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

Michael A. Insler[‡] Alexander F. McQuoid Ahmed Rahman[†]
Katherine Smith[‡]

June 2022

Abstract

This work disentangles aspects of teacher quality: the extent to which teachers set *standards* and the extent to which teachers impart *knowledge*. We exploit data from a variety of subjects that links students from randomly-assigned instructors in introductory-level courses to the students' performances in follow-on courses. We find that instructors' standards and their ability to transmit knowledge both affect learning, but standards matter more. Using this key finding, we rule out teaching-to-the-test as a central mechanism for understanding teacher value-added. Additional data on student opinions demonstrates that students' perceptions of instructor difficulty partly explain the potency of the standards effect.

JEL classification: I20; I21; I23; J24

Keywords: Soft Standards; Sequential Learning; Teacher Value-added; Rate My Professor; Higher Education

---

*We would like to thank three anonymous referees, participants at Kent State, EALE-SOLE-AASLE 2020, and WALES 2022 for comments and feedback. The views expressed in this paper are those of the authors and should not be attributed to the Department of Defense nor the United States Naval Academy. All errors are our own.

[†]Economics Department, College of Business, Lehigh University, 621 Taylor Street, Bethlehem, PA 18015

[‡]Economics Department, United States Naval Academy, 572M Holloway Road, Annapolis, MD 21402

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 why they are effective. Indeed teachers can affect students in a myriad of ways, complicating studies of teacher value-added. In this work we analyze, and decouple from each other, two channels: the extent to which teachers set *standards*, and the extent to which they impart *knowledge*. We explore these channels in a sequential learning framework where instructor treatment in an initial period influences performance in the follow-on course in the sequence. More specifically, we decompose the value-added of each first-semester instructor into standards and knowledge 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 tend to impact longer-run learning, the standards component generates the larger impact.

This work has important policy implications for higher education. In particular, college and university administrators should be able to utilize our findings to help guide policy regarding grade inflation, the proper evaluation of effective teaching, and optimal placement of instructors based on teaching styles and sequential content of courses.

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 "core" sequential courses, which includes a variety of 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 a teacher-assigned overall course grade along with an externally benchmarked final exam grade, allowing us to separately estimate multiple aspects of faculty value-added. There are both objective and subjective components of any academic performance measure. In our setting, for example, individual instructors maintain some latitude over their assignment of final course grades. For courses in our sample, however, final exams are created, administered, and graded by a committee of faculty who do not directly influence the final course grade. We proxy the *standards channel* of teacher value-added with the portion of the course grade that is subjectively controlled by the instructor, and we proxy the *knowledge channel* of teacher value-added with the final exam grade which is not directly controlled by the instructor.

In describing these channels, we adopt some additional terminology: Teachers can produce "soft standards" or "tough standards." Here we consider the standards teachers set in terms of what it requires to achieve a given grade. Those who set tough standards demand a higher level of performance for a given grade, resulting in a deflated grade distribution all else equal. Teachers who set soft standards demand less from students for a given grade, leading to grade inflation. Disentangling standards and achievement contemporaneously is challenging, but the difference shows up clearly in a sequential learning environment such as ours.

Likewise, for the knowledge channel, teachers can also produce "shallow knowledge" or "deep knowledge." Shallow knowledge conveys the idea that even those students who perform well on the final exam of the first course may not perform better in the second course, due to the instructor "teaching to the test." On the other hand, deep knowledge means that students who perform well on the final exam of the first course also perform better in the second course, suggesting applicable knowledge was transmitted and retained. This channel can only be observed with the two-course sequence; that is, we need to connect the first course final exam performance with the second course final grade performance.

We find that soft standards are bad for follow-on course performance. Importantly, while holding teacher standards constant, we estimate a *positive* effect for the knowledge channel. In other words, teachers who generate higher final exam grades in the first course impart superior performance in the follow-on course, holding the subjective component of their course grading constant. Thus in our terminology, deep knowledge generation dominates shallow knowledge, leading us to reject the teaching-to-the-test hypothesis that previous studies have cited to explain the importance of standards within the teacher value-added literature (e.g., Carrell and West (2010) and Braga et al. (2014))

To better understand how soft standards and deep knowledge impact student learning, we examine their potency across student types. While soft standards tend to harm all students in subsequent courses, deep knowledge affects students differently, based on their first course grades. Higher achieving students benefit from deep knowledge in subsequent courses, but lower achieving students actually do worse in follow-on courses when they encounter deep-knowledge-promoting faculty. The rising tide of the standards and knowledge channels lifts most boats—but some boats do sink.

Given that the soft standards channel is particularly potent, we merge data on student opinions of faculty from *ratemyprofessors.com* (RMP) to corroborate this aspect of teacher influence. 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,[1] 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 demand more effort for a given desired academic grade.

We find that instructors with higher overall ratings more severely harm sequential academic success. 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 positive effect on sequential learning than overall ratings. These findings suggest that one aspect of raising standards in the classroom may simply be raising student *perceptions* of the difficulty-level of the subject.

[1]The overall ratings in RMP tend to correlate highly with university-administered opinion forms (see for example Brown et al. (2009) and Timmerman (2008)).

Exploring the RMP data more deeply, we show there are two types of particularly harmful faculty. One type is easily identified: 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 harmful 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 standards 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: Establishing tough standards in the classroom may be penalized, while improving students' durable human capital is often unnoticed or unappreciated by the students themselves (Weinberg et al. (2009)).

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 standards and knowledge transmission, as well as to see whether sequencing of faculty types matters. We find that both 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 standards channel decays over time, the knowledge effect is more persistent.

Finally, our two-channel value-added framework provides novel insights on heterogeneous effects across both student and faculty gender. While the standards faculty set are an important part of how information is conveyed to students, the filtering of information through gender dynamics may add an additional signal extraction problem. We find for female students and/or female professors that the impact of standards channel is amplified.

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.

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 student outcomes, or are instructors merely spoon-feeding material or "teaching to the test?"

Some research measures teacher value-added using indirect outcomes like student evaluations. These evaluations focus on the students' self-professed amount of material learned, as well as students' overall rating of the instructor (Linask and Monks (2018)). While they can be a useful value-added measure, Goos and Salomons (2017) point out, however, that student evaluations in general generate biased measures of value-added.

Another approach uses alternative measures of teacher value-added, including non-cognitive skills proxied by absences, suspensions, and grade repetition (Jackson (2018)). Teaching, in other words, can have multi-dimensional effects that need to be better understood (Blazar and Kraft (2017) and Coenen et al. (2018)).

Given that faculty typically choose grades, instructor value-added as measured by contemporaneous grades distributed by the same instructors may be inflated for various reasons (Palali et al.

(2018)). To avoid this, there is a large literature that measures teaching quality in educational production 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), Hoffmann and Oreopoulos (2009), Figlio et al. (2015), and Xiaotao and Xu (2019)).

With regards to causal interpretation, there are many potential challenges, such as student sorting towards preferred instructors. This is tackled effectively by Chetty et al. (2014a) for students in grades 3-8, 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 or are they simply effective at teaching to the test. The existing 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 allow us to isolate multiple 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.

While previous work (Chetty et al. (2014b), Carrell and West (2010), and Araujo et al. (2016))

identifies that teachers matter, we are able to explain *why* teachers matter. By taking advantage of a course grade (a locally-benchmarked, subjective measure) and a final exam grade (externally-benchmarked, objective measure) we are able to distinguish professors' knowledge transmission (shallow/deep) from their standards (soft/tough). This feature allows us to challenge the assertion that the deleterious soft standards effect observed in previous work stems from a teaching-to-the-test mechanism. Additionally, our usage of RMP data enables us to more specifically classify instructor types and their disparate impacts on student learning.

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, 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.[2] 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 a diverse

[2]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).

student body from across the nation in an attempt to generate an officer corps that is reflective of the wider naval force (Kotlikoff et al. (2022)). 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. Students with strong opinions, either positive or negative, may be more likely to take the time to voluntarily respond.

