistry, calculus, social sciences (all freshman year), and physics (sophomore year).

Table 2: Randomness Checks

Test	(1) Verbal SAT	(2) Math SAT	(3) HS Rank	(4) E-I	(5) S-N	(6) J-P	(7) T-F
Empirical *p*-values (mean and st.dev.)	0.502 (0.304)	0.506 (0.301)	0.504 (0.299)	0.494 (0.288)	0.493 (0.287)	0.496 (0.291)	0.496 (0.290)
No. of obs.	4747	4747	4747	4747	4747	4747	4747
Kolmogorov-Smirnov test (no.failed/total tests)	0/126	2/126	0/126	0/126	0/126	0/126	0/126
χ^2 goodness-of-fit test (no.failed/total tests)	4/126	3/126	1/126	5/126	5/126	2/126	1/126

Notes: The empirical *p*-value of each section represents the proportion of the 10,000 simulated sections of second-semester courses with values less than that of the actual section. The Kolmogorov-Smirnov and chi-squared goodness of fit test results indicate the number of tests of the uniformity of the distribution of *p*-values that failed at the 5 percent level.

Table 3: Soft Standards and Sequential Spillovers

Current Course - Final Grade	(1)	(2)	(3)	(4)	(5)	(6)
Previous Instructor Γ_k^f	0.0171****	0.0287****	0.0380****	0.0292****	0.0251*	0.0175****
(Hard Skills Channel)	(0.0047)	(0.0056)	(0.0065)	(0.0054)	(0.0131)	(0.0049)
Previous Instructor Γ_k^c		-0.0214****	-0.0191****	-0.0581****	-0.141****	-0.0865****
(Soft Standards Channel)		(0.0051)	(0.0058)	(0.00514)	(0.0126)	(0.0047)
STEM x Prev Inst Hard Skills					0.0066	
					(0.0140)	
STEM X Prev Inst Soft Standards					0.0631****	
					(0.0133)	
Female Student X Prev Inst Hard Skills						0.0188*
						(0.0100)
Female Student X Prev Inst Soft Standards						-0.0285***
						(0.0095)
Standard Errors	Student Cluster	Student Cluster	Student Cluster	Student Cluster	Student Cluster	Student Cluster
Student Characteristics	No	No	No	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes	Yes	Yes
Section x Current Instructor FE	Yes	Yes	Yes	Yes	Yes	Yes
Courses	All Courses w/ Final Exam	All Courses w/ Final Exam	Common Final Only	Common Final Only	All Courses w/ Final Exam	All Courses w/ Final Exam
Observations	60,985	60,985	42,513	37,192	58,059	58,059

Notes: The dependent variable is the final course grade in the second course in the sequence. The primary variable of interest is the estimated instructor effect from the first course in the sequence. Columns (3) and (4) restrict the sample to just those courses with a common final exam (chemistry, calculus, and physics). Student characteristics include student fixed effects and previous course grade. All models control for year, semester, and section x current instructor fixed effects. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.

Table 4: *Ratemyprofessors.com* Ratings and Sequential Spillovers

Current Course - Final Grade	(1)	(2)	(3)	(4)
Previous Instructor - Overall Rating (RMP)	-0.0201****	0.0185****	0.0241****	0.0144**
	(0.00357)	(0.00447)	(0.0056)	(0.0069)
Previous Instructor - Level of Difficulty (RMP)		0.0775****	0.0341****	0.0204*
		(0.00553)	(0.0079)	(0.0107)
Previous Instructor Γ_k^f			0.0152**	0.0261***
(Hard Skills Channel)			(0.0066)	(0.0084)
Previous Instructor Γ_k^c			-0.0909****	-0.0588****
(Soft Standards Channel)			(0.0068)	(0.0081)
Standard Errors	Student Cluster	Student Cluster	Student Cluster	Student Cluster
Student Characteristics	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes
Section x Current Instructor FE	Yes	Yes	Yes	Yes
Courses	All Courses	All Courses	All Courses w/ Final Exam	Common Final Only
Observations	51,122	51,122	38,746	25,661

Notes: The dependent variable is the final course grade in the second course in the sequence. The primary variables of interest are the estimated instructor effects from the first course in the sequence as well as RMP student ratings of faculty diffculty and overall quality. Column (3) restricts to only courses with a final exam (excludes English). Column (4) restricts sample to only those courses with a common final exam (chemistry, calculus, and physics). Student characteristics include student fixed effects and previous course grade. All models control for year, semester, and section x current instructor fixed effects. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.

