The Dissertation Committee for Nathaniel Raley Woodward Certifies that this is the approved version of the following Dissertation:

Educational Practices in Large College Courses and Their Effects on Student Outcomes

Committee:

Germine H. Awad, Supervisor

Andrew C. Butler, Co-Supervisor

Susan Natasha Beretvas

Veronica X. Yan

Educational Practices in Large College Courses and Their Effects on Student Outcomes

by

Nathaniel Raley Woodward

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2018

Abstract

Educational Practices in Large College Courses and Their Effects on Student Outcomes

Nathaniel Raley Woodward, Ph.D.

The University of Texas at Austin, 2018

Supervisor: Germine H. Awad

Co-Supervisor: Andrew C. Butler

Part I of this study presents a large-scale characterization of normative educational practices (e.g., course structure, teaching methods, learning activities) across more than 1,000 high-enrollment undergraduate courses at a large public institution over the last 5 years. I assess the extent to which course features reflect educational best-practices by systematically reviewing course syllabi—documenting the type, quantity, and grade-weight of all work assigned in each course as well as the prevalence and variability of teaching practices such as group activities, retrieval practice, and in-class active learning. I assess the degree to which these variables have changed over time, how they differ across colleges, and whether they form distinct clusters.

I also analyze language used in the syllabus to see how instructors communicate information to students. I examine pronouns, comparisons, negations, and words related to achievement versus affiliation; I isolate words that unique to certain syllabi, courses,

departments, and colleges; and I look at how similar two syllabi from the same course are on average.

Findings revealed that high stakes exams are the norm, active learning is relatively uncommon, and students get few opportunities for spaced retrieval practice. Importantly, it was found that no one college has a monopoly on educational best-practices; different colleges had different strengths. Trends over time were mostly positive, indicating an increase in adoption of many best-practices, with a few exceptions.

Part II of this study builds directly upon Part I by combining the syllabus dataset with student records to assess how prerequisite-course features affect student performance in their subsequent courses. Specifically, introductory courses high and low in retrieval-practice requirements were compared using inverse propensity-score weighted regressions to improve causal inference. Results showed that additional retrieval practice improved students' performance in their subsequent courses. However, the average treatment effect estimates were small and somewhat sensitive to variations in model. Finally, subsequent-course performance was regressed on the full set of educational relevant variables using lasso regression, identifying several variables related to retrieval practice and spacing (including number of quizzes and cumulative exams) as important correlates of student success.

Acknowledgements

You know, I wasn't even going to include this section, but here I am at the end of it all, right before the deadline, completely overcome with emotion. First, to thank my true love Lindsay for sharing her life with me. This is just a customary gesture: She knows how essential she is to everything that I do (*et vice versa*), and I know she knows (and she knows I know), and she knows I know she knows (and I know she knows I know), *recursio ad infinitum* (because I think that's what love is)! To my mother, my sister, and my in-laws, pride of place and special thanks are due. You have encouraged and supported me every step of the way; I love you all so very much.

Now, to thank Andrew Butler. I am extraordinarily fortunate to have been advised by Andy: I have benefitted in countless ways from his example as a scholar and his guidance as a mentor, to say nothing of the many opportunities he has gone out of his way to create for me. Though we were brought together through sheer happenstance, it often feels like destiny, and I am privileged to be your first doctoral student advisee.

Finally, to thank my committee. Gigi and Tasha, you are two of the most inspiring and productive people I have ever met, and I am honored that you have given so much of your time to help me with this and various other projects over the years. Veronica, it has been so fun getting to know you these last few months! Thank you for welcoming me into your lab: I love your research and all the passion and energy you bring to it. I am so excited to see what the future holds for HDCLS, and I know that the "LS" will shine brighter than it ever has before!

Over the past five years I have been shaped personally and professionally by many wonderful people to whom I owe an immense debt of gratitude. Thank you to Marilla

Svinicki, who had enough faith in my graduate school application to invite me out to Austin to work with her, sight unseen. You brought me to academia's threshold and you welcomed me across: Because of you, I took perhaps the most important step of my life. Thank you also to Diane Schallert for being the lifeblood of the department and an inspiration to us all. I knew from the first day of Psych of Learning course that I had come to the right place.

Thank you to Jane Huk, Jamie Pennebaker, Toni Wegner, Stephanie Corliss, Siew Ang and everyone else at Project 2021. This dissertation would never have gotten off the ground without your support; because of your time, your trust, and your help, it has soared to great heights! Thank you to Vicki Keller and Michael Mahometa for helping me navigate a new department and for giving me a second home here on campus. I have relished working with you over the years and I am thrilled to be joining the Department of Statistics and Data Sciences in a faculty role. Thank you to James Pustejovsky for teaching the best courses. I have been in school for quite some time now (wow: 22 years!) and taking classes with you has been a highlight of my educational experience.

Thank you to my cohort, my classmates, my colleagues (especially Lisi Wang and Cynthia Alarcon), my students, and all the wonderful people I have met at the University of Texas over these past 5 years. Special shout-out to Longhorn Quiz Bowl with whom I've had many a fun season! I might as well thank the whole university while I'm at it: From the world-class faculty and amazing campus to little perquisites like free city bus fare, tubes at the Kickstand, cheap coffee (I see you Jester Java), and good coffee (Coffee Traders!), I am so grateful to this institution. I watched from my perch in Sanchez and the PCL as Speedway turned into the yellow-brick road twice over (repaved due to cracky masonry) and Ellsworth Kelly's colorful chapel grew up bit by bit. Let's hear it for Tower Girl, albino squirrels, Bevo XV (RIP, Bevo XIV) and the FAC cat (Domino)! I really feel like such a part of this campus now, and I am convinced that there is no place quite like it on earth.

Table of Contents

List of Tables

List of Figures

PART I: SYLLABUS REVIEW

Chapter One: Introduction

It has been more than ten years since the U.S. Department of Education's Commission on the Future of Higher Education highlighted the inadequate educational outcomes of college graduates across the country and called for greater transparency among institutions of higher learning (Spellings, 2006). However, despite this growing emphasis on accountability—particularly in the form of student gains on certain proficiency measures after completing a year or more of college—the pedagogical features of college courses that shape students' learning experiences remain largely ignored or unreported. Though more than 20 million students are enrolled in undergraduate programs in America today (NCES, 2016), very little can be said about the type of course schedules, teaching practices, classroom activities, and out-of-class assignments that constitute a typical college course, to say nothing of how these factors vary within and between departments or across universities. That the college classroom has remained something of an educational black box is a concern not just for students and parents, but also for academia at large: a lack of public understanding about what takes place in undergraduate courses and can produce potentially misleading generalizations about the standard college classroom experience (e.g., a *lecture-then-test* format and an instructor who acts as a *sage-on-the-stage*; King, 2010).

This dearth of information notwithstanding, we do know that instructional practices have arisen somewhat organically within disciplines, and that this arrangement yields certain benefits. Professors spend years becoming content experts in a specific field of study and by virtue of this experience they are poised to understand how to teach this material to others. The most effective ways of teaching specific subject matter—known as

pedagogical content knowledge—are often tacit, picked up from the approaches of their own teachers, the difficulties they themselves encountered and surmounted in their own deep study of their discipline, and, to the extent that they have teaching experience, their own practice in teaching the material (Shulman, 2013).

While this may seem to be a fine state of affairs, it is important to attend to what is missing from it. Shulman notes that "since faculty members in higher education rarely receive direct preparation to teach, they most often model their own teaching after that which they themselves received" (2005, p. 57). Because college professors typically receive little if any formal training as educators, they often lack a scientific understanding of human memory, transfer of learning, and how to design their courses in order to promote these goals (Halpern & Hakel, 2003). Furthermore, professors usually receive no formal evaluations of their teaching effectiveness beyond student ratings course evaluations, a problematic metric that may even reward certain teaching practices that undermine educational outcomes like long-term retention and transfer (Shevlin, Banyard, Davies, & Griffiths, 2010; Stroebe, 2016; Bjork & Bjork, 2011). Overall, this lack of substantive feedback leaves instructors with little actionable information about how to improve their courses and little incentive to adopt evidence-based practices, especially when those practices create more work for themselves and more challenges for their students.

If the primary goal of higher education is to teach students knowledge and skills that remain accessible to them over time and that can be flexibly applied outside of the classroom, then instruction should be designed with these goals foremost in mind. Fortunately, cognitive and educational psychologists have produced a large body of research on learning and memory, and well-replicated findings have come together to yield robust principles about how to facilitate desired student outcomes such as long-term retention, transfer of learning, metacognitive skills, and motivation. These include things

like retrieval practice, spaced/distributed learning, in-class active learning, learning from peers, setting clear learning objectives and high expectations, real-world problem solving, teacher-student relationships, and motivational factors like providing students choice (see, e.g., Dunlosky et al. 2013; Hattie, 2008; Winne & Nesbit, 2010). The extent to which stated teaching practices, assignments, course structure, activities, and other information appearing in course syllabi accord with these recommendations provides evidence for their use in the classroom; at a minimum, their presence in the syllabus suggests some knowledge of these best-practices on the part of the instructor, some consideration of their use in the classroom, and a stated intention to implement them.

To address these concerns and to provide a window into what goes on in large college courses, Part I of this study presents a large-scale characterization of normative educational practices—course structure, teaching methods, learning activities, and other educationally relevant variables—across more than a thousand high-enrollment (200+ student) undergraduate courses at a large public university over the last 5 years. I assess the extent to which these course features reflect educational best-practices drawn from research on teaching and learning by using course syllabi to document the type, quantity, frequency, and grade-weight of all work assigned to students in a given course, including in-class activities, quizzes, exams, and homework. I also document prevalence and variability of evidence-based learning activities and teaching practices such as cumulative assessments, group assignments, retrieval practice, and in-class active learning. Beyond accounting for structural features of a course such as these, which can be taken more or less directly from the syllabus, I analyze the language used in the syllabus, the size and completeness of the syllabus, and the presence or absence of stated student learning objectives to gain further insight into instructors' manner of communicating with students. I explore not only high-level summaries such as the averages of these variables across all

courses, but also how these features vary within and between colleges and departments, how they vary across different versions of the same course, the degree to which they have changed over time, and whether they differ based on course format.

As Part I of this study is descriptive in nature, no formal hypotheses are presented or tested: The aim of Part I is to characterize what goes in large college courses in terms of course structure, teaching methods, and learning activities. As such, I report overall counts and percentages, relying heavily on graphs to illustrate variability among colleges and departments. Furthermore, to give structure to the many dimensions under consideration and to aid in the presentation of results, three broad categories of variables will be used: (a) the nature, number, and grade weight of all *course work* completed by students in the course; (b) general *pedagogical approaches* used in the course; and (c) instructor *communication* to students. In addition to these categories, interrelationships among all variables and courses are examined, as are changes in these variables over time.

Part II of this study builds directly upon the work presented in Part I by exploring how these variables can impact students' learning and performance in their subsequence coursework.

EVIDENCE-BASED TEACHING PRACTICES FOR PROMOTING STUDENT LEARNING

Two of the most useful approaches for promoting long-term retention and transfer of learning are *the testing effect* (practice recalling information from memory; see Roediger & Butler, 2011 for review) and *the spacing effect* (spacing one's learning out over time rather than cramming it into a single session; see Cepeda et al. 2006 for review). The generality and effectiveness of these and other techniques emerge from meta-analyses looking at the average performance gains across many individual studies (e.g., Hattie,

2008). For an in-depth review of these techniques and their educational efficacy, see Chapter 5 in Part II of this document.

These techniques are increasingly well known, appearing as recommended best-practices in most modern textbooks about college teaching. For example, at the beginning of the first chapter of *Teaching At Its Best: A Research-Based Resource for College Instructors,* Nilson (2010) enumerates key teaching principles including spaced retrieval practice: “Build into your course plenty of assessment opportunities, including low-stakes quizzes, practice tests, in-class exercises, and homework assignments that can tell students how much they are really learning, as well as provide them with retrieval practice" (Nilson, 2010, p. 5). Indeed, representative textbooks by Nilson (2010) and by Svinicki and McKeachie (2013) both have sections devoted to facilitating in-class active learning, providing frequent assessments, giving timely feedback, setting explicit expectations of students that pose reasonable challenges, holding students accountable for their work, and teaching students real-world problem solving. As we discuss below, the course syllabus can provide valuable information about the extent to which these evidence-based practices are being taken into account in college course design.

DETERMINING EDUCATIONAL PRACTICES FROM COURSE SYLLABI

Research continues to show that the work students do for a given class (lectures, readings, assignments, activities, etc.) is more important for their learning outcomes than are professor-level variables such as charisma or teaching experience (e.g., Deslauriers, Schelew, & Wieman, 2011). Unfortunately, very few methods exist for collecting this kind of data, despite increasing demand for it from high places, such as the National Academies of Science, Engineering, and Medicine (2018).

Indeed, most colleges and universities do not measure teaching practices at all (Wieman & Gilbert, 2014); the only teaching-related data that is routinely collected are student evaluations on end-of-course surveys and evaluations, but even these are of questionable value and may even reward ineffective teaching practices (Stroebe, 2016). A more informative method is the classroom observation, and several protocols for its use in undergraduate education have been developed. However, observations are inherently limited—often to what goes on in a single day's course—and activities and teaching practices can vary greatly from day to day. Furthermore, they capture only elements of teaching that can be observed by watching a class in session (and are thus relatively silent about things like course schedules, grading policies, and out-of-class assignments). Finally, course observations are time and resource intensive, requiring trained observers to accurately characterize hours of class time.

To help address the lack of sound and efficient methodological approaches for assessing teaching in undergraduate education, Wieman and Gilbert (2014) developed the 72-item Teaching Practices Inventory (TPI), a self-report measure which instructors can use to evaluate their own teaching. These items span eight categories: the course information provided (including learning objectives), supporting material provided, in-class features and activities, assignments, feedback and testing, training and guidance of TAs, collaboration or sharing in teaching, and a general "other" category, which are mapped in various ways to learning best-practices. This sort of approach has many advantages, including ease of administration and scoring, implicit feedback to instructors about what they are doing and how they could improve, and richer data about what goes on in a course (e.g., "Do teaching assistants receive one-half day or more of training in teaching?"). The chief disadvantage is the self-report nature of the method, which could be especially problematic if it is perceived as going in one's tenure file, or really having any

professional stakes at all. The human tendency to remember one's own words and deeds (and teaching) in a personally favorable light is one to which professors are not immune; the more subjectivity is allowed and the higher the perceived stakes, the more room there is for these self-serving and social-desirability biases to skew results (e.g., Donaldson & Grant-Vallone, 2002; Williams, Walter, Henderson, & Breach, 2015).

In their article, Wieman and Gilbert (2014) note that "the TPI is not inherently a self-reporting instrument. In most cases, it is easy for another person to determine the correct responses by looking at course materials and instructor class notes" (p. 555). This suggestion was the impetus for our approach to the problem of measuring educational best-practices that bypasses self-report: using the course syllabus to document course-level variables and to infer teaching practices. As will be discussed below, public institutions of higher education in Texas are required by state law to make every undergraduate course syllabus (and instructor curriculum vitae) available to the public online through the institution's website (H.B. 2504, 2009), which makes the accessibility of such materials a non-issue.

To the extent that course syllabi provide an accurate record of what students do in a given course (e.g., the activities and assignments that students must complete), they can yield important insights into the quality of course pedagogy. Thus, it is incumbent on me to discuss the purpose of the syllabus in higher education, to provide examples of how syllabus data has been used in previous research, and to address the validity of such an approach. This approach is based on the notion that teaching practices and other course-level variables are a more accurate proxy for educational effectiveness than anything else that can be practically measured in undergraduate education (Wieman, 2015).

Instructors write syllabi to convey the content and organization of their course to students and to serve as a written record of its important features. They describe the

sequence of topics that will be covered, the assignments and assessments that students will be accountable for, and the criteria according to which students' grades will be determined. Increasingly, the syllabus has taken on the status of a legally binding document that stipulates all the conditions of the course that must be met in order to achieve a certain grade—to the degree that a popular handbook for college teaching, currently in its 14th edition and cited many thousands of times, plainly states "the syllabus is a contract between you and your students" (Svinicki & McKeachie, 2013, pg. 24). The syllabus is a key part of good course-design in undergraduate education (e.g., Svinicki & McKeachy, 2013; Nilson, 2010) and, as it turns out, good instructors tend to have good syllabi. Lough (1997) surveyed syllabi from courses given by winners of the prestigious Carnegie Professor of the Year Award, a program created to recognize outstanding undergraduate instructors. He found that almost all of these syllabi contained well-defined course objectives, detailed schedules for assignments, and descriptions of grading procedures; most also included expectations for class participation and suggestions for how to study in order to be successful. Though some syllabus content may be constrained by institutional requirements (e.g., course name, instructor contact information, office hours; see Figure 1) or influenced by departmental convention, instructors often have as much control over what appears in their syllabus as they do over what goes on in the course itself.

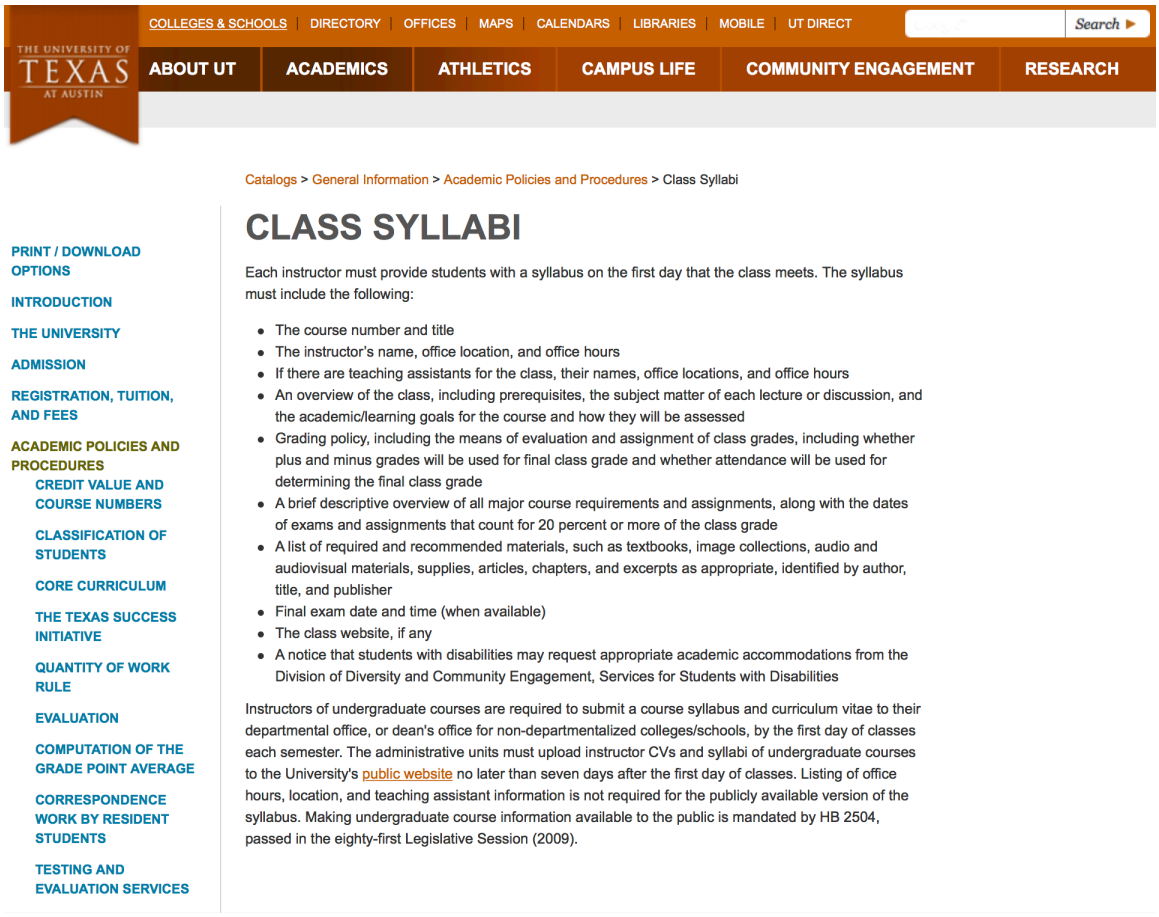

Figure 1 Required elements for class syllabi at UT Austin per the University's Academic Policies and Procedures. (Accessed 7/12/2018 at the following URL: http://catalog.utexas.edu/general-information/academic-policies-and-procedures/class-syllabi/)

PRECEDENTS FOR SYLLABUS REVIEW

Reviewing syllabi for variables of interest is often used for institutional review purposes at the university level, especially to evaluate the impact of new initiatives or programs. For example, Graves, Hyland, and Samuels (2010) collected a syllabus from every course offered over an academic year at one small liberal arts college and analyzed them to determine the frequency, variability, and characteristics of writing assignments

that students were being assigned within and across departments. This produced useful summary information, such as the average number of writing assignments per course, the average length of the assignment in pages, and the most common types of writing assignments. It also enabled these authors to examine how assignments changed across the 4-year curriculum and how they varied by discipline. In addition to their role in institutional reviews, syllabi have also been used to compare across universities: Corlu (2013) coded course syllabi to assess teaching practices across accredited and non-accredited STEM programs, finding benefits for accredited programs that were attributable to the impact of the accreditation process on instructors.

A few researchers have used syllabi to assess the prevalence of teaching best-practices at their institutions. For example, Cullen and Harris (2009) created a rubric for assessing the degree of learner-centeredness of courses from their respective syllabi. They looked at variables related to increasing contact between student and teacher, encouraging cooperation among students, employing active learning strategies, and providing feedback to students. This rubric was then used to assess the degree of learner-centeredness of courses offered by professors who were working with the Center for Teaching and Learning and was ultimately used to inform professional development at their institution.

In perhaps the most comprehensive syllabus analysis to date, Stanny, Gonzalez, and McGowan (2015) developed a rubric for coding syllabi to gather information about the extent to which a course was learning-centered. This included things like stated student learning outcomes, the extent to which these outcomes were assessed, and the presence of instructional practices that promote active learning (e.g., flipped-classroom structure, class discussions, making presentations); these variables were then combined into composite assessments of syllabus quality to look at changes over time. The authors also examined compliance with university syllabus guidelines, providing a framework for rubric

development and training of coders. They also developed guidelines for coding and inter-rater reliability: for instance, in their rubric development and training process, these authors found that coders performed more consistently when they made discrete judgments about variables (e.g., presence of group work, number of group work activities) versus judgments about a global rubric element (e.g., degree of student engagement).

These studies demonstrate how syllabi can be reliably coded according to a well-developed rubric, producing a rich dataset that can be used to answer questions about teaching and learning in college classrooms. Unfortunately, in both of these studies very little was done with this data! The authors reported only overall averages for composite variables such as "student engagement" and did not attempt to explore variability between or within departments across these variables. Furthermore, to the best of my knowledge, such course-level variability derived from syllabi has never been used to examine student-level outcomes that may be associated with such variability. Thus, the utility of such an approach has been demonstrated, providing a useful method for using syllabi to characterize normative educational practices in college courses.

SYLLABUS VALIDITY

There are several reasons to feel confident that contents of syllabi accurately reflect classroom practices at the University of Texas at Austin. First, per Texas House Bill 2504 passed in the eighty-first Legislative Session (2009), all instructors must make their syllabi publicly available online. This outside scrutiny makes it more likely that instructors will be thoughtful and thorough in constructing their syllabi. Secondly, and perhaps more importantly, the University of Texas views a syllabus as a contractual document. For example, to have a grade appeal considered by the Dean of Student Affairs, the first criterion on the Grade Appeals Form is "Instructor violated the terms of the syllabus" and

a copy of the original course syllabus is required component of the appeal. Thus, instructors have strong incentives to take syllabus-creation seriously and to describe teaching practices and learning activities that reflect the true structure of the course.

Furthermore, UT Austin requires that specific information be included on every syllabus for every course taught at the university (Figure 1). Among these requirements are grading and attendance policies, a description of exams and assignments, a calendar of due dates, and a list of course materials. More surprisingly, instructors are required to describe "the subject matter of each lecture or discussion, and the academic/learning goals for the course and how they will be assessed." Because these aspects of a syllabus are mandatory, because all syllabi are submitted to the university before the start of the course and made publicly available, and because students can ground their grade appeals in terms of deviations from the information stated in the syllabus, I feel strongly that the course structure and learning activities described in the syllabus reflect what goes on in the classroom. That being said, the syllabus is still at best an indirect measure of classroom practices; it will be relatively silent on things like the quality of instructors' assignments and activities, their overall effectiveness as teachers, or how faithfully they adhere to the calendar they developed. Still, the presence of information about educational practices in syllabi is, at minimum, indicative of professors' intentions; based on what they choose to include, it is possible to glean aspects of their teaching style, their attitude towards students, and their conceptualization of the course. By the same token, absence of such features indicates a lack of intention, and perhaps even a lack of awareness of what constitutes an effective learning environment.

Chapter Two: Procedure and Methods

Downloading Syllabi for Coding

All syllabi were downloaded from publicly accessible university webpages. All syllabi were pulled for courses of interest by searching for the course name and using a batch-downloading browser plug-in to save all PDF files locally. These were then uploaded to a project folder in UT Box for secure access. The syllabi were accessed via the following URL: https://utdirect.utexas.edu/apps/student/coursedocs/nlogon. Initially, the top 50 highest-enrollment courses at UT Austin were considered, resulting in over 4000 syllabi. Focusing coding efforts on courses with greater than 200 students resulted in a more manageable set of 1104 syllabi from 136 courses, 306 instructors, 49 departments, and 13 colleges at UT Austin, from Spring 2011 to Spring 2016. Due to lack of representation across colleges and semesters, this count was ultimately pared to 1075 (see Analysis section).

Creating the Codebook

The coding scheme, or codebook, was developed primarily through collaborative discussions between myself (the author), my advisor and co-supervisor Dr. Andrew Butler, and Dr. Stephanie Corliss, a former HDCLS graduate who worked for the UT Faculty Innovation Center and lead related initiatives for Project 2021; the final codebook was finalized by Dr. Corliss and approved by Project 2021, other members of whom will have access to some of the syllabus data for their own purposes. During development, our goal was to capture as much information as possible from the syllabus by creating variables that were exhaustive yet relatively easy to code for. We included three types of variables: forced choice (e.g., Yes/No), numeric entry (e.g., 15, .75), and text entry (e.g., a written description of an assignment). For the complete codebook, see Appendix A.

Basic course information

This category consists of standard information about the course, including items that are required by the university to appear in every course syllabus (see Figure 1). This includes information such as the course department, course number, semester, unique ID, course times and dates, room building and number, instructor's name, co-instructors or multiple sections if applicable, instructor office hours, number of TAs, and TA office hours. These variables were coded for by hand and, where appropriate, cross-checked against data pulled from a university database, from which other information was incorporated such as which flags the course carried (if any) and whether the course was a core course. Course flags are additional graduation requirements: they are carried by courses that provide enriched coverage of high-demand skills (e.g., cultural sensitivity, ethical decision-making, quantitative reasoning), while core courses are those required for all students regardless of major.

Learning objectives

The presence of learning objectives for each of three outcomes (knowledge, skills, and socio-emotional) were recorded. For each syllabus, it was required to specify whether each learning outcome was stated, suggested, or not present, and to copy the language directly from the syllabus on which these coding decisions were based. As this is one of the most subjective judgments that coders have to make (particularly, discriminating between stated and suggested objectives), we wanted to document the language from the syllabus in case review becomes necessary.

Course format and resources

The format of the course (face-to-face, online, or hybrid) was recorded. Presence of a list of course topics and whether or not the list included dates was noted. We also

coded for the presence or absence of course resources provided or assigned to students, including those focused on reading, watching, or doing. We also record descriptions of each (e.g., textbooks, articles; films, TED talks; visiting a museum, seeing a play, respectively). Other course-level variables include community learning opportunities (TA-led review sessions outside of class; support for study groups), whether social media was integrated into the course (e.g., course Facebook account, course Twitter hashtag), whether a Learning Management System (e.g., Canvas, Blackboard, Moodle) was mentioned, and whether or not the instructor described their course as a "flipped classroom."

Coursework variables

Information about coursework (homework, in-class assignments, quizzes, exams, projects/presentations, class participation, and extra-credit opportunities) is being systematically recorded from each syllabus. We document all components that contribute to overall course grade, as detailed below. In addition to graded coursework, we coded for whether in-class active learning was used (any mention of doing something during class besides listening to lecture) and the type of active-learning activities; whether informal retrieval practice was incorporated (e.g., availability of optional practice questions or old exams, iClickers or other student-response systems used informally in class) and the type of informal retrieval practice; whether students had to complete projects or make presentations; and whether there were any group assignments or collaborative participation requirements.

Exams and final exams.

Exams are defined as large assessments administered in-class that are referred to as "exams" or "tests" in the syllabus. Final exams are defined as any exam that the professor

calls a "final" exam, as well as any exam that is scheduled during the designated final exam period according to the course calendar. The number of exams, the percent that exams contribute to the overall course grade, the format of exams (multiple choice, short answer, and/or essay questions), whether or not exams were cumulative; also, coded separately was whether a final exam was given, the percent that the final exam contributes to the overall course grade, the format of the final exam, and whether or not it was cumulative.

Other pertinent information about exams was coded for: whether there was a calendar of all exam dates, whether students get to drop their lowest exam score, whether they get a re-test opportunity, and whether there is some sort of alternative exam weighting scheme (e.g., lowest exam replaced with average of highest two exams). The nature of such alternative weighting was described in the notes. A composite variable was created called *grade choice* which indicated if any of the latter three grade-choice related variables were applicable.

Quizzes

Quizzes are defined as short assessments that take place during class. The number of quizzes, the percent that quizzes contribute to the overall course grade, the type of questions featured on quizzes (multiple choice, short answer), and the delivery method of the quiz (electronic or paper-based) was also recorded. "Quizzes" that were assigned for homework were coded as homework, not as quizzes, and this was noted in the "Type of Homework Assignments" variable.

In-class assignments

In-class assignments are defined as work that students complete during class-time that receives an actual grade (i.e., not just a participation grade). We record the number of

in-class assignments, the percent that they contribute toward the overall course grade, and a brief description of the type(s) of in-class assignments.

Homework

Homework assignments were defined as any graded work that students were required to complete outside of class. This broad definition covered everything from problem-sets and on-line homework modules to large essays and take-home exams. The number of homework assignments, the percent that homework contributes to the overall course grade, and the types of homework assignments were all recorded. Care was taken to be specific when noting the types of homework assignments: for example, in a course with weekly journal responses and two important essays, it was noted that there were 15 journal assignments worth 20% of the course grade, and 2 essays worth 50% of the course grade. This allowed us to have the homework category serve as a catch-all for all work done outside the class while still preserving as much information as possible about the different types of assignments students were required to do.

Participation

Participation is defined as anything that counts toward students' overall course grade that is not individually graded. This includes attendance and grades given for completion of assigned work. No assignments graded for accuracy were coded as participation.

FINALIZING VARIABLES AND CREATING COMPOSITES

Several composite variables were created *a priori* from the raw coded variables. As stated in the codebook, syllabi were originally coded for whether learning objectives were stated outright or only suggested by syllabus language, but reliabilities for this distinction

were low. Therefore, for each type of learning objective—knowledge, skills, socio-emotional—a new binary variable was created by combining stated and suggested to indicate either whether the learning objective was present at all. Second, *Flag Course* was a binary variable created to indicate whether the course carried any UT flag status. Third, instructor office hours and TA office hours were recorded as numeric variables, but binary variables were created to indicate the presence of any instructor or TA office hours in the syllabus. Also, to reflect the amount of control students have over the grading structure, a variable called *Grade Choice* was computed to take the value 1 (zero otherwise) if the course allowed students to (a) retake exams, (b) drop their lowest exam score, or (c) exercise choice in how individual assessments/assignments would be weighted in the calculation of their final grade.

Because retrieval practice was of particular interest in both Part I and Part II of this study, individual variables were created to reflect the total number and grade percentage of all graded retrieval practice opportunities in a course by combining the number and grade percentage of exams and quizzes. As I discuss in more detail below, across all courses the median value for total graded retrieval practice opportunities was 4. If the number of graded retrieval practice opportunities was greater than or equal to 5, the new variable *High Graded RP* took the value 1 and was 0 otherwise. Similarly, a variable *High Stakes* was created to indicate when large proportion of the course grade came from relatively few assignments. The median value for exam grade percentage was 75%, and a course was considered high stakes if at least 75% of the grade came from performance on 4 or fewer exams.

Finally, for each course-grade component (homework, quizzes, exams, and in-class assignments), the percent each contributed to the total course grade was divided by the total number of each, yielding new variables reflecting the percent of grade per homework

assignment, quiz, exam, and in-class assignment. Lastly, where appropriate, nominal variables scored Yes/No/Unclear were dichotomized into binary variables taking the value 1 when "Yes" and 0 otherwise.

CODING OF SYLLABI

Two outside coders were trained to use the codebook until they reached 90% agreement on all variables. After reaching this criterion, they worked under supervision of the author, accessing syllabi through a secure online storage platform. Inter-rater reliability of syllabus coding was performed by independently re-coding a subset of the syllabi (n = 50) and assessing the degree of agreement with the primary coders. Reliability was calculated using both percent agreement and Cohen's Kappa for each categorical variable, and intraclass correlations (ICC; fixed, one-way) for each numeric variable. A sample of fifty syllabi have been re-coded in such a manner, demonstrating acceptably high reliability for variables of interest (see Table 1).

Table 1 Inter-rater reliability calculations for syllabus variables of interest

Codebook Variable	% Agreement	Cohen's Kappa	ICC (95% CI)
Course Format	1.00	1.00	-
Instructor Office Hours	1.00	1.00	-
# Office Hours	0.95	0.94	-
Course Resources: Reading	1.00	1.00	-
Course Resources: Watching	0.95	0.64	-
Course Resources: Doing	0.89	0.79	-
Social Media	1.00	1.00	-
Learning Management System	0.89	0.47	-
Community Learning Ops	0.89	0.79	-
SLOs: Knowledge	0.95	0.85	-
SLOs: Skills	0.79	0.68	-
SLOs: Socio-Emotional	0.79	0.52	-
List of Topics	1.00	1.00	-
Dates for Topics	0.95	0.85	-
# Exams	-	-	1.00
% Exams	-	-	0.996 (0.993,0.998)
Cumulative Exams	0.95	0.64	-
Drop Lowest Exam Score	1.00	1.00	-
Re-Test Opportunity	1.00	1.00	-
% Final Exam	-	-	1.00
Cumulative Final Exam	0.95	0.92	-
Alternative Weighting	0.68	0.26	-
Calendar of Exam Dates	1.00	1.00	-
Calendar of Assignment Due Dates	0.89	0.69	-
# Quizzes	-	-	1.00
% Quizzes	-	-	1.00
# In-Class Assignments	-	-	1.00
% In-Class Assignments	-	-	1.00
In-Class Active Learning	0.95	0.88	-
Informal Retrieval Practice	0.89	0.83	-
Projects or Presentations	1.00	1.00	-
% Participation	-	-	0.76 (0.579,0.865)
Attendance Enforced	0.95	0.87	-
# Homework	-	-	0.87 (0.768,0.931)
% Homework	-	-	0.95 (0.909,0.974)
Flipped Classroom	1.00	1.00	-
Extra Credit	0.95	0.90	-

ANALYSIS

As Part I of this study is descriptive in nature with few *a priori* hypotheses, significance testing was of minor importance and used primarily to narrow the scope of relationships among variables. Given our sample, it is trivial to reject the null-hypothesis of equal means (or independent groups) among colleges/departments on each of these dimensions, and given the number of relationships there are to explore, multiple comparisons would quickly become an issue. Therefore, descriptive statistics are reported (e.g., counts, proportions, means, correlations), graphs are relied upon heavily to illustrate macro-level variability and interrelationships, and whenever tests are reported, the familywise error rate is considered and robust (heteroskedasticity-consistent) standard errors are used.

After a general overview, results will be organized into sections based on groups of related variables to aid in presentation. Each section presents overall descriptive statistics for the pertinent variables as well as a breakdown of each variable by college. Notable correlations among variables are also discussed. The final three sections zoom back out to address changes across variables over time, clustering of variables and courses, and the language used in the syllabi.

Correlations.