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 and large positive correlations with institutionally based SETs (Timmerman (2008)).

Second, do teaching evaluations capture quality or easiness of faculty? In general, more lenient instructors receive better 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 course registration 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 generally 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 final course grade and final exam 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).

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.[3] All freshmen must pass a set of 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 freshman 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 course selection 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 sequence), 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 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 are uniformly distributed. We conduct 882 distinct tests (six different second-semester

[3]The course schedules for first-semester freshmen are determined unilaterally by the registrar. USNA-based sources for the information in this section are discussed in Brady et al. (2017).

courses across 21 academic years for each of the seven observable characteristics), using both a 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-values 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

We supplement our data using 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. Information collected on the site has varied over time. At present, a new rating of a professor asks for an overall rating, a difficulty rating, whether the student would take the professor again, whether the class was taken for credit, whether attendance was mandatory, what grade was received, whether the course was an online course, and the selection of three tags chosen from a predetermined list. In addition, students have the opportunity to add additional detailed comments about the course or the professor. For the present study, we rely on the overall and difficulty ratings as these have been continuously asked over the time period under consideration.

Out of 679 separate instructors who teach within the core courses of interest, 363 have RMP profiles, with an average number of student opinions per instructor of 13.2.[4] For observable characteristics, our RMP subsample resembles the overall sample in terms of course grades, final exam grades, and percentage of faculty that are female. Unsurprisingly, there is a high negative cor-

[4]Note that faculty at USNA typically teach a range of introductory, intermediate, and upper-level courses. We use the RMP scores for all courses for each instructor, thus capturing an instructor effect rather than an instructor-course effect. Our baseline estimates incorporate instructors' 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.

relation between overall recommendation scores and level of difficulty scores (with correlation coefficient of -0.6). Summary statistics for RMP data are reported in Table 1.

4 Econometric Approach

To identify aspects of faculty value-added, 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.[5] While the latter 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.[6] 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 valued-added 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:

$$Y^{1,c}_{ijks} = \Gamma^c_k + \theta^{1,c} X^1_i + \gamma_{year} + \gamma_{sem} + \gamma_s + \epsilon_{ijks} \tag{1}$$

$$Y^{1,f}_{ijks} = \Gamma^f_k + \theta^{1,f} X^1_i + \gamma_{year} + \gamma_{sem} + \gamma_s + \epsilon_{ijks} \tag{2}$$

[5] We exclude English from the analysis below because the sequence does not have required final exams, which we would need to disentangle the standards and knowledge channels. Results for the standards channel are similar when English is included, however.

[6] 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.

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. $Y^{1,f}_{ijks}$ represents the final exam grade for the same $(ijks)$-quadruplet. To account for possible cohort bias, Γ^c_k and Γ^f_k computations exclude information from the present year.[7] 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 includes information on student personalities using Myers-Briggs personality types. Fixed effects for section, semester, and year are also included.

Final exams are common for all students regardless of the instructor. Exams are written by course coordinators and are graded by external committees. Instructors have 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. On the other hand, faculty retain some control over the final course grade. They can influence the weight that the common final exam has in the course grade computation, and they maintain discretion over the assignments that make up the other portions of the grading scheme.

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^c_k and Γ^f_k, 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. Γ^c_k captures the effect stemming from the standards channel, and Γ^f_k captures the effect stemming from the knowledge 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:

[7] For simplicity, we avoid using the more accurate notation of $\Gamma^c_{k,-year}$ and $\Gamma^f_{k,-year}$ to denote the process of leaving out the current year. As discussed in Chetty et al. (2014b) and Araujo et al. (2016), this adjustment enables us to separate out classroom from teacher effects.

$$Y^{2,c}_{ijkls} = \alpha\Gamma^c_k + \beta\Gamma^f_k + \theta^2 X^2_i + \delta_{sem} + \delta_{year} + \delta_l + \delta_s + \eta_{ijkls} \quad (3)$$

Y^2_{ijkls} 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^2_i includes student-specific demographic information discussed above (or alternatively a student fixed effect). A battery of fixed effects for year, semester, section, and current instructor are also included.

Our primary interest is on the impact of the previous instructor in the sequence on follow-on course grades. The standards channel is captured through α while the knowledge channel is captured through β. Later in the analysis, we expand 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

The first three columns of Table 3 focus on the standards channel. In Column (1) we estimate α in a baseline specification that includes common fixed factors. The estimated coefficient suggests that students in courses taught by faculty who tend to give higher subjective grades hurt student performance in follow-on courses. This negative effect is what we refer to as the soft standards channel: Instructors who tend to distribute higher grades are signaling to students, inadvertently or otherwise, that sub-standard work within their discipline is sufficient. The consequences of this effect are observed in diminished academic achievement in the follow-on course. Columns (2) and (3) include student characteristics and student fixed effects, respectively. In all three estimates, we find similar negative and statistically significant effects from this measure of teacher quality.

In columns (4)-(6) we include our second measure of instructor value-added based on final exam grades. Although the standards channel continues to be negative and statistically significant, instructor effects based on more objective final exam grades are positively related to sequential learning, that is a deep knowledge effect. Importantly, even when controlling for measures of more objective teaching effectiveness (i.e., deep knowledge), the soft standards effect remains negative and significant.

Are instructors teaching to the test? They may be spoon-feeding course material, allowing students to succeed in that particular course, while failing to provide instruction that facilitates true proficiency. Table 3 findings suggest this is not the primary mechanism at work. If teaching-to-the-test was the mechanism behind the negative sign of α, then the inclusion of Γ_k^f in columns (4)-(6) would result in a negative estimate for β, and a corresponding reduction in the importance of α. If teaching-to-the-test is a prominent mechanism, faculty would be raising final exam grades at the *expense* of future learning. Instead, we find that faculty who raise final exam scores on average also improve performance in sequential courses. The fact that the negative estimate of α is statistically unchanged suggests that two distinct channels of faculty influence—neither of which is captured in the idea of teaching-to-the-test—are at play.

If teaching-to-the-test is not the culprit, what else could be driving these initial findings? To explore further we look at the potential heterogeneous effects of the soft standards and deep knowledge channels across different kinds of students. As a starting point, we examine how the standards and knowledge channels may affect students differently, based on how they performed in the first course of the sequence.

Results incorporating interactions between first-course student grades and teacher value-added effects are presented in Table 4. Strikingly, in column (1), we observe that the negative effects of soft standards are large and robust for each category of student: Students who received an A-level grade in the first course suffered just as much in the second course from soft standards as those who received a D-level grade in the first course. In these interaction models, soft standards exposure is universally bad for subsequent performance.

In column (2) of Table 4, the heterogeneous response across student types is more complicated. Interestingly, we observe that the positive estimate for the knowledge channel is isolated to A- and B-level students. The effect for A-level students is roughly three times larger than the aggregate estimate (from Table 3) , and it is roughly two times larger for B-level students. Perhaps surprisingly, C-level students are actually harmed by deep knowledge value-added instructors.[8] Results in column (3), which includes heterogeneous effects for both channels of teaching, confirm the observed pattern.

While soft standards harm all types of students, faculty who teach deep knowledge harm weaker academic students while raising the academic performance of students at the upper end of the distribution. The fundamental trade-off with instructors who truly teach challenging material and improve students' subsequent performance on average is that some students at the lower end of the distribution are harmed. In this case, the rising tide of high-quality teaching lifts most boats, even as some boats sink.

Our findings here bring forward two key tensions for administrators. First, administrators want happy students who are able and willing to pay rising (and high) price tags for college education. In addition, administrators want graduates who have developed significant human capital during their educational experience. When consumption and investment objectives are in conflict with each other, hard choices must be made about the relative balance of these two forces.

