Measuring Instructor Effectiveness in Higher Education[*]

Pieter De Vlieger, University of Michigan
Brian Jacob, University of Michigan and NBER
Kevin Stange, University of Michigan and NBER

May 15, 2016

Abstract

Professors and instructors are a chief input into the higher education production process, yet we know very little about their role in promoting student success. This is in contrast to elementary and secondary schooling, for which ample evidence suggests teacher quality is an important determinant of student achievement. Whether colleges could improve student and institutional performance by reallocating instructors or altering personnel policies hinges on the role of instructors in student success.

In this paper we measure variation in postsecondary instructor effectiveness and estimate its relationship to overall and course-specific teaching experience. We explore this issue in the context of the University of Phoenix, a large for-profit university that offers both online and in-person courses in a wide array of fields and degree programs. We focus on instructors in the college algebra course that is required by all BA degree programs. We find substantial variation in student performance across instructors both in the current class and subsequent classes. Variation is larger for in-person classes, but is still substantial for online courses. Effectiveness grows modestly with course-specific teaching experience. Our results suggest that personnel policies for recruiting, developing, motivating, and retaining effective postsecondary instructors may be a key, yet underdeveloped, tool for improving institutional productivity.

[*] We very grateful to Hinrich Eylers and Ashok Yadav at the University of Phoenix for many discussions and for providing access to the data. We also are grateful for useful feedback from Caroline Hoxby and participants at the NBER pre-conference and at the University of Michigan.

I. Introduction

Professors and instructors are a chief input into the higher education production process, yet we know very little about their role in promoting student success. There is growing evidence that teacher quality is an important determinant of student achievement in K12, with some school districts identifying and rewarding teachers with high value-added. Yet relatively little is known about the importance of or correlates of instructor effectiveness in postsecondary education. Such information may be particularly important at the post-secondary level, in which administrators often have substantial discretion to reallocate teaching assignments not only within a specific class of instructors (e.g., tenured faculty), but across instructor types (e.g., adjuncts vs. tenured faculty).

There are a number of challenges to measuring effectiveness in the context of higher education. Unlike the K12 context, there are rarely standardized test scores to use as an outcome. More generally, to the extent that college courses and majors intend to teach a very wide variety of knowledge and skills, it is harder to imagine an appropriate outcome as a conceptual matter. The issue of non-random student sorting across instructors is arguably more serious in the context of higher education because students have a great deal of flexibility in the choice classes and the timing of these classes. Finally, one might have serious concerns about the attribution of a particular skill to a specific instructor given the degree to which knowledge spills over across courses in college (e.g., the importance of calculus in intermediate microeconomics or introductory physics, the value of English composition in a history classes where the grade is based almost entirely on a term paper, etc.) For many reasons, the challenge of evaluating college instructors is more akin to the problem of rating physicians.

This paper addresses two main questions. First, is there variation in instructor effectiveness in higher education? We examine this in highly standardized setting where one would expect minimal variation in what instructors actually do. Second, what correlates with effectiveness? This informs whether teaching assignment and personnel policies could be used increase effectiveness. We examine these questions using detailed administrative data from the University of Phoenix (UPX), the largest university in the world which offers both online and in-person courses in a wide array of fields and degree programs. We focus on instructors in the college algebra course that is required for all students in BA degree programs and that often is a roadblock to student attainment.

This context provides several advantages. Our sample includes hundreds of instructors over more than a decade in campuses all across the United States. This allows us to generate extremely precise estimates, and to generalize to a much larger population than has been the case in previous studies. Students in these courses take a common, standardized assessment that provides an objective, clearly understand outcome by which to measure instructor effectiveness. And, as we describe below, student enrollment and course assignment is such that we believe the issue of sorting is either non-existent (in the case of the online course) or extremely small (in the case of face-to-face or FtF courses).

The UPX is quite different than what some might think of as the "traditional" model of higher education, in which tenured faculty at relatively elite institutions teach courses they develop themselves. UPX is a for-profit institution with contingent (i.e., non-tenured, mostly part-time) faculty, and the courses (both online and FtF) are highly standardized, with centrally designed slides, problem sets, exams, etc. So, while our findings may not generalize to all sectors of higher education, we believe it will be relevant for the growing for-profit sector and possibly less-selective 4-year and community colleges that share many of these same traits. A limitation of prior research is that it focuses on largely selective non-profit or public institutions, which are quite different from the non-selective or for-profit sectors. It is in these settings with many contingent faculty and institutions whose primary purpose is instruction (rather than, say, research) where productivity-driven personnel policies could theoretically be adapted.

We find substantial variation in student performance across instructors. A 1 SD increase in instructor quality is associated with 0.30 SD increase in student performance in current course and a 0.25 SD increase in performance in the subsequent course in the math sequence. Unlike some prior work (Carrell and West 2010), we find a strong, positive correlation between instructor effectiveness measured by current and subsequent course performance. The variation in instructor effectiveness is larger for in-person courses, but still substantial for online courses. These broad patterns and magnitudes are robust to extensive controls to address any possible non-random student sorting, using test scores that are less likely to be under the control of instructors, and other specification checks. These magnitudes are substantially larger than found in the K12 literature and in the Carrell and West (2010) study of the Air Force Academy, but comparable to the recent estimates in the DeVry University study (Bettinger at al. 2015).

Effectiveness grows modestly with course-specific teaching experience but is otherwise unrelated to tenure (time since hire). More generally, the substantial variation in instructor effectiveness at the University of Phoenix suggests that identifying, developing, and retaining highly effective instructors could be one important channel through which student performance could be improved.

The remainder of this paper proceeds as follows. We discuss prior evidence on college instructor effectiveness and our institutional context in Section II. Section III introduces our administrative data sources and our analysis sample. Section IV presents our empirical approach and examines the validity of our proposed method. Our main results quantifying instructor effectiveness are presented in Section V. Section VI examines how instructor effectiveness correlates with experience and simulates student outcomes under various hypothetical reassignment scenarios. Section VII concludes by discussing the implications of our work for institutional performance and productivity.

II. Prior Evidence and Institutional Context

A. Prior Evidence

There is substantial evidence that teacher quality is an important determinant of student achievement in elementary and secondary education (Rockoff, 2004; Rivkin, Hanushek, and Kain, 2005; Rothstein, 2010; Chetty, Friedman, Rockoff, 2014). Many states and school districts now incorporate measures of teacher effectiveness into personnel policies in order to select and retain better teachers (Jackson, Rockoff, Staiger, 2014). Yet little is known about instructor effectiveness in postsecondary education, in part due to difficulties with outcome measurement and self-selection. Standardized assessments are rare and grading subjectivity across professors makes outcome measurement difficult. In addition, students often choose professors and courses, so it is difficult to separate instructors' contribution to student outcomes from student sorting. As a consequence of these two challenges, only a handful of existing studies examine differences in professor effectiveness.

For the most part, prior studies have found that the variance of college instructor effectiveness is small compared to what has been estimated for elementary teachers. Focusing on large, introductory courses at a Canadian research university, Hoffmann and Oreopoulos (2009a) find the standard deviation of professor effectiveness in terms of course grades is no larger than 0.08. Carrell and West (2010) examine students at the U.S. Air Force Academy, where grading is standardized and students have no choice over coursework or instructors. They find sizeable differences in student achievement across professors teaching the same courses, roughly 0.05 SD, which is about half as large as in the K12 sector. Interestingly, instructors that were better at improving contemporary performance received higher teacher evaluations but were less successful at promoting "deep-learning", as indicated by student performance in subsequent courses. Braga, Paccagnella, Pellizzari (2016) estimate teacher effects on both student academic achievement and labor market outcomes at Bocconi University. They also find significant variation in teacher effectiveness, roughly 0.05 SD both for academic and labor market outcomes. They find only a modest correlation of instructor effectiveness in academic and labor market outcomes.

Bettinger, Fox, Loeb, and Taylor (2015) examine instructor effectiveness using data from DeVry University, a large for-profit institution in which the average student takes two-thirds of her courses online. Interestingly, they find a variance of instructor effectiveness that is substantially larger than prior studies in higher education. Specifically, they find that being taught by an instructor that is 1 SD more effective improves student outcomes by about 0.18 to 0.24 SD. They find somewhat less variation across instructors when courses are online, even among instructors that teach in both formats.

A few studies have also examined whether specific professor characteristics correlate with student success, though the results are quite mixed.[1] Using institutional-level data from a sample of U.S. universities, Ehrenberg and Zhang (2005) examine the effects of adjuncts (part-time faculty) on student dropout rates. They find a negative relationship between the use of adjuncts and student persistence, though they acknowledge that this result could stem, in part, from non-random sorting of students across schools. Hoffmann and Oreopoulos (2009a) find that no relationship between faculty rank and subsequent course enrollment. Two other studies find positive effects of adjuncts. Studying course-taking among students in public four-year institutions in Ohio, Bettinger and Long (2010) find adjuncts are more likely to induce students to take further courses in the same subject. Using a sample of large, introductory courses taken by first-term students at Northwestern University, Figlio, Schapiro, and Soter (2013) find that adjuncts are positively associated with subsequent course-taking in the subject as well as performance in these subsequent courses. In their study of the U.S. Air Force Academy, Carrell and West (2010) find that academic rank, teaching experience, and terminal degree are positively correlated with follow-on course performance, though negatively related to contemporary student performance.

There is also evidence that gender and racial match between students and instructors influences students' interest and performance (Bettinger and Long, 2005; Hoffmann and Oreopoulos, 2009b; Fairlie, Hoffmann, Oreopoulos, 2014). Finally, Hoffmann and Oreopoulos (2009a) find that subjective evaluations by students are a much better predictor of student performance than objective characteristics such as rank. This echoes the finding of Jacob and Lefgren (2008) that elementary school principals can identify effective teachers, but that observed teacher characteristics tend to explain little of teacher effectiveness.

B. Context: College Algebra at The University of Phoenix

We study teacher effectiveness in the context of the University of Phoenix, a large for-profit university that offers both online and face-to-face (FTF) courses. UPX offers a range of programs, including AA, BA and graduate degrees, while also offering à-la carte courses. We focus on core mathematics courses, MTH/208 and MTH/209 or College Mathematics I and II respectively, which are a requirement for most BA programs. Below we describe these courses, the process through which instructors are hired and evaluated, and the mechanism through which students are allocated to instructors.[2]

Courses

[1] Much of this evidence is reviewed in Ehrenberg (2012).

[2] This description draws on numerous conversations between the research team and individuals at the University of Phoenix.

BA-level courses at UPX are typically five weeks in duration and students take one course at a time (sequentially), in contrast to the typical structure at most universities. MTH/208 and MTH/209 focus on basic math skills. In particular, the MTH/208 curriculum focuses on setting up algebraic equations, and solving single and two-variable linear equations and inequalities. Additionally, the coursework focuses on relating equations to real-world applications, generating graphs, and the use of exponents. MTH/209 is considered a logical follow-up course, focusing on more complicated, non-linear equations and functions. As in other contexts, many students struggle in these introductory math courses and they are regarded by UPX staff as an important obstacle to obtaining a BA for many students.

Students can take these courses online or in-person. In the face-to-face (FtF) sections, students attend four hours of standard in-class lecture per week, typically held on a single day in the evening. In addition, students are required to work with peers roughly four hours per week on what is known as "learning team" modules. Students are then expected to spend 4-8 additional hours outside of class reading material, working on assignments and studying for exams.

Online courses are asynchronous, which means that a set of course materials is provided through the online learning platform, and instructors provide guidance and feedback through online discussion forums and redirect students to relevant materials when necessary. There are no classes in the standard sense, but students are required to actively participate on the online discussion by posting to the discussion board at least six times per week. This participation is the equivalent of the four hours of classes for the FTF sections.

There are substantial differences between the two course modes in terms of curriculum and grading flexibility. Both courses have standardized course curricula, assignments and tests that are made available to the instructors. Grading for these components is performed automatically through the course software. However, FTF instructors sometimes provide students with their own learning tools, administer extra exams and homework, or add other components that are not part of the standard curriculum. In contrast, online instructors mainly take the course materials and software as given, as interaction with students for these teachers is mainly limited to the online discussion forum. In both online and FtF courses, teachers are able to choose the weights they assign to specific course components for the final grade. As discussed below, for this reason we use also student performance on the final exam as an outcome measure.