Table 5: Faculty Groups and Sequential Spillovers

Current Course - Final Grade	(1)	(2)	(3)	(4)	(5)	(6)
Previous Instructor - Top 25% Overall	-0.0299***		-0.0182*	-0.00335		
	(0.0090)		(0.0093)	(0.0112)		
Previous Instructor - Bottom 25% Overall	-0.0530****		-0.0393****	-0.0446****		
	(0.0098)		(0.0095)	(0.0116)		
Previous Instructor - Top 25% Difficulty	0.0815****	0.0805****		0.0109		
	(0.0101)	(0.0103)		(0.0130)		
Previous Instructor - Bottom 25% Difficulty	-0.0709****	-0.0690****		-0.0463****		
	(0.0099)	(0.0099)		(0.0120)		
Previous Instructor - Top 10% Overall		-0.0549****				
		(0.0153)				
Previous Instructor - Top 10% to 25% Overall		-0.0228**				
		(0.0098)				
Previous Instructor - Bottom 10% to 25% Overall		-0.0569****				
		(0.0109)				
Previous Instructor - Bottom 10% Overall		-0.0451****				
		(0.0129)				
Previous Instructor - Top 10% Difficulty			0.0896****			
			(0.0110)			
Previous Instructor - Top 10% to 25% Difficulty			0.0349***			
			(0.0106)			
Previous Instructor - Bottom 10% to 25% Difficulty			-0.0310***			
			(0.0117)			
Previous Instructor - Bottom 10% Difficulty			-0.135****			
			(0.0148)			
Schmope (Low Difficulty / High Overall Rating)					-0.122****	-0.0648****
					(0.0118)	(0.0147)
Low Difficulty / Low Overall Rating					-0.155****	-0.0542
					(0.0340)	(0.0388)
High Difficulty / High Overall Rating					-0.0373	0.0358
					(0.0340)	(0.0477)
High Difficulty/ Low Overall Rating					0.0471****	-0.0155
					(0.00892)	(0.0107)
Previous Instructor Γ_k^f				0.0226****		0.0238****
(Hard Skills Channel)				(0.0065)		(0.0063)
Previous Instructor Γ_k^c				-0.0976****		-0.0953****
(Soft Standards Channel)				(0.0065)		(0.0064)
Standard Errors	Student Cluster	Student Cluster	Student Cluster	Student Cluster	Student Cluster	Student Cluster
Student Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes	Yes	Yes
Section x Current Instructor FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	51,122	51,122	51,122	38,746	51,122	38,746

Notes: The dependent variable is the final course grade in the second course in the sequence. Indicator variables are used to capture different aspects of the distribution of faculty characteristics based on overall rating and difficulty rating. Column (4) and Column (6) restrict the sample to only those courses with a final exam (history, political science, chemistry, calculus, and physics). Columns (5) and (6) use categorical variables for joint qualities in both the difficulty rating and the overall rating. "Low" is defined as bottom 25% of the relevant distribution and "High" is defined as top 25% of the distribution. Student characteristics include student fixed effects and previous course grade. All models control for year, semester, and section x current instructor fixed effects. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.