Second, administrators have to weigh the benefits and costs to different types of students when incentivizing educational approaches. Do administrators want faculty who challenge and push students towards higher achievement, knowing that some students will fall to the wayside as a result? Or would administrators prefer to hold all students back a little rather than risking harm to a few? The results here suggest that real trade-offs exist and administrators must be aware of the relative costs.

There is reason, however, to suspect that administrators may not have complete information to optimally evaluate these trade-offs. The opportunity costs of forgoing high valued-added teaching are likely to be diffuse and hard to measure, as students are typically unaware of the true costs of not learning as much as they could have under more demanding instruction. But the opportunity

[8] The interpretation here is important. Students who received a C grade and had a high deep knowledge value-added instructor in the first course do worse in the follow-on course than similar C-level students who had teachers who were less effective at deep knowledge transmission. One reasonable explanation for this somewhat counter-intuitive result is a kind of "discouragement effect," in which weaker students reduce inputs to learning—such as studying hours or effort—upon encountering a high deep knowledge value-added instructor.

costs of high value-added teaching are likely to be local and easily measured: Disgruntled students have no qualms about criticizing faculty they do not like in student opinion forms. These squeaky-wheel students may be well-known to administrators as a result, while the foregone learning loss to higher ability students has no such obvious constituency.

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 deep knowledge channel constant*. This 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. 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.

As we delve deeper into understanding these two channels of instructor influence, our preferred specification going forward is found in column (4) of Table 4. Here we economize on information by using previous course grade in lieu of the large set of interaction terms. The coefficient estimates for α and β reflect aggregate effects for the standards and knowledge channels, respectively. How large are these effects? Across freshman grades, one standard deviation is roughly one GPA point. Given that our teacher value-added measures are normally distributed, the point estimate implies that going from the 25th percentile to the 75th percentile in terms of a soft standards teacher (i.e., going from a "modestly tough" to a "modestly soft" standards teacher) lowers student performance by roughly 0.1 standard deviations in the follow-on course.

By comparison, Hanushek et al. (2012) review a wide range of papers measuring value-added for K-12 teachers in math and reading, and find that on average going from the 25th percentile to the 75th percentile of the quality distribution would mean a difference in performance of roughly 0.2 standard deviations for that year. While smaller in magnitude, our value-added estimate is reasonably close to this broad-based finding from the literature, and it is not surprising that we observe a smaller effect in the collegiate setting, where the educational growth of students is likely more anchored. In a similar university setting, Carrell and West (2010) estimate a value-added effect of

0.05 standard deviations using the same counterfactual thought experiment, somewhat smaller than our standards channel estimate here. Our knowledge channel is smaller than existing value-added studies, at least in the aggregate estimate seen in column (4) of Table 4. As we highlight, however, this is due to the heterogeneous impacts of deep knowledge: For A-level students, going from 25th to 75th percentile raises one's grade by 0.07 standard deviations, an estimate which is close to the magnitude of our standards channel estimate.

5.1 Student Perceptions of Faculty and Sequential Learning

To better understand the signals being sent by faculty, we next turn to student perceptions of faculty quality. 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. Using data from *ratemyprofessors.com*, we draw on two distinct measures of faculty as perceived by students: an overall rating; and a difficulty rating. We further use the RMP data to build profiles of perceived instructor traits that are most harmful and helpful to students' sequential learning.

In column (1) of Table 5, we start with an alternative measure of faculty effectiveness using the overall RMP rating of the faculty member who taught the initial course in the sequence.[9] 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 reduce achievement 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.

Student opinion is likely to be influenced in large part by effort demanded by faculty, which is captured in the RMP data through a measure of instructor difficulty. A stark visualization of the

[9]The results do not appear to be influenced by the number of RMP respondents. To confirm this, we considered subsamples where we restricted observations to faculty with different thresholds of RMP ratings, and weighted regressions by the number of ratings. The results were qualitatively and quantitatively unchanged (results available upon request).

negative relationship between these two measures is shown in Figure 1. While some ineffective faculty 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 nearly five 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 tends to identify as excellent those who in reality harm rather than help subsequent learning.

We next move to include both measures of RMP faculty quality along with our grade-based measures of faculty effectiveness (Γ_k^c and Γ_k^f). In order to do so, we need to limit our analysis to the subset of data for which all four measures of faculty effectiveness are available. Column (3) replicates the results from column (2), now using this subset; results do not change. Student perceptions of faculty effectiveness may not perfectly align with empirically-derived faculty value-added. Figure 2 shows the relationship between the instructor standards channel (Γ_k^c) and RMP student ratings of faculty. Consistent with our results so far, softer instructor standards are positively correlated with the RMP overall rating and negatively correlated with the RMP difficulty rating.

In columns (4) and (5), we include our grade-based measures. The overall rating and the level of difficulty continue to be highly significant, as does our measure of soft standards, while the knowledge channel provides no further information about subsequent student performance. The inclusion of grade-based measures of faculty value-added reduces the impact of the difficulty rating, confirming the relationship between Γ_k^c and student perceptions of difficulty, but also suggests that the cost of soft standards is not *fully* reflected in student opinions of course experiences. Students do not completely internalize the consequences of the standards set by faculty: The student-perceived leniency of teachers and the empirically-measured leniency of teachers each harm student performance.

We next consider heterogeneous effects in the context of the RMP data. In column (1) of Table 6, we interact the overall rating with previous course grade bins (without controlling for RMP difficulty rating at this point). The estimated negative effect found in Table 5 is observed for each of the grade bins, but once we include interactions with the difficulty rating (column (2)), we find that the aggregate positive coefficient found in Table 5 is driven by higher achieving students. The difficulty interaction terms are positive for most grade bins, except for D students, where we find no effect. The difficulty effects are particularly large for students at the top of the grade distribution. Columns (3) and (4) include the standards and knowledge channels; heterogeneous effects' estimates are robust to these inclusions. Importantly, the standards channel continues to emerge as statistically significant even after accounting for differential responses across student types and student perceptions of faculty quality.

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 "student-driven" standards (via the RMP difficulty measure) and "instructor-driven" standards (via the standards 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. Figures 3 and 4 provide illustrated representations of the impact of these four aspects of faculty quality on sequential GPA.

5.2 Non-linear Impacts of Faculty Types

In Table 7, we expand on the results by allowing for non-linear impacts of different types of faculty based on categorizations related to ratings. Table 5 suggests that difficulty and overall ratings are providing useful information about what faculty are doing in the classroom, while Table

6 indicates that 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) of Table 7, 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 the highest overall ratings are driven primarily by instructor ease: Faculty who are rated highly overall are those who are most lax on students, do not induce sufficient effort, and thus negatively impact future learning. On the other hand, the lowest rated faculty may just be ineffective teachers, who similarly fail to support future learning. Focusing on the upper and lower quartiles allows us to potentially 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. 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 the most popular instructors appear to on average harm their students' achievement in follow-on courses, even upon controlling for student-perceived instructor difficulty.

Further, to compare the non-linear groups with our empirically-derived measures of teacher quality, we include Γ_k^c and Γ_k^f (column (2)). For all four measures of RMP difficulty and overall ratings, the estimates are attenuated compared to column (1), with the top overall rating group no longer statistically significant. The significance of the bottom 25% is maintained, however, for both ratings, even after accounting for the empirically-derived channels of faculty effectiveness.

5.3 High Quality and Low Quality Faculty: Beware the Schmopes

Taken together, the results in columns (1) and (2) 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 (3) of Table 7. Faculty that are both well-liked and considered very difficult (top 25% in both RMP rating systems) 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 our 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 ("Top 25% Difficulty, Bottom 25% Overall;" "Bottom 25% Difficulty, Bottom 25% Overall;" "Bottom 25% Difficulty, Top 25% Overall"), however, are all statistically different from the excluded group of faculty (who do not fit into any of the four classifications).