Hiring and allocation of instructors

The onboarding process of teachers is managed and controlled by a central hiring committee that is hosted at the Phoenix campus, though much input comes from local staff at ground campuses. First, this committee checks whether a new candidate has the appropriate degree.[3] Second, qualified candidates then go through a five-week standardized training course they need to pass. This includes a mock lecture for FTF instructors and a mock online session for online instructors. Third, and finally, an evaluator sits in on the first class or follows the online course to ensure the instructor performs according to university standards. Salaries are relatively fixed, but do vary somewhat with respect to degree and tenure.[4]

The allocation of instructors to classes is essentially random for online classes. About 60 MTH/208 sections are started weekly and the roster is only made available to students two or three days before the course starts, at which point students are typically enrolled. The only way to sidestep these teacher assignments is by dropping the course altogether and enrolling in a subsequent week.

For FTF sections, the assignment works differently, since most campuses are too small to have different sections concurrently and students may need to wait for a couple of months if they decide to take the next MTH/208 section at that campus. While this limits the ability of students to shop around for a better teacher, the assignment of students to these sections is likely to be less random than for online sections. For this reason, we rely on value-added models that control for a host of student-specific characteristics that may correlate with both instructor and student course performance.

Evaluation and retention of teachers

UPX has in place three main evaluation tools to keep track of the performance of teachers. First, instructors need to take a yearly refresher course on teaching methods, and an evaluator will typically sit in or follow an online section every year to ensure the quality of the instructor still meets the university's requirements. Second, there is an in-house data analytics team that tracks key performance parameters. These include average response time to questions asked through the online platform, or indicators that students in sections are systematically getting too high (or too low) overall grades. For instance, if instructors consistently give every student in a section full or very high marks, this will raise a flag, and the validity of these grades will be verified. Finally, additional evaluations can be triggered if students file complaints about instructor performance. If these evaluation channels show the instructor has not met the standards of the university, the instructor receives a warning. Instructors that have received a warning are followed up more closely in subsequent courses. If the instructor performance does not improve, the university will not hire the person back.

[3] For MTH/208 sections, for instance, a minimum requirement might be having a master's degree in mathematics, or a master's degree in biology, engineering or similar coursework, along with a minimum number of credits in advanced mathematics courses and teaching experience in mathematics.

[4] For instance, all else equal, instructors with a Ph.D. can expect a higher salary than instructors with a master's degree. Additionally, tenure in this context refers to the date of first hire at the University of Phoenix. Salary differences are larger among new instructors, and tend to diminish at higher levels of experience.

III. Data

We investigate variation in instructor effectiveness using data drawn from administrative UPX records. This section describes these records, the sample selection, and descriptive statistics. While the data we analyze has very rich information about the experiences of students and instructors while at the University of Phoenix, information on outside activities is limited.

A. Data Sources

We analyze university administrative records covering all students and teachers who have taken or taught MTH/208 at least once between July 2000 and July 2014. The raw data spans 84 physical campuses (plus the online campus) and contains information on 2,343 instructors that taught 34,725 sections of MTH/208 with a total of 396,038 student-section observations. For all of these instructors and students, we obtain the full teaching and course-taking history back to 2000.[5]

Instructors

We draw on three information sources for instructor level characteristics. A first dataset provides the full teaching history of instructors that have ever taught MTH/208, covering 190,066 class sections. Information includes the campus of instruction, subject, the number of credits, and start date and end date of the section. There is typically one campus per city. For larger metropolitan areas, there may be multiple physical locations in which courses are offered.

For each instructor x section observation, we calculate the instructor's teaching load for the current year, as well as the number of sections he or she had taught in the past separately for MTH/208 and other courses. This allows us to construct a variety of different experience measures, which we will use in the analysis below. As the teaching history is censored before the year 2000, we only calculate the cumulative experience profile for instructors hired in the year 2000 or later.

The second dataset contains self-reported information on ethnicity and gender of the instructor, along with complete information on the date of first hire, the type of employment (full-time or part-time) and the zip code of residence.[6] A unique instructor identifier allows us to merge this information onto the MTH/208 sections.[7] A third dataset contains the salary information for each section, which can be merged onto the MTH/208 sections using the unique section identifier.

[5] The administrative records are not available before 2000 because of information infrastructure differences, leading to incomplete teaching and course-taking spells for professors and students respectively.

[6] This instructor dataset also contains information on birth year and military affiliation. These variables, however, have high non-response rates and are therefore not used for the analysis.

[7] The instructor identifier is, in principle, unique. It is possible, however, that an instructor shows up under two different identifiers if the instructor leaves the university and then returns after a long time. While this is a possibility, UPX administrators considered this unlikely to be a pervasive issue in their records.

Students

Student level information combines three data sources: demographics, transcript, and assessment.. The demographics dataset provides information on the zip code of residence, gender, age of the student, program the student is enrolled in, program start, and program end date. [8] A unique student identifier number allows us to merge this information onto the course-taking history of the student.

Transcript data contains complete course-taking history including the start and end date of the section, campus of instruction, grade, and number of credits. Every section has a unique section identifier that allows for matching students to instructors. Additionally, student level information includes course completion, course grade, earned credits, along with a unique student identifier that allows for merging on the student demographics.

Moreover, for sections from July 2010 to March 2014, or roughly 30 percent of the full sample, we always have detailed information on student performance separately by course assignment or assessment, which includes everything from individual homework assignments to group exercises to exams. We use this data to obtain a final exam score for each student. Because the data does not have a single, clear code for final exam component across all sections, and instructors have discretion to add additional final exam components, we use a decision rule to identify the "best" score for each student.

Ideally, this measure would capture computer-administered tests, since instructors do not have discretion over these. We therefore define a quality measure, ranging from 1 (best) to 4 (worst), that indicates how clean we believe the identification of these test scores to be. Once a student in a certain section gets assigned a test score, it is marked and not considered in later steps, so students get assigned a single quality measure and the assigned test score is of the highest quality available

Group 1 consists of the computer-administered common assessments available to all UPX instructors. To identify these assessments, we flag strings that contain words or phrases associated with the computer testing regime (e.g., "Aleks", "MyMathLab" or "MML") as well as words or phrases indicating a final exam (e.g., "final exam," "final examination," "final test"). If a student has an assessment that meets these criteria, we use the score from this assessment as the student's final exam score.[9] Specifically, we use the fraction of test items answered correctly as our measure of student

[8] Similar to the instructor dataset, these data are self-reported. While information on gender and age is missing for less than 1% of the sample, information on ethnicity, veteran status, and transfer credits exhibit much larger non-response rates and are therefore not used for the analysis.

[9] In extremely rare cases (less than 4 percent of the sample), students will have more than one assessment that meets these criteria, in which case we sum the attained and maximal score for these components, and calculate the percentage score. This is, in part, because for many cases, there was no grade component that could be clearly identified as the test score (e.g. a student may have "Aleks final exam: part 1" and "Aleks final exam: part 2").

performance. Roughly 11% of student-sections in our test score subsample have a final exam score with this highest level of quality, both for Math 208 and Math 209 test scores.

Some students have a single assessment with a word or phrase indicating a final exam (e.g., "final exam," "final examination," "final test"), but no explicit indication that the exam was from the standardized online system. If the assessment does not contain any additional words or phrases indicating that the test was developed by the instructor (e.g., "in class", "instructor generated," etc.), we are reasonably confident that it refers to the standardized online system. Hence, we use this assessment score as the student's final exam, but we consider these assessments as Group 2 for the purpose of exam quality. Another 77 percent of student-sections fall into this category for the Math 208 and Math 209 sections.

The third group looks at strings such as "final test", "final quiz", and "course exam". While quizzes and tests may sometimes refer to weekly refresher assessments, these strings identify final test scores reasonably well after having considered decision rules 1 and 2. About 9% of the student-sections fall into this category for both section types. The fourth and final group selects a grade component as a final test score if the title includes both "class" and "final". Another 2 percent of the sample gets assigned a test score of this quality for both the Math 208 and Math 209 sections.

While the analysis focuses on course grades and final test scores, it also considers future performance measures, such as grades and cumulative grade point average earned in the 180 or 365 days following the MTH/208 section of interest. Given the linear, one-by-one nature of the coursework, these measures capture the effect instructors have on moving students towards obtaining a final degree.

Census data

In addition to the UPX administrative school records, we use several census data resources to get additional variables capturing the characteristics of students' residential neighborhoods. In particular, we obtain the unemployment rate, median family income, the percentage of family below the poverty line, and the percentage with a bachelor degree or higher of students' home zip code, from the 2004-2007 five-year ACS files.

B. Sample Selection

Starting from the raw data, we apply several restrictions on the data to obtain the primary analysis sample. We restrict our analysis to the 33,200 Math 208 sections that started between January 2001 and July 2014. We then drop all students with missing data for final grade or unusual grades (0.1% of

About 3.75% of these cases have two assessments that meet the criteria. The maximum number of components for a student is five.

students) as well as students who do not show up in the student demographics file (0.3% of remaining students).[10] We then drop all cancelled sections (0.02 percent of the sections), sections with fewer than 5 enrolled students who had non-missing final grade and did not withdraw from the course (11.4 percent of the remaining sections) and sections for which the instructor is paid less than $300 (5.2 percent of remaining sections). We believe the final two restrictions exclude sections that were not actual courses, but rather independent studies of some sort, which is why we drop them. We also drop sections for which the instructor does not show up in the teacher demographics file, which is 3.5 percent of the remaining sections.

To calculate instructor experience, we use an instructor-section panel that drops observations where there is no salary information (about 3% of sections), the section was cancelled (0.04%), and with less than 5 students (21.7% of the remaining sections) or for which instructor is paid less than $300 (8.6% of the remaining sections). As above, these final two restrictions are meant to exclude independent study type courses or other unusual courses.[11] We then calculate several experience measures based on this sample. We calculate measures of experience such as number of courses taught in the previous calendar year and total cumulative experience in MTH208 specifically and in other categories of classes. The complete cumulative experience measures are only fully available for instructors that were hired after 2000, since the teaching history is not available in prior years.

Finally, we drop data from 9 campuses because none of the instructors we observe in these campuses ever taught in another physical campus or online. As discussed in the section below, in order to separately identify campus and instructor fixed effects, each campus must have at least one instructor that has taught in a different location. Fortunately, these 9 campuses represent only 2 percent of the remaining sections and 4 percent of remaining instructors.

The final analysis sample consists of 339,910 students in 26,393 sections, taught by 2,249 unique instructors. The sub-sample for which final exam data is available includes 78,865 students in 7,158 Math 208 sections taught by 1,204 unique instructors, and 62,429 students in 9,183 Math 209 sections taught by 1,474 unique instructors.

We calculate various student characteristics from the transcript data, including cumulative grade point average and cumulative credits earned prior to enrolling in MTH208, as well as future performance measures. In the rare case of missing student demographics, we set missing to zero and include an

[10] We keep students with grades A-F, I/A-I/F (incomplete A-F) or W (withdraw). Roughly 0.1% of scores are missing or not A-F or I/A-I/F (incomplete), and we drop these. These grades include AU (audit), I (incomplete), IP, IX, OC, ON, P, QC and missing values.

[11] First, there are three instructors that are first employed part-time and then employed full-time. As the part-time spells are longer than the full-time spells, we use the part-time demographics only. This restriction only impacts the employment type and date of first hire, as the other demographics are the same for the two employment spells for all three instructors.

indicator variable for missing. We merge on the FAFSA information for those students and year combinations that it is available, and set it to missing for those for who it is not.

C. Descriptive statistics

We report key descriptive statistics for the final analysis sample, spanning January 2001 to July 2014, in Table 1. We report these statistics for all sections, and for FTF and online sections separately. The upper panel of Table 1 represents section and instructor characteristics for the 26,393 Math 208 sections, while the middle panel represents student background characteristics, and the lower panel reports student performance measures. About half of all sections are taught online, and instructors are paid about $950 for teaching a course, regardless of the instruction mode.[12] Instructors are disproportionately white males and have been at the university just under five years. They typically have taught about 8.5 sections in the previous year, of which 3.5 sections were Math 208 sections.

When looking at cumulative experience measures, only available for instructors hired after 2000, highlight that instructors have typically taught about 33 sections before, of which about 25 were math sections. Across both modes of instruction, instructors seem to specialize in teaching mathematics, as this represents about 80% of their cumulative teaching experience. Nevertheless, instructors teaching online sections tend to be somewhat more experienced, and specialize more in teaching math sections compared to their counterparts teaching FTF sections. Where the former have a cumulative experience of about 35 sections at UPX, with 30 of those sections being math sections, instructors of FTF sections have a cumulative experience of about 30 sections, with only 18 of those being math sections.