As stated above, correlations among all variables of interest were computed: when both variables were numeric, Pearson's correlation coefficient was computed; when both variables were binary, a tetrachoric correlation was computed; when one variable was numeric and one was binary, a biserial correlation was computed. These pairwise correlations were computed using the *hector()* function from R package *polycor*. Because 39 variables were of interest, $\binom{39}{2} = 741$ correlations were computed, and because all

correlations were explored as potentially interesting, a conservative alpha level of 0.00005 was used to control for multiple comparisons and to narrow the space of possible relationships to consider for the purposes of this investigation. Note that this is approximately the same as considering only those correlations that more than 4 standard errors away from zero. All tests were two-sided, as the implied null hypothesis is non-directional. Because there are many degrees of freedom owing to a large sample size, even this stringent criterion leaves 174 correlations that are statistically distinguishable from zero. For the sake of brevity, in what follows only a summary of these significant effects is presented.

Factor analysis and clustering

An exploratory factor analysis (EFA) of course variables was conducted on the correlation matrix described above (i.e., including polychoric, biserial, and Pearson's correlations where appropriate) using principal-axis factoring. Note that this is technically a factor analysis of mixed data (FAMD) which extends beyond the EFA framework. However, it is not an unreasonable approach for descriptive, exploratory purposes (Pagès, 2014; Revelle, 2017). The number of factors to extract was decided upon by comparing the results of parallel analysis to an empirical scree plot. A Varimax rotation of the factor solution was performed to achieve simple structure. Note that several variables are binary

Clustering of course syllabi was performed using a subset of the variables that were used to compute correlations. An alternative to cluster analysis would have been to compute estimated factor scores for each syllabus, but this was not feasible because the variables used were of mixed type (i.e., nominal and continuous). However, the traditional *k*-means clustering algorithm cannot be used with categorical variables because they are discrete, rather than continuous, and the Euclidean distance computed on such variables is

not meaningful. A common way around this is to use Gower's dissimilarity algorithm (Gower, 1971; Podani, 1999), which computes distances appropriate for each type of variable, and then to apply a clustering method suitable for distance matrices. A widely used choice is *partitioning around medoids* (PAM), which is conceptually similar to *k*-means but uses medoids instead of centroids. To perform this analysis, the R package *cluster* was used; the function *daisy()* was used to compute the dissimilarity matrix, and the function *pam()* was used to perform the clustering. The number of clusters was chosen based on average silhouette width. Visualization was done using t-Distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction technique useful for visualizing relationships in high-dimensional data (Maaten & Hinton, 2008).

Text mining and sentiment analysis of syllabi

Syllabi were converted to plain text files with ASCII encoding using the Linux command-line utility *pdftotext*. These text files were first processed using Linguistic Inquiry and Word Count (LIWC) software (LIWC2015; see Tausczik & Pennebaker, 2010; Pennebaker, Boyd, Jordan, & Blackburn, 2015). Beyond providing basic counts of words (or prepositions, articles, etc.) in each syllabus, the software also compares text against several validated, psychologically meaningful dictionaries; in general, output from the software takes the form of the percentage of the total words in the document that also appear in each dictionary, though there are several variables that are not dictionary based, including the average number of words per sentence, the percent of words that are standard dictionary words, and the number of words that are six letters or longer.

In addition, LIWC generates four summary variable scores, reported as percentiles based on a large norming sample of various kinds of text (blogs, expressive writing, novels, natural speech, newspaper articles, tweets from Twitter). These variables are *analytical*

thinking, *clout*, *authenticity*, and *tone*. Analytical thinking is a variable that indexes logical thought and formality of writing, where lower scores indicate a more informal, narrative style (Pennebaker, Chung, Frazee, Lavergne, and Beaver, 2014). Clout is a variable that indexes confidence and leadership in speech or writing; it was developed from several studies of interpersonal interactions (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013). Authenticity is a variable that indexes personability, humility, and vulnerability in speech or writing; it was developed from several studies in which people were acting honestly or deceptively (Newman, Pennebaker, Berry, & Richards, 2003). Tone is a variable that summarizes the emotional valence of speech or writing (Cohn, Mehl, & Pennebaker, 2004), with higher scores reflecting a more positive tone.

Though the software reports some 90 variables by default, for the present analysis we are interested in those mentioned above and also the percentage of various pronouns used in each syllabus (first-person singular, first-person plural, and second person), the percentage of comparative words (e.g., *greater, best, after*) and negation words (e.g., *no, not, never*), and whether the course syllabus has an achievement focus (using words like *win*, *success*, and *better*) or an affiliation focus (using words like *ally, friend,* and *social*).

Finally, in addition to comparisons among colleges, LIWC variable scores computed on syllabi will be compared to average scores for different types of texts that are included with the LIWC2015 software (specifically, blogs, expressive writing, natural speech, novels, New York Times articles, and Twitter data). Comparisons drawn will illustrate the extent to which syllabi are similar to (or different from) various kinds of text on all dimensions under consideration. For additional details on these norms, see Pennebaker, Boyd, Jordan, and Blackburn (2015).

Because LIWC takes raw text files as input (and thus performs no pre-processing, such as the deletion of stopwords), several more analyses were conducted after cleaning

the syllabus text data. All text contained in each syllabus file was parsed into individual words and cleaned, during which punctuation, URLs, stopwords, numbers, and whitespace were removed using the R packages *tm* and *tidytext*. The remaining terms were transformed into a large document-term matrix (1,075 documents by 14,827 terms) with each cell containing a count of the total time that a given term appeared in a given document. For each of these raw counts, the *term frequency-inverse document frequency* (tf–idf) was then computed. For every pairwise comparison of syllabus tf–idf vectors, their cosine similarity was computed. These pairwise cosine similarities of tf–idf vectors measure how strongly two syllabi resemble each other in terms of the words they use and how unique those words are relative to those used in other syllabi.

Term frequency–inverse document frequency (tf–idf)

Briefly, *term frequency* ($tf_{t,d}$) is the number of occurrences of a term *t* in a document *d*, divided by the total number of terms in that document. Terms appearing frequently in a document will have higher term frequency than those appearing less frequently, and they are normalized to sum to 1 within a document. The *inverse document frequency* (idf_t) of a term is the ratio of the total number of documents *N* to the document frequency df_t (i.e., the number of documents *d* that contain the term *t*), and typically the logarithm of this quotient is used. This represents how uncommon a term is among a set of documents: the idf_t of a term appearing only in a few documents will be high, while the idf_t of a term appearing in many documents will be low. For a given term in a given document, the tf–idf is the product of each of these quantities ($tfidf = tf \times idf$). This quantity indexes how unique a term is to a specific document in a collection: High tf–idf is achieved by high term frequency (i.e., the term is common in a given document) and/or low

document frequency (i.e., the term is uncommon in the whole collection of documents). Because $idf = log(N/df)$ is always greater than 0, so too is tf–idf.

Cosine similarity

At this point, each document can be considered a vector of tf–idf scores for each term in the overall collection of terms (for terms not appearing in a document at all, tf is zero and so too is tf–idf). Similar documents will have similar tf–idf vectors; there are many ways to compute document similarity, but a common approach is to compute the cosine similarity of tf–idf vectors. In information retrieval and text analysis, cosine similarity of tf–idf vectors is a commonly used measure of how similar two or more documents are in terms of their most important words (e.g., Singhal, 2001), and in cluster analysis, cosine similarity can be used to measure how cohesive clusters of data are (e.g., Tan, Steinbach, & Kumar, 2005). The cosine similarity for a pair of vectors is computed by dividing the dot product of the two vectors by each vector's magnitude, yielding a value that corresponds to the cosine of the angle between the two vectors. For positive valued vectors such as document tf–idf vectors, the cosine similarity is bounded in [0, 1]; when the cosine is one, the angle between the two vectors is zero, indicating perfect similarity (i.e., identical documents).

After cosine similarity was computed for all pairwise comparisons of tf–idf vectors, these values could be averaged by department, by college, and over time. However, this is potentially misleading if one department or college offered many sections of each of a small number of courses (resulting in greater average similarity) while another offered fewer sections of a greater number of courses (resulting in less average similarity). One way around this is to only average the within-course cosine similarities (e.g., only those comparing a CH 301 syllabus to other CH 301 syllabi), thus resulting in departmental and

college-level averages of how similar each course is across sections. Furthermore, we look at averages of within-course cosine similarities by semester, in case there are cross-department or cross-college differences in the number of semesters a given instructor teaches a given course; if there are, then departments where the same instructors taught the same courses for more semesters would look like their syllabi were more similar.

Sentiment analysis

One common and relatively simple approach to analyzing the sentiment of a document is to break the text down into its constituent words, classify each word as positive/negative or assign each word a sentiment polarity score (e.g., an 11-point scale from -5 = *very negative* to 5 = *very positive*), and consider the sentiment of the whole document to reflect the sum of the sentiment content of the individual word.

Several sentiment lexicons exist that are commonly used to conduct sentiment analysis in this way: the *bing* lexicon (Hu & Liu, 2004), AFINN (Nielson, 2011), and the NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013). The bing lexicon consists of 6,788 words labeled as having either positive or negative sentiment. The AFINN consists of 2,476 words and phrases manually scored by the author of the lexicon from -5 (*very negative*) to 5 (*very positive*). The NRC consists of 6,458 words and their associations (0 = *not associated*, 1= *associated*) with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive), which were crowdsourced through Amazon Mechanical Turk. Validation of each sentiment lexicon was carried out either by crowdsourcing or by using text from restaurant reviews, movie reviews, or Twitter data. To the extent that syllabus text differs in kind from the sorts of texts these lexicons were validated on, results of this sentiment analysis are qualified and should be viewed as exploratory in nature.

All three of these lexicons were used to compute sentiment scores for each syllabus; these sentiment scores were then averaged across colleges, much like what was done with similarity scores. The average percentage of words per syllabus associated with each of the eight NRC categories, and the average ratio of positive-valence to negative-valence words (or average document score in the case of the AFINN) are shown for each of the lexicons.

Chapter Three: Results

Syllabi (*N*=1104) from all high-enrollment courses from 2011-2016 were collected and coded as described above. There was only a single unique syllabus from a single course the following colleges: Social Work, Information, Architecture, and Pharmacy. In each case, the course was taught by the same instructor each semester as well. Furthermore, there were anomalously few syllabi from the first, last, and summer semesters: There was only a single syllabus from Spring 2011, only three unique syllabi from Fall 2016, and only two syllabi from a single summer semester. Excluding syllabi *a priori* on these grounds left a final sample of 1075 syllabi from long semesters ranging from Fall 2011 to Spring 2016, representing 9 colleges, 45 departments, 132 unique courses, 303 instructors, and 368 unique instructor-course combinations. See Table 2 for a breakdown of syllabi counts by college, year, and semester (note abbreviations).

Courses in the syllabus dataset had an average total enrollment of 287.07 (*Mdn* = 263, *SD* = 133.38). The highest enrollment courses are all either traditional online courses or synchronous massive online courses (SMOCs), of which there are relatively few: The vast majority of all courses are traditional face-to-face format (95.63%), with the remainder being online courses or SMOCs. Additionally, 77.77% of courses were lower division, 56.74% were core courses at UT Austin, and 40.37% carried a course flag (including 16.74% with a quantitative reasoning flag and 1.49% with a writing flag).

Table 2 Number of courses (and number of unique courses) offered in each college by semester

	2011	2012		2013		2014		2015		2016	Courses	
College	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Total	Unique
BUS	13 (5)	13 (6)	13 (5)	9 (4)	17 (5)	10 (5)	18 (5)	13 (5)	15 (6)	10 (4)	131	7
CFA	4 (4)	4 (3)	5 (4)	6 (4)	5 (4)	4 (3)	3 (3)	3 (3)	6 (4)	6 (4)	46	4
CLA	42 (19)	44 (21)	46 (19)	39 (20)	40 (20)	38 (20)	34 (22)	40 (23)	42 (22)	33 (20)	398	37
CNS	27 (15)	21 (14)	28 (13)	23 (13)	33 (18)	33 (19)	27 (16)	31 (16)	33 (17)	26 (14)	282	42
COM	10 (9)	10 (9)	14 (13)	7 (6)	12 (12)	8 (7)	9 (9)	7 (6)	13 (12)	7 (6)	97	23
EDU	7 (2)	6 (2)	6 (1)	0	1	0	1	2 (2)	2 (2)	2 (2)	27	5
EGN	4 (4)	1	4 (4)	0	5 (4)	1	3 (3)	0	5 (5)	2 (2)	25	9
GEO	1	4 (3)	3 (2)	4 (3)	3 (2)	3 (3)	2 (2)	1	0	1	22	4
UGS	6 (1)	4 (1)	8 (1)	3 (1)	7 (1)	1 (1)	9 (1)	1 (1)	5 (1)	3 (1)	47	1
Sem. Total:	114 (60)	107 (60)	127 (62)	91 (51)	123 (67)	98 (59)	106 (63)	98 (57)	121 (69)	90 (54)	1075	132

Note. Here and elsewhere, BUS = Business, CFA = Fine Arts, CLA = Liberal Arts, CNS = Natural Sciences, COM = Communication, EDU = Education, EGN = Engineering, GEO = Geosciences, UGS = Undergraduate Studies.

COURSEWORK VARIABLES

The type, quantity, and grade weight of all work completed by students in a course together form an extremely useful depiction of the overall course. Below, summary statistics are provided for each coursework variable in turn (see Figure 2 for average grading rubric overall and for each college) and in the aggregate (see Figure 3 for the average number of total graded assignments per course for each college). Additionally, any correlations with other variables of interest are discussed (see Appendix B for all correlations; see Figures 4 and 5 for correlation heatmaps). As stated above, any association mentioned in the text was over 4 standard errors away from zero.

Where it is especially interesting, comparisons across colleges are noted. Figure 6 illustrates differences across colleges in their average grading rubric (grade weight per component; top graph), number of assignments per grade component (middle graph), and percent of grade per assignment (bottom graph). Keep in mind that the number of unique instructors, courses, and departments varies somewhat by college, and thus colleges differ in the extent to which they are represented by our sample of courses (see Table 3).

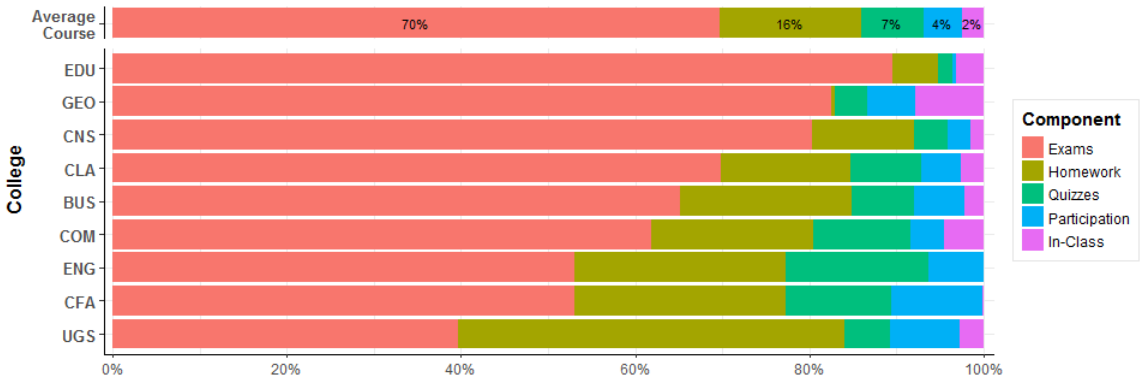

Figure 2 Average grading rubric (percent of grade assigned to each coursework component) overall (topmost bar) and by college (lower bars). See Table 2 note for college abbreviations.

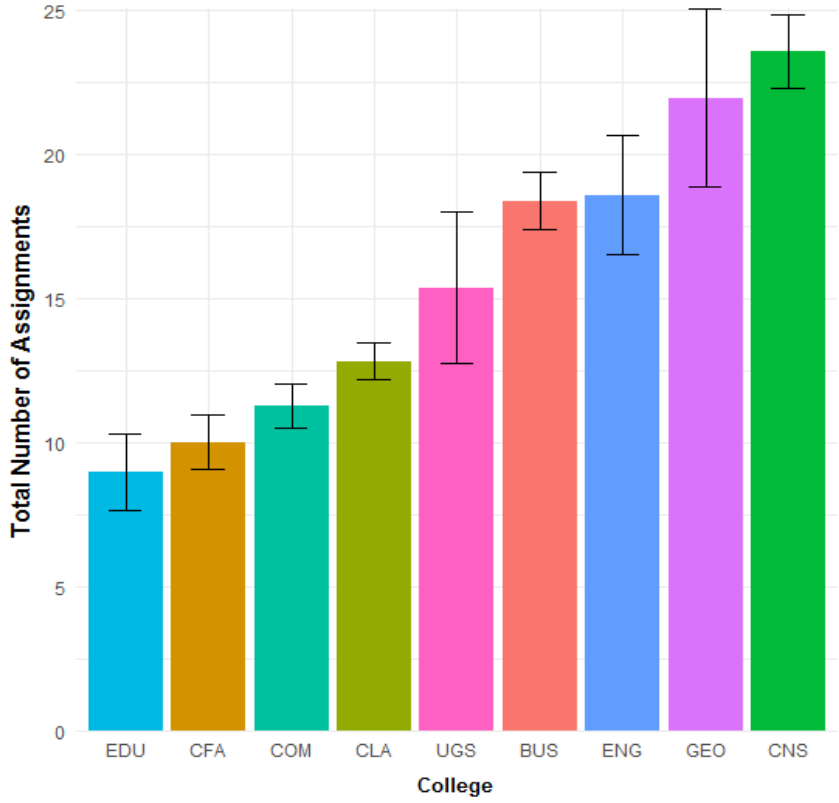

Figure 3 Mean number of total assignments (i.e., all coursework completed for a grade, including exams) by college. Error bars show bootstrapped standard errors. See Table 2 note for college abbreviations.

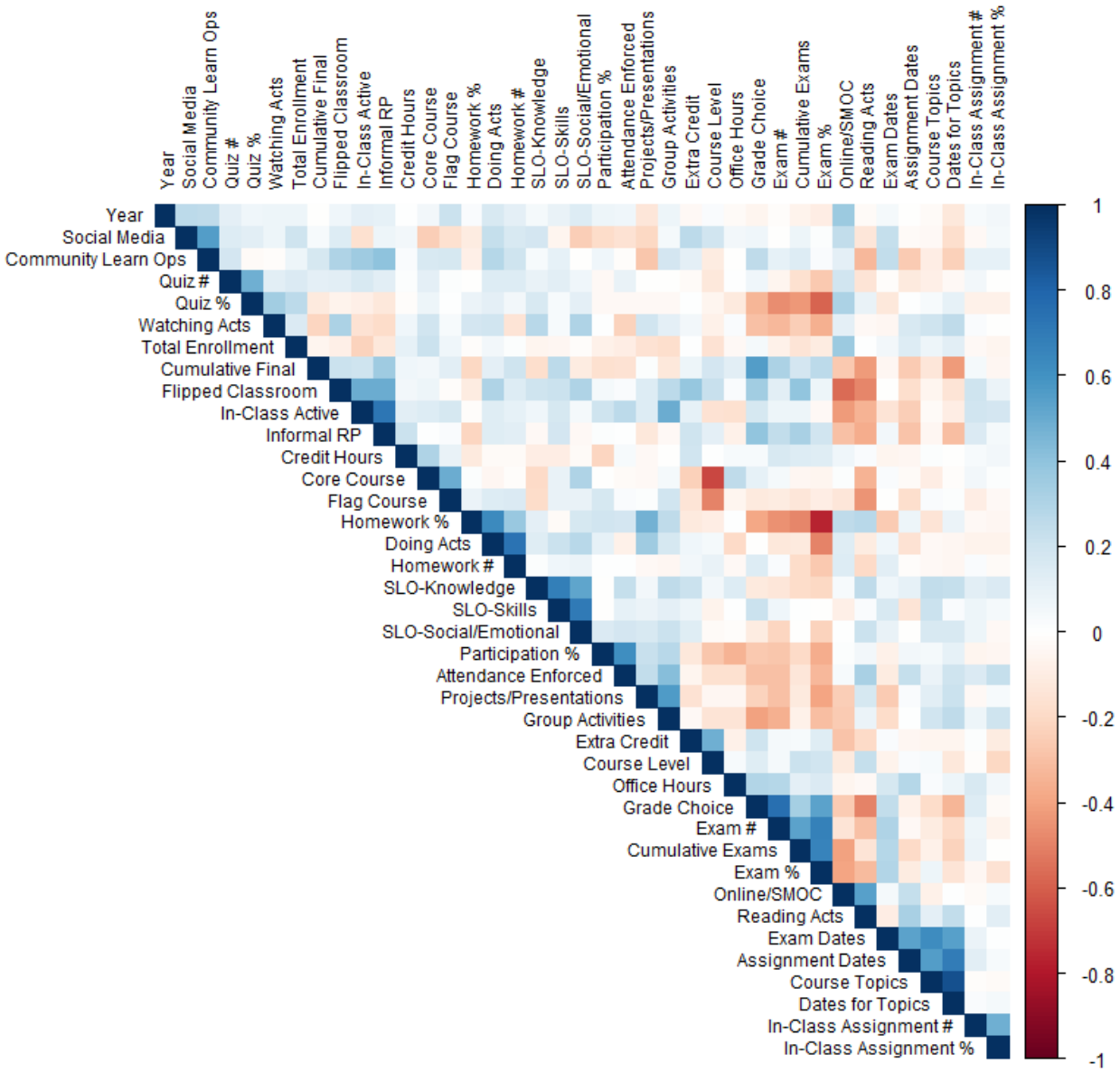

Figure 4 Heatmap of all correlation coefficients (see Appendix B for values). Color gradient ranges from dark red (perfect negative correlation) to white (no correlation) to dark blue (perfect positive correlation). All correlations are shown regardless of statistical significance.

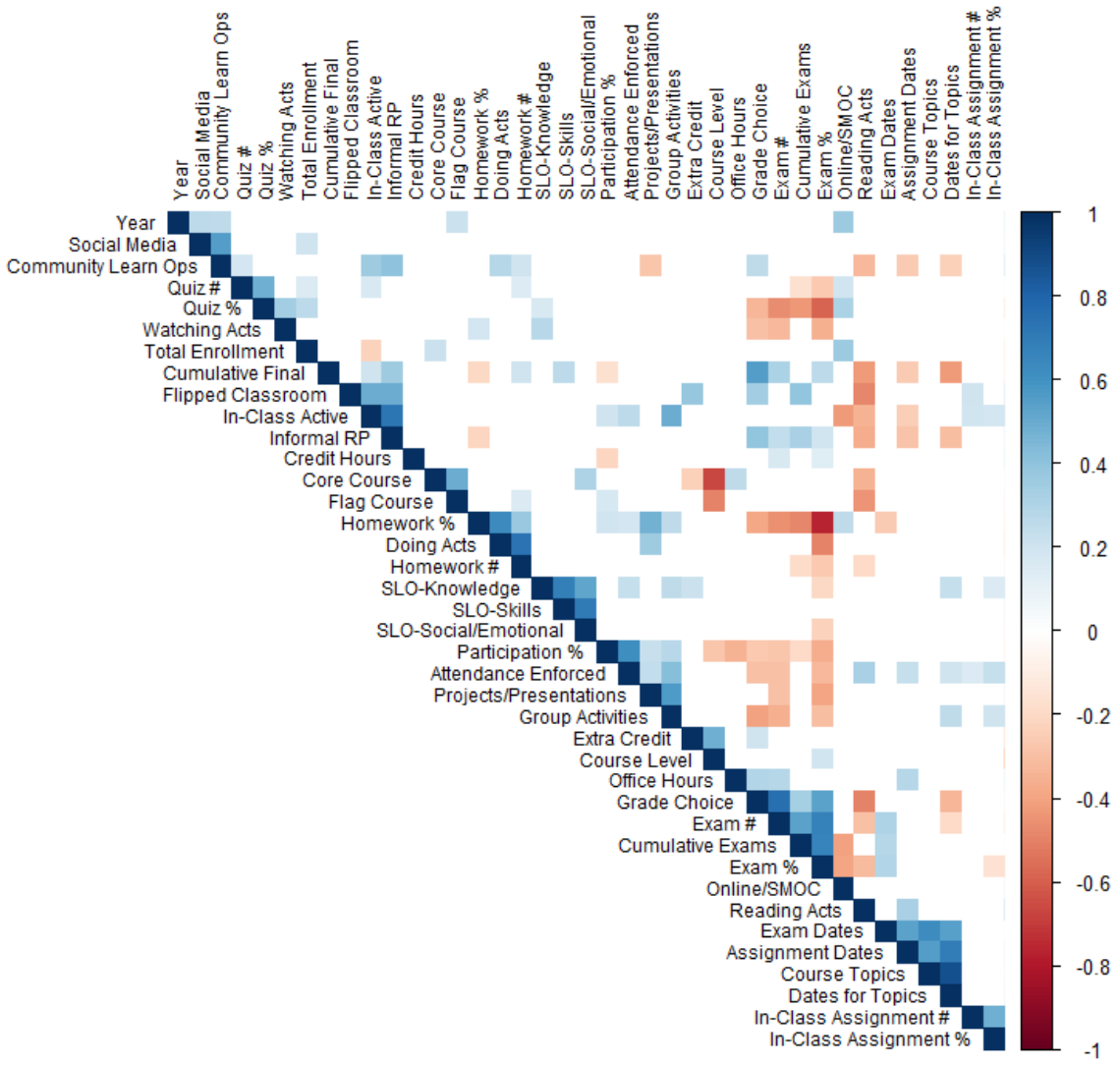

Figure 5 Heatmap of significant correlation coefficients (*p*s < .00005) for all course variables, with a color gradient from dark red (perfect negative correlation) to dark blue (perfect positive correlation). All colored cells represent significant correlations at the stated significance level; blank cells indicate insignificant correlations.

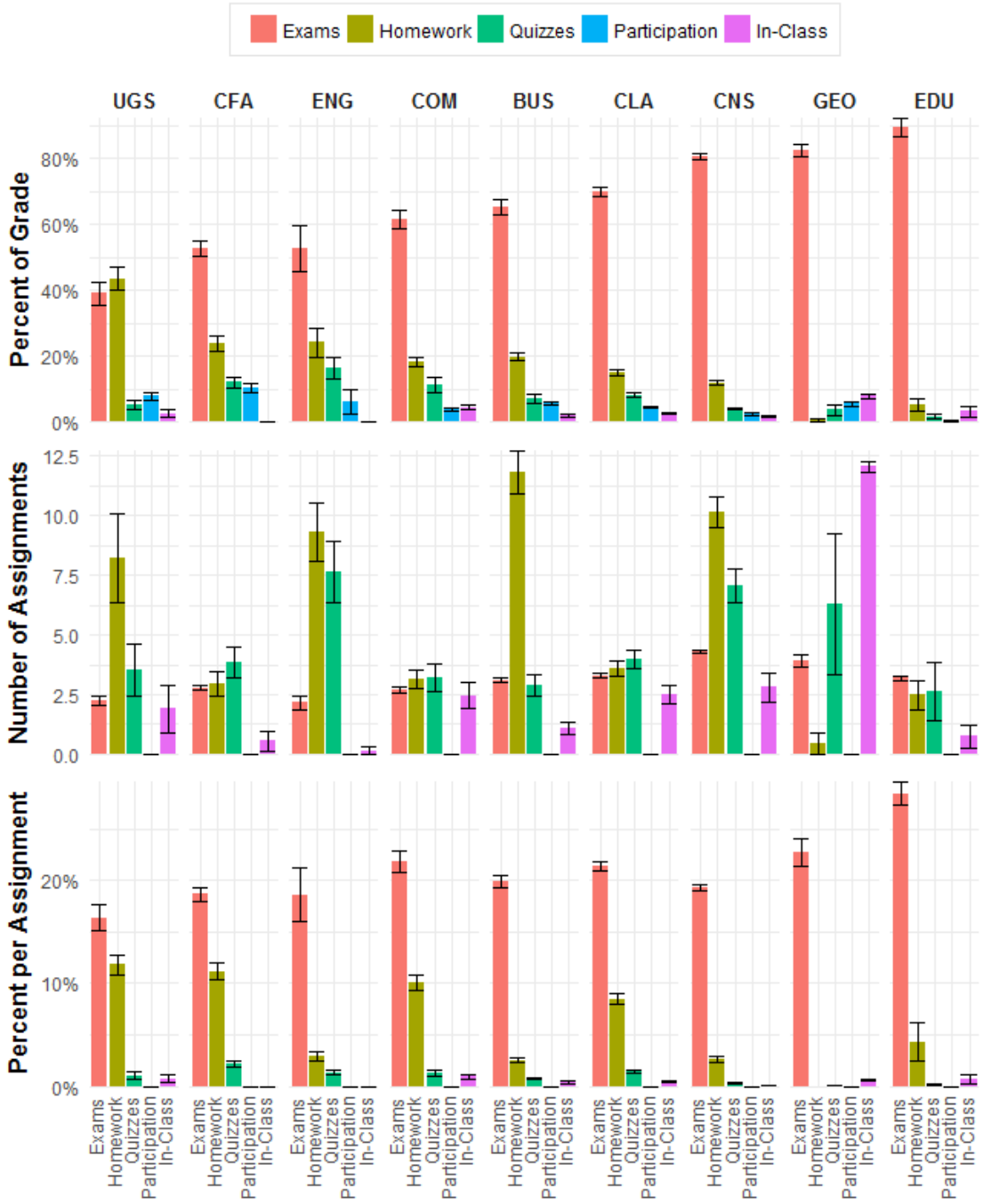

Figure 6 Breakdown of grade weight per component in grading rubric (top), number of assignments per component in the grading rubric (middle) and percent of overall grade per assignment (bottom). Plots show overall averages with bootstrapped standard errors. Note that scales differ for top, middle, and bottom plots. See Table 2 note for college abbreviations.

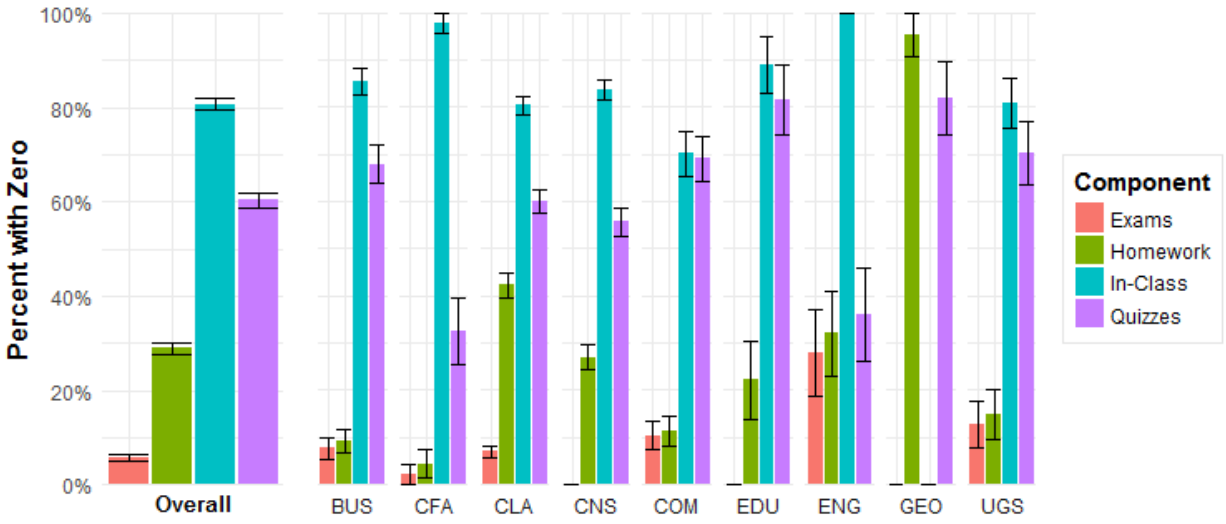

Figure 7 Percent of syllabi with zero for a given rubric component, overall (left) and by college (right). Error bars show bootstrapped standard errors.

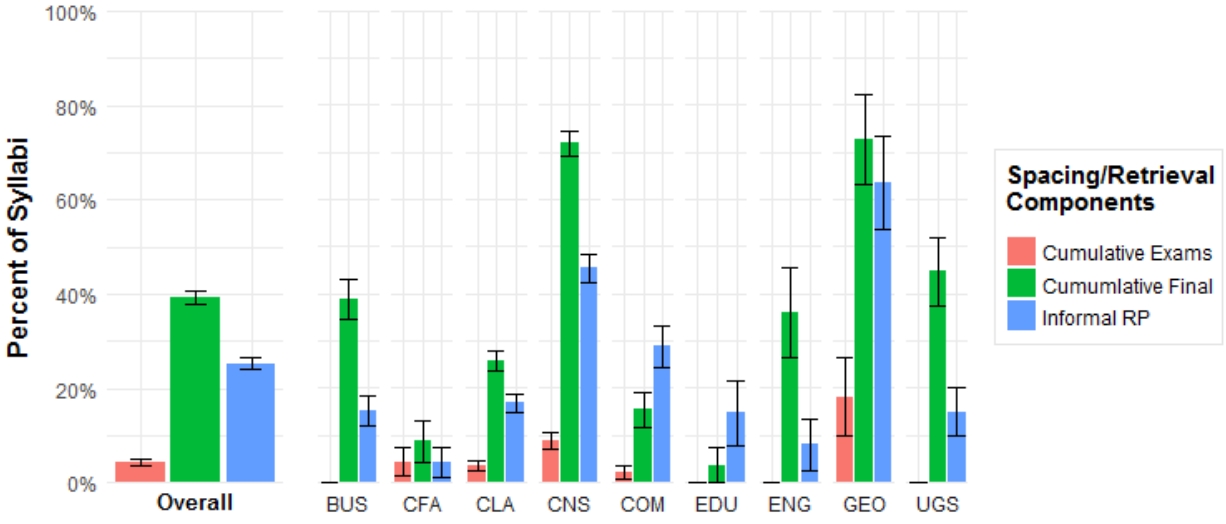

Figure 8 Percent of syllabi featuring each of the spacing/retrieval variables overall (left) and by college (right). Error bars show bootstrapped standard errors.

Table 3 Number of unique departments, instructors, courses, and combinations by college

College	Departments	Instructors	Courses	Course-instructor combinations
BUS	6	16	7	16
CFA	3	23	4	23
CLA	12	115	37	145
CNS	8	67	42	98
COM	6	34	23	38
EDU	3	7	5	7
EGN	5	17	9	17
GEO	1	8	4	8
UGS	1	16	1	16
Total:	45	303	132	368

Exams

Overall, the average number of exams (including final exams) per course was 3.40 ($Mdn = 3, SD = 1.54$), and exams accounted for 69.57% of the total course grade on average ($Mdn = 75, SD = 25.60$; see Figures 2 and 6). The average percent of students' overall grade per exam was 20.53% (SD = 8.16). Only 5.77% of courses did not have any exams (see Figure 7). Furthermore, 39.35% of courses stated that they had cumulative final exams; however, only 4.37% of courses reported that their other course exams were cumulative in nature (Figure 8).

College comparisons

Several differences in exam grade-weight are readily apparent when comparing among colleges: EDU, GEO, and CNS gave the greatest weight to exams on average (89.35, 82.30, and 80.44%, respectively), while UGS, CFA, and ENG gave the least weight

to exams (39.04, 52.40, and 52.76%; see Figure 6). Colleges who allocated a moderate percentage of the grade to exams were COM, BUS, and CLA (61.60, 65.15, and 69.85%).

In terms of cross-college differences in number of exams, the differences were less pronounced: CNS (M = 4.33, Mdn = 5) and GEO (M = 3.91, Mdn = 3) gave the greatest number, while UGS (2.26, Mdn = 2) and ENG (M = 2.16, Mdn = 3) gave the fewest. Because of this, the average percent of course grade per exam mirrors closely the overall average percent of grade from all exams (Figure 6).

The colleges most likely to have cumulative final exams were GEO (72.73%) and CNS (71.99%), followed by UGS (44.68%), BUS (38.93%) and ENG (36.00%; see Figure 8). Colleges least likely to have cumulative finals were EDU (3.70%), CFA (8.70%), and COM (15.46%). Also, GEO had the largest proportion of courses with cumulative non-final exams (18.18%), CNS was second (8.87%), and all other colleges were under 5%. Indeed BUS, EDU, ENG, and UGS had no courses with any cumulative, non-final exams.