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 ineffective teachers are perceived to be difficult by students because of lack of clarity or disorganization, and this experience pushes students to take the initiative in their own learning, which carries 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 greater achievement in the next course, but also engenders animosity towards the professor, resulting in high difficulty and low overall ratings.

Still examining Table 7 column (3), faculty who are considered very easy and poor overall do notable damage to sequential learning. These faculty are in fact likely to be broadly ineffective teachers who perhaps minimize effort themselves, resulting in a lackluster educational experience. 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.

In addition, a particularly problematic group is 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). We dub such a faculty member a *Schmope*: a "Seemingly Conscientious and Hardworking Mentor, an Otherwise Perfunctory Educator."[10] *Schmopes* may be problematic because they damage student learning but are likely elevated within the university system as role model faculty due to positive student feedback.

The perniciousness of this type of faculty is confirmed in column (4). Even after including grade-based measures of faculty value-added (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 effectiveness.

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 than student happiness, our results give the normative suggestion that *Schmopes* pose problems in university settings. In light of the positive regard students hold for these faculty, there is perhaps some role for them within the teaching component of academia. Terminal courses or those with minimal sequential content 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. This novel finding is an important consideration for university administrators.

5.4 Persistence in the Standards and Knowledge Channels

Next we exploit an additional unique feature of our setting: a third semester in the mathematics sequence. This third course is not quite as homogeneous as the course sequences described earlier, as students can choose from one of two options: a calculus course focused on vector fields (chosen

[10] 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."

by roughly 60 percent of students) or a calculus course focused more on optimization. At this point students have chosen their majors, and which course students enroll in will be dictated in part by major specialty.[11] Nonetheless, all sophomores must take a third math course, and they remain limited in their ability to select specific faculty.

Following the methodology above, we estimate a standards effect and a knowledge effect for faculty in both Calculus I and Calculus II (via appropriate analogs to equations (1) and (2)). We then explore how these faculty effects evolve as they impact the third semester of calculus. Results are shown in Table 8.

Standards and knowledge channels exist in both sequential transitions. This can be seen in the top set of coefficients in Table 8. The effects of the standards channel from the first to the third course (see column (2)) are very comparable to the effects from the first to the second course (see column (1)). This is also true for the knowledge channel. Further, both channels *persist* from the first to the third courses in the sequence, with some attenuation.

Next we examine whether the specific sequencing of instructor types leads to heterogeneous effects on student outcomes. This is motivated in part by the results shown in Column (3), which suggests that soft standards and tough standards persist at different rates over time. The effect of encountering a tough standards (Top 25%) professor in Calculus I has a similar effect on Calculus III achievement as encountering a tough standards (Top 25%) professor in Calculus II. On the other hand, encountering a soft standards (Top 25%) professor in Calculus I has less of an effect on achievement in Calculus III than encountering a soft standards (Top 25%) professor in Calculus II.

To explore this further, we estimate explicit sequencing, with results presented in column (4) and in Figure 5. If one is assigned a soft standards instructor first semester and a "neutral" standards instructor second semester, their performance will continue to suffer during the third semester. And if one is assigned a tough standards instructor first semester and a neutral standards instructor second semester, their performance will continue to benefit during the third semester. Column (4)

[11] Encountering a soft standards 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)).

reports estimates for all sequences using this classification. Thus despite evidence of persistence, we do find that standards can be reversed, but only when students encounter a notably different type of standard. Figure 5 shows similar results using a more refined classification.

5.5 Gender and Heterogeneous Effects

To conclude the analysis, we consider one important additional characteristic of students and faculty that has been closely examined within the value-added literature: gender. For example, Carrell et al. (2010) find that professor gender effects can perpetuate the gender gap in student achievement. In our setting, the existence of such heterogeneous effects could help to explain why the standards channel is so impactful.

The influences of teachers beyond just information transmission may differ across genders. For example, Rask and Tiefenthaler (2008) argue that women are more responsive to grade signals than men. We consider this thesis by interacting student gender and measures of faculty value-added (Table 9). Female students' grades are more negatively impacted by soft standard instructors than male students', although knowledge transmission doesn't vary by student gender (see column (1)). This gives us a new look into the potential differences between the learning behaviors of males and females (Giuliano (2020)). For example, females may respond differently to implicit teacher signals in ways that affect their longer-run human capital development (Goldin (2013)).

Column (2) of Table 9 focuses on female instructors. Here we observe that female instructors are somewhat more effective on average than their male counterparts in improving student performance in follow-on courses. We also see however that this positive effect is counteracted if the instructor has soft standards.[12] We next consider gender interactions for both students and faculty with teacher quality in column (3). Concerning the standards channel, it appears that cross-gender interactions are more detrimental to longer-term learning. It also appears that stu-

[12] One possible conjecture explaining these findings is that in male-centric educational environments, females exhibiting softer standards create even stronger (misleading) signals that the subject is easy. Teacher gender stereotypes can differ between male-dominated environments (such as in higher education) and female-dominated ones (such as in primary education) (Giuliano (2020)).

dent grade responses are magnified with female instructors: Softer standards produce worse grade outcomes (particularly among male students) and deeper knowledge produces better grade outcomes (again particularly among male students). One conclusion may be that female instructors in male-dominated environments need to set even higher standards than their male counterparts to produce better outcomes for their pupils.

Finally, column (4) provides the summary result for this section: The male instructor-male student "norm" at the Naval Academy (typical of many institutions of higher and technical education) is estimated to generate the least harmful effects from soft standards. Well-intentioned administrators seeking to create a more inclusive learning environment should be aware that policies incentivizing soft standards may disproportionately harm female students.

6 Conclusion

Student performance in sequential courses is impacted by the performance 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 two different phases during the course, we identify two distinct channels through which teacher quality matters: a knowledge channel and a standards channel.

Both channels are important for understanding the ways in which faculty have persistent influence on students' academic performance, but we find that the standards channel is particularly impactful. Soft instructor standards harm students in all grade levels. Knowledge transmission effects are smaller; they benefit higher-achieving students but can actually hurt C-level students. Combining these results demonstrates the striking finding that a rising tide of instructor standards raises all boats although some can sink under the pressure of more in-depth learning and knowledge accumulation.

To explore these channels more deeply, we merge our institutional data with student opinions of faculty from *ratemyprofessors.com*. We find that faculty difficulty rating is the crucial characteristic. The overall faculty rating can be misleading as it fails to account for the negative correlation with student-perceived difficulty. Difficulty matters for longer term learning as it induces effort, which students tend to dislike as effort is costly. Students' perceptions of overall quality heavily weights the up-front effort cost at the expense of longer-run educational attainment.

Using a three-semester sequence in calculus, we consider the persistence and sequential timing of faculty types on student performance. We find that both the knowledge and standards channels persist over time, though soft standards effects can be reversed as students encounter more demanding faculty standards in subsequent courses. We also find that female students tend to respond more strongly to instructor leniency, and this effect can be amplified under the instruction of female professors.

Importantly, we identify two types of faculty that are particularly damaging. One is easy to identify using traditional measures of student opinions: faculty who are perceived as easy and ineffective. 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.

What do we want higher education to achieve? Undergraduate education involves the inculcation of a variety of skills (such as good study habits), an appreciation for the time necessary to do the work, an ability to work with others, as well as 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 beyond college.

References

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., and Schady, N. (2016). Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics*, 131(3):1415–1453.

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.

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

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.

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.

Giuliano, P. (2020). Gender and culture. *Oxford Review of Economic Policy*, 36(4):944–961.

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

Goldin, C. (2013). Notes on women and the undergraduate economics major. *CSWEP Newsletter*, (Summer):4–6, 15.

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.

Hanushek, E. A., Rivkin, S. G., et al. (2012). The distribution of teacher quality and implications for policy. *Annual review of economics*, 4(1):131–157.

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.