Table A1 in the appendix reports descriptive statistics for the sample for which test scores are available (July 2010 – March 2014). The upper panel of Table A1 shows that somewhat fewer sections are taught online, and instructors are more likely to be female and hired more recently. The teaching load and experience measures based on the previous calendar year, however, are very similar to that of the full sample.

The middle and bottom panel of Table 1 provide an overview of student characteristics and performance. The students enrolled in these sections tend to be female, around 35 years old, and typically took around 23 credits worth of classes at UPX, earning a GPA of 3.35, when beginning MTH208. Students in online sections tend to have earned somewhat fewer credits than their counterparts in FTF sections, and are more likely to have taken Math 208 before. Most students, both in FTF and online sections, are enrolled in a business or general studies program.

[12] The earnings measures are deflated using the national CPI. For each year, the CPI in April was used, with April 2001 as the base.

Students across both modes of instruction are equally likely to earn a grade of A (about 32%) or B (about 27%), but students in online sections are more likely to withdraw. Students in FTF sections are more likely to earn lower grades, while students in online sections are more likely to withdraw from the section, resulting in a lower overall pass rate for online students. In terms of student performance after taking Math 208, we find that FTF students are more likely to go on and take Math 209, but conditional on taking Math 209, both online and FTF students typically take this class about a week after the Math 208 section.[13] Students in FTF sections earn about 20 credits in the year following the Math 208 section, compared to the 15 credits online students tend to earn.

The middle and lower panel of Table A2 highlights that these descriptive statistics do not really change when restricting attention to the sample for which test scores are available. Students tend to enroll in slightly different programs, and students overall tend to be somewhat more likely to get lower grades, but overall the numbers and differences across teaching modes mirror the findings of the overall sample.

IV. Empirical Approach

Our first aim is to characterize the variation in student performance across instructors teaching the same courses. Consider the standard "value-added" model of student achievement given in equation (1):

$$Y_{ijkt} = \beta_1 X_i + \beta_2 Z_{jkt} + \emptyset_t + \theta_k + e_{ijkt} \qquad (1)$$

where Y_{ijkt} is the outcome of student i in section j taught by instructor k during term t. The set of parameters θ_k quantify the contribution of instructor k to the performance of their students, above and beyond what could be predicted by observed characteristics of the student (X_i), course section (Z_{jkt}), or time period. The variance of θ_k across instructors measures the dispersion of instructor quality and is our primary parameter of interest. We are particularly interested in how the distribution of θ_k varies across outcomes and formats, and how effectiveness covaries across outcomes.

Estimation of the standard value-added model in (1) must confront three key issues. First, non-random assignment of students to instructors or instructors to course sections could bias value-added models. In the presence of non-random sorting, differences in performance across sections could be driven by differences in student characteristics rather than differences in instructor effectiveness per se. Second, outcomes should reflect student learning rather than grading leniency or "teaching to the test" of instructors. Third, our ability to make performance comparisons between instructors across campuses

[13] We report the median as the distribution for this variable is highly skewed. The mean of this variable is similar across both modes of instruction, but is a misleading measure of course-taking behavior.

while also controlling for cross-campus differences in unobserved student factors relies on the presence of instructors that teach at multiple campuses. We address each of these in turn below.

A. Course and Instructor Assignment

In many education settings, we worry about non-random assignment of instructors to sections (and students) creating bias in VA measures (Rothstein, 2009; Chetty, Friedman, Rockoff, 2014). In general, we believe that there is relatively little scope for sorting in our setting. Students do not know much about the instructor when they enroll, and instructors are only assigned to specific sections about two days before the start of the course for online sections. Students who have a strong preference with regard to instructor can choose to drop the course once they learn the instructor's identity, but this would mean that they would likely have to wait until the start of the next session to take the course, at which point they would be randomly assigned to a section again. According to UPX administrators, there is no sorting at all in online courses, which is plausible given the very limited interaction students with have with instructors in the initial weeks of the course. UPX admits the possibility of some sorting in FTF courses, but believe this is likely minimal.

To explore the extent of sorting, we conduct two types of tests. First, we test whether observable instructor characteristics correlate with the observable characteristics of students in a section. To do so, we regress mean student characteristics on instructor characteristics, where each observation is a course section. We then test the null hypothesis that the instructor characteristics are jointly equal to zero.[14] Table 2 reports the estimates from three regression models which differ in terms of the type of fixed effects that are included. Once we include campus fixed effects, we fail to reject the null hypothesis for student age, sex, incoming GPA and incoming credits. In results not reported here, but available upon request, we demonstrate similar results for subsamples limited to only online sections and to sections with final exam scores.

In addition, we follow the procedure utilized by Carrell and West (2010) to test whether the distribution of student characteristics across sections are similar to what you would get from random assignment within campus and time.). In a first step, we take the pool of students in a campus-year cell, randomly draw sections of different sizes, and compute the statistic of interest for these random sections. Similar to test 1, the statistics of interest are average age, fraction male, average prior credits, and average prior GPA. By construction, the resulting distribution of these section-level characteristics is obtained under random assignment of students to sections. In a second step, we take each actual section and

[14] An alternate approach would be to regress each student characteristic on a full set of course section dummies along with campus (or campus-year) fixed effects, and test whether the dummies are jointly equal to zero. This is equivalent to jointly testing the equality of the means of the characteristics across class sections.

compare the actual student average of each baseline characteristic to the counterfactual distribution for the relevant campus-year combination by calculating the p-value. For instance, we take a section, compute the average age, and compute the fraction of counterfactual sections with values smaller than the actual value. For each campus-year combination, we therefore obtain a number of p-values equal to the number of sections held at that campus-year combination. In a final step, we test for random assignment by testing the null hypothesis that these p-values are uniformly distributed. Intuitively, we are equally likely to draw any percentile under random assignment, which should result in these p-values having a uniform distribution. If, for instance, we have systematic sorting of student according to age, we would find we are more likely to find low and high percentiles, and the p-values would not exhibit a uniform distribution

Similar to Carrell and West (2010), we test the uniformity of these p-values using the Chi-square goodness-of-fit test, and a Kolmogorov-Smirnov test with a 5% significance level. We draw counterfactual distributions at the campus-year level, leading to 763 tests of the null hypothesis of uniformity of the p-values. We find that the null hypothesis is rejected in 56 cases using the Chi-square goodness-of-fit test, and in 51 cases using the Kolmogorov-Smirnov test, which is about 6-7%. Given that the significance level of these tests was 5%, we conclude that these tests do not reject the null hypothesis of random assignment of students to sections for these specific observables.

B. Outcomes

Unlike the elementary and secondary setting in which teacher effectiveness has been studied extensively using standardized test scores, it is harder to obtain a convincing outcome measure in the higher education sector. Following prior studies in the literature, we examine not only contemporaneous course performance as measured by a student's grade, but also enrollment and performance (measured by grades) in subsequent courses in the same subject.

An important limitation of grades as a measure of course performance is that they reflect, at least in part, different grading practices. This may be particularly worrisome in the context of FTF courses at UPX because many students have the same instructor for Math 208 and 209. Thus lenient or subjective grading practices in 208 may be correlated with the same practices in 209, meaning that the Math 209 is not an objective measure of long-run learning from Math 208. For a subset of our sample, we are able to examine student performance on the final examination for Math 208 and/or Math 209. It also might be informative to compare test-based measures to grade-based measures simply because the grade-based measures are easier for the universities to implement. It is informative to know how far from the more "objective" measures this gets you. In order to maximize sample coverage we first look at course grades and credits earned, but then also look at final exam scores (for a smaller sample).

Persistence is less susceptible to these concerns. Given that roughly one-quarter of the sample either withdraw or fail Math 208, and an equal fraction fail to take Math 209 at any point, it is interesting to look at whether students eventually pass Math 208 and/or Math 209. The number of credits accumulated in the six months following Math 208 is another outcome we examine.

C. Cross-campus comparisons

One fundamental concern about estimating instructor effectiveness is that unobservable differences between students across campuses may confound instructor differences. This is the rationale for controlling for campus fixed effects in equation (1). But separately identifying campus and instructor effects requires that a set of instructors teach in multiple campuses. This is analogous to the concern in studies that attempt to simultaneously estimate firm and worker effects as well as the literature measures teacher value-added at the K12 level. These "switchers" permit instructors across campuses to be ranked on a common scale.

The existence of the online courses, and the fact that a sizeable fraction of instructors teach both online and at a physical campus, provides the "connectedness" that allows us to separately identify campus and instructor effects. Table 2 illustrates the substantial degree of "switching" that exists across campuses in our data. About 8 percent of the exclusively FTF instructors teach in more than one campus, and about 21 percent of the online instructors also teach at a FTF.

D. Implementation

We implement the analysis with a two-step procedure. In the first step, we first estimate the standard value-added model in (1) including a host of student characteristics, campus fixed effects, and instructor FEs (θ_k). Including θ_k's as fixed effects permits correlation between θ_ks and X characteristics (including campus FEs), generating estimates of β_1, β_2, $\emptyset_t$, and δ_c that are purged of any non-random sorting by instructor (Chetty, Friedman, and Rockoff, 2014a). However, the estimated θ_k's are noisy, so their variance would be an inaccurate estimate of the true variance of the instructor effects. We then construct mean section-level residuals for each outcome

$$\tilde{Y}_{jkt} = \sum_{i \in j}(Y_{ijkt} - \widehat{\beta_1}X_i \; - \; \widehat{\beta_2}Z_{jkt} - \widehat{\emptyset_t} - \widehat{\delta_c}) \qquad (2)$$

The section-level residuals $\tilde{Y}_{jkt}$ combine the instructor effects (θ_k) with any non-mean-zero unobserved determinants of student performance at the student- or section-level. Our fully-controlled first-stage model includes student characteristics (male, age, incoming GPA, incoming credits, indicator for repeat MTH208, number of times taking MTH208, 12 program dummies, years since started program), section averages of these individual characteristics, student zip code characteristics (unemployment rate, median family income, percent of families below poverty line, percent of adults with BA degree in ZIP code, plus missing ZIP) and total section enrollment. We control for aggregate temporal changes in unobserved student characteristics or grading standards by including calendar year and month fixed effects. We include campus fixed effects to control for any unobserved differences in student characteristics across campuses. We also examine models with various subsets of these control variables and large sets of interactions between them.

In the second step, we use the mean residuals to estimate the variance of the instructor effects $\theta_k$as random effects.[15] For a single outcome, not distinguishing by mode, the model is simply $\tilde{Y}_{jkt} = \theta_k + \tilde{e}_{jkt}$. The error term $\tilde{e}_{jkt}$includes any section-specific common shocks and also any non-mean-zero student-level unobserved characteristics, both of which are assumed to be independent across instructors and time. Our preferred approach stacks outcomes and lets effectiveness vary by outcome and mode (FTF vs. online) with an unrestricted covariance matrix. For instance, for two outcomes (grade in MTH208 and MTH209) not distinguished by mode, we estimate

$$\tilde{Y}_{jkt} = {\theta_k}^{M208} \cdot M208_{jkt} + {\theta_k}^{M209} \cdot M209_{jkt} + \tilde{e}_{jkt} \qquad (3)$$

where $M208_{jkt}$ and $M209_{jkt}$ indicate outcomes for MTH208 and MTH209, respectively. The key parameters of interest are SD(θ_k), SD(${\theta_k}^{M208}$), SD(${\theta_k}^{M209}$), Corr(${\theta_k}^{M208}$, ${\theta_k}^{M209}$). Analogous models are estimated separately by mode of instruction.

V. Results on Instructor Effectiveness

A. Main Results for Course Grades and Final Exam Scores

Table 4 reports our main estimates of the variances and correlations of MTH208 instructor effects for both grade and test score outcomes. Odd columns report results for the full sample of 26,304 sections and 2,249 instructors throughout the analysis window.