Correlations

Across all courses, number and grade percent of exams were both negatively associated with having projects/presentations (r = –.29, for number; r = –.39, for grade percent), group activities (r = –.35; r = –.31), attendance requirements (r = –.29; r = –.33), and out-of-class watching (r = –.32; r = –.36) and reading (r = –.30; r = –.31) activities. Furthermore, exam percent-of-grade was negatively associated with learning objectives for knowledge and socio-emotional outcomes (r = –.21; r = –.23) as well as doing activities (r = –.50). Online and SMOC courses tended to have a smaller percent of grade from exams (r = –.39). It is unsurprising that grade choice is strongly positively associated with the number and percent of grade from exams (r = .76, r = .54) and negatively associated with all other grade components (participation, r = –.36; quiz percentage, r = –.58; homework,

r = –.77; in-class assignments, r = –.16): grade choice almost always applies to either dropping, retaking, or applying an alternative weighting scheme to exams.

Interestingly, the number of exams was negatively associated with the percent of grade from quizzes (r = –.46), homework (r = –.45), and participation (r = –.27): the fewer exams given in a course, the higher the percent of grade from quizzes and from homework assignments. It is also noteworthy that the number of exams in a course was not significantly associated with the number of homework assignments or the number of quizzes. Unlike grade percentage, which must sum to 100% and thus induce negative correlations among the grade components (i.e., more of one variable always implies less of the others), the total number of homework assignments, quizzes, and exams do not seem to trade off in the same way.

Quizzes

The average number of quizzes per course was 4.65 (*Mdn* = 0, *SD* = 8.56) and, on average, quizzes accounted for 7.15% of the total course grade (*Mdn* = 0, *SD* = 14.27; see Figures 2 and 6). The average percent of overall grade per quiz was 1.00% (*SD* = 2.16). Across all courses, 63.26% did not have any quizzes (see Figure 7). Among courses having at least one quiz, the average number of quizzes was 12.66 (*Mdn* = 10, *SD* = 9.90) and the average grade percentage from quizzes was 19.04% (*Mdn* = 15, *SD* = 18.14).

College comparisons

Among colleges, the highest grade-weights for quizzes were observed in ENG, CFA, and COM (16.35, 12.01, and 11.11%, respectively), while the lowest weights for quizzes were given by EDU, GEO, and CNS (1.66, 3.64, and 3.94%; Figure 6). Notice that grade weights for quizzes and exams seem to be strongly anticorrelated with the exception

of UGS (which has relatively little of the grade coming from either quizzes or exams). On average, the greatest number of quizzes were given by ENG, CNS, and GEO (M = 7.64, 7.07, and 6.27), although the median number of quizzes was 0 for all colleges excepting ENG and CFA (Mdn = 12 and 3). Percent of total course grade per quiz was around 1% regardless of college.

Correlations

Reinforcing a point noted above, number of quizzes was positively associated with the number of homework assignments in a course (r = .15), and the higher the percent of grade from quizzes, the more likely the course syllabus was to have learning objectives for knowledge outcomes (r = .16; Figure 5). Unlike what was observed for exams, larger courses and Online/SMOC courses tended to give more quizzes (r = .20) and to make quizzes a larger proportion of the course grade (r = .32). Additionally, courses with community learning opportunities tended to have a greater number of quizzes (r = .18).

Homework

The average number of homework assignments per course was 6.49 (Mdn = 3, SD = 8.89), and homework accounted for 16.29% of the total course grade on average (Mdn = 10%, SD = 17.63; see Figures 2 and 6). The average homework assignment was worth 6.23% (SD = 7.03) of the course grade. Out of all courses, 28.63% did not have any homework (Figure 7). Among courses having at least one homework assignment, the average number of homework assignments was 9.28 (Mdn = 6, SD = 9.33) and the average grade percentage from homework was 22.92% (Mdn = 20, SD = 17.04).

College comparisons

The colleges with the highest average grade weight for homework were UGS (*M* = 43.53%, *Mdn* = 45, *SD* = 24.38), ENG (*M* = 24.11%, *Mdn* = 25, *SD* = 21.98), and CFA (*M* = 23.87, *Mdn* = 20, *SD* = 14.97), respectively (see Figure 6). Colleges allocating very little grade weight to homework on average were GEO (M = 0.45%, *Mdn* = 0, *SD* = 2.13), EDU (*M* = 5.32%, *Mdn* = 2, *SD* = 9.99), and CNS (*M* = 11.79%, *Mdn* = 6, *SD* = 12.54).

Colleges giving the largest number of homework assignments on average were BUS (*M* = 11.83, *Mdn* = 10, *SD* = 10.50), CNS (*M* = 10.12, *Mdn* = 9, *SD* = 10.70), ENG (*M* = 9.32, *Mdn* = 11, *SD* = 6.19). Those with the smallest number of homework assignments were GEO (*M* = 0.45%, *Mdn* = 0, *SD* = 2.13), EDU (*M* = 2.48, *Mdn* = 0, *SD* = 3.20), and CFA (2.93, *Mdn* = 2, *SD* = 3.57). For CFA, COM and UGS each homework assignment was worth the most at around 10-12% of the course grade, while for GEO, ENG, CNS, and BUS each homework assignment was only worth around 1-3% of the course grade.

Correlations

The number of homework assignments offered in a course was positively associated with having community learning opportunities (*r* = .20); it was negatively associated with reading activities (*r* = –.20) and exam percentage (*r* = –.26; a larger amount of homework was associated with a smaller percentage of the grade coming from exams; Figure 5). Percent of the grade from homework was associated with watching activities (*r* = .18) and negatively associated with informal retrieval practice (*r* = –.22). However, it was positively associated with many good things, including projects/presentations (*r* = .47), group activities (*r* = .25), and attendance requirements (*r* = .18). Recall that homework was coded

to include any part of the grade that comes from work done at home, including projects and papers, thus contributing to the strong association.

In-class assignments

The average number of in-class assignments per course was 2.40 (*Mdn* = 0, *SD* = 7.73) and they accounted for 2.40% of the total course grade on average (*Mdn* = 0, *SD* = 6.32; see Figures 2 and 6). Each in-class assignment was worth 0.39% of the course grade on average (*SD* = 1.76). Out of all courses, 80.74% did not have any in-class assignments (Figure 7). Among courses having at least one in-class assignment, the average number of in-class assignments was 13.87 (*Mdn* = 10, *SD* = 13.65), worth an average grade percentage of 12.97% (*Mdn* = 10, *SD* = 9.43).

College comparisons

The colleges allocating the largest percentage of the grade to in-class activities were GEO (7.85%, *Mdn* = 6, *SD* = 2.70), COM (4.50%, *Mdn* = 0, *SD* = 8.19), and EDU (3.14%, *Mdn* = 0, *SD* = 9.11); all other colleges were less than 3% on average, with medians of 0% (Figure 6). The only college with a substantial number of in-class activities was GEO (*M* = 12.05, *Mdn* = 12, *SD* = 1.00); CNS, CLA, and UGS were around 2.5 in-class assignments on average: all others had 1 or fewer. For none of the colleges was percent of grade per in-class assignment greater than 1%.

Correlations

Unsurprisingly, the number and grade-weight of in-class assignments were positively associated with in-class active learning (*r* = .19 for both) and with attendance requirements (*r* = .15, *r* = .23), and courses described as "flipped classrooms" had a larger number of in-class assignments (*r* = .20; see Figure 5). The grade weight of in-class

assignments was also associated with group activities (r = .21) and knowledge learning objectives (r = .16).

Participation

On average, participation accounted for 4.52% of the total course grade (Mdn = 0, SD = 6.99; see Figures 2 and 6). The percentage of courses awarding no participation points was 58.76%. Among courses awarding participation points, the average grade percentage from participation rises to 10.96% (Mdn = 10, SD = 6.97).

College comparisons

Colleges awarding the most points for participation were ARC (11.67%, Mdn = 10, SD = 2.5), CFA (10.29%, Mdn = 10, SD = 9.61), and UGS (7.85%, Mdn = 5, SD = 7.35). Those awarding the fewest points for participation were EDU (0.37%, Mdn = 0, SD = 1.92) and CNS (2.50, Mdn = 0, SD = 4.88).

Correlations

Courses with a larger grade weight for participation were more likely to have projects/presentations (r = .22), to have group activities (r = .28), to be lower-division courses (r = .28), and to have fewer exams (r = –.27). All correlations are presented in Figure 5. Strangely, courses not listing office hours in their syllabi awarded more points for participation on average (r = .35).

Extra credit and grade choice

Thus far, for each course, we have only considered the individual components of the grading rubric—the weight given to various graded coursework in the final grade breakdown—but in many courses there is the possibility of earning additional points from

outside of the rubric that get added on top of the final grade, or of effectively changing the rubric by reweighting various components to result in a higher final grade. These two features are discussed below.

Extra credit

Opportunities to earn extra credit points were offered in 32.19% of courses. The proportion of courses with extra credit was highest in EDU (81.48%) and BUS (57.25%), while no courses in ENG gave extra credit and only 13.07% of CLA courses did. Extra credit tended to be offered more in upper-division courses (r = .50), flipped-classroom courses (r = .22), courses using social media (r = .28), and courses that have knowledge learning objectives in the syllabus (r = .21). It tended to be offered less in core courses (r = –.23; see Figure 5).

Grade choice

Of all courses, 37.58% had grade choice, and as expected due to their exam-heavy grading rubrics, the science-focused colleges GEO and CNS lead the way with 72.73% and 71.28% respectively. BUS (34.35%), COM (31.96%) and ENG (28.00%) were middling. Contrastingly, and despite having the largest percent of grade from exams of any college, only 3.70% of EDU courses gave students grade choice, while CFA had no courses with grade choice at all. Grade choice was positively associated with both cumulative exams (r = .34) and cumulative finals (r = .56), but negatively associated with online/SMOC courses (r = –.25) and reading activities (r = –.49). In general, its associations mirrored those of other exam-related variables (see Figure 5).

PEDAGOGICAL APPROACHES

This section presents findings related to teaching practices that are not captured by the grading rubric or the number of assignments. The following sections present groups of variables together based on their *a priori* interrelatedness (i.e., before clustering was done).

Community and collaboration

Variables under this heading (group work, community learning opportunities, projects/presentations, and social media; see Figure 9) were chosen to reflect social connections both between students in a course and also between students and their community more broadly. The correlation heatmap (Figure 5) reveals that projects/presentations and group activities form a cluster that also includes participation percentage and attendance enforcement, while social media and community learning opportunities cluster together along with year (note small blue triangles on main diagonal).

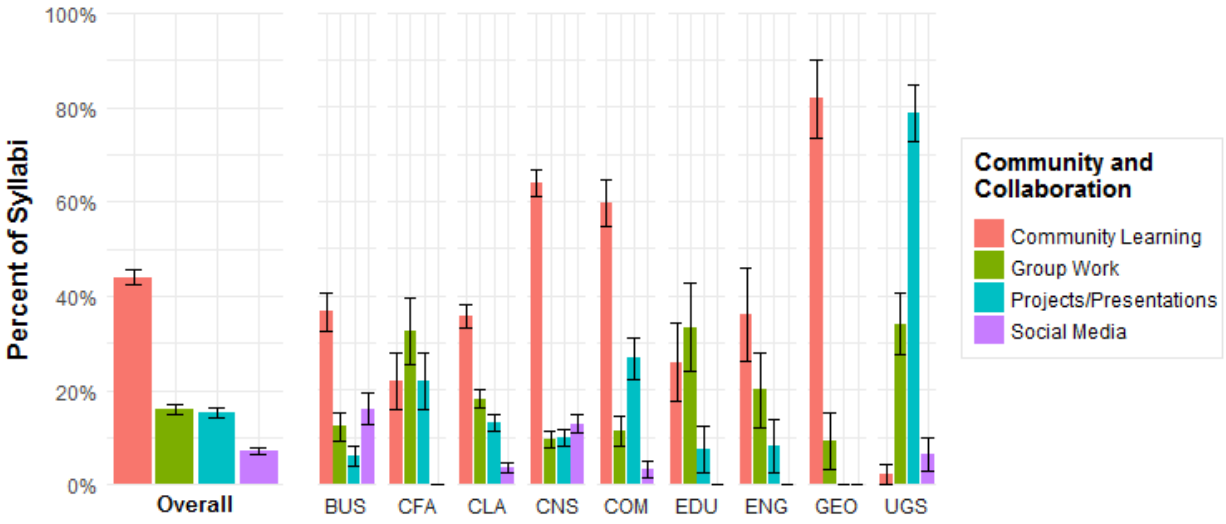

Figure 9 Percent of syllabi featuring each of the community and collaboration variables overall (left) and by college (right). Error bars show bootstrapped standard errors.

Group work

The proportion of courses having some form of group work (activities, discussions, etc.) was just 16.32%. The colleges UGS, EDU, and CFA were most likely to incorporate group work, with 34.04%, 33.33%, and 32.61% respectively. Colleges least likely to incorporate group work were GEO (9.09%), CNS (9.57%), COM (11.34%), and BUS (12.21%). See Figure 9 for a graphical display of these results.

As noted above, having group activities or discussions was positively associated with grade weight for in-class assignments (r = .21), homework (r = .25), and participation (r = .28), but negatively associated with number (r = –.35) and grade percentage (r = –.31) of exams (see Figure 5). Group work was also associated with having projects or presentations (r = .57), in-class active learning (r = .50), attendance requirements (r = .42), and stated learning objectives for knowledge (r = .25) and socio-emotional outcomes (r = .21).

Community learning opportunities

Overall, the proportion of courses having community learning opportunities was 44.00% (see Figure 9). Colleges with the greatest percentage of courses with community learning opportunities were GEO (81.82%), CNS (63.83%) and COM (59.79%). Colleges with the lowest percentage were UGS (2.13%), CFA (21.74%), and EDU (25.93%). In addition to positive associations with number of quizzes (r = .18) and homework assignments (r = .20), the presence of community learning opportunities was positively associated with social media (r = .55), in-class active learning (r = .35), informal retrieval practice (r = .40), and doing resources (r = .28); it was negatively associated with reading resources (r = –.33) as well as with projects and presentations (r = –.27; Figure 5).

Projects and presentations

Only 15.34% of all courses had students work on projects or give presentations (Figure 9). By a large margin, the college with the greatest proportion of courses with projects or presentations was UGS with 78.72%. The next greatest was in COM (26.80%), followed by CFA (21.74%) and CLA (13.07%). GEO had no courses with projects or presentations, and less than 10% of courses in CNS, ENG, EDU and BUS featured them (see Figure 9). In addition to the strong positive association with homework grade percentage and the negative association with exam variables noted above, projects and presentations were positively associated with doing resources (r = .36), participation percent-of-grade (r = .22), and attendance requirements (r = .25). The relatively strong, negative association between projects/presentations and community learning opportunities (r = –.27) is singular for variables in this otherwise positively associated grouping.

Social media

Overall, only 7.16% of courses reported incorporating social media: at the college level, 16.03% of BUS courses and 12.77% of CNS courses, and 6.38% of UGS courses used social media (Figure 7). Around 3% of courses in CLA and COM used social media, while none of the remaining colleges (CFA, EDU, ENG, and GEO) did. Use of social media in courses was associated only with community learning opportunities (r = .55), total enrollment (r = .20), and how recently the course was offered (r = .26).

In-class active learning and informal retrieval practice

Variables presented under this heading (in-class active learning, attendance requirement enforced, informal retrieval practice, and flipped classroom; see Figure 10) were chosen because they all relate to keeping students active during class-time. The correlation heatmap (Figure 5) shows a distinct cluster of positive associations between all

variables except attendance enforcement which, as noted above, clustered with group work and projects/presentations.

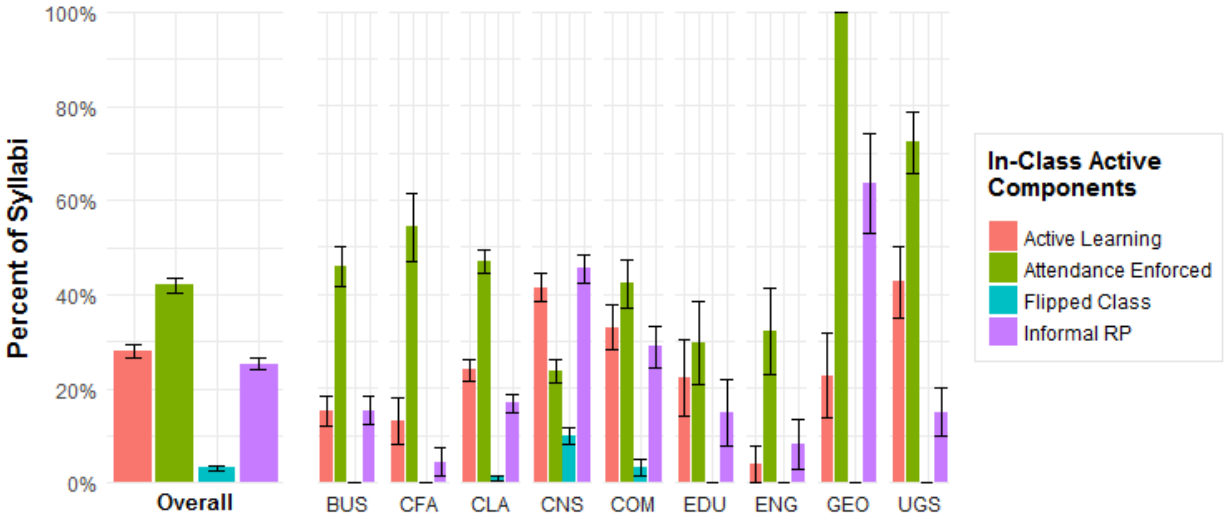

Figure 10 Percent of syllabi featuring each of the active-learning and retrieval-practice related variables overall (left) and by college (right). Error bars show bootstrapped standard errors.

In-class active learning

The proportion of courses having in-class active learning was 28.09% overall. Colleges with the greatest proportion of active-learning courses were UGS (42.55%) and CNS (41.49%); see Figure 10. Perhaps surprisingly, ENG had the lowest proportion of active-learning courses with just 4.00%; CFA and BUS were also low, with 13.04% and 15.26% respectively. In addition to positive associations with participation percent-of-grade (r = .20) and in-class assignment variables (r = .19), in-class active learning was positively associated with every other variable in this grouping (attendance enforcement, .24; informal retrieval practice, .73; flipped classroom, .48) as well as with group activities

(r = .50). It was negatively associated with total enrollment (r = –.22) and reading resources (r = –.34).

Attendance requirement enforced

Out of all courses in the dataset, 42.05% had some mechanism for enforcing attendance (see Figure 10); among GEO and UGS courses, the percentage was highest (100% and 72.34%, respectively). Attendance enforcement was lowest in CNS (23.76%), EDU (29.63%), and ENG (32.00%). Enforcing attendance was positively associated with participation (r = .61), homework (r = .18), and in-class assignment (r = .23) grade percentages; it was also associated with knowledge learning objectives (r = .24) and reading resources (r = .34). Attendance enforcement was negatively associated with all exam-related variables including grade choice.

Informal retrieval practice

The overall percentage of courses that incorporate informal retrieval practice was 25.30% (Figure 10). Here again, the exam-heavy science colleges GEO and CNS had the greatest proportions: 63.64% and 45.90%, respectively. COM was also relatively high with 28.87% while CLA, BUS, UGS, and EDU hovered around 15%. Colleges with the lowest proportion of informal retrieval practice were CFA (4.35%) and ENG (8.00%). Informal retrieval practice was associated with number and grade-weight of exams (r = .24, .19), as well as with cumulative exams and finals (r = .33, .35); it was negatively associated with homework grade weight (r = –.22) and reading resources (r = –.37).

Flipped classroom

Only 3.26% of courses mention being flipped classrooms, and most of these were in CNS, where 9.93% of courses were described that way (Figure 10). In COM 3.09% of

courses were flipped, and in CLA only 1.01% were. No other colleges had any courses with a flipped-classroom format. Flipped classrooms were more likely to give more in-class assignments (r = .20), extra credit (r = .38), grade choice (r = .34), and cumulative exams and finals (r = .40, .22); they were less likely to require reading resources (r = –.48).

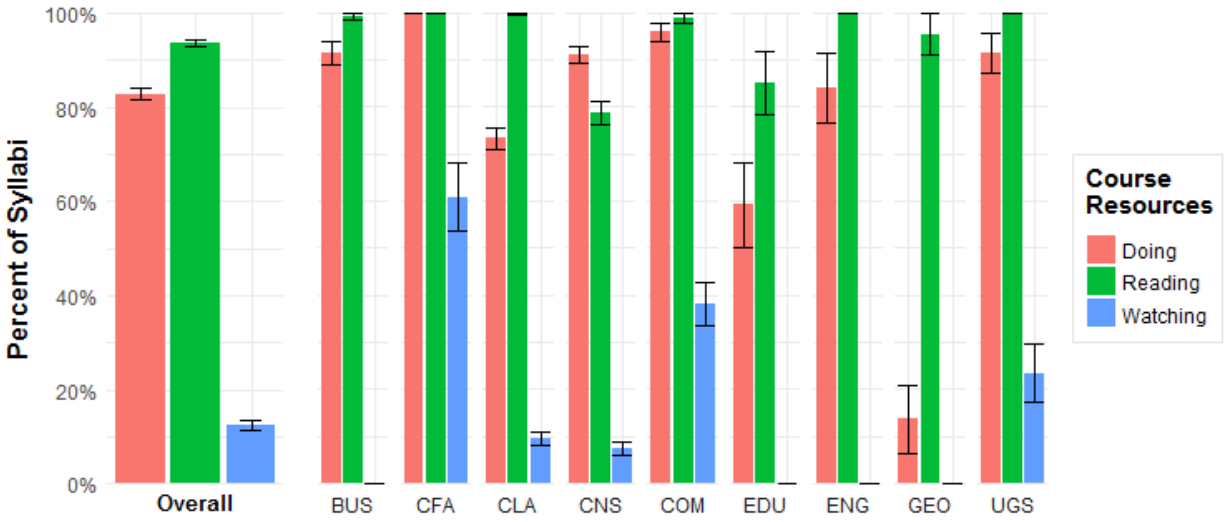

Figure 11 Percent of syllabi featuring each of the course resources variables overall (left) and by college (right). Error bars show bootstrapped standard errors.

Types of resources or activities

The proportion of courses having required reading resources was 93.67%, including 99-100% of all courses in CFA, BUS, CLA, ENG, UGS, and COM (see Figure 11). For CNS, 78.73% of courses had reading resources, while 85.19% of EDU courses did. The proportion of courses with watching resources was 12.45%. The college with the highest percentage of courses with watching resources was CFA (60.87%), followed by COM (38.14%), UGS (23.40%), CLA (9.55%) and CNS (7.45%). ENG, GEO, and BUS, had no courses with watching resources. Overall, 82.56% of courses had doing resources. Colleges with a large proportion of courses with doing resources were CFA (100%), COM (95.88%),

BUS (91.60%), UGS (91.49%) and CNS (91.13%). Colleges with few courses listing doing resources were GEO (13.63%) and EDU (59.26%).

All course resources were negatively associated with exam variables. Requiring reading resources was negatively associated with being a core course ($r = -.34$) and being a flag course ($r = -.44$), as well as several other good things, including community learning opportunities ($r = -.33$), in-class active learning ($r = -.34$), informal retrieval practice ($r = -.37$), and being a flipped classroom ($r = -.48$). Watching resources were associated with homework and quiz grade weights ($r = .18$, $r = .34$) as well as with knowledge learning objectives ($r = .27$). Doing activities/resources were strongly associated with homework variables (number, $r = .74$; grade weight, $r = .64$) as well as with projects and presentations ($r = .36$).

INSTRUCTOR EXPECTATIONS

In this section, descriptive results are presented for the presence of learning objectives in syllabi and the degree to which a syllabus is organized and complete. Later, results of syllabus text mining are presented which, together with learning objectives, help inform our understanding of instructor communication to students via the syllabus.

Stated learning objectives

Learning objectives for knowledge, skills, and socio-emotional outcomes were all positively associated with each other (knowledge with skills, $r = .68$; knowledge with socio-emotional, $r = .52$; skills with socio-emotional, $r = .71$), indicating that instructors who include one type of learning objective are likely to include others as well.

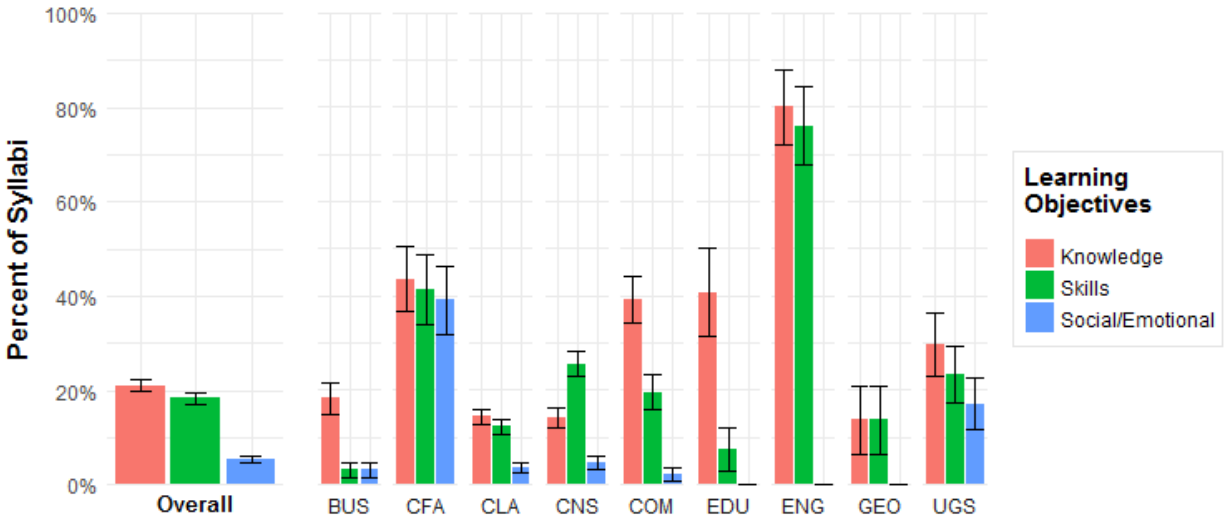

Figure 12 Percent of syllabi featuring each of the learning objectives variables overall (left) and by college (right). Error bars show bootstrapped standard errors.

Learning objectives for knowledge outcomes

The percentage of courses that listed knowledge learning objectives in their syllabus was 21.11% overall (Figure 12). By college, 80.00% of ENG courses, 43.48% of CFA courses, 40.74% of EDU courses, and 39.2% of COM courses had learning objectives for knowledge. In contrast, only 13-14% of courses in each of GEO, CLA, and CNS had them. As noted above, knowledge learning objectives were positively associated with attendance enforcement, group activities, extra credit opportunities, watching activities, and grade weight for in-class assignments and quizzes.

Learning objectives for skills outcomes

For skills learning objectives, 18.41% of courses had provided them in the syllabus (Figure 12). The pattern by college is somewhat different than it was for knowledge learning objectives: ENG and CFA were still among the most likely to list them (76.00%

and 41.30%, respectively), but CNS jumped into third position with 25.5% having learning objectives for skills; in contrast, only 3.05% of BUS courses and 7.41% of EDU courses did. Learning objectives for skills were associated with cumulative final exams (r = .26) but with no other variables.

Learning objectives for socio-emotional outcomes

A much lower proportion of courses had socio-emotional learning objectives overall (5.49%; Figure 12). Indeed, in only two colleges was the percentage of courses with these learning objectives greater than 5%: in CFA 39.13% of courses had them and in UGS 17.02% of courses had them (CNS was third overall with 4.61%). No socio-emotional learning objectives were listed in any EDU, ENG, or GEO course. Core courses were also more likely to include socio-emotional outcomes (r = .31).

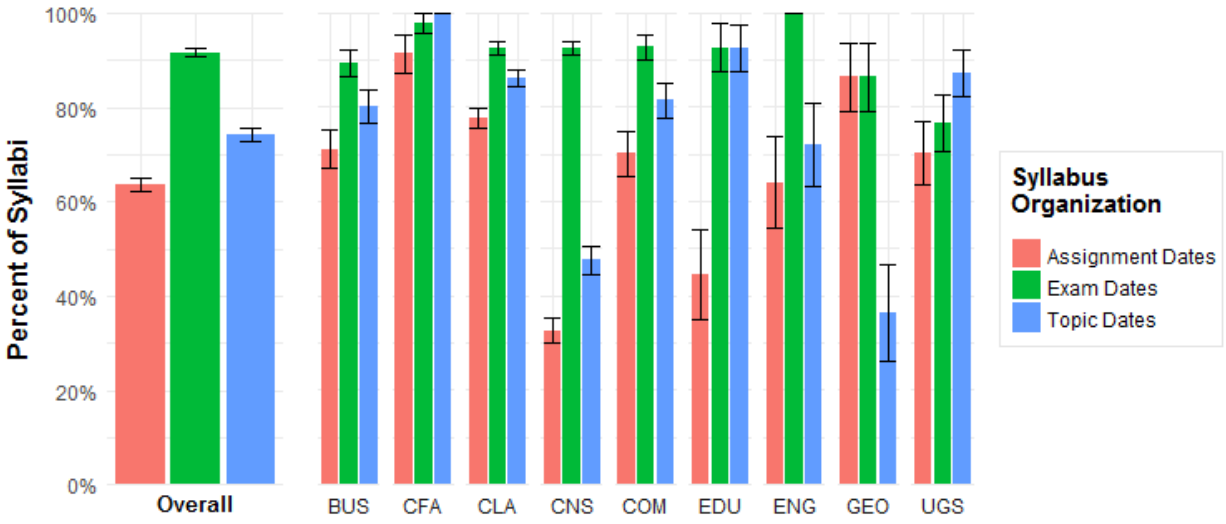

Figure 13 Percent of syllabi featuring each of the syllabus organization variables overall (left) and by college (right). Error bars show bootstrapped standard errors.

Syllabus organization and completeness

Across all syllabi, 85.49% of courses have a list of course topics while 74.32% provide dates for course topics (see Figure 13). Colleges with noticeably low rates of providing dates for course topics are GEO (36.36%) and CNS (47.52%). The vast majority of courses had dates listed for exams (91.72%). Only UGS was low in this regard (76.60%). However, only 63.63% of courses had a calendar of due dates for all course work. CFA (91.30%), GEO (86.36%), and CLA (77.64%) were good in this respect, while CNS (32.62%) and EDU (44.44%) faired more poorly. Additionally, 78.32% of instructors mentioned office hours in their syllabus. The colleges with the lowest proportion of courses listing office hours was EDU (48.15%), while CFA and GEO had the highest proportion (95.65% and 95.45%, respectively).

Syllabus-completeness variables were all positively associated with each other (all rs > .5), though instructor office hours only significantly related to dates for all assignments (r = .27). Listing exam dates was related to exam variables, but nothing else. However, listing dates for all assignments was negatively associated with several of the active learning and community/collaboration variables discussed above, including community learning opportunities (r = –.25), in-class active learning (r = –.25), informal retrieval practice (r = –.29), and cumulative finals (r = –.26); on the other hand, providing assignment due dates was positive associated with attendance enforcement (r = .25), office hours (r = .27), and reading resources (r = .33). The same pattern of correlations was exhibited by courses listing dates for course topics, which was also positively associated with group activities (r = .25) and knowledge learning objectives (r = .24) while being negatively related to number of exams (r = –.20).

CHANGES IN COURSE VARIABLES OVER TIME

How recently a course was offered is associated with several course variables, indicating change over time across the six years for which we have data (see Figure 14 for linear trends and regression coefficient estimates). Specifically, the proportion of online/SMOC courses, flag courses, courses using social media, and courses offering community learning opportunities have increased significantly over time since 2011. The average number of homework assignments and quizzes has also increased over time (by roughly half an assignment per year for both), while the percent of course grade from exams appears to be in decline (all *p*s <.001).

Other trends of note include an increase in in-class active learning and informal retrieval practice (increasing linearly by 2-3% per year on average), but a decline in projects/presentations over time. Total enrollment per course appears to be increasing (by around 5 students per year on average), as does the use of learning objectives for skills and socio-emotional outcomes. All other variables appear relatively steady across this time period. Finally, one exciting trend that is not explicitly depicted in Figure 14 is a significant increase in the total number of assignments (all exams, quizzes, homework, and in-class assignments): On average, the average number of assignments students completed for a course increased by 1.3 each year, $t(1073) = 4.158, p < .001$.

Notice that there are several variables for which a linear trend does not appear appropriate (e.g., flipped classroom, online/SMOC courses, stated learning objectives, in-class assignments). Still, all of the best-practice variables highlighted above do appear to be increasing linearly with time (i.e., number of quizzes and homework assignments, in-class active learning, informal retrieval practice), and others (such as percentage of the grade from exams) appear to be decreasing linearly with time.

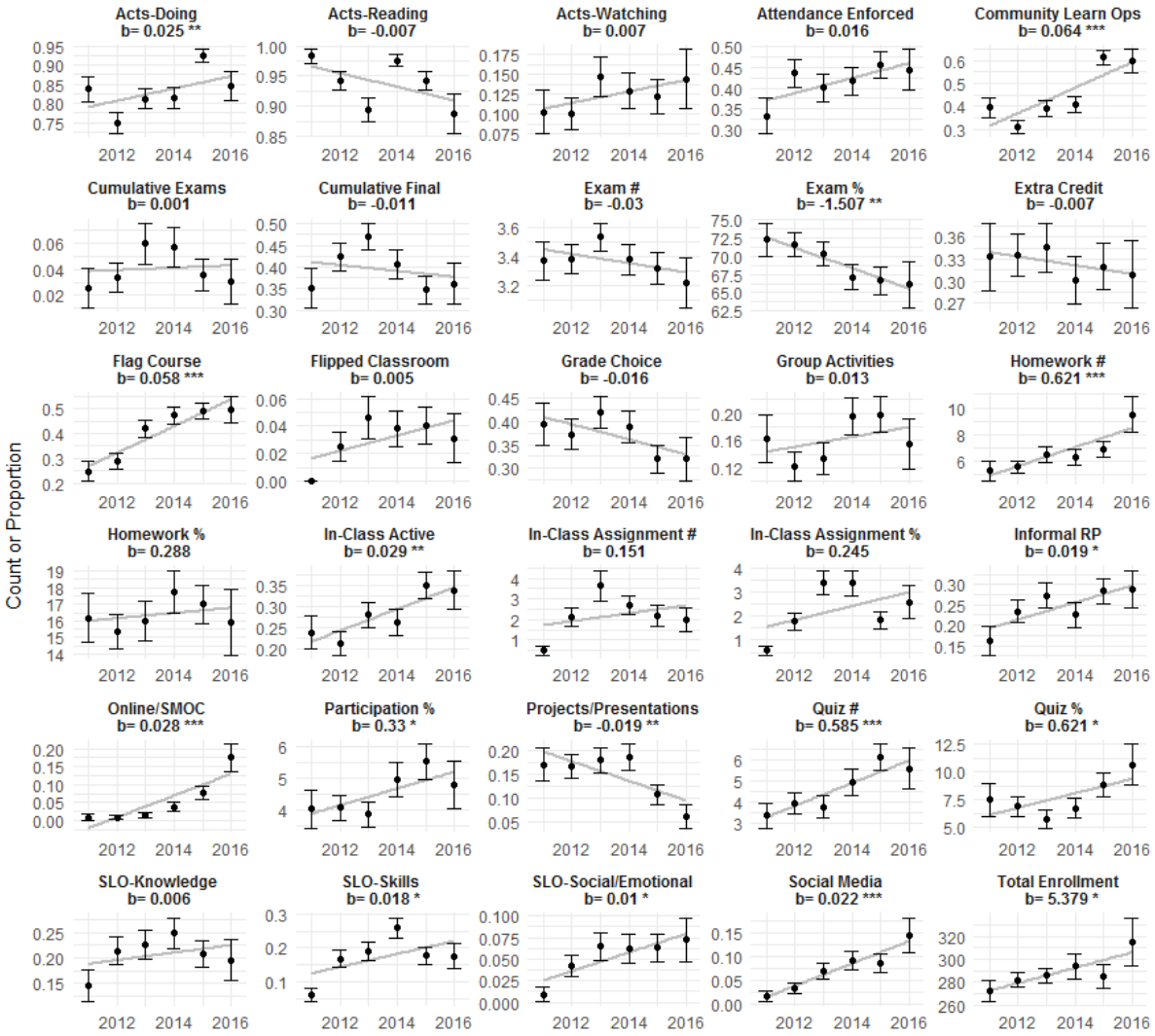

Figure 14 Trends over time for 30 syllabus variables. Note that vertical axis scales differ for each panel, and that the coefficient estimate is unstandardized. Error bars show bootstrapped standard errors. Unadjusted significance indicators for slopes are as follows: *$p < .05$; **$p < .01$; ***$p < .001$.