Kotlikoff, P., Rahman, A. S., and Smith, K. A. (2022). Minding the gap: academic outcomes from pre-college programs. *Education Economics*, 30(1):3–28.

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.

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.

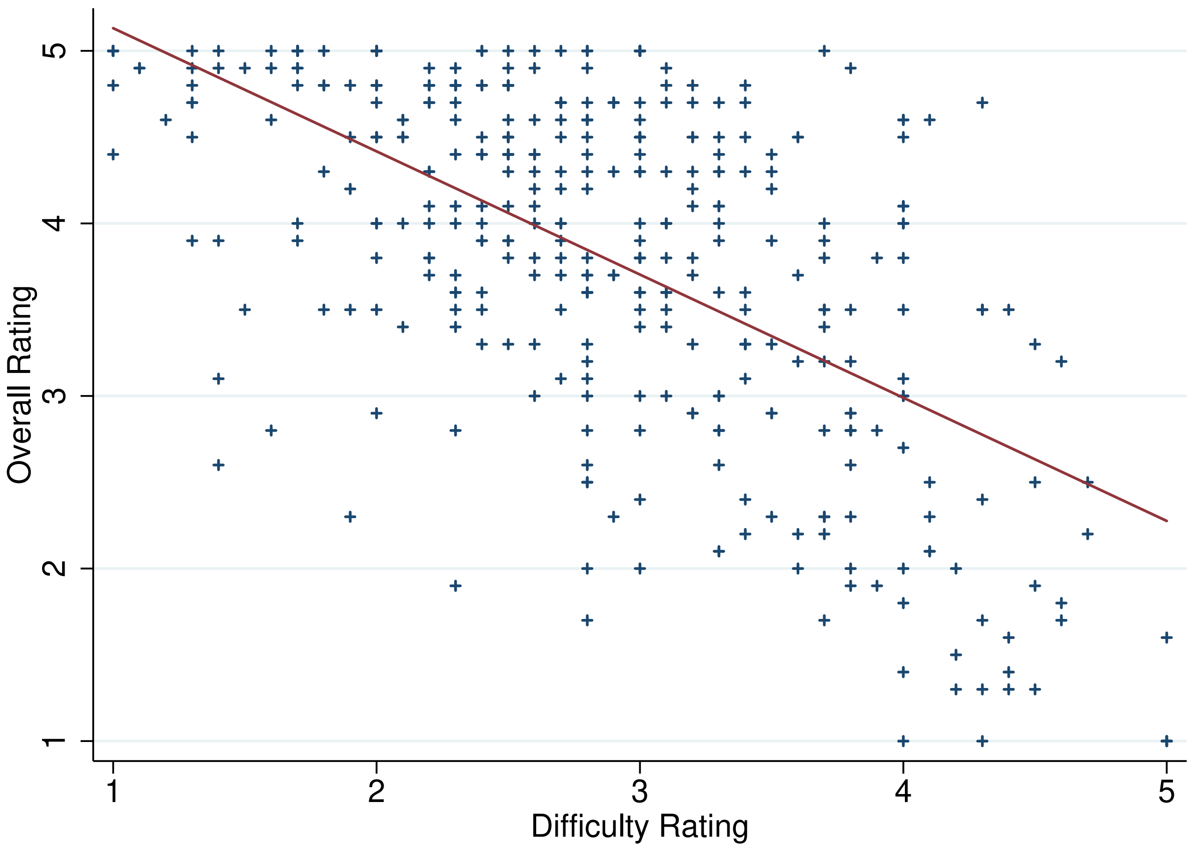
Timmerman, T. (2008). On the validity of ratemyprofessors. com. *Journal of Education for Business*, 84(1):55–61.

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.