For the full sample, a one-standard deviation increase in MTH208 instructor quality is associated with a 0.30 and 0.25 standard deviation increase in student course grades in MTH208 and MTH209, respectively. In course grade points, this is a little larger than one grade step (going from a “B” to “B+”). Thus MTH208 instructors substantially affect student achievement in both the introductory and follow-on math courses. These estimates are statistically significant and quite a bit larger than effects found in prior research in postsecondary (e.g. Carrell and West, 2010) and elementary schools (Kane et al. 2008). We also find that instructor effects in MTH208 and MTH209 are highly positively correlated (correlation coefficient = 0.56). This tells us that MTH208 instructors that successfully raise student performance in MTH208 also raise performance in follow-on courses. Thus we do not observe the same negative tradeoff

[15] Second stage models are estimated with Stata’s “mixed” command.

between contemporaneous student performance and “deep learning” highlighted by Carrell and West (2010).

Columns (3) and (5) split the full sample by whether the MTH208 section was held at a ground campus (face-to-face) or the online campus. Though slightly more than half of sections are held at ground campuses, they make up three-quarters of the instructors in the full sample. Instructor quality is more variable at ground campuses (0.31 SD vs. 0.24 SD for online for MTH208), particularly as measured by follow-on course performance (0.31 SD vs. 0.08 SD for online for MTH209). There are a number of reasons that online instructors may have less variation in quality than face-to-face instructors. First, ground instructors have more discretion over course delivery and are more likely to modify the curriculum. Ground instructors also have more direct interaction with students. Both of these factors may magnify differences in their effectiveness in a ground setting. Second, personnel management is centralized for online sections, while many aspects of hiring, evaluation, and instructor training are done by individual campuses for ground sections. Finally, since faculty are not randomly assigned to section formats (FTF vs. online), variance differences across formats could reflect differences in instructor characteristics. For instance, if teaching experience relates to effectiveness and ground campuses have a greater variance of instructor experience, then this will be reflected in the variance of instructor quality. In addition, if there is less non-random sorting of students to instructors (conditional on our extensive control variables) in online sections than in ground sections, this will inflate the estimated variance of instructors at ground campuses.

Interestingly, instructor quality in contemporaneous and follow-on course performance are positively correlated for face-to-face sections, but negatively correlated for online sections. Later we present evidence that online instructors whose students perform better in MTH208 are also better at getting their students to enroll in MTH209, pushing down average student performance in MTH209 relative to the students of instructors that do not get their students to enroll in MTH209. This creates a negative sample bias in the correlation between contemporaneous and follow-on course performance that disappears once we account for the sample selection. One potential explanation is that students taking FtF courses are more engaged in and committed to their degree, and there is therefore less scope for instructor influence in this dimension.

Course grades are problematic as a measure of student achievement to the extent that systematic differences across instructors reflect different grading policies or standards rather than student learning. We address this by examining student performance on normalized final course exams.[16] Even columns in

[16] Since exams differ in maximum point values across sections and for MTH208 and MTH209, the outcome is the fraction of points earned (out of the maximum). This fraction is then standardized to mean zero and standard deviation one for the individuals with scores across the entire sample.

Table 4 restrict analysis to sections that start between June 2010 and March 2014, for which we have such exam scores. For FtF sections, the variance of instructor effects is actually larger when using final exam score rather than course grades: 0.44 compared with 0.29. This is consistent with less effective teachers grading more easily than more effective teachers. In contrast, in online sections, the variance of instructor effects is smaller when using final exam score, consistent with less effective teachers grading more harshly. Effectiveness is also highly positively correlated (correlation = 0.61) between contemporaneous and follow-on course performance. The negative correlation between contemporaneous and follow-on course performance for online MTH208 sections is also observed with final exam scores, though it is imprecisely estimated and generally not robust (in magnitude or sign) across alternative specifications.

One candidate explanation for the high positive correlation between instructor effects in contemporaneous and follow-on courses (particularly for course grade outcomes) is that many students have the same instructors for MTH208 and MTH209 at ground campuses. Fully 81% of students in ground sections have the same instructor for MTH208 and MTH209, while fewer than 1% of students taking MTH208 online do. This difference in the likelihood of having repeat instructors could also possibly explain differences between online and face-to-face formats. Having the same instructor for both courses could generate a positive correlation through several different channels. First, instructor-specific grading practices or tendency to "teach-to-the-test" that are similar in MTH208 and 209 will generate correlated performance across classes that does not reflect true learning gains. Alternatively, instructors teaching both courses may do a better job of preparing students for the follow-on course.

To examine this issue, Table 5 repeats our analysis on the subset of MTH208 face-to-face sections where few (< 25%) or no students take MTH209 from the same instructor. While instructor quality may influence some students' choice of MTH209 instructor, it is unlikely to trump other considerations (such as schedule and timing) for all students. Thus we view these subsamples as identifying situations where students had little ability to have a repeat instructor for other reasons. Though the number of sections is reduced considerably and the included instructors are disproportionately low-tenure, the estimated instructor effects exhibit a similar variation as the full sample, both for course grades and exam scores. The correlation between MTH208 and 209 instructor effects is reduced substantially for grades and modestly for test scores, but remains positive and significant for both, even with the most restricted sample.

B. Robustness and Other Outcomes

Table 6 examines the robustness of our test score results to different first stage models. Our preferred model includes numerous student characteristics, section averages of these individual characteristics, total section enrollment, campus fixed effects, instructor fixed effects, calendar year fixed effects, and month fixed effects. Even models with only time controls exhibit patterns that are qualitatively similar to our base model, with substantial instructor quality variation, particularly for face-to-face sections. In fact, the extensive controls have little impact on estimates of instructor quality, suggesting minimal systematic non-random sorting of students to instructors based on observed characteristics (and possibly unobserved characteristics too). The only consequential controls we include are campus fixed effects when combined with instructor fixed effects, which increase the estimated variance of instructor effects on MTH208 and MTH209 exam scores and reduce their correlation. For online sections, estimates of instructor effects do not change at all across first stage specifications, but the estimated correlation is not robust, changing signs and often insignificant.

Table 7 addresses sample selection issues in two ways. In Panel A, we examine the likelihood of having an exam score for MTH208 or 209 and likelihood of enrolling in MTH209 as outcomes and also how effectiveness in these dimensions correlates with measured performance. We find that there is a reasonable amount of variability in instructor impacts on students' likelihood of having test scores or MTH209 grades. In fact, for online sections the effects on likelihood of taking MTH209 is more positively correlated with effects measured by MTH208 performance (grades or test scores) than it is for face-to-face sections. This implies that the highest quality online instructors in particular (measured by MTH208 performance) get more of their students to take MTH209, which could create a sample selection bias problem if the marginal students induced to take MTH209 are lower-achieving. This could at least partially explain the negative correlation we find between instructor effects when measured by MTH208 and MTH209 performance for online instructors.

In Panel B we assign zeros for final exam scores or MTH209 grades for students that withdraw before taking the exam or who do not enroll in MTH209, respectively. Our main model excludes these individuals when calculating section-level mean residuals in the first stage. Consistent with the interpretation above, correlations between effects on MTH208 and MTH209 performance turn positive for online sections when the students who did not enroll in MTH209 are included as zeros (contrast the positive 0.52 correlation between MTH208 and adjusted MTH209 effects for online sections to the negative 0.70 correlation between MTH208 and unadjusted MTH209 effects in Table 4). The correlations do not change with this adjustment for face-to-face sections, as these exhibit a weaker correlation with likelihood of taking MTH209.

Table 8 presents estimates of instructor effects for several different outcomes, both for the full sample and the restricted sample for which test scores are available. There is substantial instructor variability in students' likelihood of taking MTH209 and in the number of credits earned in the six months following MTH208. Both of these are important indicators of students' longer-term success at UOPX. A one-standard-deviation increase in MTH208 instructor quality is associated with a five percentage point increase in the likelihood a student enrolls in MTH209 (on a base of 76%), with the variability twice as large for face-to-face MTH208 sections as it is for online ones. A similar increase in instructor quality is associated with a 0.13 SD increase in the number of credits earned in the six months following MTH208, again with face-to-face instructors demonstrating more than twice as much variability as online sections. Total credits earned after MTH208 is an important outcome for students and the university which is unlikely to be manipulated by individual instructors. Table 9 reports correlations between instructor effects measured with these different outcomes for the test score sample, overall and separately by format.[17][18] Most of the outcomes are positively correlated overall and for face-to-face sections. Interestingly, value-added measured by likelihood of taking MTH209 or total credits earned after MTH208 is only weakly correlated with value-added measured by final exam scores

VI. Correlates of Instructor Effectiveness

Having demonstrated substantial variation in instructor effectiveness along several dimensions of student success, particularly for face-to-face sections, we now consider how teaching experience correlates with effectiveness. Teaching experience- both course-specific and general - may be an important factor in instructor performance given results found in other contexts (e.g. Ost, 2014; Cook & Mansfield, 2015).

For this analysis, we focus on instructors hired since 2002 so that we can construct a full history of courses taught across all courses and in MTH208 specifically, not censored by data availability. This results in 18,418 sections (5,860 in the test score sample). Our main approach is to regress section-level residuals $\tilde{Y}_{jkt}$ on observed instructor characteristics at the time the section was taught:

$$\tilde{Y}_{jkt} = f\left(Exp_{MTH208,t}\right) + \theta_k + e_{jkt} \tag{3}$$

Where f(.) is a flexible function of experience teaching MTH208. Our preferred model includes instructor fixed effects, θ_k, isolating changes in effectiveness as individual instructors gain experience. This model controls for selection into experience levels based on fixed instructor characteristics, but does not control for time-varying factors related to experience and effectiveness. We also include other dimensions of experience.

Figures 1 and 2 present estimates of (3) for a non-parametric version of f(.), regressing section mean residuals on a full set of MTH208 experience dummies (capped at 20) along with year, month, and

[17] Correlations are quite similar for the full sample.

[18] These correlation matrices are formed by predicting the BLUP instructor effects for different outcomes one at a time and correlating these using section-level data. It would be more efficient to estimate all the effects and the correlations simultaneously as we did for pairs of outcomes (e.g. grades in MTH208 and MTH209 in Table 4), but these models did not converge. This is also why the correlations reported in Table 7 differ from those in Table 4.

(when noted) instructor fixed effects.[19] Figure 1 depicts results for course grade outcomes. Effectiveness increases very modestly the first few times instructors teach MTH208, as measured by MTH208 and MTH209 course grades. Interestingly, including instructor fixed effects stabilizes the effectiveness-experience profile, suggesting that less effective instructors are more likely to select into having more MTH208 teaching experience. Figure 2 repeats this analysis but for final exam test scores on the restricted test score sample. Estimates are quite imprecise, but do suggest modest growth in MTH208 exam scores as instructors gain experience. Improvement with experience is not as clear-cut for MTH209 test score performance.

To gain precision, Table 10 presents estimates from parametric specifications for f(.), while also including teaching experience in other courses (in Panel C) . We find that teaching MTH208 at least one time previously is associated with a 0.03 to 0.04 SD increase in effectiveness (measured by MTH208 grade), but that additional experience improves this outcome very little. This holds even after controlling for additional experience in other subjects. Instructor impact on follow-on course grades is more modest and gradual. Test score results are much less precise, but do suggest that instructor effectiveness increases with experience for final exams in contemporaneous courses and (very modestly) in follow-on courses. We find that general experience in other subjects has little association with effectiveness in MTH208 (not shown). Finally, we find no systematic relationship between teaching experience and instructors' impact on the number of credits their students earn subsequent to MTH208.

Though not reported in the table, we also found that whether the instructor was hired in the past year and the number of years since first hire date had no association with instructor effectiveness (after controlling for MTH208 experience) nor did including them as controls alter our conclusions. This is important as years since first hire is the one consistent predictor of the salary instructors are paid for MTH208 courses (Table 11). Instructors receive approximately $70 more per course for each ten years of tenure (approximately 7% higher pay) after fixed instructor differences are accounted for.

VII. Conclusion and Discussion

In this study, we document substantial differences in effectiveness across instructors of required college algebra at the University of Phoenix. A one-standard-deviation in instructor quality is associated with a 0.25 SD increase in course grades and a 0.40 SD increase in final exam scores in the follow-on course, as well as a 0.13 SD increase in the number of credits earned within six months. Variation is much smaller for online sections, yet still measurable and larger than that found in other contexts.

[19] Approximately one quarter of the sections are taught by instructors that have taught MTH208 more than 20 times previously. Nine percent have not previously taught MTH208.