FACTOR ANALYSIS OF COURSE VARIABLES

In an effort to further identify patterns in the data presented above, an exploratory factor analysis was conducted to summarize the interrelationships among course variables (which were of mixed type; polychoric, biserial, and Pearson's correlations as described above alongside caveats about the novelty of the method) with a smaller number of interpretable, orthogonal factors; these can be thought of as latent factors that give rise to the observed course variables. Principal-axis factor extraction and Varimax rotation were performed using the *fa()* function in the R package *psych* to determine a robust factor solution in the absence of multivariate normality and to aid in description. Horn's parallel analysis suggested 13 factors, and indeed 13 factors had eigenvalues greater than 1. However, examination of the scree plot of eigenvalues by rank revealed that 5 factors were appropriate. The five-factor solution was favored for parsimony and ease of interpretation.

The communalities of variables were mostly large, though there were a few exceptions: *Year, Office Hours, Credit Hours, In-Class Assignment #, and In-Class Assignment %* were notably low, with communalities less than 0.2, but the other 34 variables had satisfactory communalities and thus shared common variance with other items (see Table 4 for all loadings and communalities.). The five-factor solution explained 43% of the variance: each of the five factors (PA1 through PA5) accounted for 11%, 10%, 8%, 8%, and 6% of the variance, respectively. Though principal-axis factoring and an orthogonal rotation were used for the reasons mentioned above, maximum likelihood extraction and oblique rotation were compared, yielding very similar loadings for the five-factor solution and producing no notable correlations among the factors (all rs < 0.10 in magnitude). Furthermore, our sample size to variable ratio was quite high (about 30:1), much higher than that reported in many studies and more likely to produce an accurate factor solution (Costello & Osborne, 2005).

Table 4 Factor loadings and communalities for factor analysis with principal axis factoring after varimax rotation

	PA1	PA2	PA3	PA4	PA5	Communality
Group Activities	**0.65**	**0.3**	0.14	-0.03	0.08	0.54
Projects/Presentations	**0.61**	0.1	0.04	-0.09	-0.03	0.39
Homework %	**0.57**	-0.21	-0.09	0.36	-0.02	0.51
Attendance Enforced	**0.52**	0.06	0.14	-0.09	0.06	0.3
Participation %	**0.47**	0.03	0.02	-0.03	0.18	0.26
Reading Acts	**0.33**	**-0.52**	0.21	-0.1	**-0.38**	0.59
SLO-Social/Emotional	**0.32**	0.23	**0.32**	0.24	0	0.31
In-Class Active	0.28	**0.7**	-0.12	0	0.21	0.63
SLO-Knowledge	0.27	0.17	**0.4**	0.26	-0.29	0.41
Doing Acts	0.25	0.22	-0.06	**0.65**	-0.09	0.55
Dates for Topics	0.24	-0.18	**0.89**	-0.12	0.07	0.9
Watching Acts	0.21	-0.08	0.23	**0.3**	0.03	0.2
Quiz %	0.18	-0.28	0.02	**0.46**	0.02	0.32
In-Class Assignment %	0.13	0.09	0.08	0	0.07	0.04
Flag Course	0.11	0.13	-0.05	0.12	**0.62**	0.43
SLO-Skills	0.11	**0.39**	0.27	0.23	-0.03	0.28
Course Topics	0.07	0.02	**0.82**	-0.08	-0.02	0.68
Flipped Classroom	0.07	**0.78**	0.01	0.17	-0.13	0.67
Quiz #	0.01	0.06	-0.04	**0.4**	0.11	0.18
In-Class Assignment #	0.01	0.2	0.12	0.02	0.01	0.05
Homework #	0	0.14	-0.05	**0.5**	-0.03	0.27
Assignment Dates	-0.01	-0.3	**0.67**	-0.07	-0.05	0.55
Year	-0.03	0.04	0	**0.33**	0.06	0.12
Core Course	-0.05	0.09	0.04	0.13	**0.77**	0.63
Online/SMOC	-0.09	**-0.66**	0.05	**0.5**	0.01	0.7
Extra Credit	-0.11	**0.3**	0.01	0.12	**-0.39**	0.27
Credit Hours	-0.12	0.09	-0.01	0.01	0.1	0.03
Informal RP	-0.14	**0.65**	-0.16	0.09	0.01	0.48
Course Level	-0.15	0.1	-0.07	-0.03	**-0.86**	0.78
Total Enrollment	-0.18	-0.24	0.11	**0.31**	0.16	0.22
Community Learn Ops	-0.2	**0.38**	-0.06	**0.44**	0.14	0.4
Office Hours	-0.27	0.01	0.19	-0.04	0.04	0.11
Social Media	-0.28	0.04	0.01	**0.45**	-0.26	0.35
Cumulative Final	-0.29	0.41	-0.26	0.06	0.02	0.32
Cumulative Exams	**-0.39**	**0.43**	-0.02	**-0.34**	-0.12	0.47
Exam Dates	**-0.43**	0.11	**0.71**	0.09	0.05	0.71
Grade Choice	**-0.64**	**0.49**	-0.09	0.01	0	0.66
Exam #	**-0.66**	**0.32**	-0.01	-0.21	0.03	0.59
Exam %	**-0.72**	**0.3**	0.04	**-0.57**	-0.06	0.95

Note. PA1 was labeled "Groups, Projects, and Participation"; PA2 was labeled "Active Classroom, Cumulative Tests"; PA3 was labeled "Course Planning/Organization"; PA4 was labeled "Supportive High Workload"; PA5 was labeled "Required Lower-Division".

Simple structure was achieved through Varimax rotation and informative labels were created to reflect the pattern of factor loadings (see Figure 15 for a graphical representation of the factor loadings). Note again that this factor analysis has not been validated and labels are given simply to aid in the interpretation of interrelationships among so many variables. Specifically, the first factor extracted (PA1) was given the label "Groups, Projects, and Participation" based on large positive loadings for Group Activities (.65), Projects/Presentations (.61), Participation % (.47), and Attendance Enforced (.52) and large negative loadings for exam variables (around –.70). The second factor, PA2, was labeled "Active Classroom, Cumulative Tests" on the basis of large loadings for Flipped Classroom (.78), In-Class Active Learning (.70), Informal Retrieval Practice (.65), Cumulative Exams (.43), and Cumulative Final (.41); "classroom" was emphasized by a large negative loading for Online/SMOC (–.66).

The third factor extracted, PA3, was labeled "Course Planning/Organization"; it had large loadings for Course Topics (.82), Dates for Topics (.89), Exam Dates (.71), Assignment Dates (0.67), and Learning Objectives for both Knowledge (.40) and Socio-emotional outcomes (.32); most other loadings were very close to zero. The fourth factor, PA4, was labeled "Supportive, High Workload." It had large loadings for Doing Activities (.65), number of homework assignments (Homework #; .50), number and grade weight of quizzes (Quiz # and Quiz %; .46 and .40, respectively), Community Learning Opportunities (.44), and Social Media (.45). Finally, PA5 was labeled "Required Lower-Division." This last factor had large loadings for Core Course (.77), Flag Course (.62) and a large negative loading for upper division Course Level (–.86). Note that several course variables—including instructor office hours, number of credit hours, number of in-class assignments, and grade percentage of in-class assignments—did not load appreciably on any of the five factors.

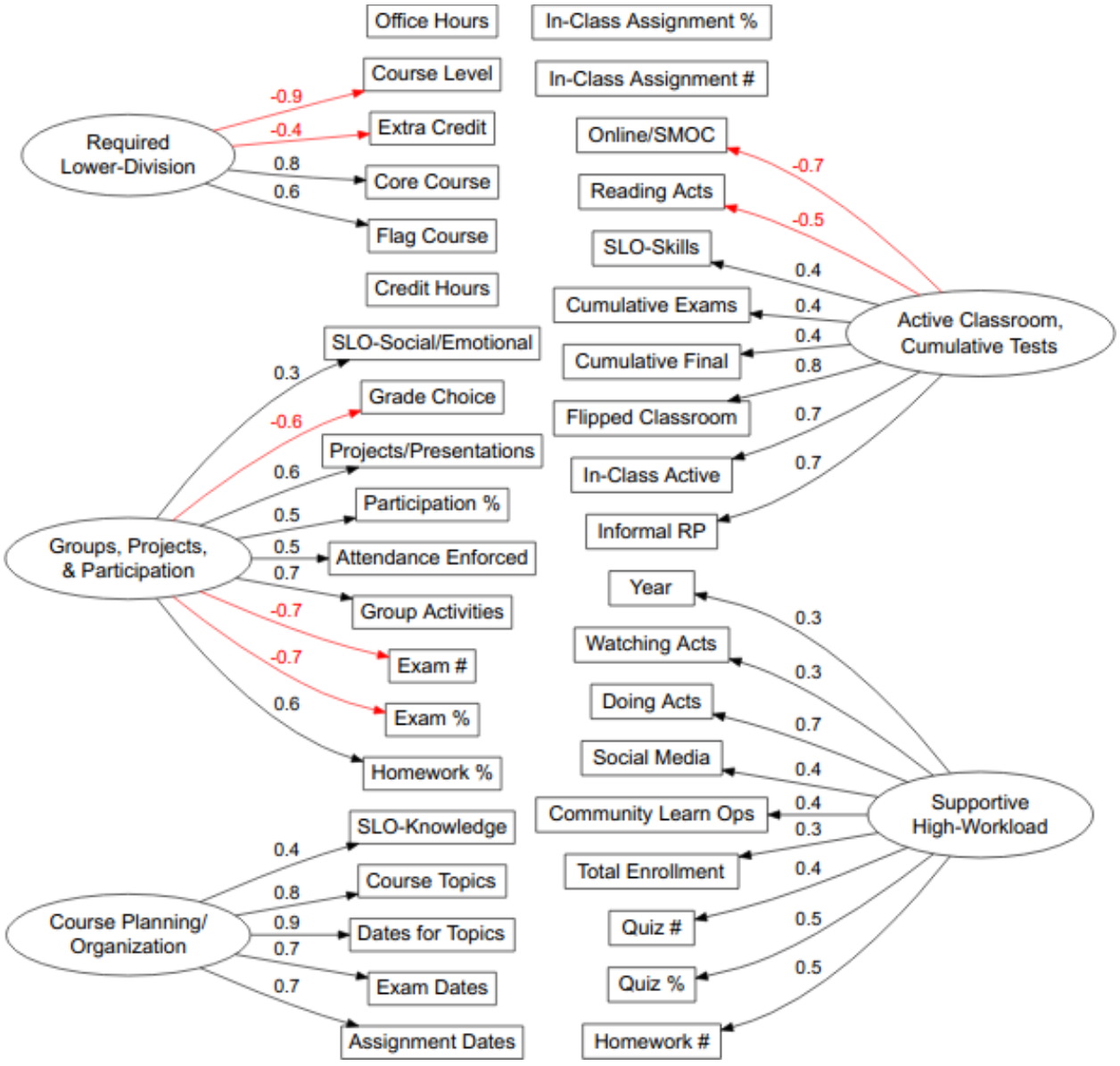

Figure 15 Visual depiction of Varimax rotated five-factor solution. Factors extracted using principal-axis factoring. Loadings 0.3 or greater in magnitude are depicted; negative loadings are shown in red. Note that this visualization is for descriptive, exploratory purposes only; see Table 4 for loadings and communalities.

Overall, these factors are readily interpretable and help to more simply explain many of the correlations discussed in the previous section. Two factors were structural in nature: Required Lower-Division and Course Planning/Organization had large loadings for course attributes (e.g., core course, lower-division) and syllabus structure (e.g., list of course topics, assignment dates) respectively. The other three factors give an interesting picture of how course-level variables tend to co-occur across a range of classes. The Groups, Projects, and Participation factor is characterized by participation grades, group activities, socio-emotional learning objectives, a large percentage of the grade coming from relatively few assignments or projects completed outside of the classroom, and few if any exams. On the other hand, the Active Classroom, Cumulative Tests factor represents classrooms with a great emphasis on in-class active learning, frequent quizzing, cumulative exams, and learning objectives for skills, but a lack of emphasis on readings or work done out-of-class. Finally, a third type of classroom emerged in factor labeled Supportive High-Workload: this factor is characterized by lots of individual assignments (e.g., quizzes and homework), the availability of community learning opportunities (such as TA-led study sessions), and the incorporation of social media into the classroom. Furthermore, courses with high scores on this factor also tended to have higher enrollment and to have been offered more recently than other courses, indicating that course model may be becoming more prevalent. A final important caveat to this section is that certain unwanted dependencies among variables may arise because most course instructors (approximately 60%) are represented more than once in our data set. Indeed, in the dataset there are 3.45 courses per instructor on average ($Mdn = 2$, $SD = 4.1$).

Cluster analysis of course variables

Due to the fact that course variables were of mixed type (i.e., nominal and continuous), factor scores could not be easily estimated. Instead, a cluster analysis was performed using the PAM algorithm on a matrix of Gower dissimilarities created for a subset of course variables (attendance enforcement, projects/presentations, in-class active learning, informal retrieval practice, group activities, cumulative exams, cumulative final, and flipped classroom). A 6-cluster solution was chosen based on highest average silhouette width (though going higher than 10 clusters results in even higher average silhouette widths).

The results of a t-Distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction technique (Maaten & Hinton, 2008), are visualized in Figure 16 depicting both cluster assignment and college. Cluster cohesiveness appears to be best for Clusters 1, 3, and 5. Note that the t-SNE algorithm preserves relationships between points in a high-dimensional space, but because the absolute point position is arbitrary, axes are given a quantitative interpretation (the technique merely serves to visualize non-linear cluster separation). The top panel of Figure 17 presents bar plots indicating, for each college, the percentage of courses assigned to each cluster. The bottom panel presents, for each cluster, the percentage of courses having each of the pedagogical approaches related to spacing and retrieval practice (Figure 8), community and collaboration (Figure 9), and in-class active learning (Figure 10) discussed above. Note that color coding of clusters facilitates comparisons within the top and bottom panels of Figure 17 and is consistent with the colors used to denote cluster assignment in Figure 16.

Cluster 1 is characterized by high attendance enforcement and low frequencies of all other variables; CFA had a large percentage of courses grouped into this cluster (37%), followed by CLA, COM, and GEO (all ~23%). Cluster 3 is characterized by having low

frequencies of all course variables; almost 70% of EDU courses were grouped into Cluster 3, followed by ENG and COM (~38%). Cluster 6 is characterized by high frequencies of attendance enforcement and cumulative final exam, but low frequencies of all other course variables. Most GEO courses (55%) were assigned to this cluster, with UGS second (26%) and BUS third (21%).

Cluster 2 is characterized by high frequencies of informal retrieval practice, in-class active learning, and cumulative finals; CNS was the only college to be significantly represented in this cluster (44% of courses), followed by GEO (18%). Notably, no courses from CFA were assigned to this cluster. Cluster 5 is characterized by high frequencies of cumulative final exam, but low frequencies for all other variables. CNS and BUS had the largest proportion of courses assigned to this cluster (27% and 20%, respectively). Cluster 4 is characterized by high frequencies of group activities, attendance enforcement, and in-class active learning. Note that it is the only cluster in which group activities appear appreciably. UGS has the highest proportion of courses in this cluster (30%), with EDU second (18%) and CFA third (13%).

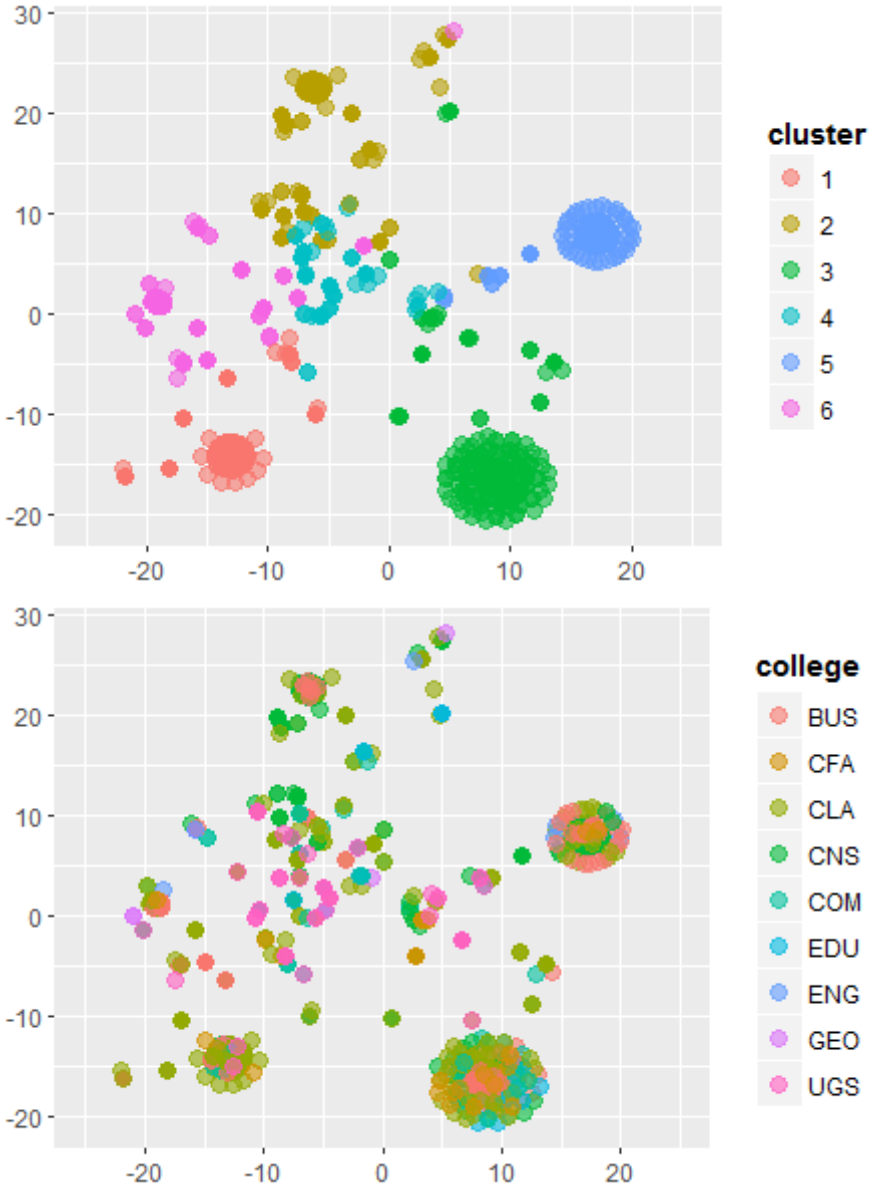

Figure 16 Visualization of cluster separation using t-SNE, colored by cluster assignment (top) and by college (bottom). Note that because axes are not easily interpretable, they are given arbitrary units and remain unlabeled.

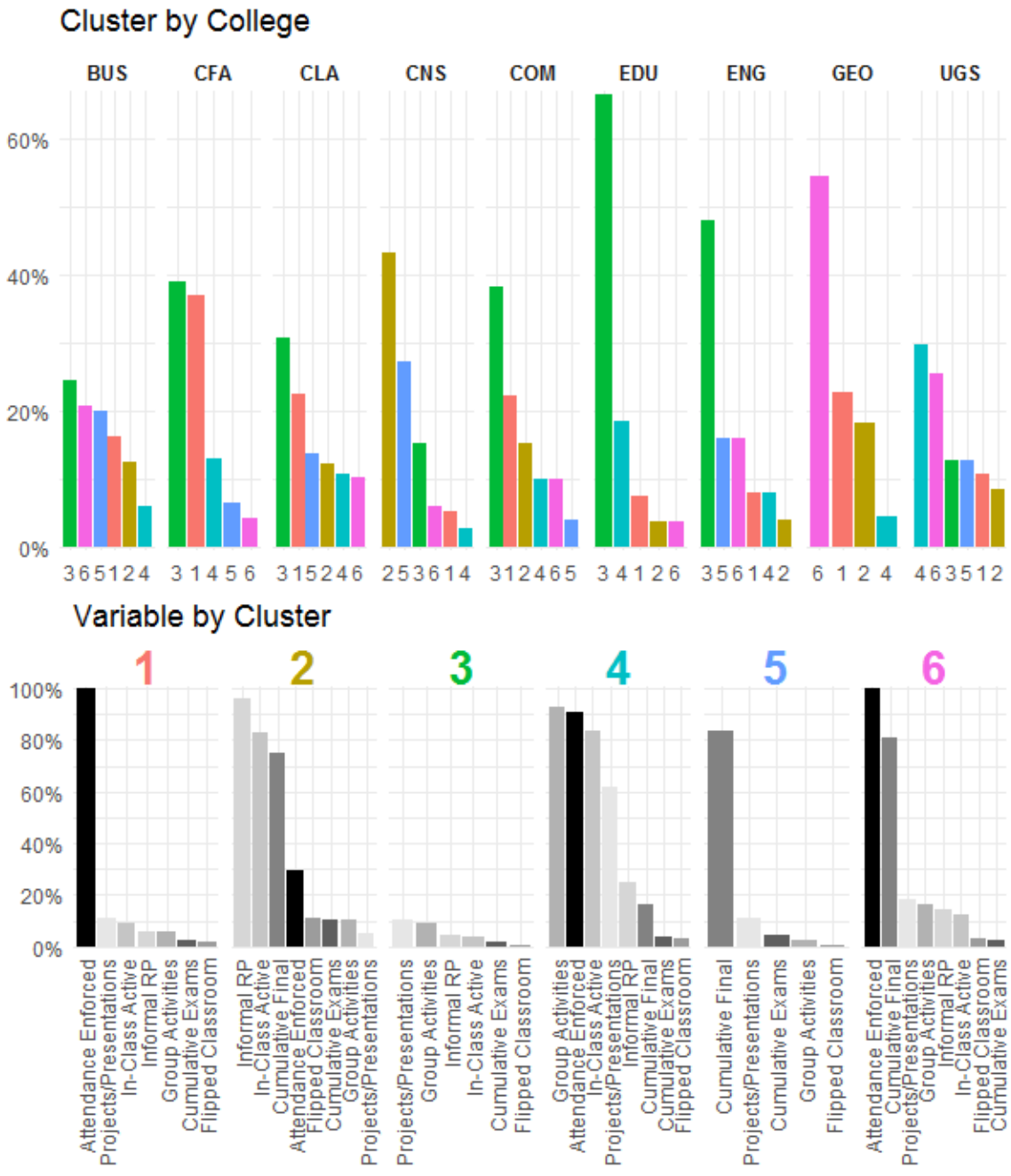

Figure 17 Percentage of courses that were assigned to each cluster by college (top). Percentage of courses having each pedagogical variable (bottom; variables not appearing within a cluster are omitted).

SYLLABUS TEXT MINING

Communication

Instructor communication to students was assessed in two broad ways. First, all syllabi were processed using Linguistic Inquiry and Word Count software (LIWC2015) described above. For each college, scores on LIWC variables—either percentile scores, word counts, or average percentages of total syllabus words related to each category, depending on variable type—are presented in Tables 5 and 6, which also include comparison scores for each variable by norming sample. Note that certain variables have very small standard deviations within colleges (e.g., Analytical Thinking), while others are rather large (e.g., Emotional Tone). Each LIWC variable will be discussed in detail below.

Table 5 Mean (*SD*) of syllabus word-count and LIWC summary variables by college

	Analytic		**Clout**		**Authentic**		**Tone**		**Words/Sent.**		**6+ Letters**		**Dictionary**	
	M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
College														
BUS	91.09	(3.08)	73.42	(5.49)	21.48	(7.48)	50.50	(9.10)	23.03	(5.59)	27.17	(2.61)	75.70	(4.82)
CFA	89.74	(5.66)	76.18	(7.80)	19.95	(8.16)	47.11	(10.81)	22.39	(8.96)	23.57	(3.31)	71.08	(10.57)
CLA	91.98	(4.36)	67.26	(7.91)	20.48	(10.20)	37.91	(12.55)	20.85	(10.07)	25.22	(3.11)	69.68	(7.84)
CNS	90.01	(3.86)	69.75	(7.12)	23.04	(7.06)	36.91	(10.41)	22.17	(11.91)	23.05	(2.29)	75.36	(6.57)
COM	91.89	(3.44)	73.93	(7.27)	21.60	(8.80)	44.10	(11.36)	18.60	(3.60)	25.93	(2.84)	72.73	(4.75)
EDU	88.11	(5.37)	77.78	(7.54)	24.73	(10.84)	54.98	(8.15)	18.19	(3.64)	25.54	(2.58)	77.90	(4.73)
EGN	95.55	(3.09)	54.50	(7.68)	18.82	(6.06)	41.60	(7.59)	20.47	(6.57)	31.61	(3.78)	68.55	(6.98)
GEO	89.90	(1.90)	69.45	(5.02)	23.82	(9.27)	39.59	(11.80)	20.95	(3.90)	25.65	(1.65)	69.32	(3.95)
UGS	90.27	(9.98)	71.06	(9.04)	19.41	(9.17)	42.17	(14.61)	17.30	(4.22)	25.93	(2.76)	70.51	(6.45)
Grand Mean	91.14	(4.66)	70.00	(8.14)	21.44	(8.87)	41.19	(12.63)	21.00	(9.30)	25.10	(3.23)	72.51	(7.42)
Norms for comparison[a]														
Blogs	49.89	-	47.87	-	60.93	-	54.50	-	18.40	-	14.38	-	85.79	-
Expressive Writing	44.88	-	37.02	-	76.01	-	38.60	-	18.42	-	13.62	-	91.93	-
Novels	70.33	-	75.37	-	21.56	-	37.06	-	16.13	-	16.30	-	84.52	-
Natural Speech	18.43	-	56.27	-	61.32	-	79.26	-	-	-	10.42	-	91.60	-
NYT	92.57	-	68.17	-	24.84	-	24.84	-	21.94	-	23.58	-	74.62	-
Twitter	61.94	-	63.02	-	50.39	-	50.39	-	12.10	-	15.31	-	82.60	-
Grand Mean	56.30	(17.58)	58.00	(17.51)	49.20	(20.92)	54.20	(23.27)	17.40	(16.38)	15.60	(3.76)	85.18	(5.36)

Note. Analytic = Analytical Thinking (Pennebaker et al., 2014; indexes formality of writing: lower scores mean a more informal, narrative style), Clout indexes confidence (Kacewicz et al., 2012); Authentic = Authenticity (indicates more personal /honest language; Newman et al., 2003);
[a] Normed averages are average percentiles from a large sample of text from blogs, expressive writing, novels, natural speech, newspaper articles, and twitter (included for comparison only).

Table 6 Mean (*SD*) of Pronouns, Comparisons, Negations, Affiliation, and Achievement by college

	I (1st sing.)		**We (1st pl.)**		**You (2nd)**		**Comparisons**		**Negations**		**Affiliation**		**Achievement**	
	M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
College														
BUS	0.696	(0.52)	0.197	(0.18)	2.833	(0.68)	1.381	(0.46)	0.888	(0.29)	0.863	(0.36)	1.870	(0.70)
CFA	0.403	(0.26)	0.691	(0.57)	2.549	(1.22)	1.458	(0.37)	0.851	(0.30)	1.655	(0.72)	1.183	(0.41)
CLA	0.526	(0.53)	0.373	(0.31)	1.914	(1.07)	1.393	(0.56)	0.754	(0.35)	1.059	(0.55)	1.278	(0.47)
CNS	0.512	(0.42)	0.511	(0.31)	2.815	(1.04)	1.992	(0.42)	1.087	(0.32)	1.132	(0.41)	1.521	(0.47)
COM	0.433	(0.45)	0.364	(0.24)	2.359	(1.01)	1.539	(0.38)	0.934	(0.35)	1.521	(0.75)	1.457	(0.52)
EDU	0.428	(0.29)	0.742	(0.58)	2.385	(0.86)	1.316	(0.24)	1.006	(0.34)	1.962	(0.77)	1.384	(0.60)
EGN	0.536	(0.63)	0.105	(0.15)	0.652	(1.05)	1.310	(0.31)	0.591	(0.40)	0.735	(0.43)	2.208	(0.41)
GEO	0.325	(0.35)	0.447	(0.28)	2.375	(0.46)	1.448	(0.18)	0.859	(0.20)	0.957	(0.35)	1.024	(0.41)
UGS	0.420	(0.50)	0.381	(0.36)	1.957	(1.29)	1.268	(0.41)	0.650	(0.37)	1.348	(0.67)	1.513	(0.59)
Grand Mean:	0.510	(0.48)	0.400	(0.35)	2.330	(1.12)	1.560	(0.53)	0.880	(0.36)	1.150	(0.59)	1.450	(0.56)
Norms for comparison														
Blogs	6.26	-	0.91	-	1.32	-	2.17	-	1.81	-	2.2	-	1.27	-
Expressive Writing	8.66	-	0.81	-	0.68	-	2.42	-	1.69	-	2.45	-	1.37	-
Novels	2.63	-	0.61	-	1.39	-	2.13	-	1.68	-	1.39	-	0.91	-
Natural Speech	7.03	-	0.87	-	4.04	-	2.35	-	2.42	-	2.06	-	0.99	-
NYT	0.63	-	0.38	-	0.34	-	2.39	-	0.62	-	1.69	-	1.82	-
Twitter	4.75	-	0.74	-	2.41	-	1.89	-	1.74	-	2.53	-	1.45	-
Grand Mean	4.99	(2.46)	0.72	(0.83)	1.7	(1.35)	2.23	(0.95)	1.66	(0.86)	2.05	(1.28)	1.30	(0.82)

Note. Normed averages are average percentiles from a large sample of texts including blogs, expressive writing, novels, natural speech, newspaper articles, and twitter (included for comparison only).

LIWC summary variables

Figure 18 depicts syllabus scores on LIWC's four summary variables: *analytical thinking*, *clout*, *authenticity*, and *tone*. First, college syllabi in general can be compared to averages based on various kinds of text. Note that in general, compared to overall averages, syllabi tend to be high in analytical thinking (more formal, less narrative) and lower in authenticity (less personal and open) relative to the norming sample of texts. They also tend to be higher in clout (expressing more confidence and leadership) but less warm in terms of emotional tone than the norming sample, though here there are exceptions by individual college here: ENG courses in our sample tend to be low in clout, while EDU courses are high in emotional tone.

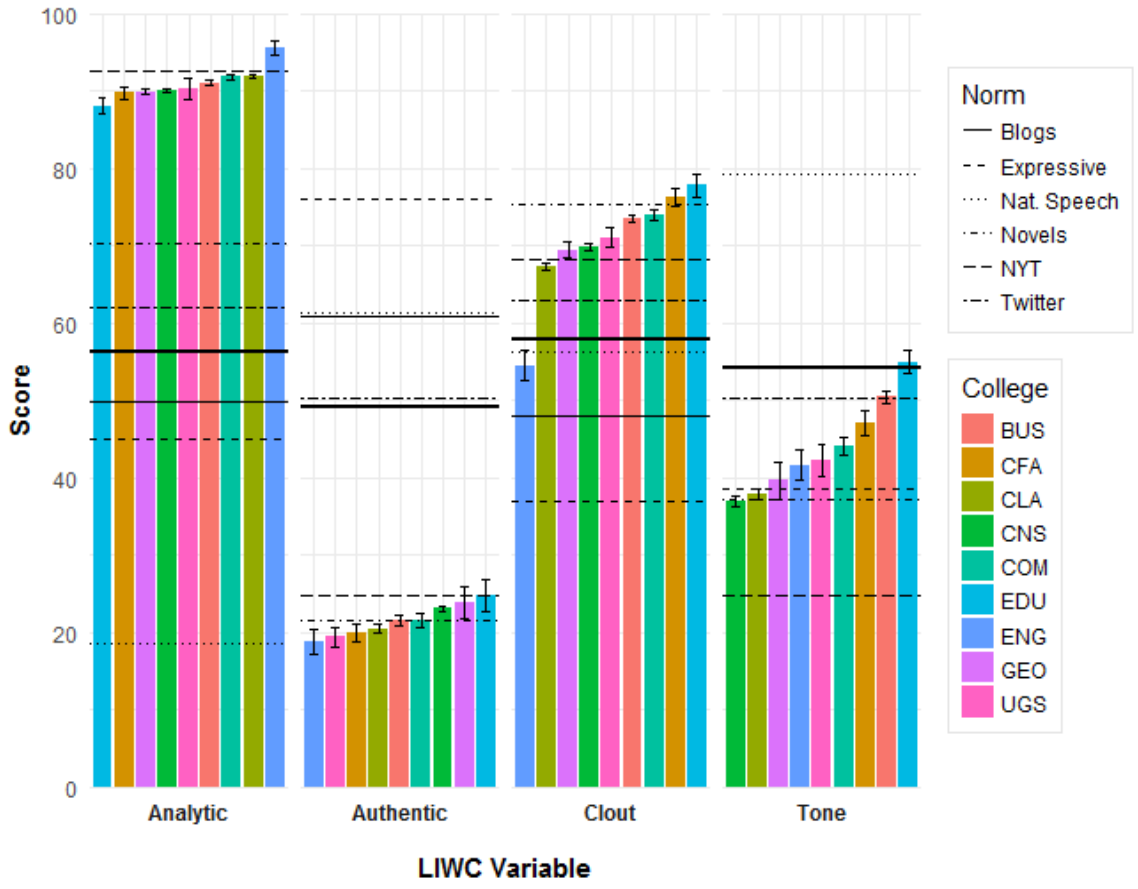

Figure 18 Mean LIWC summary variable scores by college. Error bars show bootstrapped standard errors. Horizontal lines show normed averages for each variable

Shifting focus to differences among colleges on each of these four variables, it can be seen that ENG and EDU continue to be opposites on each dimension: EDU is the college with the lowest mean score for analytical reasoning, while ENG has the highest by a large margin; in contrast, EDU has the highest scores for authenticity on average, while ENG has the lowest (though the standard errors are larger here). ENG has the lowest average clout score by a significant margin, but EDU has the highest average clout score among colleges. Emotional tone was the only variable for which the two colleges did not have

scores at opposite ends of the scale: while EDU courses in our sample score highest on emotional tone, ENG was fourth lowest (with CNS taking the bottom spot).

Interestingly, it is not just EDU and ENG courses that show polarity across the four summary variables: certain colleges tend to cluster together in such a way that certain variables appear correlated with others. For example, colleges scoring high in clout tend to score high in emotional tone but low in analytical thinking: the top four colleges for clout and tone are the same (EDU, CFA, COM, and BUS). Lowest in both tone and clout were the science colleges CNS, GEO, ENG, and also CLA. Additionally, the top two colleges and bottom two colleges are completely switched between clout and analytical thinking (EDU and CFA lowest in analytical thinking but highest in clout; ENG and CLA highest in analytical thinking but lowest in clout). If we compute the correlation across all syllabi, we find that clout and tone are slightly correlated ($r = .12$, adjusted $p < .001$), while clout and analytical thinking are strongly negatively correlated ($r = -.39$, adjusted $p < .001$). Conversely, colleges scoring high on authenticity tend to score lower on analytical thinking ($r = -.30$, adjusted $p < .001$), but also score slightly higher on clout ($r = .13$, $p < .001$). There was no relationship between tone and either authenticity or analytical thinking.

Finally, when compared to averages for various types of text (blogs, expressive writing, natural speech, novels, New York Times articles, and Twitter), syllabi tend to closely resemble NYT articles in terms of analytic thinking, authenticity, and clout; novels are also close to syllabi on these dimensions. Furthermore, on these three variables, syllabi are very dissimilar to natural speech, expressive writing, and blogs (see horizontal lines in Figure 18). The sole exception to this pattern is ENG, whose syllabi tend to have similar clout scores as natural speech. However, in terms of tone, syllabi tend to be much more positive than NYT articles, bearing a resemblance to novels, Twitter, blogs, and expressive writing. Syllabi from EDU, BUS, and CFA were more similar in tone to blog posts and

Twitter tweets (more positive), while CNS, CLA, and UGS were more similar in tone to novels and expressive writing.

Text-level descriptives

Figure 19 displays the average proportion of dictionary words, proportion of words containing six or more letters, and number of words per sentence for each college. There are no extremely compelling patterns to observe here, other than to note that EDU has the highest proportion of dictionary words and the lowest number of words per sentence on average: these syllabi seem to use language that is more straightforward. Another observation is that ENG has the lowest percentage of dictionary words but the highest percentage of words with six or more letters. This suggests that ENG is using a greater proportion of specialized terms that do not appear in the dictionary, but that tend to have more letters. However, the same might be expected of CNS, which does not follow this trend.