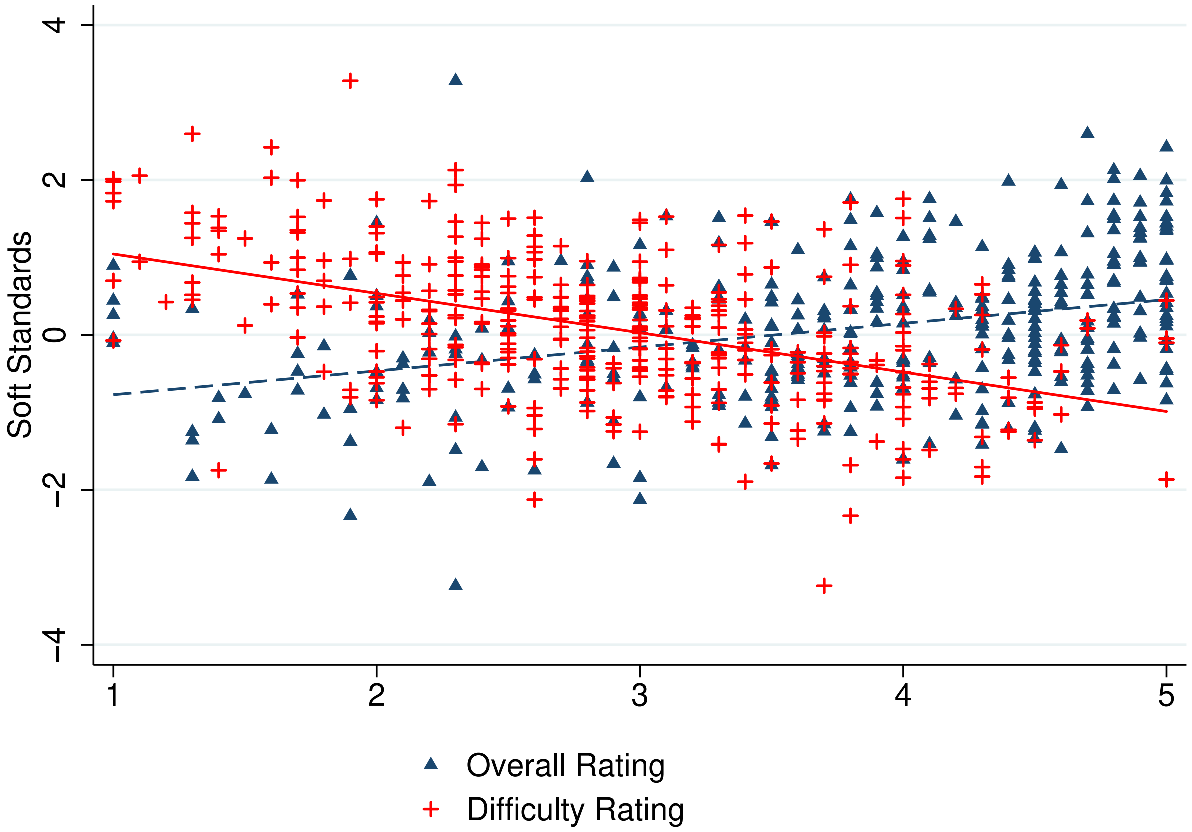
Figures

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



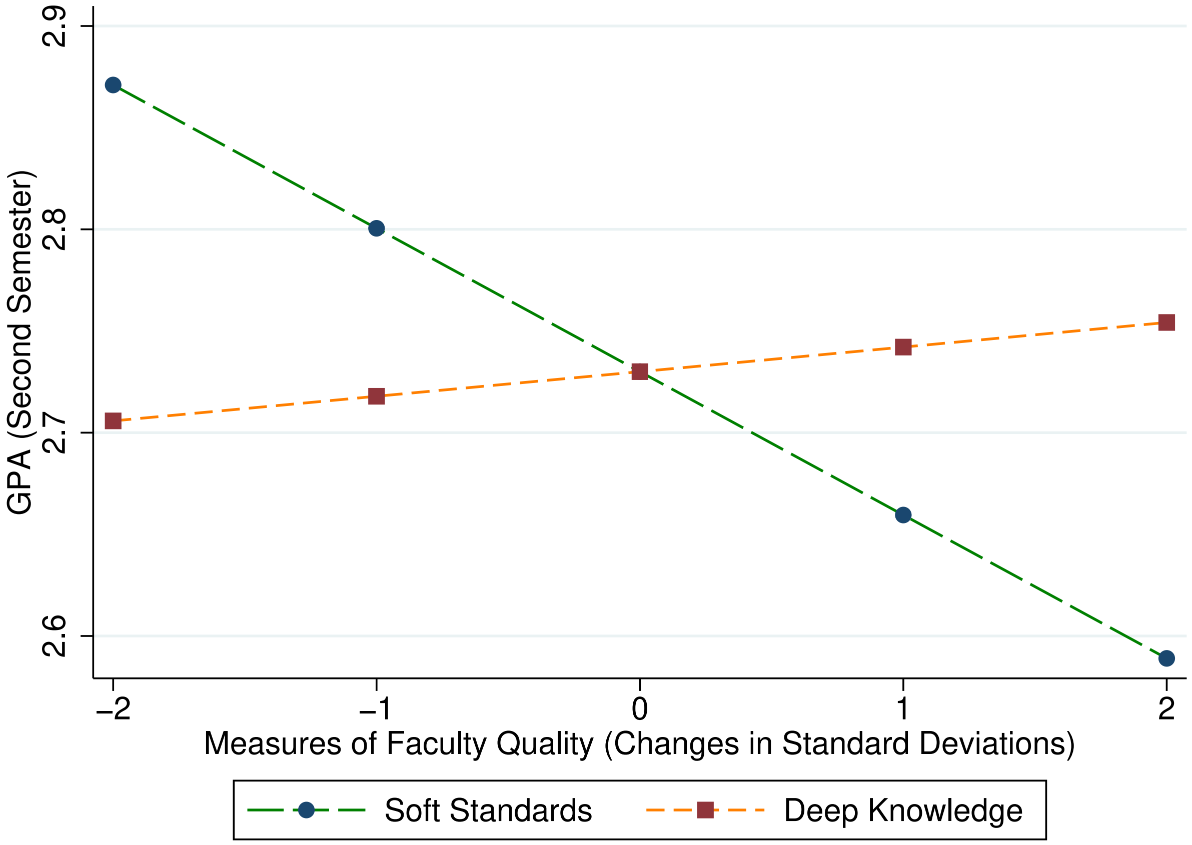
Note: Figure shows the relationship between overall rating and difficulty rating for instructors who teach the first course in a core sequence. Data was drawn from *ratemyprofessors.com*. Ratings for these two categories are recorded on a 1 (poor) to 5 (excellent) scale.

Figure 2: Instructor Standards Effect (Γ_k^c), Difficulty Rating, and Overall Rating



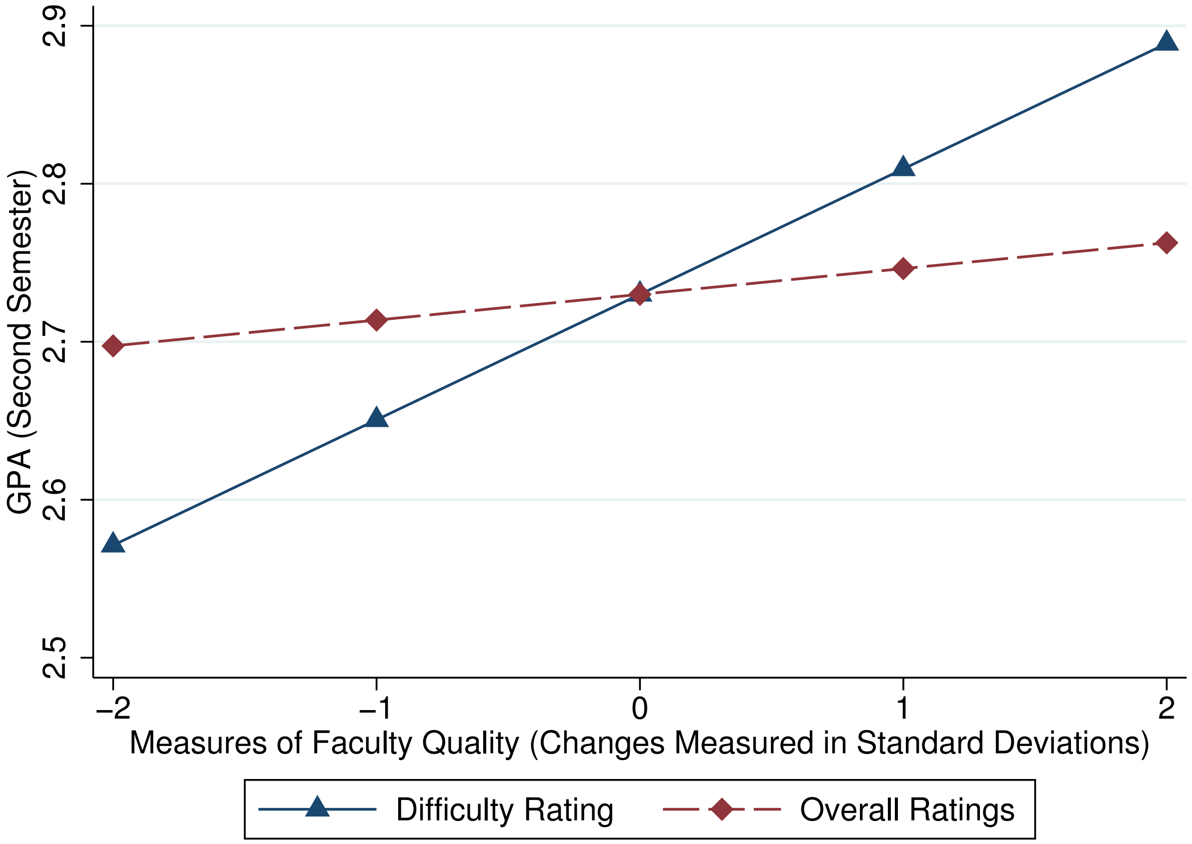
Note: Figure shows the relationship between the standards effect (Γ_k^c) and *ratemyprofessors.com* ratings (overall rating and difficulty rating). Moving vertically along the y-axis is associated with softer instructor standards.

Figure 3: Counterfactual Effects of Aspects of Teacher Quality on Sequential GPA (Γ_k^c and Γ_k^f)