It is worth considering what institutional factors may contribute to such large differences across instructors, particularly in contrast to other settings. Prior work in postsecondary has focused on selective and research-oriented public and non-profit universities, courses taught by permanent or tenure-track faculty, institutions operating in a single geographic location, and serving “traditional” students. It is possible that instructors are a more important factor in the success of “non-traditional” students or that there is more variation in instructor quality among contingent and adjunct faculty than among permanent or tenure-track faculty.   The one prior study that finds instructor variation comparable to ours (Bettinger, Fox, Loeb, and Taylor, 2015) shares all of these traits with our study institution. Having a better understanding of the importance of faculty at less selective institutions and in settings where most faculty are contingent is important, as these institutions serve a very large (and growing) share of postsecondary students in the U.S..

This substantial variation across instructors suggests potential to improve student and institutional performance via changes in how faculty are hired, developed, motivated, and retained. Institutions like UPX reflect the sector-wide trend towards contingent faculty (e.g. adjuncts and lecturers), which aimed to save costs and create flexibility (Ehrenberg, 2012). Debate about whether adjuncts are better or worse for instruction than permanent faculty obfuscates the feature that contingent arrangements create opportunities for improving student performance via personnel policies that are not available when faculty are permanent. However, instructor evaluation and compensation systems have not kept up with these changes; our study institution has an evaluation system (student course evaluations) that is similar to that at elite research universities and a salary schedule that varies primarily with tenure and credentials. Of course the potential for improvement through changes in personnel policies – and how these policies should be designed – depends critically on the supply of instructors available (e.g. Rothstein, 2015). Online and ground campuses likely face quite different labor markets for instructors, the former drawing on instructors across the country, suggesting that personnel policies should differ between them. Better understanding the labor market for postsecondary faculty – particularly at less selective institutions – is an important area for future attention.

Finally, we have focused on the role of individual faculty in promoting the success of students. In fact, differences in instructor effectiveness is one potential explanation for cross-institution differences in institutional performance and productivity that has yet to be explored. Our study suggests it should.

References

Bettinger, E., Fox, L., Loeb, S., & Taylor, E. (2014). *Changing distributions: How online college classes alter student and professor performance*. Working Paper, Stanford University.

Bettinger, E. P., & Long, B. T. (2005). Do faculty serve as role models? The impact of instructor gender on female students. *The American Economic Review, 95*(2), 152-157.

Bettinger, E. P., & Long, B. T. (2010). Does cheaper mean better? The impact of using adjunct instructors on student outcomes. *The Review of Economics and Statistics, 92*(3), 598-613.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). The academic and labor market returns of university professors. *IZA Discussion Papers*.

Carrell, S. E., & 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., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *The American Economic Review, 104*(9), 2593-2632.

Cook, J. B., & Mansfield, R. K. (2014). Task-Specific Experience and Task-Specific Talent: Decomposing the Productivity of High School Teachers. Working Paper

Ehrenberg, R.G. (2012). American Higher Education in Transition. *Journal of Economic Perspectives, 26(1): 193-216.*

Ehrenberg, R. G., & Zhang, L. (2005). Do tenured and tenure-track faculty matter? *Journal of Human Resources*, *40*(3), 647-659.

Fairlie, R. W., Hoffmann, F., & Oreopoulos, P. (2014). A Community College Instructor Like Me: Race and Ethnicity Interactions in the Classroom. *The American Economic Review*, *104*(8), 2567-2591.

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

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

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

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. *Annual Review of Economics*, *6*, 801-25.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of labor Economics*, *26*(1), 101-136.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education review*, *27*(6), 615-631.

Ost, B. (2014). How Do Teachers Improve? The Relative Importance of Specific and General Human Capital. *American Economic Journal: Applied Economics*, *6*(2), 127-51.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

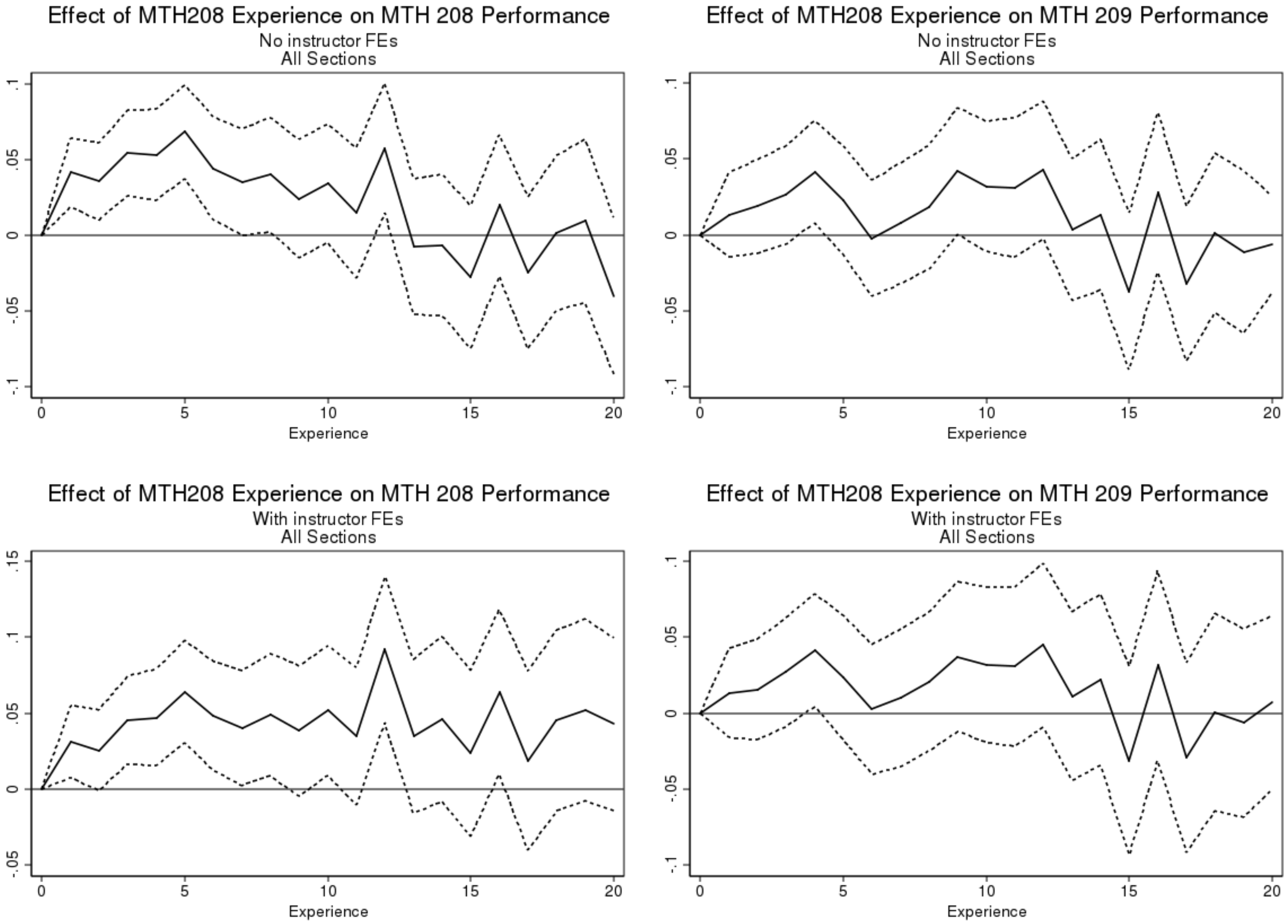
Rothstein, J. (2009). "“Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables.” *Education Finance and Policy* 4(4), Fall 2009: 537-571

Rothstein, J. (2010). "“Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement.” *Quarterly Journal of Economics* 125(1), February 2010: 175-214.

Rothstein, J. (2015). “Teacher Quality Policy When Supply Matters.” *American Economic Review* 105(1), January 2015: 100-130.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247-252.

Figure 1. Relationship between Instructor Effectiveness (Grades) and Teaching Experience


Effect of MTH208 Experience on MTH 208 Performance
No instructor FEs
All Sections

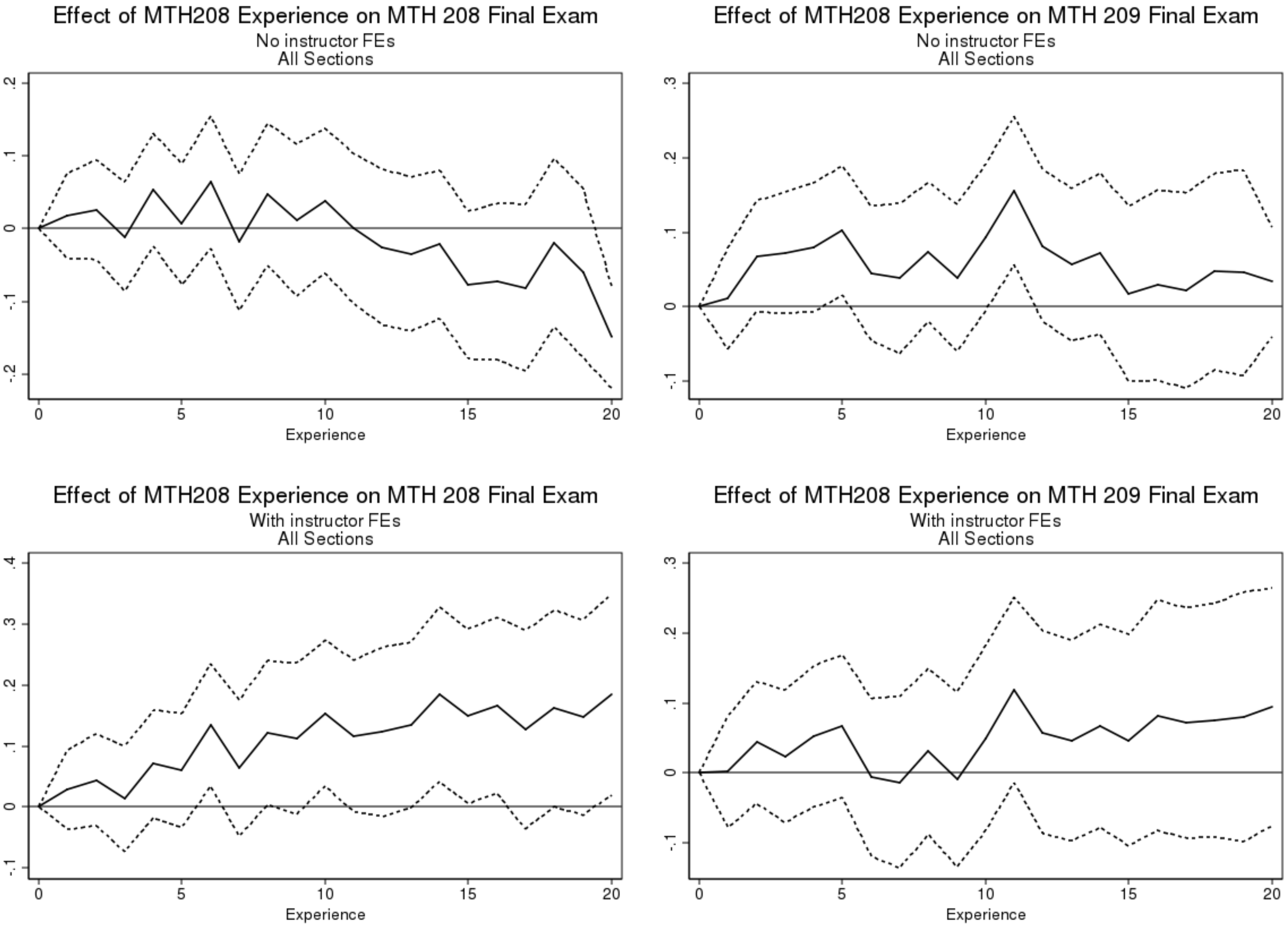
Effect of MTH208 Experience on MTH 209 Performance
No instructor FEs
All Sections

Effect of MTH208 Experience on MTH 208 Performance
With instructor FEs
All Sections

Effect of MTH208 Experience on MTH 209 Performance
With instructor FEs
All Sections

Notes: Dashed lines denote 95% CI with standard errors clustered by instructor. Section mean residuals are regressed on MTH208 teaching experience (capped at 20), instructor fixed effects, and year and month fixed effects. Sample restricted to 18,418 sections taught by instructors hired since 2002. First stage model includes full controls (see text).