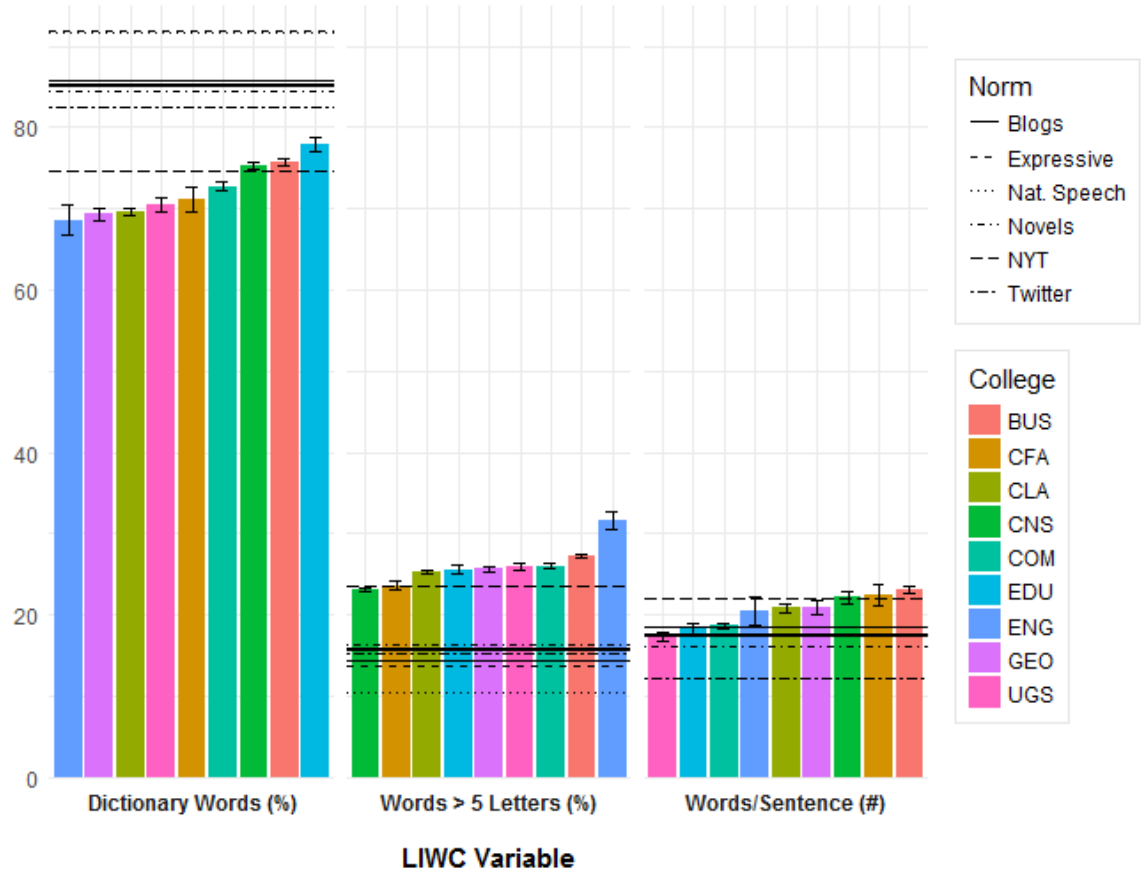

Figure 19 Mean counts for text-level descriptive variables. Error bars show bootstrapped standard errors. Horizontal lines show normed averages for each variable

Compared to averages for various types of text, syllabi resemble NYT articles in having a relatively low proportion of dictionary words; this is likely a function of many specialized terms and proper nouns in each case. Syllabi also closely resemble NYT articles in their proportion of words with over five letters and in their average number of words per sentence, although for this last we see some differences among colleges: UGS, ENG, and COM have fewer words per sentence on average and more closely resemble expressive writing and blog posts.

While exploring these coarse-grained text variables, it is sensible to also consider average syllabus length by computing total word count. This has been done in two different ways (Figure 20). First, LIWC generated a raw word count for each syllabus based on spaces between character strings. This estimate is raw in the sense that it does not remove numbers (e.g., those appearing in the grading rubric), punctuation, or stopwords. The second plot of average syllabus word count was computed manually after such text cleaning had taken place. In general, BUS (raw: 3770; clean: 2144), CFA (raw: 3854; clean: 1908), and CNS (raw: 3687; clean: 1739) have longer syllabi on average, while ENG (raw: 1798; clean: 1039), GEO (raw: 1909; clean: 1018), and UGS (raw: 2114; clean: 1064) have shorter syllabi on average, leaving CLA (raw: 2270; clean: 1146), COM (raw: 2420; clean: 1274), and EDU (raw: 2401; clean: 1508) in the middle. Comparing the two word-count plots, it can be seen that removing the numbers, stopwords, punctuation, and foreign characters significantly reduced average syllabus word count, sometimes by nearly half: this drop was especially pronounced in CNS.

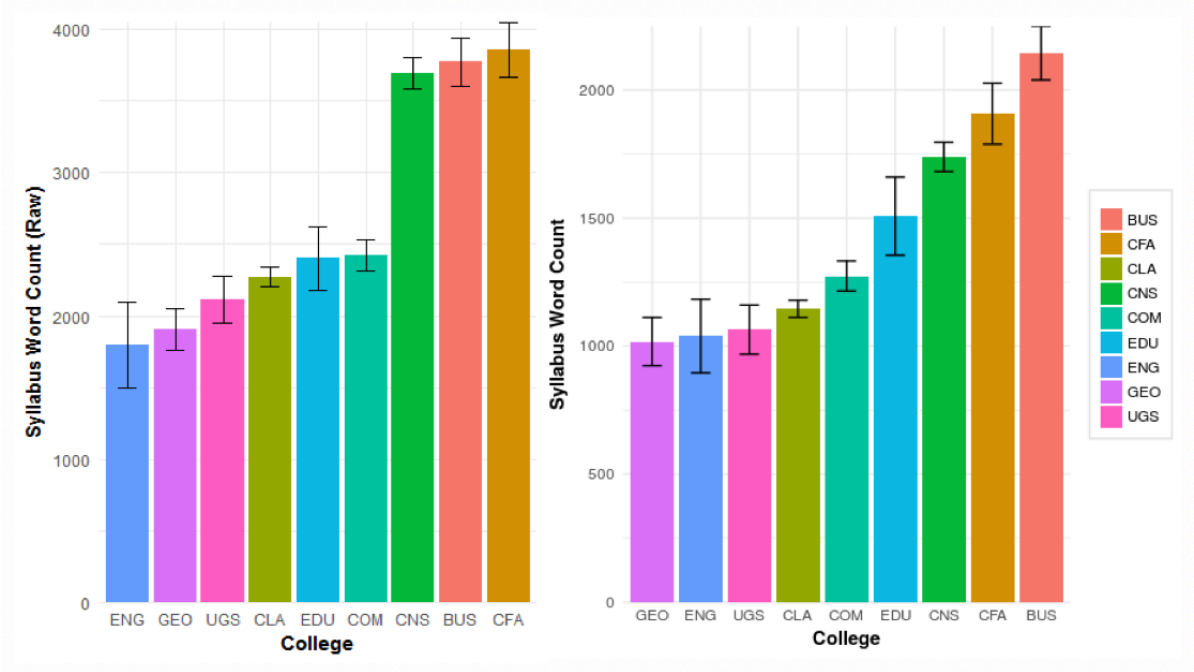

Figure 20 Mean syllabus word count by college. Left panel shows raw counts before text cleaning; right panel shows counts after cleaning text to remove non-words. Note that y-axes differ between panels. Error bars show bootstrapped standard errors.

Pronouns

Examination of differential pronoun use in syllabi by college revealed several trends, including one that polarizes EDU and ENG yet again: EDU had the highest proportion of first-person plural pronouns (e.g., *we*, *our*) per syllabus, while ENG had the lowest proportion (Table 5; Figure 21). BUS was not far behind with the second least, but they topped the charts for use of first-person singular (e.g., *I, my*) and second-person singular (e.g., *you, your*) pronouns. In general, second-person singular pronouns were used far more than the other pronouns, with the sole exception of ENG, which used far fewer of them than other colleges. Pronoun use was somewhat correlated: syllabi using more first-

person singular pronouns and first-person plural pronouns also tended to use more second-person singular pronouns (r = .33 and r = .37, respectively; adjusted ps < .001). However, using a higher proportion of first-person singular pronouns was not correlated with greater use of first-person plural pronouns (r = .05; adjusted p = .13).

Looking at how syllabi compare with other types of text in their use of pronouns, interesting differences can be seen (Figure 21). For use of first-person pronouns, syllabi are most similar to NYT articles on average, with relative few of both. However, it can be seen that while a low rate of first-person plural pronoun use is relatively common across different types of texts, the rate of first-person singular pronouns is quite variable; indeed, they are used 7-8 times more often in expressive writing, natural speech, and blogs than they are in syllabi, which are actually lower than NYT articles on average. A different trend emerges for second-person pronouns: for most colleges, frequency of use looks a lot like it does for Twitter, blogs, and novels. Here again, ENG is an exception, with low levels of use more closely resembling NYT articles and expressive writing.

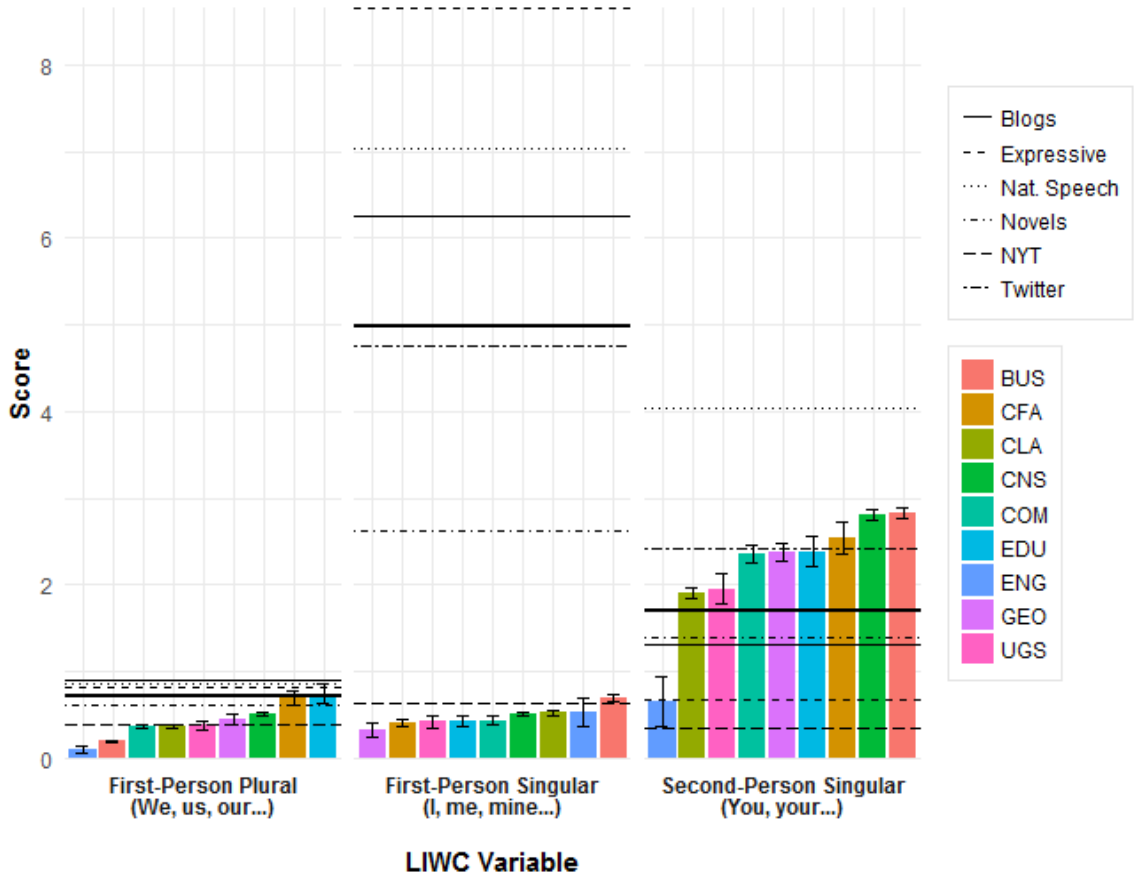

Figure 21 Mean syllabus pronoun counts by college. Error bars show bootstrapped standard errors. Horizontal lines show normed averages for each variable

Comparisons and Negations

The use of comparison words (e.g., *greater*, *after*) and negation words (e.g., *no*, *never*) was examined to see whether different colleges put more or less emphasis on comparison and whether colleges differed in their use of restrictive, negative language (e.g., "no late work, no exceptions"). From the right side of Figure 22 we can see that, in general, syllabi contained a greater proportion of comparison words than negation words, but use of negations and use of comparisons were correlated ($r = 0.34$, $p < .001$). Courses

in CNS made far greater use of comparisons than did courses in other colleges, with BUS and COM tied for a distant second. BUS and CNS made the greatest use of negations as well: they were at the top in both categories. Consistent with the moderate correlation, the colleges with the fewest comparisons (UGS and ENG) were also the colleges with the fewest negations.

Syllabi made use of negations with a relatively low frequency, similar to NYT articles and very different from natural speech. However, in terms of making comparisons, syllabi were quite low relative to all comparison text means; the closest comparison text sample in terms of comparison frequency was Twitter tweets. As noted above, CNS had by far the highest frequency of comparison words, surpassing the average for tweets.

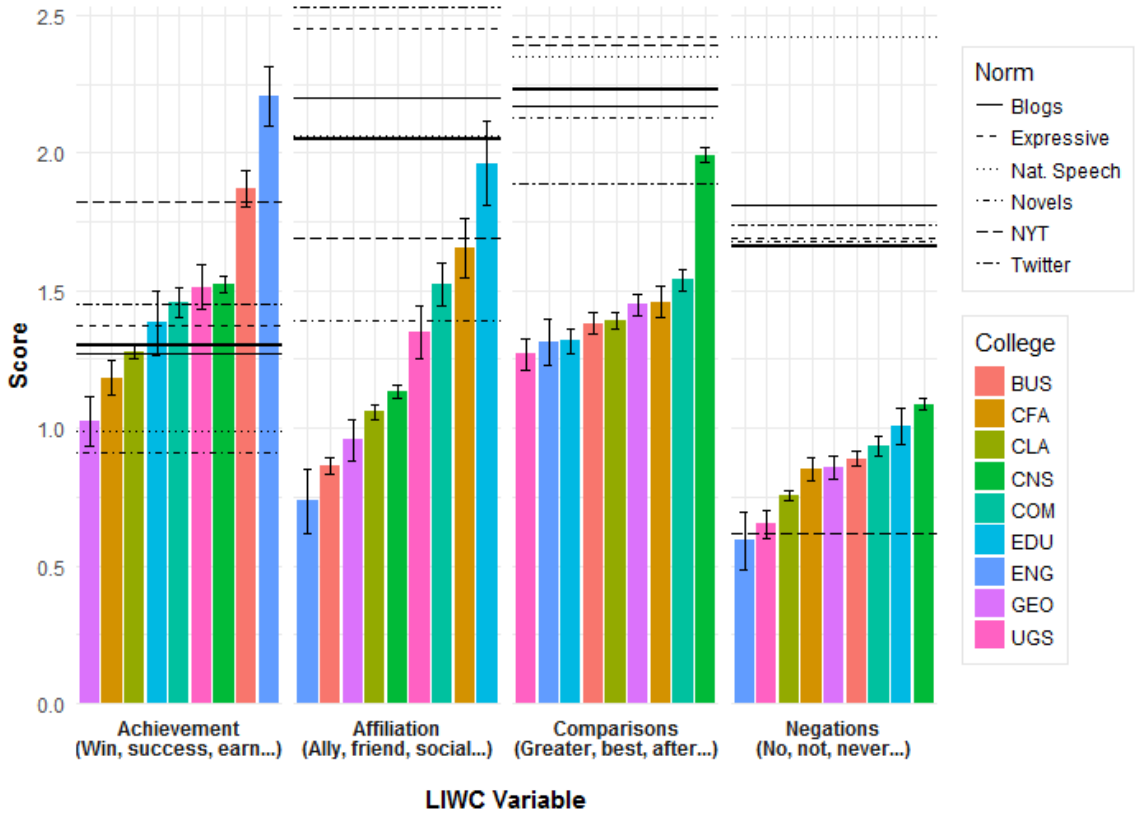

Figure 22 Mean syllabus counts for words related to achievement, affiliation, comparisons, and negations by college. Error bars show bootstrapped standard errors. Horizontal lines show normed averages for each variable.

Achievement and Affiliation

Words in the achievement and affiliation categories reflect an attention to, or an awareness of, achievement (e.g., *success*, *better*) or affiliation (e.g., *friend*, *social*), respectively, on the part of the speaker or writer (Figure 22, left two panels). Words emphasizing achievement were especially common in ENG and BUS; words emphasizing affiliation were especially common in EDU and CFA. Thus, for these dimensions as well, ENG and EDU can be observed on opposite ends of the spectrum. Interestingly, the colleges with the highest use of achievement words (ENG and BUS) were also the colleges

with the lowest use of affiliation words. Some colleges, such as UGS and COM, used a moderate amount of words related to both achievement and affiliation. Despite their differential use across colleges, the proportion of words related to achievement and to affiliation within syllabi were not significantly correlated.

With respect to comparison texts, syllabi diverged in their resemblance based on college: ENG and BUS used achievement-related words frequently, at a rate similar to NYT articles, while UGS used them very infrequently, similar to novels and natural speech. For all other colleges, the use of achievement words was middling, similar to Twitter, blogs, and expressive writing. Use of affiliation-related words was lower in syllabi than for other types of text media, most closely resembling that of novels and NYT articles on average. However, EDU was higher than the rest, showing greater similarity to natural speech in use of affiliation terms.

Sentiment analysis

While LIWC computes a tone summary that indexes emotional valence, it does so using the full text from each syllabus. Thus, no cleaning of text data takes place before the computation (e.g., no removal of stopwords or punctuation), which could skew the results. Furthermore, there are many extant lexicons for emotional words that may have different strengths and weaknesses; furthermore, some of these explore different types of emotional terms (e.g., trust, anticipation, fear) rather than just labeling them positive or negative (see Analysis section).

Emotional valance

Figure 23 shows average scores for each colleges' syllabi on emotional valence (higher scores positive), computed using three lexicons popular for sentiment analysis as

described above. The first thing to note is that there is consistency, as well as discrepancy, among the various lexicons in their sentiment scores. CNS, BUS, and UGS are consistently in the bottom five (more negative), while EDU, ENG, and CFA are consistently in the top four (more positive). However, certain colleges jump around depending on which lexicon is used: for example, GEO is given the highest score using the *bing* lexicon, but the third lowest score by both the NRC and AFINN lexicons. Though not as extreme, CLA also shows some inconsistency across lexicons, though tending mostly toward the middle of the scale.

Comparing these sentiment analysis results to average LIWC tone scores by college (cf. Figure 18) raises similar issues to those above: compared to rankings from the three lexicons, CNS stays put at the bottom, UGS and COM remain in the middle, and CFA and EDU continue to achieve very high scores. However, BUS is rated much more positively by LIWC than by the three lexicons, where it was among the lowest in emotional tone, and ENG was given a far more negative rating by LIWC than it had gotten using each of the three lexicons.

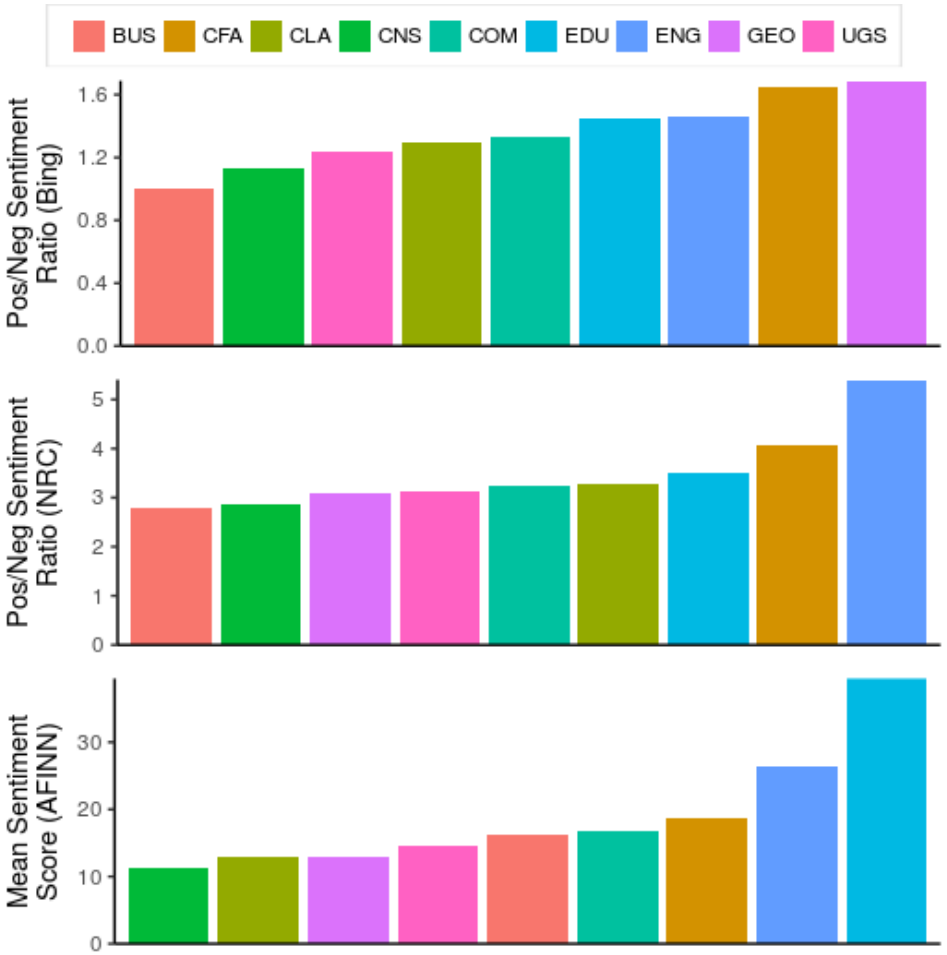

Figure 23 Mean syllabus emotional valence scores by college for each of three common sentiment lexicons

The differences between LIWC scores and the lexicon-based sentiment scores are likely due to the fact that very different methods and data were used to compute them: LIWC scores were calculated based on whole-text syllabi using a proprietary scoring system, while the lexicon scores were calculated after cleaning the text by removing stopwords, proper nouns, numbers, URLs, etc. The differences between lexicons are more difficult to account for because the data was the same in each case; however, the lexicons were created by different people for different purposes using different methods (see

Analysis section). Also, the summary variable for one of the lexicons was the sum of positive and negative word scores of different magnitudes, while for the others it was the ratio of words labeled *positive* to words labeled *negative*. For each pairwise combination of emotional valence scores (positive-negative ratio or sum), the correlation across all 1075 syllabi was 0.56 for *bing* and NRC, 0.43 for *bing* and AFINN, but only 0.17 between NRC and AFINN (all ps <.0001).

Fortunately, in spite of all this variability, certain things can be said with confidence: relative to courses in other colleges, CNS courses in our sample have a more negative emotional tone while EDU and CFA have a more positive emotional tone, regardless of the scoring method used. Certain colleges such as UGS and COM have a consistently neutral average for emotional tone among the courses in this sample, with COM being slightly more positive. Colleges with particularly inconsistent results across methods were BUS and ENG, which were near the top using some methods and near the bottom using others. Certain words are inherently ambiguous with respect to sentiment in the absence of context (e.g., *well*, *kind, like*); further work should be done to develop methods that account for such differences in context.

Sentiment across eight emotions

A richer picture of the emotional content of a document can be gotten by using sentiment labels that give detail beyond *positive* and *negative*. To achieve such a picture, a lexicon of eight potentially overlapping sentiment categories was used to label words in syllabi (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust). Figure 24 presents the average proportion of syllabus words falling into each category by college; Figure 25 presents the category word counts, but as a proportion of all *emotion-labeled words* rather than the total number of words in the syllabus as before. In the former, it can

be seen that UGS is the most emotional college by far, ranking first in anger, disgust, sadness, and fear, but also in joy and trust; UGS is also second highest in anticipation words and third highest in surprise words. CLA also ranks near the top in every category.

The least emotional colleges appear to be GEO, CNS, BUS, and GEO, scoring at the bottom on almost every sentiment category. Note that these are also the colleges that tended to score most negatively on emotional valence. There are three exceptions to this generally low-emotion set: CNS is relatively high in fear, GEO is relatively high in surprise, and ENG is relatively high in trust. COM, EDU, and CFA tended to rank near the middle across the board with just a few exceptions.

When conditioning on overall emotionality (i.e., taking only the *emotion* words in each syllabus and calculating the proportion of these words in each emotion category; Figure 25), a slightly different picture emerges. First, note that trust words make up about 28% of all syllabus emotion words, followed by anticipation words (19%); fear, joy, and sadness words are the next most common (11.3%, 11.0%, and 10.6%, respectively), with the least common emotions being surprise, anger, and disgust (7.6%, 7.0%, and 4.9%, respectively). UGS still gets the top score for anger, disgust, and sadness, but anticipation is led by ENG, COM, GEO, and CNS (emotion words in these colleges are likely to be anticipation words) while fear is dominated by CNS, CLA, and BUS. On the flip side, CNS, ENG, BUS, and GEO used few joy words relative to other emotion words, but ENG, GEO, and BUS were especially high in trust-related words (when one of these colleges uses an emotion word, there is greater than 30-35% chance it will be a trust word) while CFA and UGS were at the bottom, having more emotion words in other categories: these two colleges led the way in joy, sadness, and anger. Finally, surprise words make up a large proportion of the emotion words used in GEO, COM, and CLA compared to other colleges. The ranking of the proportion of each emotion is also interesting to note. While trust and

anticipation were the first and second most frequently occurring emotions, fear was thirdmost in CNS, sadness was thirdmost in UGS, and joy was thirdmost in CFA, CLA, EDU, COM, GEO, and ENG.

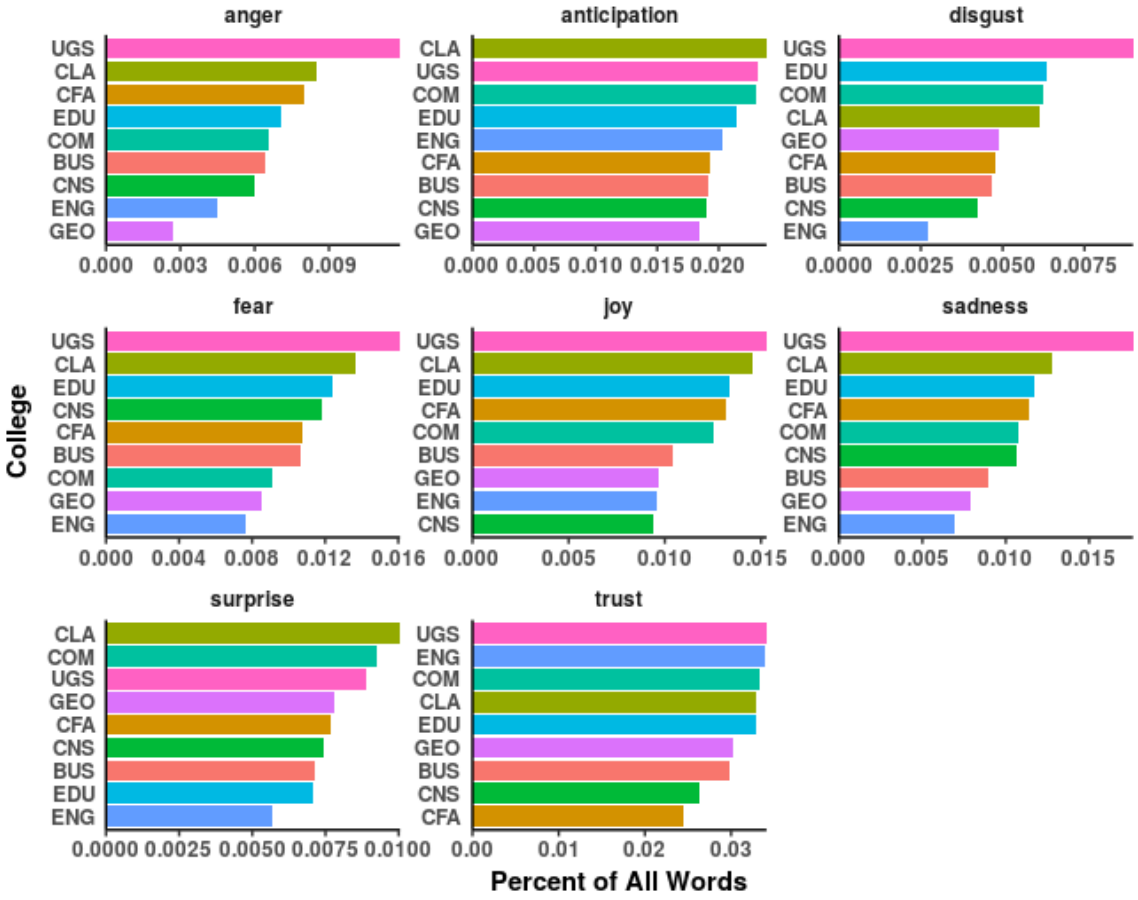

Figure 24 Mean proportion of all syllabus words falling into each of eight emotional categories by college

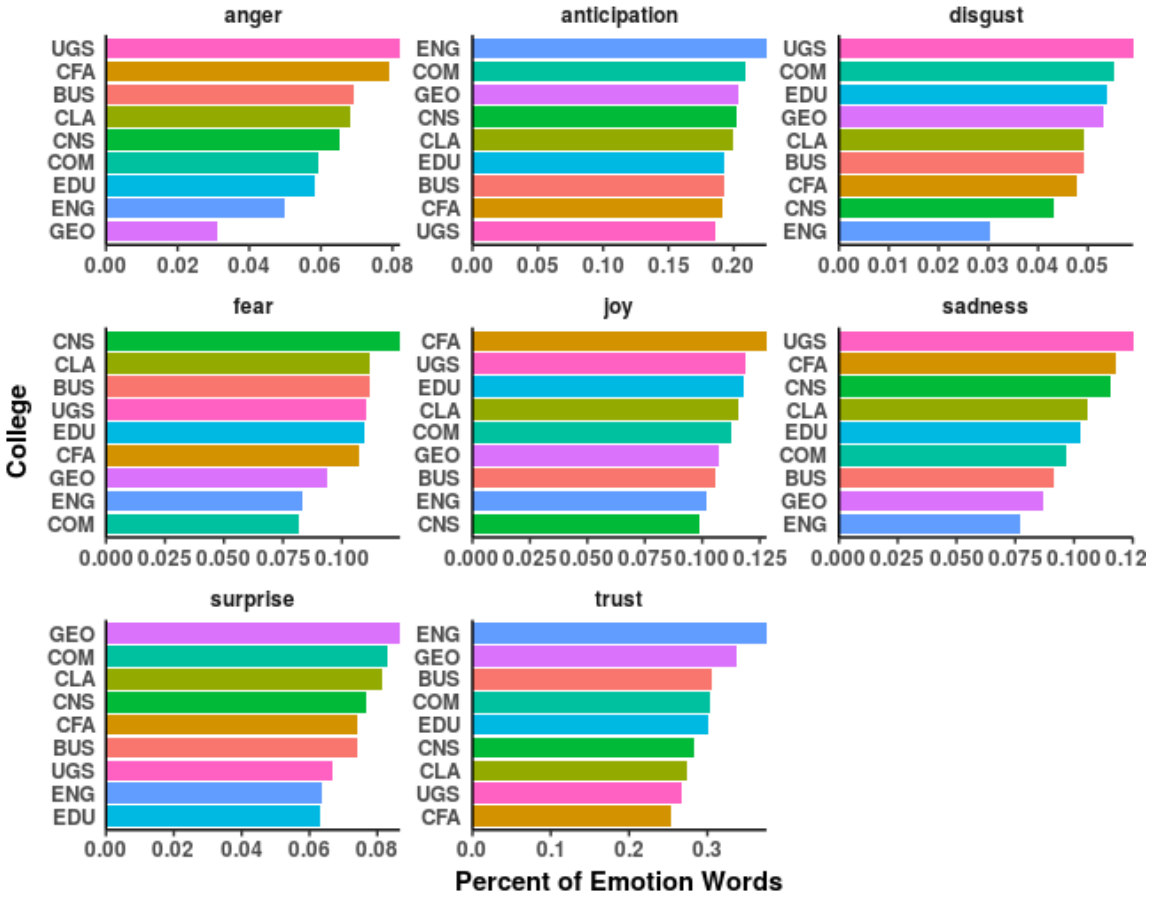

Figure 25 Mean proportion of emotion-labeled syllabus words falling into each of eight emotional categories by college

Syllabus-level tf–idf

As described in the Analysis section, tf–idf scores were computed for each word in each syllabus. For each college, plots of the 10 highest tf–idf words and 10 lowest tf–idf words are given in Figures 26 and 27, respectively. Notice that the lowest tf–idf words are indeed the most common words across all syllabi ("course," "time," "academic," "students," "disabilities," "accommodations"). It is more interesting to compare the words that are most unique to a given syllabus by college. Aside from the college-specific terms

("marketing," "aerospace," "algorithms," "rocks"), we can see others that are suggestive of pedagogical differences. For example, among the top ten highest tf–idf terms for each college, CFA has "creative," BUS has "activity," "scenario," and "speaker," CLA has "images" and "handbook", CNS has "programming," ENG has "clicker" and "analysis," COM has "interview," GEO has "lab," and UGS has "capstone."

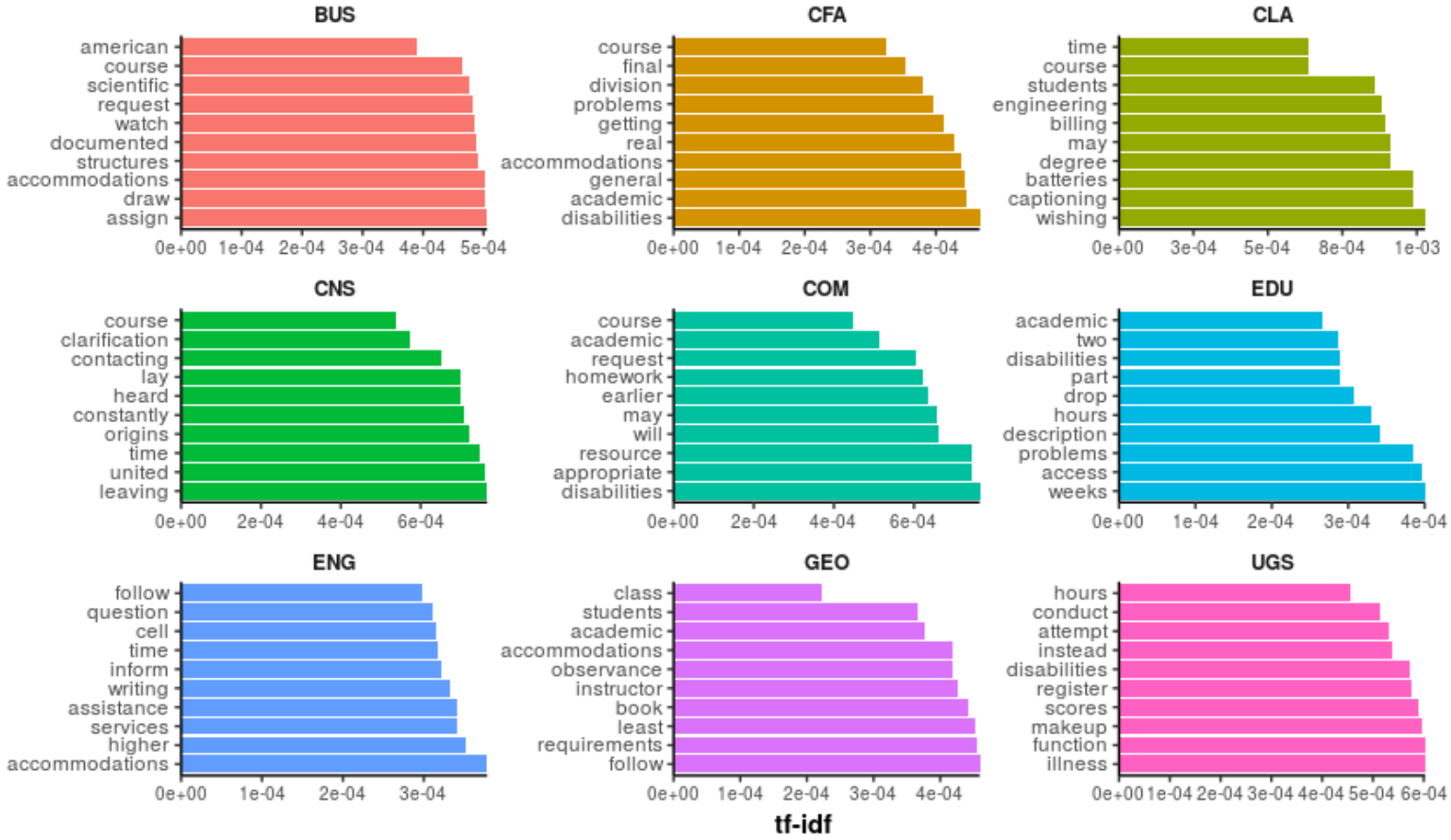

Figure 26 Words with the lowest syllabus-level term frequency, inverse document frequency (tf–idf) scores by college.