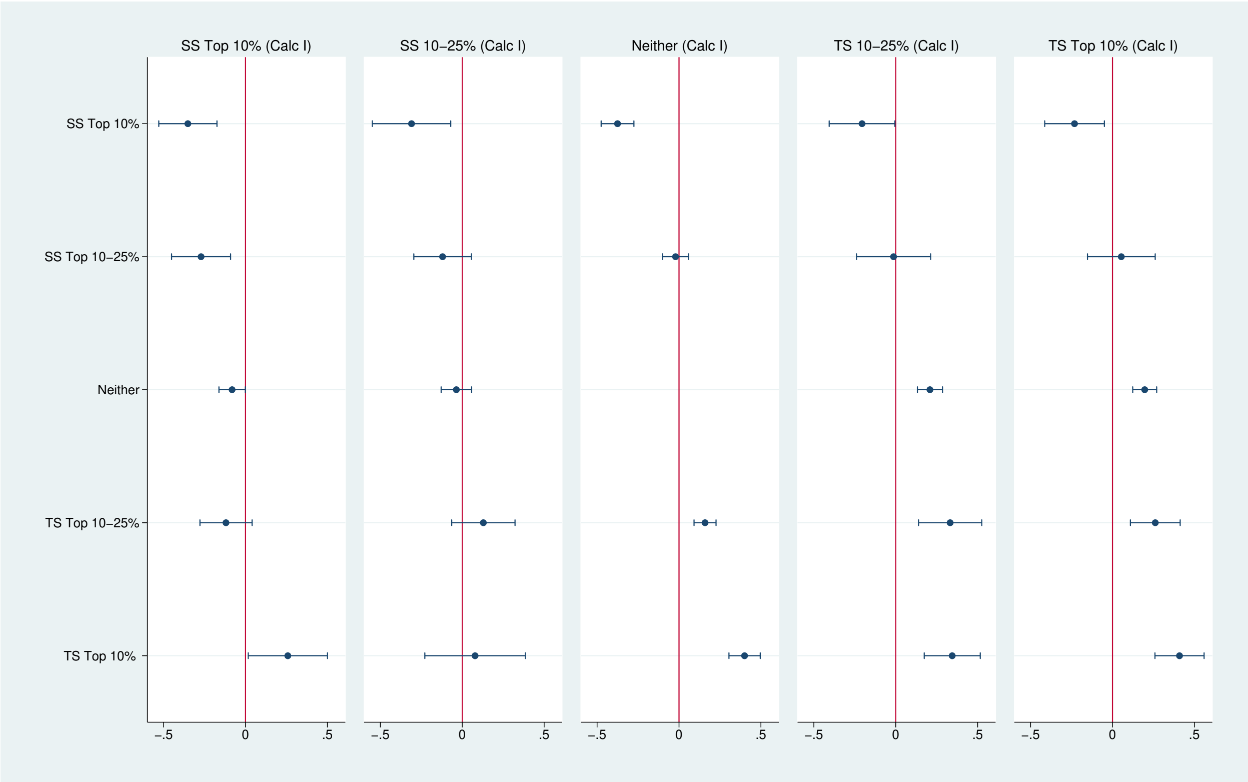
Note: Figure shows the predicted effect on second semester GPA from encountering different types of faculty in the first semester. Measured in standard deviation changes in faculty value-added based on estimates from Column (4) in Table 4. Moving along the x-axis is associated with softer standards and deeper knowledge instructors, respectively.

Figure 4: Counterfactual Effects of Aspects of Teacher Quality on Sequential GPA (RMP Measures)



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 ratings based on estimates from Column (3) in Table 5.

Figure 5: 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, next 15% of standards distribution, middle 50% of the distribution, next 15% of the standards distribution, and the bottom 10% of the standards distribution. The top of the distribution, denoted SS, refers to softer standards, while the bottom of the distribution, denoted TS, refers to tougher standards.

Table 1: Student and Faculty Characteristics - Summary Statistics

	Mean	Standard Deviation	Min	Max
Student Characteristics (N=22,980)				
Pre-College Characteristics:				
SAT Verbal	666.1	71.6	410	800
SAT Math	643.4	76.9	410	800
High School Standing Proxy	546.4	143.3	137	800
Female	0.197	0.398	0	1
Minority	0.266	0.442	0	1
Athlete	0.270	0.444	0	1
Myers-Briggs Personality:				
Extroversion / Introversion	3.0	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=77,490)				
First Course	2.83	0.87	0	4
Second Course	2.73	0.95	0	4
Final Exam Grade (N=61,314)				
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=679)				
Course Grade	2.91	0.84	1	4
Final Exam Grade	2.52	1.17	0	4
Female	0.265	0.442	0	1
Rate My Professor Sample (N=363)				
Number of Ratings	13.1	12.2	1	91
Overall Rating	3.8	1.01	1	5
Difficulty Rating	2.9	0.85	1	5
Course Grade	2.92	0.85	1	4
Final Exam Grade	2.48	1.15	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 sequential course sequences are: 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: Standards, Knowledge, and Sequential Learning

Current Course - Final Grade	(1)	(2)	(3)	(4)	(5)	(6)
Previous Instructor Γ_k^c	-0.0125***	-0.0104***	-0.0082**	-0.0182****	-0.0165****	-0.0151****
(Standards Channel)	(0.0044)	(0.0039)	(0.0038)	(0.0053)	(0.0047)	(0.0044)
Previous Instructor Γ_k^f				0.0117**	0.0126**	0.0143***
(Knowledge Channel)				(0.0056)	(0.0050)	(0.0048)
Student Characteristics	No	Yes	No	No	Yes	No
Student Fixed Effects	No	No	Yes	No	No	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Semester Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Section Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Current Instructor Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	51,379	51,379	51,379	51,379	51,379	51,379

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. All models control for year, semester, section, and current instructor fixed effects. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.

Table 4: The Heterogeneous Effects of the Standards and Knowledge Channels

Current Course - Final Grade	(1)	(2)	(3)	(4)
Previous Instructor Γ_k^c		-0.0713****		-0.0705****
(Standards Channel)		(0.00432)		(0.00432)
Previous Instructor Γ_k^f	0.0123***			0.0121***
(Knowledge Channel)	(0.00462)			(0.00461)
Prev Course A X Previous Instructor Γ_k^c	-0.0661****		-0.0839****	
	(0.00734)		(0.00814)	
Prev Course B X Previous Instructor Γ_k^c	-0.0614****		-0.0650****	
	(0.00619)		(0.00682)	
Prev Course C X Previous Instructor Γ_k^c	-0.0854****		-0.0701****	
	(0.00717)		(0.00787)	
Prev Course D X Previous Instructor Γ_k^c	-0.0731****		-0.0706****	
	(0.0139)		(0.0154)	
Prev Course A X Previous Instructor Γ_k^f		0.0393****	0.0467****	
		(0.00783)	(0.00877)	
Prev Course B X Previous Instructor Γ_k^f		0.0225****	0.0190***	
		(0.00653)	(0.00718)	
Prev Course C X Previous Instructor Γ_k^f		-0.0191**	-0.0196**	
		(0.00751)	(0.00821)	
Prev Course D X Previous Instructor Γ_k^f		0.00653	0.00623	
		(0.0144)	(0.0159)	
Prev Course Grade: A (relative to D)	1.009****	1.007****	1.008****	
	(0.0183)	(0.0180)	(0.0185)	
Prev Course Grade: B (relative to D)	0.701****	0.700****	0.700****	
	(0.0158)	(0.0153)	(0.0160)	
Prev Course Grade: C (relative to D)	0.346****	0.347****	0.347****	
	(0.0144)	(0.0138)	(0.0146)	
Previous Course - Final Grade				0.337****
				(0.00550)
Student Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Semester Fixed Effects	Yes	Yes	Yes	Yes
Section Fixed Effects	Yes	Yes	Yes	Yes
Current Instructor Fixed Effects	Yes	Yes	Yes	Yes
Observations	51,379	51,379	51,379	51,379

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. For columns (1)-(3), the indicators for overall grades A, B, and C are relative to the excluded category of D grades, which is estimated as the constant in the model. All models control for year, semester, section, and current instructor fixed effects. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.

Table 5: *Ratemyprofessors.com* Ratings and Sequential Learning

Current Course - Final Grade	(1)	(2)	(3)	(4)	(5)
Previous Instructor - Overall Rating (RMP)	-0.0210****	0.0160****	0.0163***	0.0202****	0.0203****
	(0.0035)	(0.0044)	(0.0056)	(0.0056)	(0.0057)
Previous Instructor - Level of Difficulty (RMP)		0.0753****	0.0794****	0.0550****	0.0555****
		(0.0054)	(0.0071)	(0.0075)	(0.0081)
Previous Instructor Γ_k^c				-0.0546****	-0.0540****
(Standards Channel)				(0.0053)	(0.0066)
Previous Instructor Γ_k^f					-0.0010
(Knowledge Channel)					(0.0066)
Student Fixed Effects	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
Semester Fixed Effects	Yes	Yes	Yes	Yes	Yes
Section Fixed Effects	Yes	Yes	Yes	Yes	Yes
Current Instructor Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	51,562	51,562	36,038	36,038	36,038

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 difficulty and overall quality. Columns (1) and (2) use the entire sample of faculty with RMP ratings. Column (3) restricts the sample to that used in Table 4) for faculty with RMP ratings. All models control for year, semester, section, current instructor fixed effects, and previous final grade. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.