Table 6: Sequential Learning Over Three Semesters - Calculus

	(1)	(2)	(3)	(4)	(5)
	Calc II	Calc III	Calc III	Calc III	Calc III
Calc I Γ_k^f	0.0439****	0.0362***	0.0374***	0.0364***	0.0373***
(Hard Skills Channel)	(0.0124)	(0.0122)	(0.0120)	(0.0121)	(0.0125)
Calc II Γ_k^f		0.0631****	0.0394***	0.0623****	0.0423****
(Hard Skills Channel)		(0.0127)	(0.0124)	(0.0128)	(0.0125)
Calc I Γ_k^c	-0.155****	-0.0809****			
(Soft Standards Channel)	(0.0124)	(0.0123)			
Calc II Γ_k^c		-0.153****			
(Soft Standards Channel)		(0.0117)			
Soft Standards Faculty Calc I (Top 25%)			-0.0958****		
			(0.026)		
Soft Standards Faculty Calc II (Top 25%)			-0.182****		
			(0.0270)		
Tough Standards Faculty Calc I (Top 25%)			0.165****		
			(0.026)		
Tough Standards Faculty Calc II (Top 25%)			0.170****		
			(0.025)		
Soft Standards Faculty Calc I (Top 10%)				-0.119****	
				(0.033)	
Soft Standards Faculty Calc I (Top 10-25%)				-0.0492	
				(0.037)	
Tough Standards Faculty Calc I (Top 10-25%)				0.163****	
				(0.033)	
Tough Standards Faculty Calc I (Top 10%)				0.158****	
				(0.031)	
Soft Standards Faculty Calc II (Top 10%)				-0.357****	
				(0.039)	
Soft Standards Faculty Calc II (Top 10-25%)				-0.071**	
				(0.033)	
Tough Standards Faculty Calc II (Top 10-25%)				0.124****	
				(0.028)	
Tough Standards Faculty Calc II (Top 10%)				0.309****	
				(0.038)	
Sequence (SS , SS)					-0.250****
					(0.051)
Sequence (TS , TS)					0.318****
					(0.046)
Sequence (SS , TS)					0.0205
					(0.055)
Sequence (TS , SS)					-0.0898
					(0.055)
Sequence (SS , Neither)					-0.0662**
					(0.033)
Sequence (TS , Neither)					0.207****
					(0.030)
Sequence (Neither , SS)					-0.146****
					(0.034)
Sequence (Neither , TS)					0.215****
					(0.030)
Year FE	Yes	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes	Yes
Section FE x Current Instructor FE	Yes	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes	Yes
Student Personality	Yes	Yes	Yes	Yes	Yes
Observations	7,084	7,089	7,089	7,089	7,089

Notes: The dependent variable is the final course grade in either Calculus II (Column 1) or Calculus III (Colums (2)-(5)). Indicator variables are used to capture different aspects of the distribution of faculty characteristics based on estimated instructor effects. Indicator variables for sequential faculty draws based on 25% threshold of soft standards and tough standards. All models control for year, semester, and section x current instructor fixed effects. The sample difference between Column (1) and Colunns (2)-(5) is due to five Calculus II course having only a single section, which is dropped because of the section x instructor fixed effect. Student characteristics include math and verbal SAT scores, high school standing proxy, gender, minority status, athlete, feeder source indicators, and previous course grade. All models control for year, semester, and section x current instructor fixed effects. Standard errors are in parentheses. Significance: * 10 percent; ** 5 percent; *** 1 percent; **** 0.1 percent.

Table 7: Mechanism Robustness

Current Course - Final Grade	(1)	(2)	(3)	(4)	(5)
Previous Instructor Γ_k^f	0.0213****	0.0212****	0.0211****	0.0191***	0.0210****
(Hard Skills Channel)	(0.0045)	(0.0045)	(0.0045)	(0.0060)	(0.0045)
Previous Instructor Γ_k^c	-0.0922****	-0.0943****	-0.0920****	-0.0858****	-0.0885****
(Soft Standards Channel)	(0.0043)	(0.0049)	(0.0043)	(0.0054)	(0.0050)
Extroversion / Introversion X Prev Inst Γ_k^c	-0.0095***				
	(0.0034)				
High Introversion X Prev Inst Γ_k^c		0.0181**			
		(0.0092)			
High Extroversion X Prev Inst Γ_k^c		-0.0055			
		(0.0089)			
Thinking / Feeling X Prev Inst Γ_k^c			0.00645*		
			(0.0034)		
Previous Instructor - Military Officer				0.00614	
				(0.0068)	
Prev Inst Military Officer X Prev Inst Γ_k^f				0.00895	
				(0.0091)	
Prev Inst Military Officer X Prev Inst Γ_k^c				-0.0187**	
				(0.0086)	
Female Student X Prev Inst Γ_k^c					-0.0240***
					(0.0092)
Prev Inst Female					0.0301****
					(0.0085)
Prev Inst Female X Prev Inst Γ_k^c					-0.0024
					(0.0100)
Female Student X Prev Inst Female					-0.0061
					(0.0170)
Female Student X Female Prev Inst X Prev Inst Γ_k^c					0.0248
					(0.0248)
Standard Errors	Student Cluster	Student Cluster	Student Cluster	Student Cluster	Student Cluster
Student Characteristics	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes	Yes
Section FE x Current Instructor FE	Yes	Yes	Yes	Yes	Yes
Observations	58,059	58,059	58,059	56,664	58,041

Notes: The dependent variable is the final course grade in the second course in the sequence. Student characteristics include student fixed effects and previous course grade. All models control for year, semester, and section x current instructor fixed effects. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.