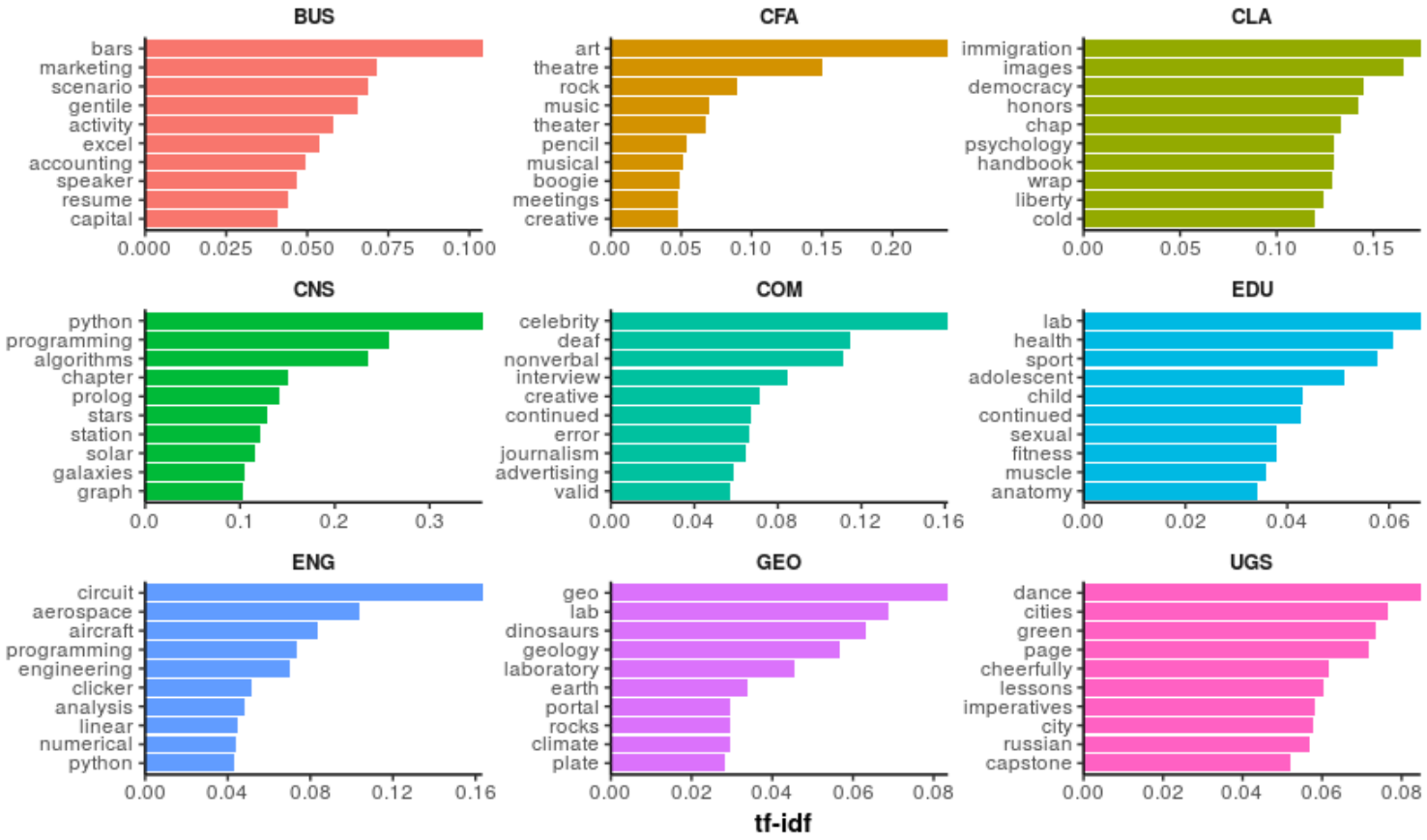

Figure 27 Words with the highest syllabus-level term frequency, inverse document frequency (tf–idf) scores by college.

College-level tf–idf

Many of the above terms disappear when tf–idf is computed by college instead of by syllabus (i.e., computing term frequency across all terms in a given college, and inverse-document frequency by number of colleges containing that term in at least once syllabus; see Figure 28). However, other interesting terms appear in their place when taking this more coarse-grained approach. In CLA, the term "benchmarks" has the highest college-level tf–idf: it is the most important and discriminating word for CLA syllabi relative to all other colleges' syllabi. The terms "scripts" and "showcase" are the most important terms for syllabi in UGS; the term "expressing" is very important for CFA, and "scenario" is still important for BUS courses. There are also some apparent issues when colleges have very distinctive courses (e.g., exercise physiology courses) that have terms in the syllabus that appear only in that college and nowhere else. Because of this, for example, the most important term in EDU in terms of college-level tf–idf was "fitness," because this college includes the Kinesiology and Health Education department (this is also apparent from the document-level tf–idf above). Caution must therefore be exercised when examining the terms with the highest tf–idf, and it is sensible to calculate department-level and course-level tf–idf for comparison.

Department-level tf–idf

Calculating tf–idf at the department level (i.e., pooling words from all syllabi in a single department and treating each department as a "document") and then taking the top ten highest tf–idf within each college provides another view of how unique certain terms are to a given discipline (Figure 29). This gives a better picture of EDU (terms such as "adolescent," "literacy," and "educator" have risen to the top), but perhaps a more skewed

view of CNS (which appears dominated by terms from departments related to marine science and nutrition). Calculating the most important and distinctive terms across departments yields a more flattering and traditional view of CLA: here, "judgement," "aesthetics," "skepticism," and "virtue" are all among the top ten in the college. Thus, there are some departments within CLA that use very distinctive liberal-arts terminology, and these words are found to be relatively uncommon in other departments.

Course-level tf–idf

Finally, we can calculate tf–idf by pooling terms from every syllabus on file for a given course (effectively collapsing across different instructors and semesters, but keeping the tf–idf at the course level) and then take the top ten from each college (Figure 30). In BUS, the terms "mock," "budgeting," "resume," and "peer" among the top terms, and in UGS we see "capstone" and "annotated bibliography" (the latter appearing as separate words). However, CNS and ENG are dominated by computer programming terminology that is very specific to a single course and unlikely to appear elsewhere, and terms from EDU are again swamped by words related to exercise physiology. For COM, "celebrity" and "publicity" are near the top, just as they were for college-level tf–idf.

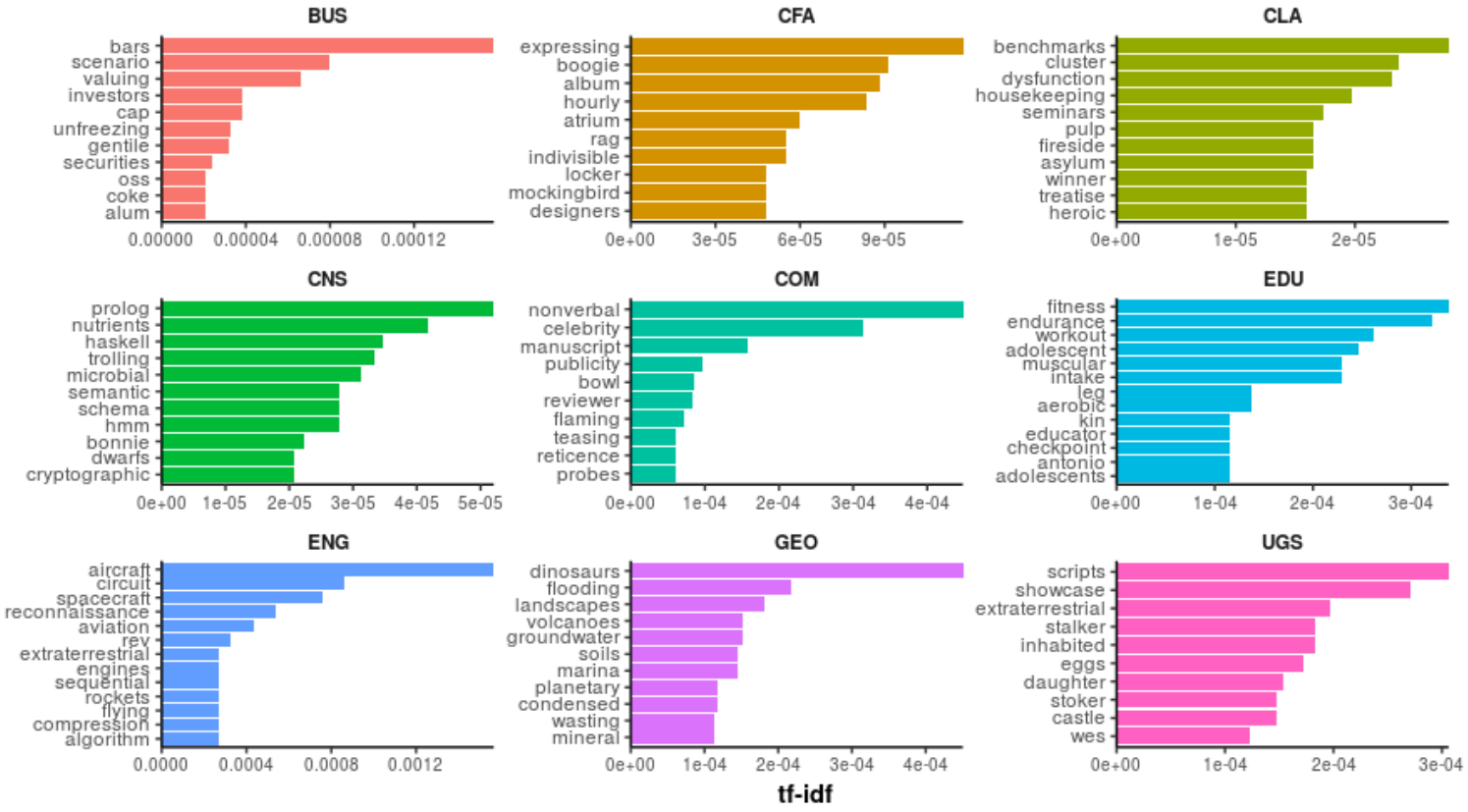

Figure 28 Words with the highest college-level term frequency, inverse document frequency (tf–idf) scores by college.

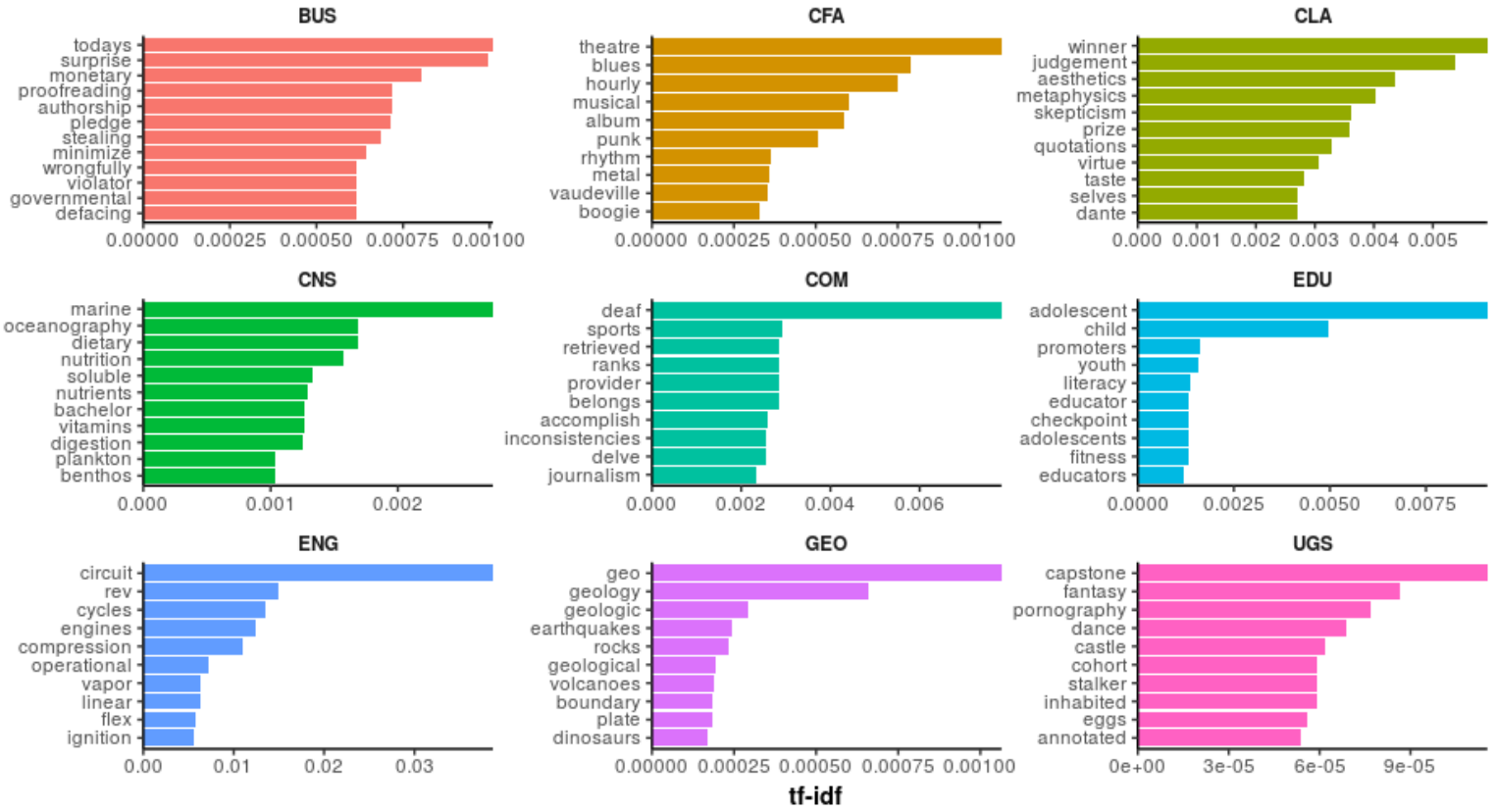

Figure 29 Words with the highest department-level term frequency, inverse document frequency (tf–idf) scores by college.

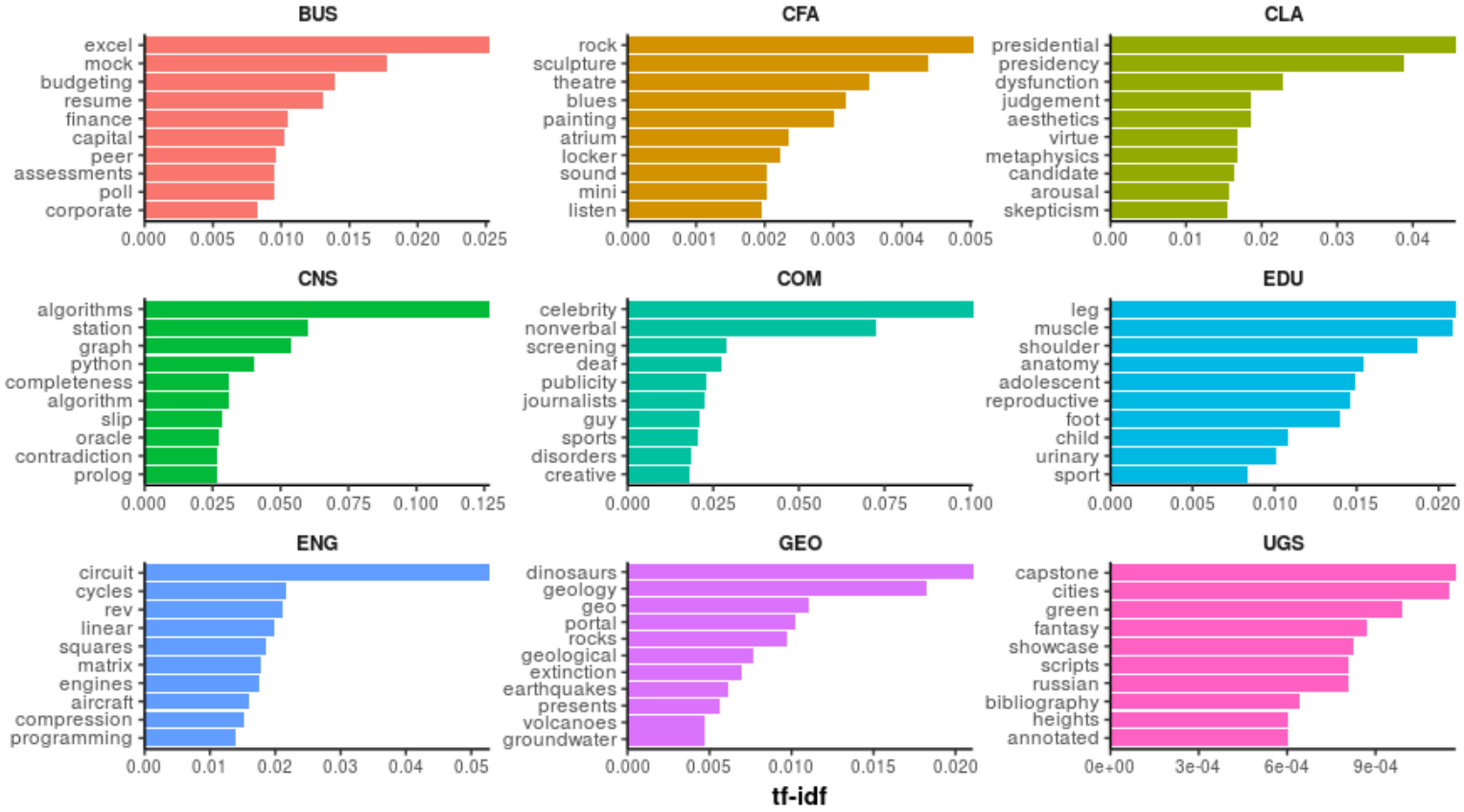

Figure 30 Words with the highest course-level term frequency, inverse document frequency (tf–idf) scores by college.

Within-course syllabus similarity

What accounts for the differences discussed in the foregoing analysis? One possibility is that syllabi are more homogenous in some courses than they are in others, courses are more homogenous in some departments than they are in others, and departments are more homogenous in some colleges than they are in others. It is easy to explore the relative homogeneity of syllabi at each of these levels by computing cosine-similarity scores for all pairs of syllabi for a given course and, separately, all pairs of syllabi from different courses within the same department or college, and then averaging these within- and between-course similarities. Using within-course similarity, it can be asked to what extent syllabi remain similar when different instructors teach the same course (i.e., how much different sections of a course are individualized) or how much a course changes over time (by computing similarity scores for same-course syllabi across several semesters). Computing between-course similarities at the department level provides a measure of how much variability there is among syllabus content within a given department, but it is inherently less precise because of our syllabus sampling methods which picked up large courses at the expense of small ones—courses which are overrepresented in some departments (e.g., CLA, CNS) and underrepresented in others (e.g., UGS, GEO).

Figure 31 presents the mean cosine similarity across all same-course pairs for each department. Thus, each bar depicts the average within-course similarity for a single department, collapsing across semester (though specific departments are not identified by name, the color of the bar indicates the college the department belongs to). It is apparent that departments in CNS, ENG, GEO, and BUS tend to have higher within-course similarity, COM, CFA, and EDU tend to have more moderate levels of within-course

similarity, and CLA and UGS have lower levels of within-course similarity. UGS is particularly low, which makes sense: several courses in these departments are signature courses for first-year students to take so they get exposure to new ideas and ways of thinking. As such, each course varies considerably from the others, crossing disciplinary boundaries and depending on current events and faculty expertise. Likewise, many of the lowest within-course similarity CLA courses are in departments like Government, English, and History, where topics and readings tend to be more variable than they are in other disciplines. However, this overall pattern is not perfectly reflected by all departments: notice, for example, that ENG has one department with the second highest within-course similarity and one department with relatively low within-course similarity.

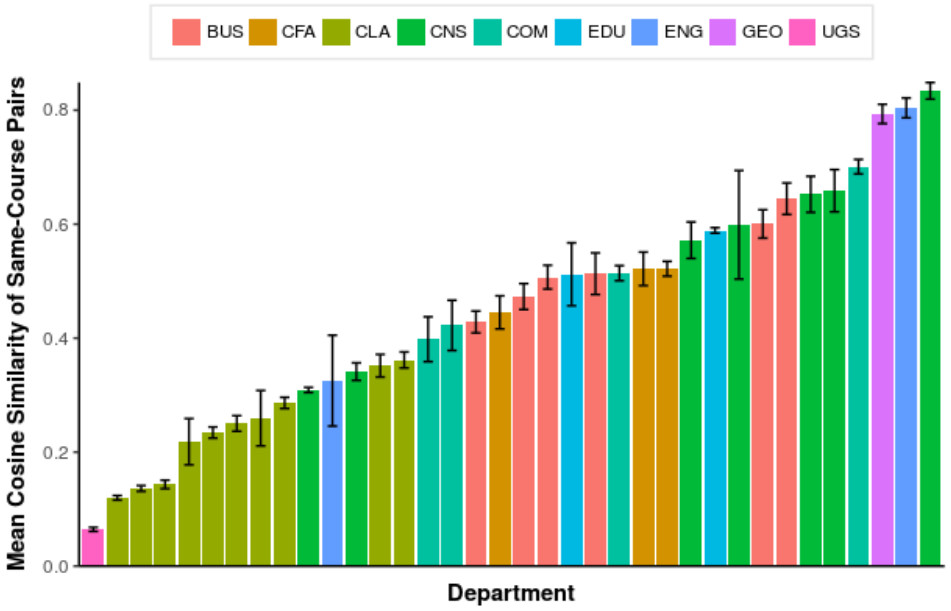

Figure 31 Mean cosine similarity scores of all same-course pairs by department, colored by college to maintain department anonymity

Because the last plot collapses across semester, it is not possible to tell whether within-course similarity is due to the same instructor offering the same course repeatedly over time rather than actual similarity of syllabi for various versions the same course. One way around this is to limit consideration to same-course pairs offered in the same semester (Figure 32). Doing so reveals the same general pattern as seen in the previous analysis. Each semester, UGS and CLA courses tend to have lower within-course similarity, while CNS, BUS, and ENG tend to have higher within-course similarity. Interestingly, departments in CFA tend to have very similar syllabi within a given semester. Note that GEO is no longer represented because they never offered the same course more than once per semester. The bottom of Figure 30 shows the overall average within-course, within-semester similarity for each course, collapsing across semester; note that only courses offered multiple times in at least one semester are included. Thus, this plot presents the average within-course similarity for each course when compared to other sections of the course that were offered in the same semester. The divide between CLA and CNS courses is still apparent, but BUS courses appear to vary considerably in terms of their within-course similarity for a given semester.

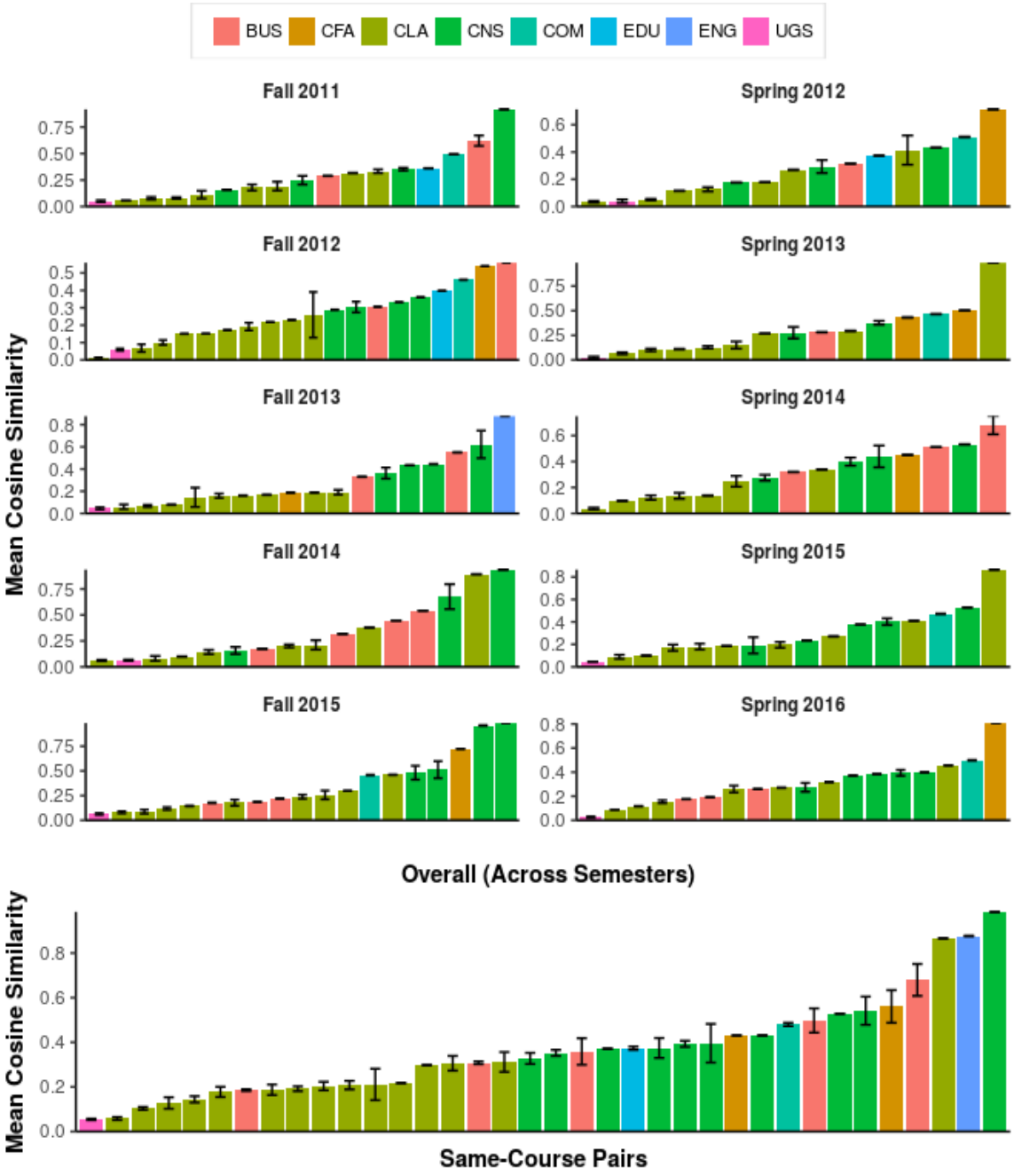

Figure 32 Mean cosine similarity scores of all same-course pairs offered in a given semester by department, colored by college to maintain department anonymity

Another view of within-course similarity can be achieved by averaging same-course cosine similarity scores at the college level (Figure 33). The only major change between this view and the department-level view is that CNS has dropped considerably; this is because a large number of CNS courses are from a single department that had very low within-course similarity, while the CNS departments with the highest within-course similarity were based on relatively few courses. UGS has the lowest within-course similarity, followed in increasing order by CLA, CNS, CFA, COM, BUS, EDU, ENG, and GEO.

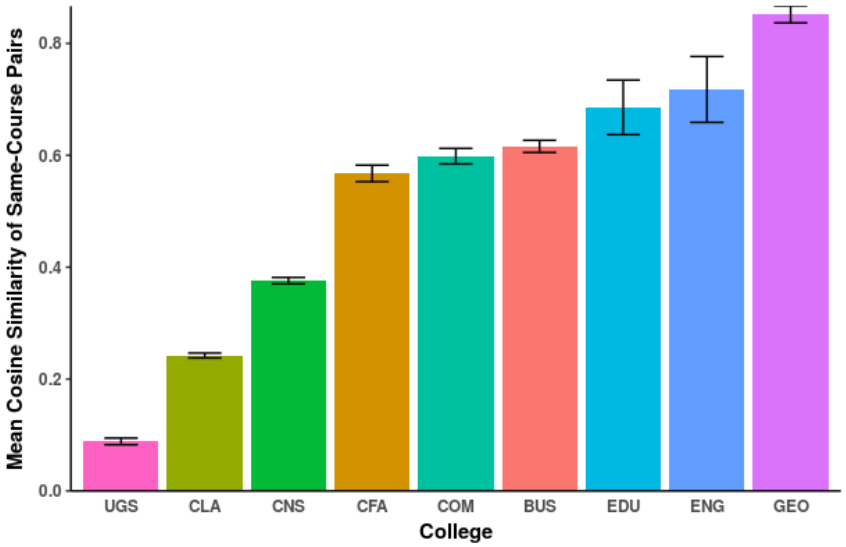

Figure 33 Mean cosine similarity scores of same-course pairs by college

Though the focus thus far has been on within-course similarity, it is also interesting to look at *between*-course similarity: it can be asked, for example, to what extent are different courses in the same college similar, and to what extent are different courses in the same department similar (Figures 34 and 35, respectively). Comparing different courses in the same college, it can be seen that CLA courses have the lowest between-course

similarity on average; thus, CLA courses have very great within-course differences and very great between-course differences. On the other hand, GEO courses have very great within-course similarity and very great between-course similarity for the syllabi in our data set. The only big change to be observed is EDU, which has very high within-course similarity (courses don't change much between professor or semester), but very low between-course similarity (different courses are very different relative to other colleges). This makes sense in this instance because EDU includes the Department of Kinesiology, whose courses contain subject matter that is quite different from typical EDU courses. When we look at average different-course, same-department pairs by college (Figure 35), the picture remains largely the same with a single exception: CFA has relatively dissimilar different-course, same-college pairs but highly similar different-course, same-department pairs. This is because the departments within CFA are themselves homogenous in terms of their courses (e.g., Art History, Theater), yet very different from each other. Thus, similarity for different courses in the same department is high for CFA, but because of large differences between departments, similarity for different courses at the college level is low.

Overall, it is clear that syllabi from certain courses are more similar to each other than others, and that these differences are reflected even when averaging by department or by college. This is largely attributable to the fact that certain large introductory courses—often with very many sections taught simultaneously—tend to have a common syllabus structure, including the same assignments and readings.

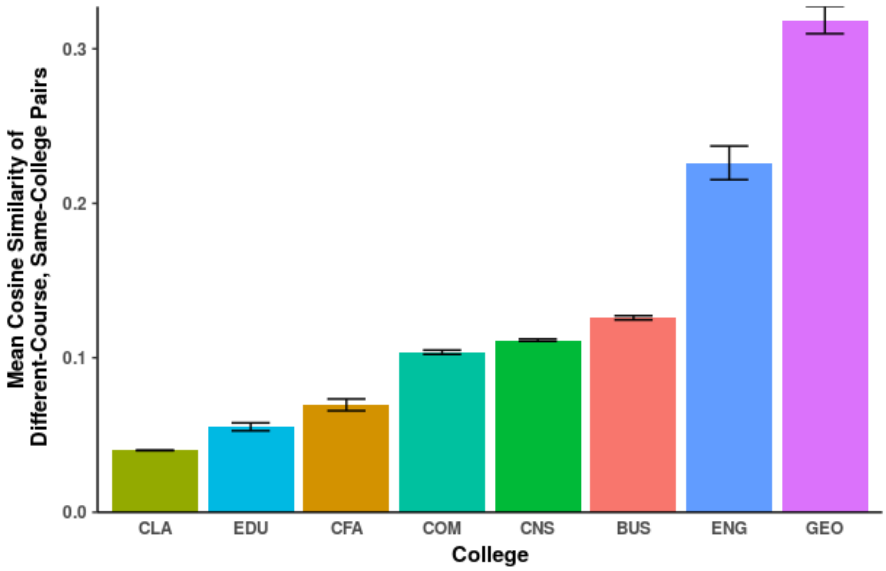

Figure 34 Mean cosine similarity of different-course, same-college pairs by college.

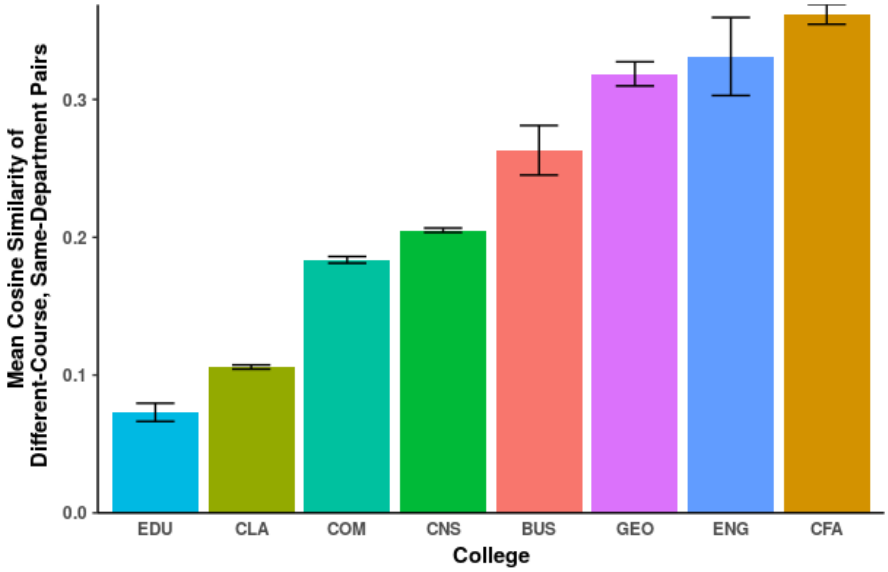

Figure 35 Mean cosine similarity of different-course, same-department pairs by college

Chapter Four: Discussion Part I

This line of inquiry has provided much-needed information about what goes on in large undergraduate courses, shedding light on the prevalence and variability of educational practices within and across disciplines. To date, such data have been extremely limited: when Wieman and Gilbert (2014) polled members of the Association of American Universities (AAU), the American Public and Land Grant Universities, and Presidents and Chancellors of the Association of American Colleges and Universities, not a single institution reported collecting data on teaching practices. At a time when the value of higher education is being increasingly called into question by academics and public figures (e.g., Lacy, 2011; Caplan, 2018), it is important to gather data about what is going on in large undergraduate courses to be able to justify their pedagogical soundness or make necessary improvements.

To reiterate, the literature on how to improve desirable student learning outcomes, such as long-term retention and transfer, makes clear recommendations for teaching practice: in short, we know "what works." There have been so many empirical articles about effects on student achievement that meta-analyses abound: indeed, a full decade ago Hattie (2008) was able to synthesize over 800 meta-analyses of influences on student achievement—including over 50,000 studies encompassing over 80 million subjects—in his work *Visible Learning*. Less generally, there have been meta-analyses specific to undergraduate education yielding consistent recommendations for evidence-based practices such as informal retrieval practice, peer discussion and small group activities, graded homework assignments, and frequent low-stakes testing rather than infrequent, high-stakes testing (e.g., Freeman et al., 2014; Dunlosky et al. 2013).

Where courses are revealed not to have kept pace with educational best-practices, the findings presented herein can help raise awareness among instructors who might

otherwise receive little or no feedback about their own teaching and how it squares with evidence-based practices and course design. Results have largely been constrained to the level of individual colleges; in part, this was done to make the scope of the project more manageable, but another consideration for this choice of grain-size—and one no less important—is respect for the instructors themselves and their individual freedom as educators. We feel that a far more effective approach is to present these results in the aggregate, allowing individuals to reflect on them in relation to their own teaching, which they are left to improve on their own terms in the ways they see most fitting.