Figure 2. Relationship between Instructor Effectiveness (Test Scores)and Teaching Experience



Effect of MTH208 Experience on MTH 208 Final Exam
No instructor FEs
All Sections

Effect of MTH208 Experience on MTH 209 Final Exam
No instructor FEs
All Sections

Effect of MTH208 Experience on MTH 208 Final Exam
With instructor FEs
All Sections

Effect of MTH208 Experience on MTH 209 Final Exam
With instructor FEs
All Sections

Notes: Dashed lines denote 95% CI with standard errors clustered by instructor. Section mean residuals are regressed on MTH208 teaching experience (capped at 20), instructor fixed effects, and year and month fixed effects. Sample restricted to 5860 sections taught by instructors hired since 2002. First stage model includes full controls (see text).

Table 1. Descriptive Statistics (Full Sample)

	All		Face-to-Face		Online	
	N	Mean	N	Mean	N	Mean
Section and Instructor Characteristics						
Online section	26,393	0.480	13,800	0.000	12,593	1.000
Male	24,234	0.730	12,235	0.750	11,999	0.710
White	21,641	0.650	10,280	0.630	11,361	0.660
Instructor Compensation per Section ($)	26,393	955.14	13,800	949.39	12,593	961.45
Instructor Teaching Load (Sections) during calendar year	26,393	10.62	13,800	11.11	12,593	10.07
Instructor Teaching Load (Math Sections) during calendar year	26,393	4.90	13,800	3.63	12,593	6.28
Years since first hire	26,393	4.78	13,800	5.00	12,593	4.54
# sections instructor taught in past year	26,393	8.55	13,800	8.89	12,593	8.18
# Math 208 sections the instructor taught in the past year	26,393	3.45	13,800	2.49	12,593	4.50
Total sections instructor taught prior to this section	20,493	33.23	9,254	30.82	11,239	35.21
Total math sections instructor taught prior to this section	20,493	24.71	9,254	18.47	11,239	29.85
Share of prior sections taught that were math courses	20,493	0.790	9,254	0.690	11,239	0.880
Share of prior sections taught that were business courses	20,493	0.040	9,254	0.050	11,239	0.020
Share of prior sections taught that were other courses	20,493	0.170	9,254	0.260	11,239	0.100
Student Background Characteristics						
Male	338,090	0.360	191,948	0.370	146,142	0.340
Age	339,358	34.82	192,549	34.27	146,809	35.54
Baseline GPA (0-4)	339,910	3.35	192,813	3.35	147,097	3.35
Credits earned prior to start of Math 208	339,910	23.39	192,813	25.71	147,097	20.33
Took Math 208 before	339,910	0.090	192,813	0.070	147,097	0.120
BS in Business	339,910	0.500	192,813	0.590	147,097	0.390
BS (general studies)	339,910	0.210	192,813	0.210	147,097	0.210
BS in Nursing	339,910	0.050	192,813	0.030	147,097	0.080
BS in Management	339,910	0.040	192,813	0.020	147,097	0.070
BS in Criminal Justice Administration	339,910	0.030	192,813	0.050	147,097	0.020
BS in Health Administration	339,910	0.030	192,813	0.030	147,097	0.030
BS in Human Services	339,910	0.030	192,813	0.020	147,097	0.050
BS in Information Technology	339,910	0.030	192,813	0.030	147,097	0.030
BS in Education	339,910	0.020	192,813	0.010	147,097	0.030
Outcomes						
Performance in Math 208						
A / A-	339,910	0.320	192,813	0.320	147,097	0.310
B+ / B / B-	339,910	0.270	192,813	0.270	147,097	0.260
C+ / C / C-	339,910	0.170	192,813	0.190	147,097	0.150
D+ / D / D-	339,910	0.070	192,813	0.080	147,097	0.070
F	339,910	0.040	192,813	0.040	147,097	0.050
Withdrawn	339,910	0.120	192,813	0.100	147,097	0.160
Passed Math 208	339,910	0.830	192,813	0.870	147,097	0.790
Final exam score available	339,910	0.240	192,813	0.280	147,097	0.190
Performance following Math 208						
Took Math 209	339,910	0.760	192,813	0.820	147,097	0.660
Days before taking Math 209 (Median)	253,169	7	157,516	7	95,653	7
Credits earned in following year	339,910	17.74	192,813	19.38	147,097	15.60

Table 2. Randomization Check

	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Outcome = Average Age			**Panel B. Outcome =Fraction Male**		
Years since first hire	-0.00649	0.00723	0.00245	0.000813	-0.00124***	-0.000527
	(0.012)	(0.009)	(0.008)	(0.001)	(0.000)	(0.000)
< 1 year since hire	-0.183*	-0.0458	0.044	0.0120**	-0.000802	-0.00151
	(0.100)	(0.090)	(0.093)	(0.006)	(0.005)	(0.005)
Sections taught last year	-0.0145***	0.00169	-0.00224	0.000526**	0.000286	0.000539***
	(0.004)	(0.004)	(0.003)	(0.000)	(0.000)	(0.000)
Math sections last year	0.0649***	0.00864	-0.00182	-0.00241***	-0.00130***	-0.000949*
	(0.010)	(0.009)	(0.008)	(0.000)	(0.000)	(0.000)
R-squared						
	0.041	0.119	0.175	0.033	0.104	0.167
	Panel C. Outcome =Incoming GPA			**Panel D. Outcome =Incoming credits**		
Years since first hire	0.00209***	0.000217	0.0001	0.0646*	0.0418*	-0.00398
	(0.001)	(0.000)	(0.000)	(0.033)	(0.022)	(0.013)
< 1 year since hire	0.0214***	0.0131**	0.00362	-0.321	-0.595***	-0.187
	(0.006)	(0.005)	(0.005)	(0.295)	(0.228)	(0.181)
Sections taught last year	0.000335	0.0000	-0.000196	0.0972***	0.0197*	-0.000559
	(0.000)	(0.000)	(0.000)	(0.017)	(0.011)	(0.007)
Math sections last year	-0.0001	0.0000	-0.000223	-0.307***	0.0344	0.00148
	(0.001)	(0.000)	(0.000)	(0.027)	(0.023)	(0.016)
R-squared	0.324	0.383	0.427	0.123	0.276	0.422
Observations	23298	23298	23298	23298	23298	23298
FE	None	campus	campus-year	None	campus	campus-year

Notes: Each panel-column is a separate regression of section-level student average characteristics on instructor and section characteristics. All specifications also include year and month fixed effects. Robust standard errors clustered by instructor in parenthesis.

Table 3. How much switching is there between online and FTF campuses?

Number of MTH208 faculty by online and FTF participation

	Total FTF campuses taught at					
	0	1	2	3	4	Total
Never online	0	1,498	110	10	1	1,619
Taught online	534	126	14	3	0	677
Total	534	1,624	124	13	1	2,296

Table 4. Main Course Grade and Test Score Outcomes

	FTF and Online Combined		FTF only		Online only	
	Full sample	Test score sample	Full sample	Test score sample	Full sample	Test score sample
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Outcome = Standardized Course Grade						
SD(MTH208 effect)	0.2957	0.2794	0.3108	0.2920	0.2412	0.2259
	(.0058)	(.0079)	(.0069)	(.0094)	(.0085)	(.0126)
SD(MTH209 effect)	0.2490	0.2651	0.3069	0.3065	0.0790	0.0657
	(.0055)	(.008)	(.0071)	(.0099)	(.0052)	(.0109)
Corr (MTH208, MTH209)	0.5606	0.5749	0.7020	0.6967	-0.6427	-0.7031
	(.0212)	(.0301)	(.019)	(.0284)	(.0538)	(.124)
SD (Residual)	0.3510	0.3292	0.3285	0.3317	0.3711	0.3245
	(.0011)	(.0021)	(.0015)	(.0027)	(.0017)	(.0034)
Panel B. Outcome = Standardized Test Score						
SD(MTH208 effect)		0.4061		0.4399		0.1245
		(.0108)		(.0134)		(.0114)
SD(MTH209 effect)		0.4063		0.4673		0.1021
		(.011)		(.014)		(.0132)
Corr (MTH208, MTH209)		4.0000		0.6007		-0.2976
		(.0268)		(.0304)		(.1294)
SD (Residual)		0.4050		0.4274		0.3634
		(.0026)		(.0035)		(.0038)
Observations (sections)	26,304	7,271	13,749	4,711	12,555	2,560
Number of Instructors	2,249	1,203	1,716	944	676	292

Notes: Random effects models are estimated on section-level residuals. First stage models include instructor, campus, year, and month fixed effects in addition to individual controls, section average controls, and ZIP code controls. Residuals are taken with respect to all these variables other than instructor fixed effects. Individual controls include male, age, incoming GPA, incoming credits, indicator for repeat MTH208, number of times taking MTH208, 12 program dummies, years since started program. Section average controls include section averages of these same characteristics plust total enrollment in section. ZIP controls include the unemployment rate, median family income, percent of families below poverty line, percent of adults with BA degree in ZIP code from 2004-2007 ACS (plus missing ZIP). Robust standard errors clustered by instructor in parentheses.

Table 5. Robustness to Having Same Instructor for MTH208 and MTH209, FTF Sections

	All FTF sections		FTF Sections with < 25% same instructor		FTF sections with 0% same instructor	
	Full sample	Test score sample	Full sample	Test score sample	Full sample	Test score sample
	(3)	(4)	(3)	(4)	(3)	(4)
Panel A. Outcome = Standardized Course Grade						
SD(MTH208 effect)	0.3108	0.2920	0.2804	0.2565	0.2604	0.2334
	(.0069)	(.0094)	(.0165)	(.0261)	(.0174)	(.0304)
SD(MTH209 effect)	0.3069	0.3065	0.3235	0.3829	0.3266	0.3739
	(.0071)	(.0099)	(.0181)	(.0275)	(.0189)	(.0301)
Corr (MTH208, MTH209)	0.7020	0.6967	0.1606	0.2434	0.1836	0.2888
	(.019)	(.0284)	(.075)	(.1071)	(.0805)	(.1239)
SD (Residual)	0.3285	0.3317	0.4036	0.3757	0.4058	0.3881
	(.0015)	(.0027)	(.0071)	(.013)	(.0078)	(.0148)
Panel B. Outcome = Standardized Test Score						
SD(MTH208 effect)		0.4399		0.4224		0.3958
		(.0134)		(.0316)		(.0351)
SD(MTH209 effect)		0.4673		0.5247		0.4923
		(.014)		(.0322)		(.0353)
Corr (MTH208, MTH209)		0.6007		0.4230		0.4511
		(.0304)		(.0828)		(.0986)
SD (Residual)		0.4274		0.4826		0.4995
		(.0035)		(.0159)		(.0183)
Observations (sections)	13,749	4,711	1,587	574	1,403	514
Number of Instructors	1,716	944	806	372	764	352

Notes: Random effects models are estimated on section-level residuals. First stage models include instructor, campus, year, and month fixed effects in addition to individual controls, section average controls, and ZIP code controls. Residuals are taken with respect to all these variables other than instructor fixed effects. Individual controls include male, age, incoming GPA, incoming credits, indicator for repeat MTH208, number of times taking MTH208, 12 program dummies, years since started program. Section average controls include section averages of these same characteristics plust total enrollment in section. ZIP controls include the unemployment rate, median family income, percent of families below poverty line, percent of adults with BA degree in ZIP code from 2004-2007 ACS (plus missing ZIP). Robust standard errors clustered by instructor in parentheses.