Table 6: Heterogeneous Effects and RMP Ratings

Current Course - Final Grade	(1)	(2)	(3)	(4)
Prev Course A X Overall Rating	-0.0476****	0.0355****	0.0369****	0.0366****
	(0.00842)	(0.0107)	(0.0107)	(0.0107)
Prev Course B X Overall Rating	-0.0234****	0.0126	0.0157*	0.0155*
	(0.00675)	(0.00870)	(0.00870)	(0.00876)
Prev Course C X Overall Rating	-0.0142*	0.00746	0.0130	0.0128
	(0.00732)	(0.00970)	(0.00966)	(0.00971)
Prev Course D X Overall Rating	-0.0277**	-0.0267	-0.0173	-0.0176
	(0.0140)	(0.0192)	(0.0192)	(0.0193)
Prev Course A X Level of Difficulty		0.151****	0.122****	0.121****
		(0.0121)	(0.0124)	(0.0128)
Prev Course B X Level of Difficulty		0.0707****	0.0464****	0.0457****
		(0.0105)	(0.0108)	(0.0112)
Prev Course C X Level of Difficulty		0.0421****	0.0240*	0.0231*
		(0.0126)	(0.0127)	(0.0132)
Prev Course D X Level of Difficulty		-0.00341	-0.0155	-0.0166
		(0.0261)	(0.0262)	(0.0265)
Previous Instructor Γ_k^c			-0.0516****	-0.0526****
(Standards Channel)			(0.00533)	(0.00658)
Previous Instructor Γ_k^f				0.00169
(Knowledge Channel)				(0.00656)
Previous Semester Course Grade: A (relative to D)	1.034****	0.291*	0.404***	0.403***
	(0.0604)	(0.155)	(0.155)	(0.155)
Previous Semester Course Grade: B (relative to D)	0.648****	0.294**	0.375**	0.373**
	(0.0562)	(0.150)	(0.150)	(0.150)
Previous Semester Course Grade: C (relative to D)	0.277****	0.0552	0.0972	0.0968
	(0.0557)	(0.155)	(0.155)	(0.155)
Student Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Semester Fixed Effects	Yes	Yes	Yes	Yes
Section Fixed Effects	Yes	Yes	Yes	Yes
Current Instructor Fixed Effects	Yes	Yes	Yes	Yes
Observations	36,038	36,038	36,038	36,038

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. The indicators for overall grades A, B, and C are relative to the excluded category of D grades, which is estimated as the constant in the model. All models control for year, semester, section, and current instructor fixed effects. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.

Table 7: Faculty Groups and Sequential Spillovers

Current Course - Final Grade	(1)	(2)	(3)	(4)
Previous Instructor - Top 25% Overall	-0.0232**	-0.00424		
	(0.0112)	(0.0115)		
Previous Instructor - Bottom 25% Overall	-0.0470****	-0.0379***		
	(0.0121)	(0.0121)		
Previous Instructor - Top 25% Difficulty	0.0851****	0.0382***		
	(0.0125)	(0.0135)		
Previous Instructor - Bottom 25% Difficulty	-0.0775****	-0.0621****		
	(0.0121)	(0.0123)		
Schmope (Top 25% Difficulty / Top 25% Overall Rating)			-0.130****	-0.0863****
			(0.0146)	(0.0151)
Low Difficulty / Low Overall Rating			-0.172****	-0.119***
			(0.0412)	(0.0420)
High Difficulty / High Overall Rating			0.0447	0.0654
			(0.0479)	(0.0474)
High Difficulty/ Low Overall Rating			0.0539****	0.0171
			(0.0102)	(0.0110)
Previous Instructor Γ_k^c		-0.0641****		-0.0619****
(Standards Channel)		(0.00629)		(0.00621)
Previous Instructor Γ_k^f		0.00840		0.00881
(Knowledge Channel)		(0.00637)		(0.00630)
Student Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Semester Fixed Effects	Yes	Yes	Yes	Yes
Section Fixed Effects	Yes	Yes	Yes	Yes
Current Instructor Fixed Effects	Yes	Yes	Yes	Yes
Observations	36,038	36,038	36,038	36,038

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. Columns (3) and (4) 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. All models control for year, semester, section, current instructor fixed effects, and previous final grade. Standard errors are clustered by student. Significance: * 10 percent; ** 5 percent; *** 1 percent; ****0.1 percent.

Table 8: Sequential Learning Over Three Semesters - Calculus

	(1) Calc II	(2) Calc III	(3) Calc III	(4) Calc III
Calc I Γ_k^c	-0.155****	-0.0809****		
(Standards Channel)	(0.0124)	(0.0123)		
Calc II Γ_k^c		-0.153****		
(Standards Channel)		(0.0117)		
Calc I Γ_k^f	0.0439****	0.0362***	0.0374***	0.0373***
(Knowledge Channel)	(0.0124)	(0.0122)	(0.0120)	(0.0125)
Calc II Γ_k^f		0.0631****	0.0394***	0.0423****
(Knowledge Channel)		(0.0127)	(0.0124)	(0.0125)
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)	
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 Fixed Effects	Yes	Yes	Yes	Yes
Semester Fixed Effects	Yes	Yes	Yes	Yes
Section Fixed Effects	Yes	Yes	Yes	Yes
Current Instructor Fixed Effects	Yes	Yes	Yes	Yes
Student Characteristics	Yes	Yes	Yes	Yes
Student Personality	Yes	Yes	Yes	Yes
Observations	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 (Columns (2)-(4)). 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, section, and current instructor fixed effects. Student characteristics include math and verbal SAT scores, high school standing proxy, gender, minority status, athlete, feeder source indicators, and previous course grade. Standard errors are in parentheses. Significance: * 10 percent; ** 5 percent; *** 1 percent; **** 0.1 percent.

Table 9: Heterogeneous Effects: Faculty and Student Gender

Current Course - Final Grade	(1)	(2)	(3)	(4)
Previous Instructor Γ_k^c	-0.0659****	-0.0672****	-0.0613****	-0.0904****
(Standards Channel)	(0.00476)	(0.00472)	(0.00519)	(0.00725)
Previous Instructor Γ_k^f)	0.0115**	0.00928*	0.00731	0.0219***
(Knowledge Channel)	(0.00508)	(0.00497)	(0.00549)	(0.00786)
Female Student X Prev Instr Γ_k^c	-0.0222**			
	(0.00996)			
Female Student X Prev Instr Γ_k^f	0.00346			
	(0.0106)			
Previous Instructor - Female		0.0235***		
		(0.00829)		
Female Instructor X Prev Instr Γ_k^c		-0.0240**		
		(0.0117)		
Female Instructor X Prev Instr Γ_k^f		0.0171		
		(0.0128)		
(Male Instr and Female Student) X Prev Instr Γ_k^c			-0.0291***	
			(0.0108)	
(Female Instr and Male Student) X Prev Instr Γ_k^c			-0.0330**	
			(0.0129)	
(Female Instr and Female Student) X Prev Instr Γ_k^c			-0.0168	
			(0.0238)	
(Male Instr and Female Student) X Prev Instr Γ_k^f			0.00967	
			(0.0115)	
(Female Instr and Male Student) X Prev Instr Γ_k^f			0.0265*	
			(0.0143)	
(Female Instr and Female Student) X Prev Instr Γ_k^f			-0.00893	
			(0.0258)	
(Male Instr and Male Student) X Prev Instr Γ_k^c				0.0292****
				(0.00866)
(Male Instr and Male Student) X Prev Instr Γ_k^f				-0.0147
				(0.00933)
Student Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Semester Fixed Effects	Yes	Yes	Yes	Yes
Section Fixed Effects	Yes	Yes	Yes	Yes
Current Instructor Fixed Effects	Yes	Yes	Yes	Yes
Observations	51,370	51,370	51,370	51,370

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