These findings are derived from a multi-year process of coding and text-mining 1075 unique course sections—spanning 10 long semesters, 303 instructors, 45 departments, and 9 colleges—at a large, public R1 university with an average enrollment of 287 students (*SD* = 133.38). Over 95% of these courses were taught using a traditional face-to-face format rather than a web-based format, and only 3.3% mention being flipped classrooms (within CNS, this rises to almost 10%). Most of these courses were lower division (77.8%), and just over half were core courses (56.7%). Keep in mind that all findings are limited to syllabi from large courses at a single large public university, thus limiting generalizability to higher education at large, though around 74% of American undergraduates do attend public colleges (NCES, 2016).

Amount of course work is low but increasing

Some of the most important (and reliable) information that can be taken directly from a syllabus is contained in the course grading rubric. This is because each syllabus must describe in some detail all the work that students will do in the course and how they will be evaluated for it. We know, for example, that graded homework assignments support student learning outcomes in K-12 settings (especially among older students; for a meta-

analysis, see Cooper et al., 2006), and in undergraduate courses (e.g., Cheng, Thacker, Cardenas, & Crouch, 2004; Richards-Babb, Drelick, & Robertson-Honecker, 2011). Based on course grading rubrics alone, we can tell that in our sample, the median number of homework assignments per course was 3, and 30.1% of courses gave no homework at all, which is a pretty poor showing. In this way, the course rubric has provided many valuable observations for the present study about the nature of coursework in large undergraduate classes.

In the aggregate (averaged across all 1075 courses), students complete 6.5 homework assignments, 4.7 quizzes, 3.4 exams, and 2.4 in-class assignments, for an average of 16.4 total assignments per course (summed before rounding; $Mdn = 12$, $SD = 15.7$). Notice that the median is considerably smaller than the mean, indicating that the bulk of the observations are low: indeed, 30% of courses had 6 of fewer assignments, and 13% of courses had 3 or fewer assignments (and in most cases, these are all exams). Because students learn better when they engaged in course work throughout the semester, this dearth of activity is concerning. The skew is also evident when looking at the long right tail where there are few courses assigning large quantities of work: the course at the 75th percentile had 19 assignments, and the course at the 99th percentile had 71 assignments, while the maximum number of assignments in any course was 105, almost ten times as many as the median course!

Thus, while it is clearly possible to have students engaging with many assignments in a single course, the majority of instructors give students very little to do for which they will be held accountable. This likely reflects two factors: (a) the predominant incentive structure in higher education under which research is more highly rewarded than teaching (Chen, 2015; Young, 2006), and (b) the extremely limited amount of time faculty members

have to devote to their many responsibilities (Jacobs, 2004). This strain is summarized nicely by Kuh (2003):

> "The more pages students write, the more pages faculty members have to read and give feedback about. And the more often that we do, the more likely it is that students will make appointments during office hours to talk with us about that feedback. In terms of student engagement, all this is generally positive. But it becomes problematic in terms of allocating time across multiple faculty priorities." (p. 28)

Because students and faculty both have reason to avoid a heavy workload, this convergence of interests leads to a tacit and mutually satisfactory agreement between teachers and students, to wit: I like to assign less work because I have less to grade, and you like it too because you have less to do!

Furthermore, students tend to reward low workloads with high course evaluations, the most commonly used metric of teacher effectiveness in higher education (for review, see Stroebe, 2016). For example, ratings of professor quality and course easiness were found to have a correlation of .62 for a sample of professors taken from RateMyProfessor.com (Felton, Koper, Mitchell & Stinson, 2008). Findings like this are legion, but it is worth mentioning one here based on actual grades and student evaluations of teaching: when Wellesley College introduced a maximum average grade of B+ to help combat grade inflation in certain departments, professors in those departments (and only those departments) received significantly worse course evaluations than they had before (Butcher, McEwan, & Weerapana, 2014). A similar rule at Princeton was implemented but repealed in 2014 amid concerns that students' post-graduation opportunities would be affected because their GPAs were lower than those of students from peer institutions (Windemuth, 2014).

Perhaps if students were studying on their own time and in their own way, this relatively light workload would not be so problematic. But they are not. In fact, a large

longitudinal survey of undergraduates at 29 four-year colleges and universities from 2005 to 2009 found that students spend only about 15 hours per week studying or doing class work (Arum & Roksa, 2011; Pascarella, Blaich, Martin, & Hanson, 2011). An historical comparison can help put this number into perspective: in 1961, 67% of college students reported studying more than 20 hours per week (around 3 hours per day); in 1981, this number had dropped to 44%, and by 2011 only 20% of students reported studying this much (Babcock & Marks, 2011).

Thus, it is unlikely that students are doing much academic work of their own volition, making it all the more important that educators motivate students with frequent assignments to keep them engaged with their coursework. Later, it will be argued that this and other issues related to the logistics of implementing educational best-practices in large college courses could be readily addressed by, for example, increasing the number of teaching assistants to better manage all of the grading that necessarily results from high workloads and student accountability.

Few retrieval practice opportunities and little spacing

Across all courses, the median number of graded retrieval practice opportunities (total number of exams and quizzes) was 4. That is, over half of courses had 4 or fewer graded opportunities for retrieval practice (this is still the case when considering only those courses who gave at least one exam or quiz). Looking at quizzes specifically, the median number per course was 0; indeed, 63% of courses had no quizzes at all! The power of retrieval practice (e.g., coming up with an answer during a quiz in class) to improve memory for course material has been shown in countless studies (see Roediger & Butler, 2011), and it is clear from these findings that the technique is woefully underutilized in large college courses.

Spacing is more difficult to quantify from syllabus variables, but the degree of cumulativity in a course is one potential indicator to examine (e.g., making exams cumulative holds students accountable for previously learned material so that they will have strong incentives to revisit it at multiple points in time). Unfortunately, not including the final exam, only 4% of courses reported having cumulative exams in their respective syllabi. When including the final exam, this number was quite a bit higher: 39% of courses reported giving cumulative finals. This, together with the limited quantity of work students are assigned for their courses discussed above, provides evidence that students are not being held accountable for previously learned material to the extent that they could be and that the spacing effect is not being used widely or effectively these courses. Unfortunately, this lack of spacing in the classroom is extremely common and has been for some time (Dempster, 1988). This could be due, in part, to either the perceived lack of value in returning to previous material over time, or to the difficulty inherent in spacing/interleaving and to the temptations of procrastination and cramming. Regardless, educators would do well to build such spacing into their courses to serve as external checks on temptations to cram: keeping students engaging with material over time (e.g., by assigning frequent graded homework assignments) would likely improve student retention of the material dramatically (e.g., Rawson & Kintsch, 2005; see below).

Different colleges have different strengths

The findings presented here suggest that educational best-practices are being implemented heterogeneously across colleges, with some appearing at much higher rates in certain colleges than in others. However, no one college or department has a monopoly on any of them: by way of illustration, note that while some colleges were much higher in this regard, at least 9% of syllabi in each college had collaborative assignments, and there

were even a few courses in CFA that incorporated informal retrieval practice. This is an encouraging situation, because it shows that applying each best-practice is feasible across the disciplinary gamut of large undergraduate courses. There is no reason to believe that any of these best-practices are mutually exclusive or rigidly bound to the subject matter of a specific discipline; indeed, research suggests that many if not all of the best practices considered here are domain general (e.g., Dunlosky et al. 2013; Hattie 2008).

Let's take a moment to review this heterogeneity. Compared to other colleges, CFA, EDU, and UGS had a high percentage of courses that incorporated group activities, while in the hard sciences (GEO and CNS) there were relatively few. Overlapping somewhat with group activities, there was a high percentage of courses with projects and presentations in UGS, COM, and CFA: however, this number was very low in GEO, BUS, EDU, and CNS. For in-class active learning, UGS, CNS, and COM incorporated it into their courses most often; strangely it was lowest in ENG, followed by CFA and BUS. Among all colleges, courses in CNS were the least likely to enforce attendance.

Conversely, the science colleges fared more favorably with respect to spacing and retrieval practice measures. CNS and GEO had a relatively large proportion of courses with cumulative finals and informal retrieval practice; these variables were quite low in CFA and EDU. However, relative to others, CFA had a high frequency of socio-emotional learning objectives, while ENG had a very high frequency of learning objectives for knowledge and skills. CFA, BUS, and COM had the lowest proportion of courses that did not assign homework, and all GEO courses had graded in-class assignments. In terms of total number of graded assignments (including exams, homework, quizzes, and in-class activities), CNS and GEO had the most on average, but BUS was the highest non-science college, tied with ENG for third.

The overall pattern that emerges can perhaps be seen most clearly in the cluster membership of each college (Figure 17, top panel). For instance, Cluster 4 (characterized by group activities, in-class active learning, and projects/presentations) was almost entirely composed of courses from UGS, EDU, CFA. On the other hand, Cluster 2 (which has informal retrieval practice, in-class active learning, and cumulative final exams) was primarily made up of courses from CNS and GEO (notably, CLA had a similar number of courses in Cluster 2 and Cluster 4). Indeed, CLA, BUS, COM, and UGS were more variable in their cluster assignments, suggesting a wider array of course practices.

On the other hand, it is also clear that certain colleges are overrepresented in the "do-nothing" clusters (those characterized by few best-practices, e.g., Clusters 1, 3, and 5). Though all colleges had sizable fractions of courses in these clusters, in CFA about 75% of courses were assigned to one, and over 65% of EDU courses were as well. In summary, there appear to be two dominant patterns of best-practice use, both making use of in-class active learning: one focusing on group activities, projects, and presentations (and more typical of UGS, CFA, EDU), and the other making use of informal retrieval practice and cumulative exams (more typical of CNS and GEO). However, there are many courses who lack these elements entirely, and this group cuts across all colleges. Overall, these findings are indicative of a diversity of best-practices, but several pedagogical gaps can be seen across colleges and departments. Thus, there is ample room to adopt best-practices where they do not yet exist and to discover and utilize those that may be relatively uncommon within a given college or department.

STEM and Business versus Arts divide apparent in syllabus language

Text-mining of syllabi yielded additional insights into differences among colleges. First, course syllabi for the same course (but different instructors) were found to be more

similar to each other in science courses and less similar in liberal arts courses, regardless of semester. This indicates that liberal arts faculty customize and individualize their syllabi to a greater degree than their counterparts in the sciences (who tend to use the same textbook, assignments, etc.). Indeed, CLA and UGS had anomalously low similarities relative to all other colleges. However, this pattern could also reflect the nature of the material taught in these courses: science courses tend to cover the same fundamental material, but liberal arts courses (e.g., English) have greater latitude with respect to the specific subject matter taught, thus leading to greater inter-syllabus differences at the college level.

Results of the sentiment analysis were somewhat inconsistent, but there was an overall stable tendency for BUS and CNS syllabi to be rated as negatively emotionally valenced compared with EDU and CFA syllabi, which were consistently among the most positive (e.g., Figure 23). This pattern was observed across several analyses using three different lexicons as well as LIWC scores for emotional tone, using both cleaned and uncleaned text. Beyond the positive-negative continuum, there were differences among colleges with respect to other emotions (Figure 24).

Differences in linguistic content of syllabi were also observed between colleges, though syllabi were comparable on the whole. In terms of sheer word count, the three colleges with the longest syllabi on average were CNS, CFA, and BUS, while those with the shortest were ENG, GEO, and UGS (Figure 20; both before and after cleaning the text by removing stopwords, punctuation, numbers, etc.). In addition to sentiment differences, there were also differences in other LIWC variables (Tables 4 and 5). Syllabi in EDU were lowest in analytical language, but highest in authenticity (honest, personable) and clout (confidence). The opposite pattern was observed ENG courses: these syllabi were highest in analytical language, lowest in authenticity, and lowest in clout (Figure 18).

It is also interesting to compare colleges with respect to the achievement or affiliation orientation of their syllabi (Figure 22). On these dimensions too an inverse relationship was observed for science and business relative to fine arts, liberal arts, and education. For words relating to affiliation (e.g., *ally*, *friend*, *social*) EDU and CFA had the most such words while ENG had the least (with BUS not far behind). However, for words relating to achievement (e.g., *win*, *success*, *earn*), ENG and BUS syllabi had the greatest proportion (with CNS thirdmost), while UGS had the least, followed by CFA, CLA, and EDU.

The proportion of comparison words (e.g., greater, after) and negation words (e.g, not, never) used in syllabi was also of interest a priori (Figure 22). It was found that CNS syllabi used far more comparison words than any other colleges, but otherwise there was little variation across colleges. With respect to negation words, CNS also came out on top, but EDU, COM, and BUS were not far behind. The colleges with the smallest proportion of negation words in their syllabi were ENG, UGS, CLA, and CFA.

High-stakes exams are the norm

Across all courses, the median number of exams (including the final exam) was 3 and the median grade percentage for exams was 75%. About 94% of courses in our sample gave exams (versus, say, large take-home papers or projects). Among the 6% of courses giving no exams, the mean percent of the grade from homework assignments was 46%, and the mean number of homework assignments in these courses was 10 (*Mdn* = 5). Note that this large percentage is due to the fact that the homework category was defined to capture all assignments completed outside of class, including papers or "take-home" exams, which are featured more often in courses without formal exams.

One conservative definition of high-stakes exams is having four or fewer exams accounting for at least 75% of your grade; by this metric, in our sample 40% of courses have a grading rubric based on high-stakes exams (see Figure 36 for high-stakes exams—according to this definition—by college and department). Indeed, in 24% of courses, 3 or fewer exams accounted for at least 75% of the grade. Note that this includes courses not giving any exams at all. Among courses that gave at least one exam, 43% were high-stakes using these criteria. Using a less conservative cut-off of 50% of the total grade from exams, 65% of courses qualify as grading based on high-stakes exams (69% among courses giving at least one exam). Whatever the definition, high-stakes exams are clearly the norm in large college courses: recall that the median number of exams (including finals) was 3, and the median percent of grade from exams was 75%.

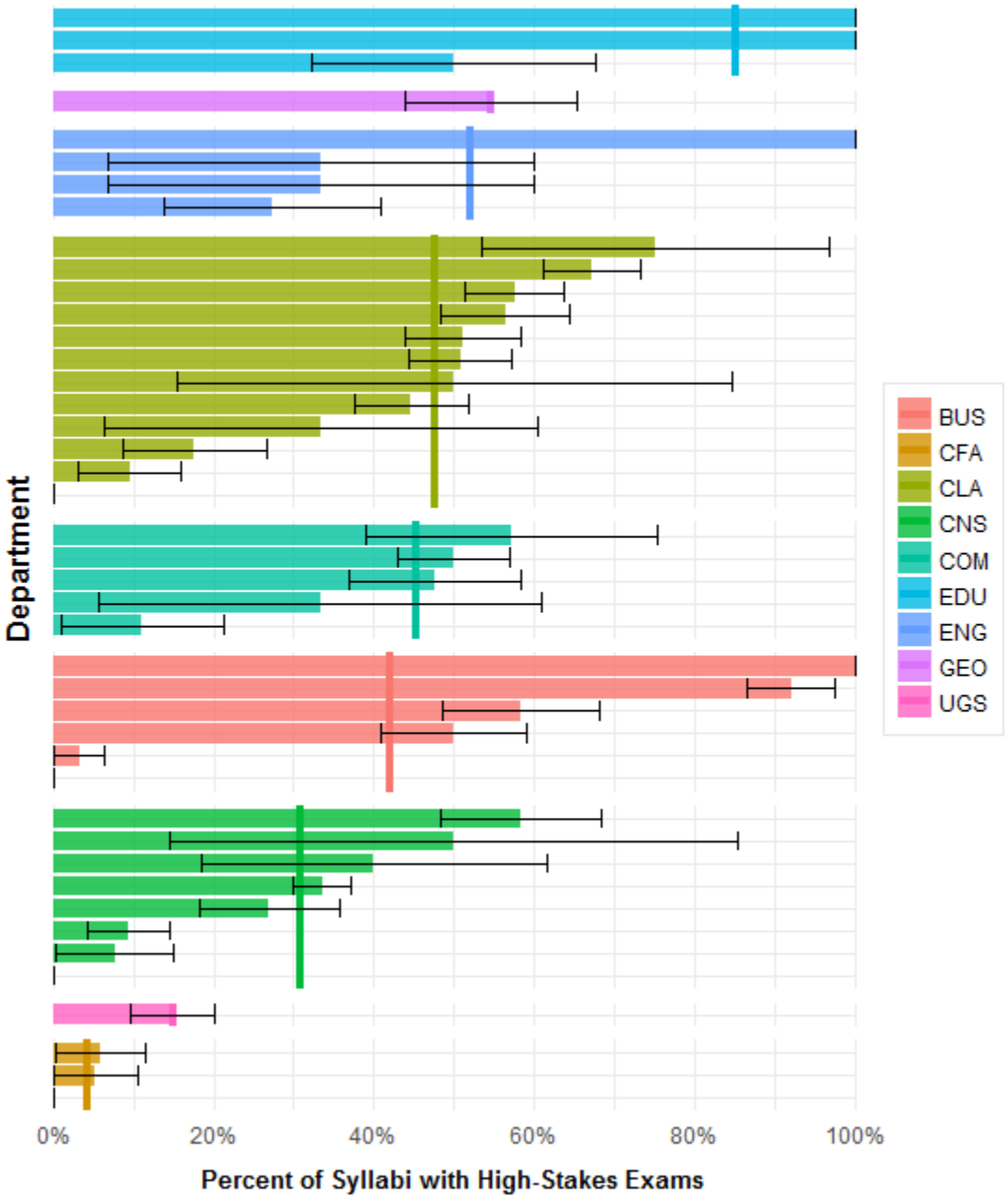

Figure 36 Percent of syllabi with high-stakes exams (defined here as 4 or fewer exams accounting for 75% or more of the final grade) by department, grouped by college (color legend). Departments are unlabeled to maintain anonymity. Colored vertical bars spanning horizontal bars of the same color indicate college means. Error bars show bootstrapped standard errors.

How did high-stakes exams vary across colleges? More than 40% of courses in BUS, CLA, COM, EDU, ENG, and GEO had high-stakes exams using the stricter criteria of 4 or fewer exams, together worth 75% or more of the final course grade (Figure 36). Interestingly, only 31% of CNS courses gave high-stakes exams, only 15% of UGS courses did, and 4% of CFA courses did. One counterintuitive observation to be made here is that large liberal arts courses use high-stakes testing more often (47.5%) than do large natural science courses. Notice, however, that there is significant variability within a college (i.e., across departments)

High-stakes exams are traditional and convenient, but while they have a semblance of validity, they are in actual fact relatively poor assessments of what students have learned both in terms of their long-term retention of course information and their ability to apply it in new situations. Furthermore, they encourage poor study strategies, such as cramming, which are ultimately responsible for these impoverished learning outcomes (e.g., Bahrick, 2000; Custers, 2010; Pennebaker, Gosling, & Ferrell, 2013). Research has firmly established that, with respect to encouraging retention, transfer, and good study habits, more tests is better than fewer (for review, see Roediger & Butler, 2011; Karpicke & Roediger, 2008), and more frequent testing necessarily creates spacing by eliminating the possibility of one-shot cram sessions (e.g., Rohrer, 2015; Kang, 2016).

Cause for optimism in trends over time

The foregoing discussion relies on statistics obtained by averaging syllabi over time. Another approach is to look at averages within a given year and examine the trend. In this study, we have examined syllabi ranging from Fall 2011 to Spring 2016, and several trends observed during this time period appear positive. There are significant increases in such positive course variables as homework assignments, quizzes, quiz grade percentage,

in-class active learning, learning objectives, informal retrieval practice, and community learning opportunities. There were also significant decreases in more negative course features such as percent of grade from exams. However, variables reflecting certain best practices were found not to have increased (group work, cumulative exams/final, flipped classroom) and in some cases even to have decreased (projects/presentations). Overall, however, the story is a positive one: though a few lag behind, best-practices are generally on the rise.

Comparisons with recent classroom observations.

Recently, the journal *Science* published an article that used systematic classroom observations to assess over 2,000 classes in the STEM fields across 25 different colleges and universities (Stains et al., 2018). The protocol they used documented the occurrence of 13 student behaviors and 12 instructor behaviors observed in 2-min intervals of class time (for more about the protocol used, COPUS, see Smith, Jones, Gilbert, & Wieman, 2013). In their sample 71.4% of courses were lower-division, making it comparable to our own (77.8%), though it was limited only to courses in STEM fields. However, it is important to note that class sizes were considerably smaller on average than those in our dataset. For example, around 30% of their courses had less than 50 students, and only 44% had more than 100 students, while all of our courses had more than 100 students (indeed, 84% had more than 200 students).

These classroom observations revealed that during 75% of the 2-min intervals, instructors were lecturing, and during 87% of the intervals, students were listening to the instructor, though the variability of these estimates was large. The remaining time was taken up by students answering questions (occurring in 21% of intervals) and asking questions (10% of intervals). The authors then conducted a latent profile analysis on the

observations, revealing seven unique instructional profiles. Of these, 55% were classrooms in which lecturing took place at least 80% of the time, with little if any interactivity (a *didactic* style); 27% were classrooms that supplemented lecture with questions and activities (*interactive lecture* style); and 18% were classrooms in which group work and inquiry learning featured more prominently (*student centered* style).

Interestingly, these three broad profiles map onto the three factors that emerged from our CFA rather well. Though there were no variables in the factor analysis that captured lecturing per se, the factor Groups, Projects, & Participation (based on syllabus data from a variety of courses) looks a lot like the *student-centered* style as determined from classroom observations of STEM courses, and the factor Active Classroom, Cumulative Tests looks like the *interactive lecture* style. Perhaps courses with a *didactic* style would be characterized by having low factor scores on these two factors. The third factor we observed that could be characterized as a course profile was Supportive, High-Workload. Syllabus variables like number of quizzes and number of homework assignments had high loadings on this factor. Because there is certain information that can only be obtained through observations (e.g., time spent lecturing) and other information that can only be gotten at through review of syllabi or other course materials (e.g., number and grade percent of assignments), perhaps a useful approach to assessing instructional efficacy lies in some synergy of the two approaches.

Finally, it is interesting to look at how our course clusters map on to those found by Stains et al. (2018). Our analysis favored a six-cluster solution (comparable to their seven) but differed somewhat in the distribution of courses belonging to each cluster type. For example, in CNS (a college including all STEM fields except engineering and geology), the cluster assignments analogous to their *interactive lecture* style made up around 50% of courses (cf. 27%), though the proportion of *didactic* courses was

comparable (47% vs. 55% of their courses). On the other hand, the number of *student-centered* courses in CNS was much lower than their average (around 3% vs. their 18%). These differences are almost certainly attributable to class size; when Stains et al. (2018) condition on large class sizes, the percentage of *student-centered* courses (those featuring group work, inquiry learning, or student-instructor interactions) was closer to 10%, and this was especially true of courses held in lecture halls with fixed seating. The pattern for our ENG and GEO courses, however, revealed an exaggerated version of their overall findings: well over 50% of courses were didactic in nature (i.e., lacking group activities, student projects, or in-class active learning), with only 10-20% being assigned to *interactive* or *student-centered* clusters.

A new way forward for large introductory courses

There is a widespread belief that when college students matriculate, they should be equipped with good study-skills and unflagging motivation—essentially, that from day one they are responsible for all of their learning in a course and they must take it into their own hands. The instructor furnishes the information, either via a lecture or through the assigned readings, and it is incumbent upon the student to assimilate this information for themselves. Evidence that learning has taken place is traditionally assessed with one or two midterm exams covering a unit's worth of lectures and readings, and also a final exam.

However, in the internet age, the role of the instructor is changing dramatically: modern educators can no longer afford to be mere lecturers—content delivery systems. Now information about even the most obscure academic subjects is readily available online through primary and secondary texts, often entirely free of charge. The same is true for high-quality video lectures given by world-class faculty (e.g., MIT OpenCourseWare; Abelson, 2008; Carson, 2009) and now also for self-paced online courses with well-

sequenced material and high quality assessments (e.g., Coursera, edX; Breslow et al., 2013). Indeed, even free online textbooks, written and edited by subject-matter experts and leaders in their field, are becoming the norm (e.g., OpenStax; Pitt, 2015).

The structure of traditional education will need to accommodate these high-tech changes, just as the music, newspaper, and retail industries have. Fortunately, this instability can be an impetus for reinvention, fostering positive change by forcing relatively static systems of education everywhere to revisit long-held assumptions and reshape courses and curricula in ways that best serve the interests of students, creating the human capital needed for individual success and societal productivity.

In higher education, the role of the instructor must be redefined. Failure to do could leave the entire collegiate enterprise vulnerable to criticism and competition. There is a rising current of skepticism about the ultimate value of our hallowed educational institutions, and many of the criticisms are hard to parry: under the status quo, students seem to remember very little from their formal education. Bahrick and Hall (1991) found that half of all introductory mathematics was forgotten within five years for those who don't continue on to higher mathematics, and virtually all of it was forgotten within 25 years. This finding isn't just limited to mathematics: after the equivalent of a single semester of a Spanish language course, grammar and vocabulary recall is close to zero after 5-6 years without additional practice (Bahrick, 1984).

Indeed, basic facility with written English is poor among adults: In a study of 18,000 randomly selected Americans, the National Assessment of Adult Literacy found that just over half of Americans are considered "Intermediate" or "Proficient" in prose and document literacy (with less than half receiving these scores for quantitative literacy; Kutner, Greenberg, & Baer, 2006). To put this into context, a score of Intermediate on prose literacy required respondents to summarize a newspaper job advertisement; for

document literacy, the Intermediate task was using a TV guide to find out when a certain program ended. Compared to literacy and numeracy, other subjects fared worse still. The Intercollegiate Studies Institute has found that 71% of American adults fail tests covering basic American history and government (Cribb & Bunting, 2008), and when Newsweek gave a sample of American adults the citizenship test, 38% scored too low to become citizens of their own country (Romano, 2011).

These dismal outcomes are the product of the educational status quo, and I would argue that they result from failure to make evidence-based reforms to improve long-term learning outcomes. Instead, policies favor quick fixes and interventions that produce illusory, short-term performance gains masquerading as accountability in the eyes of a short-sighted public (e.g., annual statewide testing and teaching-to-the-test; propping up graduation rates by lowering academic standards). As we have shown, college curricula, course structure, and teaching practices reflect this state of affairs, and employers have begun to chafe at how poorly prepared many recent graduates are for the workforce. A 2010 survey of over 300 employers found that only 28% felt that 4-year colleges and universities adequately prepared students to fulfill workplace demands (Hart Research Associates, 2010). A more recent, less formal survey of 64,000 managers revealed that 60% find new graduates lacking in critical thinking and 44% find them lacking writing proficiency (2016 Workforce Skills Preparedness Report). This agrees with data from the Program for the International Assessment of Adult Competencies (PIAAC) which shows that more than half of adults born after 1980 (i.e., millennials) ranked among the lowest worldwide in literacy, numeracy, and problem-solving in technology rich environments at all levels of educational attainment (Coley, Goodman, & Sands, 2015). Ten years earlier, the US Department of Education (2006) issued this warning:

> At a time when we need to be increasing the quality of learning outcomes and the economic value of a college education, there are disturbing signs that suggest we are moving in the opposite direction. As a result, the continued ability of American postsecondary institutions to produce informed and skilled citizens who are able to lead and compete in the 21st-century global marketplace may soon be in question. (p. 12)

One proposal is to further increase the quantity of education that students consume, but this solution runs counter to the well-established finding that while the number of students with college degrees continues to skyrocket, learning outcomes continue to stagnate or decline. Focus needs to shift dramatically to the quality of education. A recent report by the Educational Testing Service cautions that "simply providing more education may not be the answer. There needs to be a greater focus on skills—not just educational attainment—or we are likely to experience adverse consequences that could undermine the fabric of our democracy and community" (ETS, 2016, p. 5). I agree with this, but I interpret "skills" broadly to mean any concrete learning outcomes that students carry with them into the future and apply outside of the classroom: the focus should be on what students can do in the future, not merely how many credit hours they accumulated in the past.

Owing to these deficits, an entire job-training industry has sprung up in recent years in the form of technology "bootcamps" and other alternative postsecondary programs, many of which are for-profit, and some of which offer industry-recognized credentials and job guarantees upon graduation (Crispe, 2017). Indeed, the number of certificates awarded by postsecondary institutions increased by 73% from 2000 to 2013—faster than the rate of bachelor's degrees—with almost half awarded by community colleges (Brown & Kurzweil, 2017). Among the for-profit, postsecondary institutions not eligible for government funding, around 700,000 students were enrolled in short-term certificate programs in 2012. In March 2016, approximately 35 million students were enrolled in Massive Open Online Courses (MOOCs), and almost 20,000 students were projected to

graduate from coding bootcamps (Brown & Kurzweil, 2017). While many MOOC and bootcamp participants already hold a degree, these programs tend to cost less, take less time, offer more flexible formats, and align better with employer-defined skills than do traditional degree programs. If these new formats make better use of learning science than traditional higher education, producing graduates with demonstrably better learning outcomes, then the pedagogical reputation of the academy will continue to suffer.

To improve the quality of undergraduate education, research-based best-practices must be implemented. One way to achieve this is to increase educational support staff in college courses to help manage and maintain an effective learning environment. For example, Talbot, Hartley, Marzetta, and Wee (2015) make use of learning assistants—undergraduate facilitators who have previously taken the course—to incorporate active learning and increase academic-task engagement in high-enrollment undergraduate science courses and to help with the additional grading that this increased activity inevitably produces.

In my own teaching as an advanced graduate student, I receive similar support. I am the instructor of record for an introductory statistics course at UT Austin developed by an educational psychologist and former graduate of my doctoral program. To qualify for such a teaching position as a doctoral student, it was required to take a college teaching methodologies course surveying the latest research into effective educational practices and how to implement them (arguably more training in these topics than your average professor receives). My 100-student course section has a graduate-student TA who holds office hours, grades exams, and facilitates the laboratory component. In addition, the course receives support in the form of three undergraduate TAs, recruited from previous students who were successful in the course and showed exceptional potential to help their peers. These undergraduate TAs attend class each day and facilitate active learning (e.g., by

circulating among small-group discussions after a question is posed to give feedback) and they grade students' weekly homework assignments and laboratory exercises. This format allows me to focus my attention on improving my lectures and assignments, individual students who might be experiencing difficulties in the course, and student projects that require specific, individualized feedback. I do not have any data showing improvements to students' long-term outcomes, but the course encourages student engagement and makes full use of evidence-based recommendations for such outcomes.

This arrangement could be extremely beneficial for professors who may be too strapped for time, or otherwise lack the incentive, to redesign their courses in ways that make lectures more effortful and grading more onerous. Professors have too many students to teach as it is: While the percentage of 18- to 24-year-olds enrolled in post-secondary, degree-granting institutions in the US has increased dramatically from 25.7% in 1970 (~7 million) to 40.5% in 2015 (~17 million; Snyder, deBrey, & Dillow, 2018, chap. 3), during roughly the same time period the percentage of full-time instructional faculty at US colleges and universities fell from 78% in 1970 to only 52% in 2005 (Snyder, 1993). This means that today, most professors have to balance a full teaching load including enormous survey courses with research and many other professional or administrative responsibilities (Jacobs, 2004). Considering that research is much more important for tenure decisions across the board, it is easy to see why courses continue to be taught with easier traditional approaches like passive lectures and few quizzes or assignments (e.g., Remler & Pema, 2009; Cadez, Dimovski, & Groff, 2017).

Furthermore, relative to other aspects of their job, faculty (especially those at large, research-focused institutions) often do not enjoy teaching multiple sections of large undergraduate courses year after year (e.g., Alpay & Verschoor, 2013). This is largely because it eats into the little time they have to extend their own research program (on which

performance evaluations are based), to teach upper-division courses in their specialist area, and to mentor more advanced students as they get ready to embark on their own careers. An arrangement in which faculty receive additional help from facilitators in their large undergraduate courses would lighten these burdens while improving the overall quality of their courses for students by keeping them active and engaged with the course content in and out of the classroom!

My recommendations are no less applicable in cases where the professors are stellar educators. We have seen that for large undergraduate courses, instructor "quality" (as indicated by student evaluations, academic credentials, years of experience, or scientific productivity) matters much less for students' learning than what they are actually doing in the classroom: what processes they engage as they grapple with the material that result in better encoding, longer retention, and improved odds of transfer (see Deslauriers. Schelew, & Wieman, 2011). It is my hope that this study will call attention to the need for better alignment between "what works" and "the way things are done" in large college courses, spurring changes in course structure and classroom dynamics to support the outcomes that we as educators care about most: our students' success.

After a century of incredible social and technological progress, education in America remains fundamentally unchanged. Charles W. Eliot, president of Harvard from 1869 to 1909 and one of the chief historical architects of our modern higher education system (who, among other sweeping innovations, introduced standardized course credits), wrote of his many reforms that "It is not well, that a house should last a century—it becomes unsuited to the improved habits of succeeding generations" (Gerhard, 1955, p. 652). The time has again come to rebuild our house: as instructors and administrators at all levels, we ignore this admonition at our peril.

PART II: SUBSEQUENT-COURSE ANALYSIS

Chapter Five: Introduction

OVERVIEW

In Part I of this study, a large-scale syllabus review of normative educational practices (e.g., course structure, teaching methods, learning activities) was conducted across more than 1,000 high-enrollment undergraduate courses at a large public institution. Based on these findings, I now use student outcome data to conduct a subsequent-course analysis: an assessment of the extent to which certain prerequisite-course variables affect student performance in their subsequent course over the same subject. Specifically, taking a section of an introductory courses with many retrieval practice requirements is compared with taking a section of the same course with few retrieval practice opportunities, and the causal effect of taking a prerequisite course high in retrieval practice is estimated using inverse propensity-score weighted regressions. Variations in model specification (fixed effects, random effects, cluster-robust error models) are examined to assess robustness. Finally, student subsequent-course performance was regressed on the full set of educational relevant variables using lasso-regularized regression in an attempt to identify additional variables related to retrieval practice and spacing as important prerequisite-course predictors of subsequent student success.

RATIONALE AND LITERATURE REVIEW

All systems of education are predicated on assumptions about how people learn. Though rarely stated outright, these assumptions are always implicit in *what* exactly is being taught and *why, how,* to *whom*, *where*, and *when*. In some instances, these assumptions are conscious choices on the part of educators and administrators; in the ideal

case, they will have been intentionally grounded in scientific research. Often, however, these assumptions are unconscious byproducts of political considerations, practical expedients, or accidents of history. Take, for example, the typical 9-month school year with a break for the summer. This became standard at the end of the 19th century when fully 85% of Americans worked in agriculture; today, the number of Americans who do agricultural work is less than 3% (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996). What if this arrangement—adaptive in its original context but now an arbitrary feature of modern life—is suboptimal for learning?

Zooming in on the standard school-year reveals further subdivisions, usually semesters or trimesters, during which full-time students are required to take a minimum number of courses. College courses are self-contained units of instruction covering a circumscribed topic which meet for a set number of hours each week. Since 1910, for reasons of bureaucratic efficiency, the amount of time students' courses meet for each week (the number of semester "credit hours") has been the primary measure of attainment in American higher education (Wellman, 2005). At the University of Texas, for example, students need 120 semester-hours to receive an undergraduate degree, provided they maintain a 2.0 GPA (General Requirements, 2017).

From very early on, this policy of treating units of class-time as a valid measure of student learning was met with resistance. Norman Foerster (1937), for example, likens the credit system to "purchasing a diploma on the installment plan" and complains that "once a credit was earned... it would be deposited and indelibly recorded in the registrar's savings bank, while the substance of the course could be, if one wished, happily forgotten" (p. 97). However, despite questionable assumptions, this system of accounting has become so entrenched that it has persisted relatively unchallenged for over one hundred years.

Zooming in further, we see that within a typical college course, subject matter is broken down into several stand-alone units of material. Students attend lectures and are given readings over each unit's material, which they are then tested over before moving on to the next unit and repeating the process (King, 1993). This non-cumulative tendency is exacerbated by textbooks, which present material one topic at a time instead of periodically returning to prior topics (Rohrer & Taylor, 2007).

Since unit exams often receive more weight than other assignments, they thus account for the greatest proportion of the final course grade and largely determine advancement. (In Part I, a survey of 1075 large undergraduate courses revealed that in the median course, 3 exams accounted for 75% of the course grade.) Notice that here learning has been implicitly equated with passing exams. In general, students are considered to have "learned" the material if they have "passed" each unit in a given course; if they have passed enough courses, they are assumed to have learned enough to merit a degree. We care about students passing exams and courses because we believe that it indicates what students will be able to remember and use at a future time, in a different place. Are the assumptions we make about learning and transfer well founded?