Table 6. Robustness of Results to First-stage Model

	No instructor FE in first stage		Instructor FE included in first stage				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. All Sections (just test score sample)							
SD(MTH208 test effect)	0.2676	0.2534	0.2705	0.2707	0.2709	0.2927	0.4061
	(.0088)	(.0084)	(.0088)	(.0088)	(.0086)	(.0088)	(.0108)
SD(MTH209 test effect)	0.2625	0.2351	0.2663	0.2679	0.2642	0.2775	0.4063
	(.0089)	(.0083)	(.009)	(.009)	(.0087)	(.0087)	(.011)
Corr (MTH208, MTH209)	0.9480	0.9722	0.9256	0.9266	0.9231	0.9514	0.6094
	(.0168)	(.0157)	(.0178)	(.0176)	(.0171)	(.014)	(.0268)
SD (Residual)	0.4245	0.4075	0.4238	0.4231	0.4078	0.4067	0.4050
	(.0027)	(.0026)	(.0027)	(.0027)	(.0026)	(.0026)	(.0026)
Panel B. FTF Sections (just test score sample)							
SD(MTH208 test effect)	0.3101	**	0.3118	0.3126	0.3106	0.3129	0.4399
	(.0109)		(.0109)	(.0109)	(.0108)	(.0108)	(.0134)
SD(MTH209 test effect)	0.3090		0.3130	0.3142	0.3078	0.3055	0.4673
	(.0111)		(.0112)	(.0112)	(.0109)	(.0108)	(.014)
Corr (MTH208, MTH209)	0.9700		0.9597	0.9606	0.9701	0.9732	0.6007
	(.0163)		(.0168)	(.0166)	(.0158)	(.0156)	(.0304)
SD (Residual)	0.4357		0.4352	0.4347	0.4280	0.4277	0.4274
	(.0035)		(.0035)	(.0035)	(.0034)	(.0034)	(.0035)
Panel C. Online Sections (just test score sample)							
SD(MTH208 test effect)	0.0935	0.0992	0.1000	0.0998	0.1080	0.0992	0.1245
	(.0113)	(.0101)	(.0115)	(.0115)	(.0106)	(.0101)	(.0114)
SD(MTH209 test effect)	0.0760	0.0707	0.0854	0.0845	0.0822	0.0723	0.1021
	(.0145)	(.0126)	(.0144)	(.0143)	(.0126)	(.0125)	(.0132)
Corr (MTH208, MTH209)	0.2991	0.2342	0.0941	0.0918	-0.0058	0.2150	-0.2976
	(.1997)	(.1814)	(.182)	(.1824)	(.1625)	(.1791)	(.1294)
SD (Residual)	0.3971	0.3641	0.3957	0.3949	0.3637	0.3636	0.3634
	(.0041)	(.0038)	(.0041)	(.0041)	(.0038)	(.0038)	(.0038)
Controls in First Stage Model							
indiv controls	yes	yes	yes	yes	yes	yes	yes
zip controls	no	yes	no	yes	yes	yes	yes
section avg controls	yes	yes	yes	yes	yes	yes	yes
year FE, month FE	no	yes	no	no	yes	yes	yes
campus FE	no	yes	no	no	no	online only	yes

Notes: Indiv controls include male, age, incoming GPA, incoming credits, indicator for repeat MTH208, number of times taking MTH208, 12 program dummies, years since started program. Section average controls include section averages of these same characteristics plust total enrollment in section. ZIP controls include the unemployment rate, median family income, percent of families below poverty line, percent of adults with BA degree in ZIP code from 2004-2007 ACS (plus missing ZIP). Robust standard errors clustered by instructor in parentheses. ** Indicates that model failed to converge.

Table 7. Selection into Test Scores and MTH209

First stage model with full controls

Panel A. Instructor Effects on Selection into Test Scores and MTH209

	Have MTH208 test score	Take MTH209	Have MTH209 test score	Have MTH208 test score	Take MTH209	Have MTH209 test score	Have MTH208 test score	Take MTH209	Have MTH209 test score
	Overall			FTF Sections			Online Sections		
SD (instructor effect)	0.1475	0.0591	0.1419	0.1657	0.0695	0.1638	0.0553	0.0318	0.0153
	(.0039)	(.0029)	(.0041)	(.0047)	(.0035)	(.0052)	(.0046)	(.0044)	(.0066)
Correlation with MTH208 grade effect	0.1875	0.3911	0.3289	0.1640	0.3453	0.3358	0.3789	0.5327	0.3934
Correlation with MTH208 test score effect	-0.0839	0.0150	0.2820	-0.0949	-0.0237	0.2910	-0.0411	0.1475	0.1209

Panel B. Replace Missing Test Scores and MTH209 Grade With Zeros

	Adjusted Course Grade			Adjusted Standardized Test		
	All	FTF	Online	All	FTF	Online
	(2)	(4)	(6)			
SD(MTH208 effect)	0.29062	0.30021	0.24401	0.3122	0.3434	0.1032
	0.00776	0.00923	0.01277	(.0087)	(.0106)	(.0104)
SD(MTH209 effect)	0.2150	0.2499	0.0308	0.3967	0.4530	0.0803
	(.0068)	(.0087)	(.0121)	(.0104)	(.0129)	(.0124)
Corr (MTH208, MTH209)	0.6775	0.7396	0.5190	0.6835	0.6892	-0.5695
	(.0256)	(.026)	(.2459)	(.0241)	(.0262)	(.1557)

Notes: Random effects models are estimated on section-level residuals. First stage models include instructor, campus, year, and month fixed effects in addition to individual controls, section average controls, and ZIP code controls. Residuals are taken with respect to all these variables other than instructor fixed effects. Individual controls include male, age, incoming GPA, incoming credits, indicator for repeat MTH208, number of times taking MTH208, 12 program dummies, years since started program. Section average controls include section averages of these same characteristics plust total enrollment in section. ZIP controls include the unemployment rate, median family income, percent of families below poverty line, percent of adults with BA degree in ZIP code from 2004-2007 ACS (plus missing ZIP). Robust standard errors clustered by instructor in parentheses.

Table 8. Instructor Effects for Alternative Outcomes

First stage model with full controls

	Outcome		
	Pass MTH208	Take MTH209	Credits earned 6mo
Panel A. Full Sample			
SD (instructor effect) - overall	0.0732	0.0507	0.1265
	(.0017)	(.0017)	(.0038)
SD instructor effect - FTF	0.0805	0.0625	0.1552
	(.0021)	(.0022)	(.0051)
SD instructor effect - online	0.0586	0.0309	0.0591
	(.0023)	(.0022)	(.0045)
Panel B. Test Score Sample			
SD (instructor effect) - overall	0.0727	0.0591	0.1323
	(.0024)	(.0029)	(.0062)
SD instructor effect - FTF	0.0780	0.0694	0.1529
	(.003)	(.0035)	(.0076)
SD instructor effect - online	0.0564	0.0319	0.0399
	(.0036)	(.0044)	(.0114)

Notes: Random effects models are estimated on section-level residuals. First stage models include instructor, campus, year, and month fixed effects in addition to individual controls, section average controls, and ZIP code controls. Residuals are taken with respect to all these variables other than instructor fixed effects. Individual controls include male, age, incoming GPA, incoming credits, indicator for repeat MTH208, number of times taking MTH208, 12 program dummies, years since started program. Section average controls include section averages of these same characteristics plust total

Table 9. Correlation across Outcomes (Restricted to Test Sample)

First stage model with full controls

All sections, restricted to test sample (N =7048 sections)

	Test MTH208	Test MTH209	Grade MTH208	Grade MTH209	Credits earned 6mo	Pass MTH208	Take MTH209
Test MTH208	1.00						
Test MTH209	0.55	1.00					
Grade MTH208	0.52	0.23	1.00				
Grade MTH209	0.34	0.39	0.41	1.00			
Credits earned 6mo	0.05	0.03	0.41	0.29	1.00		
Pass MTH208	0.30	0.10	0.83	0.25	0.54	1.00	
Take MTH209	0.02	-0.02	0.39	0.17	0.52	0.53	1.00

FTF sections, restricted to test sample (N = 4531 sections)

	Test MTH208	Test MTH209	Grade MTH208	Grade MTH209	Credits earned 6mo	Pass MTH208	Take MTH209
Test MTH208	1.00						
Test MTH209	0.57	1.00					
Grade MTH208	0.55	0.29	1.00				
Grade MTH209	0.42	0.39	0.59	1.00			
Credits earned 6mo	0.14	0.07	0.46	0.32	1.00		
Pass MTH208	0.30	0.15	0.79	0.40	0.60	1.00	
Take MTH209	-0.02	0.00	0.35	0.23	0.51	0.51	1.00

Online sections, restricted to test sample (N = 2517 sections)

	Test MTH208	Test MTH209	Grade MTH208	Grade MTH209	Credits earned 6mo	Pass MTH208	Take MTH209
Test MTH208	1.00						
Test MTH209	0.09	1.00					
Grade MTH208	0.29	-0.32	1.00				
Grade MTH209	0.01	0.53	-0.37	1.00			
Credits earned 6mo	0.12	-0.16	0.54	-0.01	1.00		
Pass MTH208	0.17	-0.35	0.91	-0.41	0.62	1.00	
Take MTH209	0.15	-0.25	0.53	-0.22	0.66	0.59	1.00

Table 10. Correlates of Instructor Effectiveness
First stage model with full controls
All sections, faculty hired since 2002

	Outcome: Section-level mean residual for				
	MTH208 grade	MTH209 grade	MTH208 test	MTH209 test	Credits earned 6 months
	(1)	(2)	(3)	(4)	(5)
A. Linear, Only MTH208 Experience, Instructor FEs					
Taught MTH208 previously	0.0385***	0.0210	0.0284	0.00911	-0.0167
	(0.0108)	(0.0129)	(0.0311)	(0.0374)	(0.0104)
Times taught MTH208	0.00001	-0.000268	-0.000155	-0.00572	0.000541
	(0.0008)	(0.0007)	(0.0039)	(0.0039)	(0.0006)
B. Piecewise, Only MTH208 Experience, Instructor FEs					
Times taught MTH208 = 1	0.0309**	0.013	0.0254	0.000672	0.00000
	(0.0121)	(0.0149)	(0.0332)	(0.0408)	(0.0121)
Times taught MTH208 = 2 to 5	0.0413***	0.0251*	0.0382	0.0418	-0.0198*
	(0.0121)	(0.0147)	(0.0364)	(0.0412)	(0.0114)
Times taught MTH208 = 6 to 10	0.0404***	0.0163	0.101**	-0.00128	-0.00584
	(0.0156)	(0.0180)	(0.0483)	(0.0507)	(0.0140)
Times taught MTH208 = 11 to 15	0.0412**	0.0143	0.114**	0.0576	-0.00166
	(0.0200)	(0.0211)	(0.0578)	(0.0598)	(0.0169)
Times taught MTH208 = 16 to 20	0.0393*	-0.00229	0.127*	0.0694	0.0177
	(0.0236)	(0.0240)	(0.0688)	(0.0713)	(0.0191)
Times taught MTH208 > 20	0.0342	0.00409	0.135*	0.0809	0.0425*
	(0.0278)	(0.0282)	(0.0768)	(0.0795)	(0.0225)
C. Linear, Control for MTH209, other math, non-math experience linearly, Instructor FEs					
Taught MTH208 previously	0.0295**	0.0198	0.0633	-0.00178	-0.0267**
	(0.0133)	(0.0148)	(0.0437)	(0.0516)	(0.0116)
Times taught MTH208	0.000152	-0.000201	0.00021	-0.00491	0.00088
	(0.0008)	(0.0008)	(0.0038)	(0.0045)	(0.0006)

Notes: Section mean residuals are regressed on teaching experience, instructor fixed effects, and year and month fixed effects. Sample restricted to 18,418 sections (5860 for test scores) taught by instructors hired since 2002. First stage model include instructor, campus, year, and month fixed effects in addition to individual controls, section average controls, and ZIP code controls. Residuals are taken with respect to all these variables other than instructor fixed effects. Individual controls include male, age, incoming GPA, incoming credits, indicator for repeat MTH208, number of times taking MTH208, 12 program dummies, years since started program. Section average controls include section averages of these same characteristics plust total enrollment in section. ZIP controls include the unemployment rate, median family income, percent of families below poverty line, percent of adults with BA degree in ZIP code from 2004-2007 ACS (plus missing ZIP). Robust standard errors clustered by instructor in parentheses.

Table 11. Correlates of Instructor Effectiveness

All sections, faculty hired since 2002

	Total Salary Paid for MTH208 Section ($1,000) (mean = 1.077)					
	(1)	(2)	(3)	(4)	(5)	(6)
Total sections taught previously	0.00092***	0.00086***	0.00002			
	(0.0002)	(0.0002)	(0.0001)			
Taught MTH208 previously				0.01657***	0.01331***	0.00392
				(0.0036)	(0.0037)	(0.0036)
Times taught MTH208				0.00021	-0.00012	-0.00049*
				(0.0003)	(0.0003)	(0.0003)
Times taught MTH209				0.00035	0.00026	0.00024
				(0.0003)	(0.0003)	(0.0003)
Times taught other math courses				0.00070**	0.00065*	-0.00015
				(0.0004)	(0.0004)	(0.0003)
Times taught nonmath courses				0.00085***	0.00085***	0.0001
				(0.0001)	(0.0001)	(0.0002)
Years since first hire date	0.02542***	0.02316***	0.00643*	0.02431***	0.02307***	0.00701*
	(0.0017)	(0.0017)	(0.0038)	(0.0021)	(0.0021)	(0.0039)
First hire more than one year ago	0.00201	0.00028	0.0059	-0.00181	-0.00273	0.00495
	(0.0041)	(0.0039)	(0.0037)	(0.0042)	(0.0039)	(0.0038)
Fixed effects	None	Campus	Instructor	None	Campus	Instructor

Notes: Sample restricted to 18,418 sections taught by instructors hired since 2002.All specifications also include year and month fixed effects. Robust standard errors clustered by instructor in parentheses.

Appendix Table A1. Descriptive Statistics (Restricted Sample)

	All		Face-to-Face		Online	
	N	Mean	N	Mean	N	Mean
Section and Instructor Characteristics						
Online section	7,276	0.350	4,716		2,560	
Male	5,430	0.680	3,386	0.700	2,044	0.660
White	4,371	0.640	2,551	0.630	1,820	0.650
Instructor Compensation per Section ($)	7,276	950.77	4,716	933.13	2,560	983.25
Instructor Teaching Load (Sections) during calendar year	7,276	9.30	4,716	9.79	2,560	8.39
Instructor Teaching Load (Math Sections) during calendar year	7,276	3.92	4,716	3.32	2,560	5.04
Years since first hire	7,276	6.27	4,716	5.90	2,560	6.94
# sections instructor taught in past year	7,276	8.48	4,716	8.43	2,560	8.57
# Math 208 sections the instructor taught in the past year	7,276	3.36	4,716	2.54	2,560	4.86
Total sections instructor taught prior to this section	6,275	45.77	3,850	38.97	2,425	56.58
Total math sections instructor taught prior to this section	6,275	32.77	3,850	22.93	2,425	48.39
Share of prior sections taught that were math courses	6,275	0.790	3,850	0.720	2,425	0.900
Share of prior sections taught that were business courses	6,275	0.040	3,850	0.050	2,425	0.020
Share of prior sections taught that were other courses	6,275	0.170	3,850	0.230	2,425	0.070
Student Background Characteristics						
Male	94,461	0.380	59,720	0.420	34,741	0.320
Age	94,698	34.32	59,792	33.57	34,906	35.60
Baseline GPA (0-4)	94,811	3.21	59,853	3.19	34,958	3.23
Credits earned prior to start of Math 208	94,811	24.53	59,853	25.26	34,958	23.29
Took Math 208 before	94,811	0.100	59,853	0.090	34,958	0.130
BS in Business	94,811	0.380	59,853	0.470	34,958	0.240
BS (general studies)	94,811	0.160	59,853	0.160	34,958	0.170
BS in Nursing	94,811	0.040	59,853	0.020	34,958	0.090
BS in Management	94,811	0.060	59,853	0.030	34,958	0.100
BS in Criminal Justice Administration	94,811	0.100	59,853	0.120	34,958	0.060
BS in Health Administration	94,811	0.090	59,853	0.090	34,958	0.090
BS in Human Services	94,811	0.040	59,853	0.040	34,958	0.060
BS in Information Technology	94,811	0.040	59,853	0.050	34,958	0.040
BS in Education	94,811	0.030	59,853	0.010	34,958	0.050
Outcomes						
Performance in Math 208						
A / A-	94,811	0.280	59,853	0.280	34,958	0.300
B+ / B / B-	94,811	0.280	59,853	0.280	34,958	0.270
C+ / C / C-	94,811	0.190	59,853	0.200	34,958	0.170
D+ / D / D-	94,811	0.090	59,853	0.100	34,958	0.080
F	94,811	0.050	59,853	0.050	34,958	0.050
Withdrawn	94,811	0.110	59,853	0.090	34,958	0.140
Passed Math 208	94,811	0.840	59,853	0.860	34,958	0.810
Final exam score available	94,811	0.850	59,853	0.890	34,958	0.780
Performance following Math 208						
Took Math 209	94,811	0.780	59,853	0.830	34,958	0.690
Days before taking Math 209 (Median)	73,014	7	49,423	7	23,591	1
Credits earned in following year	94,811	18.53	59,853	19.77	34,958	16.41

Table A2. First Stage Results for Full Sample

	Outcome: MTH 208 Grade			Outcome: MTH209 Grade		
	(1)	(2)	(3)	(4)	(5)	(6)
male	0.0911***	0.0849***	0.0849***	0.0760***	0.0711***	0.0710***
	(0.00375)	(0.00374)	(0.00374)	(0.00450)	(0.00448)	(0.00448)
age	-0.0116***	-0.0116***	-0.0116***	-0.000873***	-0.000864***	-0.000869***
	(0.000264)	(0.000265)	(0.000265)	(0.000292)	(0.000292)	(0.000292)
incoming_gpa	0.751***	0.739***	0.740***	0.739***	0.730***	0.730***
	(0.00480)	(0.00479)	(0.00479)	(0.00567)	(0.00566)	(0.00565)
incoming_credits	0.00342***	0.00335***	0.00335***	0.00203***	0.00196***	0.00196***
	(0.000122)	(0.000121)	(0.000121)	(0.000142)	(0.000141)	(0.000141)
repeat_MTH208	-0.0915***	-0.0949***	-0.0950***	-0.117***	-0.122***	-0.122***
	(0.0114)	(0.0113)	(0.0113)	(0.0172)	(0.0171)	(0.0171)
N_taken_MTH208	-0.172***	-0.168***	-0.168***	-0.175***	-0.170***	-0.170***
	(0.00904)	(0.00900)	(0.00900)	(0.0143)	(0.0142)	(0.0142)
years_since_pstart	0.000480	-0.000468	-0.000449	0.00486***	0.00438***	0.00437***
	(0.00140)	(0.00140)	(0.00140)	(0.00167)	(0.00167)	(0.00167)
savg_age	0.00805***	0.00548***	0.00554***	0.00450***	0.00382***	0.00377***
	(0.000850)	(0.000819)	(0.000811)	(0.000944)	(0.000914)	(0.000899)
savg_male	0.0493***	0.0131	0.00875	0.0610***	0.0225	0.0173
	(0.0142)	(0.0137)	(0.0137)	(0.0176)	(0.0168)	(0.0168)
savg_gpa	-0.00397	-0.0163	-0.0191	0.0589***	-0.0124	-0.0115
	(0.0150)	(0.0139)	(0.0140)	(0.0162)	(0.0160)	(0.0160)
savg_credits	0.000418	0.000631	0.000617	-0.000239	0.000319	0.000356
	(0.000403)	(0.000403)	(0.000406)	(0.000501)	(0.000491)	(0.000492)
savg_repeat	-0.0130	-0.00105	0.000528	0.0387	0.0562	0.0532
	(0.0427)	(0.0410)	(0.0409)	(0.0514)	(0.0492)	(0.0490)
savg_ntaken	-0.0142	0.0128	0.0101	-0.101**	-0.0611	-0.0584
	(0.0356)	(0.0342)	(0.0341)	(0.0409)	(0.0404)	(0.0402)
savg_pstart	-0.0173***	-0.00743	-0.00809	-0.000746	0.0122*	0.0109*
	(0.00587)	(0.00556)	(0.00555)	(0.00658)	(0.00652)	(0.00658)
enrollment3	1.27e-05	-0.000582	-0.000532	-0.000687	-0.000854	-0.000951
	(0.000607)	(0.000576)	(0.000572)	(0.000650)	(0.000617)	(0.000614)
zip_punemp		-0.593***	-0.570***		-0.331***	-0.312***
		(0.0648)	(0.0634)		(0.0767)	(0.0749)
zip_medfamy		0.0133***	0.0130***		0.00901***	0.00900***
		(0.00158)	(0.00158)		(0.00192)	(0.00193)
zip_fambpl		-0.373***	-0.373***		-0.326***	-0.328***
		(0.0387)	(0.0387)		(0.0470)	(0.0473)
zip_perdeg_bdh		-0.181***	-0.176***		-0.0673***	-0.0652**
		(0.0217)	(0.0215)		(0.0261)	(0.0260)
online		0.0305			-0.228***	
		(0.0235)			(0.0297)	
Constant	-2.185***	-1.859***	-1.947***	-2.476***	-1.982***	-2.190***
	(0.0701)	(0.0723)	(0.107)	(0.0725)	(0.0766)	(0.124)
Basic Controls	yes	yes	yes	yes	yes	yes
zip controls	no	yes	yes	no	yes	yes
year FE	no	yes	yes	no	yes	yes
campus FE	no	online only	yes	no	online only	yes
Observations	337516	337516	337516	251618	251618	251618
R-squared	0.206	0.215	0.215	0.168	0.173	0.174

Appendix – test score data and identification.

In order to identify the test scores, we use transcript records from the University of Phoenix for 99,406 students in 9,033 MTH/208 sections, and 104,281 students in 11,776 MTH/209 sections. The data provides an overview of the different components that make up the test score. In particular, we have the unique student and section identifier that allow us to match the grade components to the other data files, an attendance report (that is constant for a student within a section), and a string describing the grade component (the title). For each component, the maximal grade attainable and the actual attained grade are available.

The grade components are not identical across sections, as instructors have discretion over several outcomes. First, they are free to use the online materials that are provided and available to all instructors and add extra assignments, homework, or tests. Second, they can adjust the weights that are given to these different components. Third, and finally, they are also free to change the titles of the course components. Going through the data reveals that instructors exercise this discretion: there are many different string describing course components, exams are graded on many different scales, and some sections have grade components indicating instructors include take-home or written exams.

Therefore, a decision rule ideally identifies computer-administered exams that are made available to all instructors, since these are standardized across instructors. Given the differences in scores and titles, the decision rule we use is based on a variety of string combinations. As some string combinations provide a cleaner identification of computer-administered tests than other ones, we include a quality measure that indicates the reliability of the procedure.

The table below gives an overview of the decision rule that was used and indicates the associated quality measure. Quality measure 1 identifies a combination of different IT systems that were used in UPX (Aleks and MyMathLab) and words that indicate a final exam or test. The second quality measure gets at titles that indicate a final exam, while the third quality measure looks for strings that are related with final quizzes or similar. Finally, the fourth quality measure tries to identify in-class finals.

Every quality measure consists of several steps. For instance, quality measure 1 consists of two decision rules. These rules are hierarchical: if a student in a section gets assigned a test score in the very first step, this observation is marked and not considered for any of the next steps. This ensures that the test score identifications of good quality are not contaminated with lesser quality ones.

After applying this set of rules, the fraction of students with MTH/208 test with quality measure 1, 2, 3 or 4 is 1%, 80%, 8%, and 3% respectively. For MTH/209 grades, the fractions are 0.5%, 79%, 8%, and 3% respectively. About 8% of MTH/208 students and 10% of MTH/209 students don’t get assigned any test score. Going through these cases reveals two cases. First, about two thirds of these students withdrew from the section, providing a possible explanation of why no test score could be found. Second, going through the remainder of these cases shows test score components where there are no clear indications of what the final test score would be.

The final sample covers 81,162 student-sections in MTH/208 sections and 67,045 student-sections in MTH/209 sections. The MTH/208 final exam sample covers 78,865 students and 1,204 instructors in 7,158 sections. The MTH/209 final exam sample covers 62,429 students and 1,474 instructors in 9,183 sections.

Quality	**Decision rule**
1	1. The title contains: “mml final”, “final mml exam”, “final in mathlab”, “mymathlab final”, “my math lab final”, “mathlab final”, “aleks final”, “aleks quiz – final”, “alek final”, “final (aleks)”, “final (mml)”, “exam (mml)”, “final in mymathlab”, “online final” 2. The title contains at least 1 word of each of the following 2 lists: a. “exam”, “final”, “test” b. “mathlab”, “econlab”, “econ lab”, “online”, “math lab”, “mml”, “alek”
2	1. The title equals: “final exam” or “final examination” 2. The title contains: “final exam”, “final ex”, “fnl ex”, “fianl ex”, “fnial ex”
3	1. Title contains “final test” 2. Title equals “final” 3. Title equals “exam” 4. Title contains “final quiz” 5. Title contains “course exam” 6. Title equals “test” or contains any of the following strings”: “final – exam”, “quiz5/final”, “individual final”, “final~individual”, “finals”, “test”, “wk5- final”
4	1. Title contains both “class” and “final”