Accountability for student learning outcomes in higher education, where it exists at all, is very decentralized: whereas high school students must pass minimum-competence exams by law in order to graduate (Popham, 1978), there are no such comprehensive assessments of learning in higher education. Furthermore, professors usually receive no formal evaluations of their teaching effectiveness beyond student ratings on end-of-course surveys, a problematic metric at best (Shevlin, Banyard, Davies, & Griffiths, 2010; Stroebe, 2016). Thus, not only do instructors have to teach course material in a way that promotes durable, transferable learning; they are also responsible for creating cumulative performance metrics that accurately reflect this learning. A grade in the gradebook thus

takes on added importance: it is a final verdict reached by a professor about students' mastery of the course material. To be effective, this judgment should do more than document a student's past performance: it should act as a promissory note, vouching for what a student knows about a subject and what a student is capable of doing on their own. After all, what is important is not what students know in the classroom on the day of the final exam, but what they can remember and apply long after the course has ended.

For course performance to reflect these durable learning outcomes, the type of work that students produce and are graded on must engage cognitive processes that support them. Indeed, there is reason to believe that what students do in a given class (lectures, readings, assignments, activities, etc.) is far more important for retention and transfer than professor-level variables such as teaching experience. For example, Deslauriers, Schelew, and Wieman (2011) compared learning in two large (270-student) undergraduate Physics course sections taught concurrently at the same university: one was taught in a "traditional lecture" style by an engaging lecturer with high course evaluations and many years of experience teaching; the other, experimental group was taught "using instruction based on research in cognitive psychology" by a post-doctoral researcher who had never taught a course before. The instructional intervention lasted only a week and consisted of short pre-class readings, reading quizzes, in-class clicker questions with pair discussions, and small-group learning tasks; there was no lecturing, but guidance and feedback was provided by the intervention instructor. A 12-question test over material covered during the experimental week was developed collaboratively by both instructors, who "had agreed to make this a learning competition" (p. 863). This test was given to students during the next class period one week later. Students in the traditional lecture format averaged 41%, while students in the intervention condition averaged 74%, an enormous 2.5 standard deviation effect in favor of the intervention with pre-class quizzes, in-class retrieval practice, and

group learning. Unfortunately, it is difficult to know which of these elements were effective and whether they would have each been effective in isolation. One aim of the present project is to compare across courses that are otherwise similar, but that differ on important variables (e.g., group assignments, in-class retrieval practice) in order to determine the individual efficacy of these variables as well as the best "package" or combination of such variables for success in subsequent courses. Furthermore, this was a one-shot learning paradigm; in the present study, I am interested in the impact of these course-level variables over time, particularly on long-term retention and transfer.

If the primary goal of education is to teach students knowledge and skills that remain accessible to them over time and that can be flexibly applied outside of the classroom, then instruction should be designed with these goals foremost in mind. Cognitive psychology has produced a large body of research on learning and memory, and well-replicated findings have come together to yield robust principles about how to facilitate such long-term retention and transfer of learning. From this perspective, three of the most insidious assumptions made in education are (1) that testing is for assessment purposes only—that a test is a learning-neutral event for measuring what a given student knows; (2) that lectures and reading assignments determine what is learned—that passive and active forms of learning are equivalent; and (3) that performance on a test is proof of learning that will persist unaided and automatically transfer to new contexts.

A course based on these assumptions would have a small number of high-stakes exams to measure what knowledge has been acquired, but no more, because this would be time better spent learning. Since learning is thought to occur through lectures and assigned readings, in-class activities and group assignments would be considered inefficient make-work and thus avoided. Further, in such a course, previously covered content would not be

revisited after the exam, because students' scores are evidence that the material has already been learned—no need to waste time repeating ourselves!

It is likely that most large college courses fit this pattern. In Part I of this study, a sweeping syllabus review of 1075 large undergraduate courses revealed that the median number of exams (including final exams) was 3 and the median grade-weight for exams in the course rubric was 75%. More than 80% of courses had no graded in-class assignments and the average grade weight courses gave for such assignments was 2% (72% had no in-class activities at all besides listening to lecture). There was also little evidence that students are reviewing material after unit exams: only 4% of courses reported having cumulative exams, though 39% reported having cumulative final exams.

This design would be very sensible if the foregoing assumptions were to hold. And since the course abruptly ends after 15 weeks, they *appear* to hold! Unless a professor were to actively seek disconfirming evidence—say, by following up with students later to assess their retention of course material and their ability to apply it—the final course grades that they submit are readily accepted by all parties (professors, students, administrators, and ultimately employers) as proof that teaching has happened and that learning has happened.

Unfortunately, these assumptions are not only wrong, they are completely at odds with two of the most powerful, dependable techniques in experimental psychology for promoting long-term retention and transfer of learning—retrieval practice (the "testing effect"; Roediger & Butler, 2011 for review) and spaced repetition (the "spacing effect"; Cepeda et al. 2006 for review). Indeed, Hattie's (2008) sweeping synthesis of over 800 meta-analyses found that, of the 138 variables associated with achievement, the third largest effect (average d = 0.88 across 78 effects) was formative assessment (i.e., low-stakes retrieval practice with feedback), and spaced practice was not far behind at thirteenth (d = 0.71, 112 effects). In the following section, I will review the evidence for these

principles with respect to retention and transfer, followed by a discussion of how they can be applied in higher education, even in large classes. After this, I will describe a technique for assessing the prevalence of these and other educational practices in college classrooms; finally, I will propose methods for testing whether these course-level variables actually increase student success in subsequent courses—a clearer indication of retention and transfer than a course grade can offer.

Testing and Spacing for Retention and Transfer

The notion that testing could have beneficial effects on learning has been gradually making inroads into modern educational practice. For example, a modern textbook on college teaching methodology makes a clear distinction between summative assessment ("a performance evaluation") and formative assessment ("intended to furnish helpful feedback"; Nilson, 2010, p. 281). But even in the absence of feedback, testing can be a potent learning tool: the act of retrieving information from memory increases the likelihood that the information will be retrievable in the future, a finding known as the testing effect (Carrier & Pashler, 1992; Roediger & Butler, 2011). Indeed, after initial learning, being tested produces better retention of the material than an equivalent amount of time spent restudying it. Furthermore, the benefit of testing over restudy becomes larger as the delay before the final test grows longer: relative to restudying, retrieving information from memory results in slower forgetting over time and thus better long-term retention (Kornell, Bjork, & Garcia, 2011), an effect which holds both in the laboratory and in the classroom (e.g., McDaniel et al., 2011).

The memory retrieval required by testing is thought to enhance learning by directly modifying the retrieved content (e.g., by elaborating upon the representation of this content in memory, increasing its availability and accessibility; Bjork & Bjork, 1992). Though

retention is a goal in its own right, it is also a crucial first step for being able to apply the knowledge one has recalled to new contexts. But retrieval practice has recently been found to have a more direct role than just retention: compared to restudying, repeated testing can lead to better transfer performance on novel inference questions (Butler, 2010; Rohrer, Taylor, & Snoler, 2010) and on spatial learning tasks (Carpenter & Kelly, 2012). What's more, introducing variability during retrieval appears to enhance the transfer of learning to novel problems above and beyond the benefits of repeated testing. Specifically, practicing retrieval with *different* questions that tap the same underlying concept makes it more likely that one can apply knowledge of the underlying concept to novel questions, relative to retrieval practice with the same question (Butler, Black-Maier, Raley, & Marsh, 2017).

Related to the testing effect—and perhaps even more well-known—is the spacing effect: the finding that spacing out one's studying or testing sessions produces superior learning relative to an equivalent amount of studying or testing in a single sitting or in sessions occurring closer together in time (Cepeda et al., 2006 for review). That is, students who spread their practice out over time enjoy greater long-term retention of that information than those who practice for the same amount of time but do not space it out. The benefit of spaced practice over massed practice on retention holds across learners of all ages and subject-matter of all kinds, including learning grammar, spelling, reading skills, advanced mathematics, motor skills, foreign language vocabulary, history, and more (Carpenter et al., 2012).

Further, this finding is neither new nor unusual (Pyle, 1913; Austin, 1921; Gordon, 1925). Almost a century ago, Austin (1921) found that reading a text five times in one day was just as effective as reading the text once a day for five days on tests of immediate recall. However, the spaced readings resulted in much better performance on delayed retention tests, and the effect grew with the size of the delay. The magnitude, robustness,

and consistency of findings related to the spacing effect led Dempster (1988) to call it "one of the most remarkable phenomena to emerge from laboratory research on learning"...but tellingly, this quote appears in an article entitled "The Spacing Effect: A Case Study in the Failure to Apply the Results of Psychological Research."

Teachers often admonish their students to study a little bit each day instead of cramming right before the test, and thus appear to have some intuitive understanding that spaced-out is better than massed-together. However, this principle is seldom used in the classroom or reflected in course structure (Dempster, 1988). As mentioned above, the standard format of a college course is still such that a few high-stakes tests covering distinct units of material account for the bulk of students' grades. If cramming right before these exams can produce equivalent (sometimes better) performance, then students have little reason to space out their studying. Since spacing results in superior long-term retention, this format unintentionally rewards behaviors that lead to transitory learning while penalizing those that lead to durable learning. For example, Rawson and Kintsch (2005) had college students study expository texts and take tests over them. One group of students read the text only once, while the other two groups read it twice: one of these two groups (the massed-study condition) read the text twice in a row, while the other (the spaced-study condition) read it two times with a week in between. The final test was given either immediately or two days after the final reading and consisted of a recall component plus 12 short-answer comprehension questions. Two experiments using two different texts produced the same results: massed study produced significantly better performance on the immediate test, but spaced study produced significantly better performance on the delayed test. Thus, cramming can be an effective way to get high marks on exams, but it is clearly a poor way to achieve durable learning.

Studies like these demonstrate improved retention over days or weeks, but how do we know that spaced retrieval practice is better for long-term retention? And just how long is long-term retention: can we achieve indefinite retention, and if not, what's the best we can hope for? Harry Bahrick's pioneering research into long-term retention has offered many exciting answers to these questions. With respect to the first question posed, he conducted a 9-year longitudinal study using members of his own family as participants (Bahrick et al., 1993). They learned and relearned 300 English–foreign-language word pairs, varying both the number of relearning sessions (13 vs. 26) and the interval between sessions (14, 28, or 56 days) within subjects. After the training, retention was tested 1, 2, 3, and 5 years later. Bahrick found strong main effects on retention for both the additional sessions and the longer spacing intervals. In fact, just 13 retraining sessions spaced 56 days apart produced the same retention benefit as 26 sessions spaced 14 days apart. But while the longer spacing intervals resulted in much better retention 5 years later, they hindered initial learning during the training sessions. Thus, we are again cautioned against the dangers of judging learning from performance on tests given soon afterwards: immediate and long-term performance are often inversely related.

The second question (exactly how long *is* long-term retention?) has proven more difficult to answer definitively. However, many important insights have been offered by analyzing people's memory for things like basic Spanish vocabulary (material covered in an Introductory Spanish course) and basic algebra rules (those taught in an Algebra I course) years after they last encountered the material (Bahrick, 1993, 1984a; Bahrick, Bahrick, & Wittinger, 1975). These cross-sectional studies survey hundreds of people about their background in a given subject—when their most recent course in it was, how many classes total they took in it, what grades they received in those classes, and to what extent they have used the material since they quit learning it. This results in a sample of

participants who vary widely in their time since content acquisition, how long it had been since they stopped using the content (the "retention interval"), and how well the content was learned initially (e.g., number of courses), all naturalistically acquired. Then, participants are tested over their retention of the basic introductory material in these subjects (e.g., a test of introductory Spanish vocabulary, a test of basic algebra skills).

From this data, researchers then generate a regression equation which can be used to plot retention over time for different degrees of initial learning (i.e., different amounts of retrieval practice, spacing, and elaboration that occurs when taking additional courses in these subjects), while controlling for rehearsals during the retention interval (see Figure 37). One very interesting finding from these analyses is that in general, retention declines exponentially for the first 3 to 6 years after learning has ceased, but that after this time it remains largely unchanged, even after periods of up to 50 years. Concretely, 3 years after taking a single semester of Spanish, almost all of the basic Spanish–English vocabulary covered in the course will be lost without any subsequent practice. However, those who took 5 semesters of Spanish recall around 60% of their basic Spanish vocabulary more than 25 years later, controlling for subsequent practice (Bahrick, 1984a).

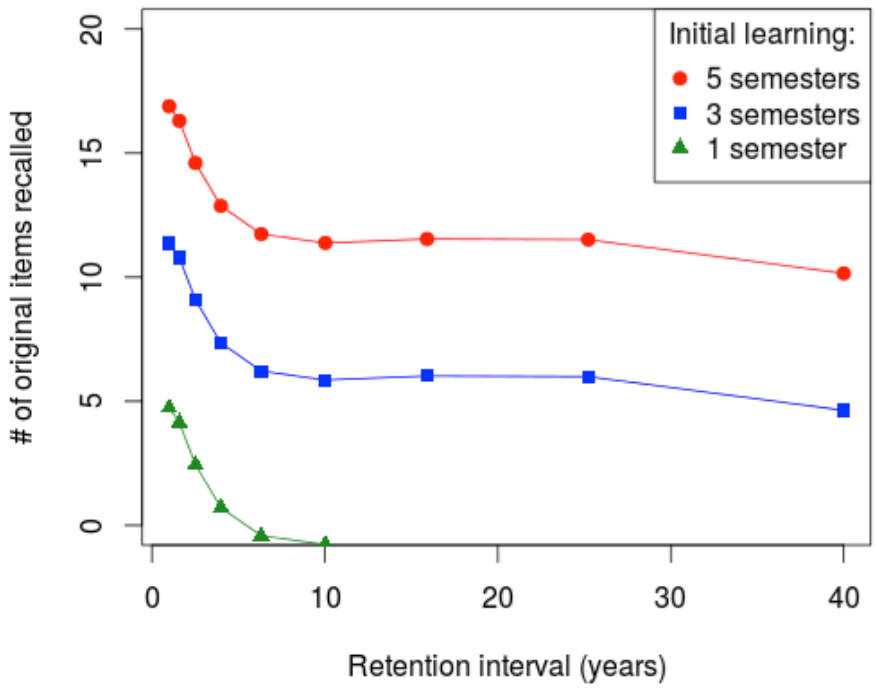

Figure 37 Retention of basic Spanish-English vocabulary (recall) by level of initial learning (number of semesters), with zero subsequent rehearsals. Figure adapted from Bahrick (1984a; Fig. 6) using the regression equation given in his Table 8.

Still more robust findings are observed with retention of basic math: it has been shown that people who take several mathematics courses in college show *no significant declines* in their retention of introductory algebra or geometry content during a 50-year retention interval, even if they have not used or in any way rehearsed the material during that time (Bahrick & Hall, 1991). On the other hand, students who performed equally well in their high school math courses but took no additional math in college were found to have forgotten almost everything during the same time period. Note that these studies are unable to separate out the specific effects of spacing and retrieval-practice on retention. They do

show, however, that if your initial learning was high (i.e., multiple learning sessions spaced out over time), your long-term retention stabilizes at a higher level than would be the case if your initial learning was low; if the content is acquired over a very short period, retention tends to decline rapidly and may be lost altogether. Since taking multiple semesters of mathematics or a foreign language requires you to repeatedly access this basic knowledge over a long period of time, the retention benefits for this material are most parsimoniously attributable to spaced retrieval practice.

TWO CASE STUDIES OF SPACED RETRIEVAL PRACTICE

Case Study 1: Benefits in Real-World Medical Settings

Even very simple spaced-practice interventions can have a large impact on both retention of knowledge and on transfer into real-world contexts. Dolan, Yialamas, and McMahon (2015) conducted a randomized controlled study with medical students who were completing their residency: after receiving a 1-hour lesson on osteoporosis care and fracture prevention, students in the control group received one email containing a 25-item multiple choice self-assessment, while students in the intervention group received the same 25 multiple choice questions, but instead of being delivered all at once, 1–3 questions were emailed at a time over a 3–6 month period. Items answered correctly were repeated once 28 days later, while items answered incorrectly were repeated twice at 14-day intervals (the variability in the length of time for treatment was due to differences in the number of incorrect responses among students). Ten months after the start of the intervention, the treatment group significantly outperformed the control not only on a bone-health knowledge assessment, but also on real clinical outcome measures: they screened more patients for low bone density, screened them more accurately, and effectively treated more who were at risk for fracture. This is not an isolated result: other experiments using similar

spaced interventions have observed improved retention and transfer to clinical practice (Price et al., 2010). Studies like these are especially important, given that medical students have been shown to forget a substantial portion of basic knowledge by the time they begin clinical rotations (Butler & Raley, 2015).

Case Study 2: Benefits in Large College Classrooms

In 2011, two professors in the Department of Psychology at the University of Texas at Austin decided to integrate spaced retrieval practice into their large Introductory Psychology course (Pennebaker, Gosling, & Ferrell, 2013). These professors had been teaching this course every year from 2006 to 2011 using "traditional approaches" in which 4 in-class multiple choice exams were given across the semester and accounted for 86% of students' final grade. In their new course format, tests were eliminated altogether; instead, students took in-class quizzes each day, for a total of 26 quizzes that together accounted for 86% of students' final grade. Quizzes consisted of 8 items: seven covered material from the previous lecture and readings, while one item was personalized, repeating a question that they had previously missed. Quizzes were made available through an online platform and students completed them using their laptops. Other aspects of the course, including the lecture format, content, and sequence, were intentionally kept constant.

The results of their comparison revealed that performance on items common to both old course exams and daily quizzes was marginally better in the new frequent-quizzing format (77.1% vs. 71.2%, $p = .06$). They also found that students' average performance in other classes taken concurrently, as well as their average performance in classes taken the following semester, was significantly better for those in the frequent-quizzing course, controlling for year and SES. The authors suggested that the frequent-quizzing format may have encouraged self-regulatory skills and study habits which generalized to other courses,

though this hypothesis was not directly tested. Finally, they examined socio-economic status (SES) disparities and class performance, finding a significant interaction between SES and course format on course grades. Specifically, grade differences between higher and lower SES students were twice as large in the traditional course than in the frequent-quizzing course.

This study is of immediate relevance to the present project for several reasons. First, it was conducted in an authentic undergraduate classroom setting: a high-enrollment, lecture-based introductory course at a large public university. Second, it shows that implementing a system of frequent testing is feasible, even in large (~500 student) introductory courses. Third, it provides still more evidence that spaced retrieval practice is associated with better outcomes on several levels, including better performance in concurrent and subsequent courses, as well as reductions in the SES achievement gap. Fourth, it required *very little change* to instructors' teaching or course materials: the same lecture format was observed and the same topics were covered in the same way. Finally, by examining how one course can influence performance in concurrent and subsequent courses, this study sets an important precedent for studying the broader, longer-term, cross-curriculum effects of college learning experiences.

RETENTION AND TRANSFER AS PREPARATION FOR FUTURE LEARNING

Transfer of learning occurs when something previously learned is applied to new situations. However, conceptions of how to facilitate transfer through instruction have differed greatly over the years. In most of the research literature, transfer of learning has been measured by scores on a final transfer problem in tightly controlled experimental settings: no seeking help from other resources, no opportunity to test possible solutions and revise in light of feedback (e.g., Gick & Holyoak, 1980). While this "sequestered problem

solving" paradigm is a valid approach to measuring transfer, it restricts the definition of the phenomenon by limiting the type of evidence admissible for it: such designs implicitly equate transfer with the ability to directly apply a previously learned concept to a new problem of a specific type (Bransford & Schwartz, 1999).

A more fruitful characterization of transfer is one that captures the benefits of previous learning not just on task performance, but also on the speed and quality of new learning. This idea is as old as memory research itself (Ebbinghaus' famous forgetting curve plotted "savings" in the time required to relearn a list; Ebbinghaus, 1885/1913) and it permits much finer measurements of transfer. Importantly, to conceive of transfer as "preparation for future learning" (Bransford & Schwartz, 1999) reinforces the notion that transfer is not an all-or-none proposition: the influence of past learning on the present is always a question of degree.

In the present study, I will avoid the issue of "sequestered problem solving" by measuring transfer through performance on subsequent related coursework. Measuring transfer from a prerequisite course by examining students' achievement in the next-in-series course provides a summative measure of their performance on a variety of relevant transfer tasks over time (such as graded student work) in a naturalistic setting where students are free to make use of resources that facilitate such learning. That being said, some transfer questions worth posing will require a narrower criterion to adequately assess, such as how the aspects of writing-intensive coursework influence students' future writing ability in other courses. This sort of transfer question would be better addressed by direct-application methodology, such as collecting samples of students' writing and assessing their quality. In the present study, because our operationalization of transfer presupposes retention, and because direct tests of retention for material covered in previous courses would be prohibitively difficult, retention will not be examined in isolation.

THE PRESENT STUDY

To what extent are the research-based principles discussed above being used in higher education? Despite widespread advances data collection, educational practices in college courses remain largely unmeasured: beyond broad generalizations about the traditional approaches (*lecture-then-test*, *sage-on-the-stage*; King, 2010), little is known about the scope of course schedules, teaching practices, classroom activities, and out-of-class assignments in large college courses. This dearth of information means that little research has been done to examine how course features such as opportunities for retrieval practice and cumulative exams are associated with student outcomes such as long-term retention and transfer of learning. Though there have been previous attempts to characterize what goes on in college courses using syllabi (e.g., Graves, Hyland, & Samuels 2010; see below), there is virtually no research connecting these course-level variables to student outcomes.

However, because of the large-scale syllabus review undertaken in Part I of this study, I am now well positioned to explore questions of this nature. Having developed a rich corpus of data about course designs, teaching practices, and learning activities in large college courses, and having mined this descriptive data for associations of interest, the all-important issue of student outcomes can finally be addressed: Specifically, what course-level variables predict future student achievement? Does taking a course section with more retrieval practice opportunities lead to improved transfer?

Research Questions and Hypotheses

I approach these two broad questions by examining course sequences (i.e., two courses in the same academic subject in which one course is an immediate prerequisite for the other) to estimate effects of certain course variables hypothesized *a priori* to positively impact subsequent-course performance and to explore what other features of prerequisite courses are predictive of success in the next course in the sequence. Based on previous research on the efficacy of educational practices reviewed above and leveraging our previous descriptive results, two main questions of interest are posed for investigation:

1. Does taking a prerequisite course with more spaced retrieval practice result in better performance in the subsequent course? Specifically, it is hypothesized that compared to students in that subsequent course whose prerequisite course featured little graded retrieval practice, those in the prerequisite course with more graded retrieval practice will have higher subsequent-course achievement. Additionally, I expect to find an association between subsequent-course performance and variables related to both formal spacing and retrieval practice opportunities (e.g., number of quizzes/exams, cumulative exams) as well as more informal, ungraded opportunities (e.g., classroom response systems, availability of practice quizzes).
2. Is taking a prerequisite course with more active learning during class-time (as indicated by number of in-class assignments, descriptions of a flipped classroom, or other in-class active-learning features) predictive of better performance in the subsequent course? It is hypothesized that prerequisite courses with active-learning elements will be associated with higher average subsequent-course grades than prerequisite courses with less active learning.

Additionally, because students are not randomly assigned to their prerequisite courses, I must control for any differences between treatment groups (e.g., any preexisting

demographic differences between students in prerequisite courses with retrieval practice and students in prerequisite courses without retrieval practice). This is done in an effort to rule out pre-existing differences between groups as an alternative explanation for any observed effects that would otherwise confound the relationship between treatment (specifically, high retrieval practice) and the outcome (subsequent course performance). This will be achieved using propensity score methods and covariate balance checks as detailed in the Analysis section below.

Chapter Six: Research Design

For important background about the syllabus review, see the Procedure and Methods section in Part I. In what follows, familiarity with this section is taken for granted.

SUBSEQUENT-COURSE ANALYSIS WITH OBSERVATIONAL DATA

To preview, four different modeling approaches will be used to address the research questions in the present study—regression with fixed effects of course section, regression with random effects of course section, regression with cluster-robust standard errors, and individual regressions within a given courses. As I cannot randomly assign students to college courses, these methods are inherently quasi-experimental, but I employ propensity-score weighting techniques to ensure that treatment and control groups have very similar distributions of covariates. Within such homogenous groups, differences in outcome are more plausibly attributed to the treatment, since the groups are otherwise as equivalent as possible (a situation which, in an experimental study, is achieved with a random treatment-assignment mechanism). This approach is akin to a retrospective cohort design in medical research; we want to assess the impact of course-level variables on performance in subsequent courses, so student records (here, institutional and demographic data) are used to establish two groups of subjects who are as alike as possible on potentially confounding covariates but who differ from each other on the independent variable of interest. These two groups are then compared with respect to the outcome as if the independent variable had been randomly assigned (Mann, 2003).

For example, one characteristic of interest in the present study is the presence of retrieval practice opportunities in a student's prerequisite course (say, Chemistry 301), and the outcome of interest is a student's grade in the subsequent course (Chemistry 302, the next course in the sequence). A basic approach would be to use a linear regression

framework to explore whether there was a significant difference in subsequent-course grades for students whose prerequisite courses had, for example, a large retrieval-practice component, versus those whose prerequisite courses did not. However, the presence or absence of retrieval practice in a prerequisite course is unlikely to be the only variable with respect to which courses differ, so I aggregate results across a variety of courses, I try to account for all such variability by explicitly modeling it wherever possible, and I eliminate systematic confounds using propensity-score methods to achieve covariate balance.

Despite the importance of this topic, very little research exists about how students' experiences in one course are associated with their performance in subsequent courses, and thus there is no standard approach for conducting such an analysis. However, in the economics literature addressing teacher quality there are several studies that set a precedent for using students' subsequent course grades as a measure of learning that took place in the prerequisite course (e.g., Carrell & West, 2010; Weinberg, Hashimoto, & Fleisher, 2010). For example, value-added modeling (VAM) approaches seek to measure teacher quality by assessing the unique contributions of teachers to students' academic attainment (Hanushek & Rivkin, 2010; for review, see Koedel, Mihaly, & Rockoff, 2015). Such approaches isolate the effects of teachers on student achievement while controlling for other factors that could explain differences in student performance, such as student demographic information, socioeconomic characteristics, and previous academic achievement. Often using standardized test scores rather than subsequent course performance, these approaches have identified great variability in teacher effectiveness, even within the same school.

I look to value-added methods for modeling methodology, but my subsequent course analysis departs from such approaches in several ways. First, whereas most of the teacher quality literature focuses on primary and secondary education, we examine teacher

effects in a post-secondary setting. Only a small number of papers have attempted to use such techniques in a higher-education setting (e.g., Carrell & West, 2010; Hoffman & Oreopoulos, 2009; Weinberg, Fleisher, & Hashimoto, 2009). Furthermore, studies in the VAM tradition rely on education production functions which estimate the overall contribution of each teacher to student outcomes, disregarding individual teacher characteristics which are of interest in the present study. Furthermore, such studies typically use an outcome variable that measures short-term gain (achievement test score, final grade in current course), controlling for previous grades or achievement scores (Hanushek & Rivkin, 2010).

In contrast, the present study seeks to estimate the effect of course characteristics (rather than overall teacher effects) on undergraduate students' learning as indicated by their subsequent course performance, controlling for background variables. One approach to modeling such a relationship is shown below.

$$g_{ips} = \tau_p + \tau_s + \mathrm{X}_i'\beta_i + e_{ips} \quad (1)$$

Here, g_{ips} represents the grade of student i who took prerequisite course p and subsequent course s; τ_p represents the unique effect of the prerequisite course p on the outcome; τ_p represents the unique effect of the subsequent course s on the outcome (to control for unmeasured aspects of the environment that differ across classrooms); β_i is a vector representing the contributions of each background covariate for student i; and e_{ips} is an error term that captures unmeasured influences on the outcome. In this formulation, τ_p indicates how well students who took prerequisite course p do in their subsequent course s after adjusting for student characteristics X_i and differences in grading or difficulty among subsequent courses τ_s.

In Equation 1, the separate intercepts for prerequisite course and subsequent course appear as fixed effects, but they could also be modeled as random effects. Indeed, the data

are inherently clustered: attributes of courses do not vary across students within each course, and thus the observations within each course are unlikely to be independent, which violates standard regression assumptions. This modeling choice assumes that the random effects are uncorrelated with all other explanatory variables, an assumption which holds if strong ignorability holds (see below; Carrell & West, 2010). Importantly, this approach also allows us to account for student-level and course-level variation when estimating course-level regression coefficients. Specifically, in fixed-effects models it would not be admissable to include course-level indicators along with course-level predictors (e.g., Gelman & Hill, 2007, p. 246). Since the present study is concerned with estimating the effect of course-level variables on student outcomes, a multi-level approach is more appropriate. That being said, random-effects models do rely on additional assumptions that may be untenable, including that the random effects are normally distributed and that there are sufficient observations at each level for the asymptotic theory for the test statistics to be justified, but given the size of our sample and the number of group-level effects to be estimated, these assumptions should hold.

In addition to explicitly modeling course section effects as fixed or random, a third way to model this relationship is to use ordinary least squares (OLS) with cluster-robust standard errors to account for the clustered structure of the data (e.g., Primo, Jacobsmeier, & Milyo, 2007; Cheah, 2009). This approach has the benefit of relying on fewer assumptions but would not give separate estimates for the individual courses. In the present study this is not an issue, since the objective is not to examine the impact of individual teachers on student outcomes.

Potential outcomes and propensity scores

Assignment of students to courses is very rarely random. In K-12 education, for example, student assignment to classes can be influenced by student tracking, by parental requests for certain teachers, or by an effort on the part of administrators to separate students with behavioral problems or ensure a certain teacher gets certain students. In higher education, students register for courses on the basis of many factors including scheduling considerations and the reputation of the professor. Because students are not randomly assigned to specific pre-requisite courses and instead self-select into them, any variable that could have influenced both their choice of initial course and their performance in the subsequent course could lead to a spurious treatment effect, a pervasive issue in observational research known as *selection bias* (see Rothstein, 2010 for a discussion of this problem in value-added modeling). Because treatment subjects may differ systematically from control subjects at the outset, the causal effect of treatment cannot be estimated simply by comparing outcomes between groups.

Modern approaches to estimating causal effects from observational studies depend heavily on the potential outcomes framework (Rubin, 1974; 2005). Within this framework, an individual causal effect is defined as the difference in potential outcomes for a given subject (i.e., the difference between the subject's outcome if he or she had received treatment and the subject's outcome if he or she had *not* received treatment). For example, let Y_i represent the outcome of interest for student i (e.g., grade earned in the subsequent course), and let T_i be a binary variable indicating the treatment condition of student i (1 = *high retrieval practice in prerequisite course*, 0 = *low retrieval practice in prerequisite course*). Writing the outcome as a function of treatment, we can see that if student i were assigned to treatment, the outcome would be $Y_i(T_i = 1)$; if instead student i had been assigned to control, the outcome would be $Y_i(T_i = 0)$. Thus, with a binary treatment

variable, each subject has two potential outcomes, and the difference between these outcomes is the individual-level causal effect of treatment, $Y_i(1) - Y_i(0)$. But because a given individual can only be assigned to either the treatment condition or the control condition (but not both), we can only ever observe a single potential outcome per subject (the other remains a hypothetical, counterfactual outcome). The fact that the universe only allows us to observe a single potential outcome per subject is known as the "fundamental problem of causal inference" (Holland, 1986).

The average treatment effect (ATE) across all subjects is our quantity of interest. It is obtained by averaging all individual-level causal effects, which can be written

$$ATE = \mathrm{E}[Y_i(1) - Y_i(0)] = \mathrm{E}[Y_i(1)] - \mathrm{E}[Y_i(0)] \quad (2)$$

If treatment is randomly assigned, those subjects who are assigned to treatment represent a random subset of the entire sample (those assigned to control represent a random sample as well). We do not observe every subject's potential outcome under treatment $Y_i(1)$, but we do observe them for the random subset assigned to treatment. Likewise, we do not observe every subject's potential outcome under control $Y_i(0)$, but we do observe them for the random subset assigned to control. Therefore, the expected potential outcome under treatment, for the treatment group is the same as the expected potential outcome under treatment for the entire sample (and equivalently for the control condition):

$$\mathrm{E}[Y_i(1) | T_i = 1] = \mathrm{E}[Y_i(1)]$$
$$\mathrm{E}[Y_i(0) | T_i = 0] = \mathrm{E}[Y_i(0)] \quad (3)$$

Notice that these equalities imply that treatment assignment is independent of potential outcomes, $Y(1), Y(0) \perp T$. Importantly, the quantities on the left-hand side of Equation 3 can be estimated based on observable data. Thus, under random assignment, an unbiased estimate of the ATE can be obtained by computing the mean difference in outcome between subjects in the treatment group and subjects in the control group,

$$\begin{aligned}\widehat{ATE} &= \mathrm{E}[Y_i(1)|\, T_i = 1] - \mathrm{E}[Y_i(0)|\, T_i = 0] \\ &= \frac{1}{\sum T_i}\sum Y_i T_i - \frac{1}{\sum(1 - T_i)}\sum Y_i(1 - T_i) \\ &= \bar{Y}_T - \bar{Y}_C \end{aligned} \quad (4)$$

where $\bar{Y}_T$ represents the mean outcome in the treatment group and $\bar{Y}_C$ represents the mean outcome in the control group. Thus, under random assignment, standard statistical methods such as *t* tests are appropriate for detecting causal effects and regression coefficients for treatment condition can be given a causal interpretation.

If treatment is not randomly assigned to subjects, then treatment assignment is not necessarily independent of potential outcomes. Because $\mathrm{E}[Y_i(1)|\, T_i = 1] \neq \mathrm{E}[Y_i(1)]$, the ATE cannot be estimated as shown above (under randomization). The best way forward in the absence of random assignment is to identify and measure all possible factors that could confound the relationship between treatment receipt and outcome and then estimate causal effects by comparing treated and untreated subjects conditional on the measured confounders. If all confounding pre-treatment covariates have been measured in $\boldsymbol{X}$, then potential outcomes are conditionally independent of treatment given $\boldsymbol{X}$,

$$[Y(1), Y(0)] \perp T|\boldsymbol{X} \quad (5)$$

If the conditional independence in Equation 5 holds, and there are some treated and untreated subjects in each subgroup defined by $\boldsymbol{X}$, then we have satisfied the so-called *strong ignorability* (or *unconfoundedness*) assumptions and can proceed as to estimate causal effects as we would under true random assignment (Rosenbaum & Rubin, 1983). Indeed, this is the rationale for using techniques such as ANCOVA to adjust for covariates when estimating the effect of interest.

Though regression adjustment has long been the primary research tool used to "control for" potentially confounding factors in an effort to recover unbiased estimates of

treatment effects, this traditional approach has several shortcomings. Extrapolation beyond observed data is common in regression approaches and can be difficult to detect (e.g., hidden within interactions among variables). Such extrapolations serve to increase the sensitivity to the specific regression model specification (e.g., choice of functional form, interactions, higher-order terms), leading to large variations in effect estimates (Ho, Imai, King, & Stuart, 2007). Crucially, this sensitivity to model specification comes with great potential for misuse. If an initially specified model produces a non-significant effect, the temptation to alter the specification (e.g., by adding, removing, or interacting covariates) will be high; through such tinkering, "significant" effects can often be produced that are in line with one's favored hypothesis by capitalizing on chance alone. As will be discussed below, propensity-score methods can be used in conjunction with any regression model to make causal effect estimates more accurate and far less model-dependent by improving covariate balance between comparison groups (e.g., treatment and control).

A propensity score is an individual's predicted probability of receiving the "treatment" (i.e., the dependent variable of interest, whether observational or experimental) based on that individual's background covariates and other relevant characteristics. Typically, propensity scores are computed by measuring all covariates thought to influence both the treatment and the outcome, regressing the treatment indicator on those covariates in a logistic regression, and then using the fitted model to predict the probability of treatment receipt for each subject. This predicted probability of treatment receipt given a set of covariates is called a propensity score, denoted $p(x)$. Notice that a propensity score model does not involve the outcome data and is used for purely predictive purposes to model the treatment-assignment process.

The propensity score is an example of a *balancing score*—a univariate summary that preserves all information about the relationship between treatment and covariates—

representing subjects' probability of being treated (Rosenbaum & Rubin, 1983). Conditioning on the estimating propensity score is done to remove observed systematic differences between treated and control subjects. Specifically, within subgroups defined by the propensity score, treatment status is independent of baseline covariates:

$$\boldsymbol{X} \perp T \mid p(\boldsymbol{X}) \tag{6}$$

Given that all confounding variables have been measured and included in the propensity score model, then adjusting for differences in propensity score between treatment and control conditions removes all bias in the treatment effect estimate associated with differences in covariates. Importantly, if strong ignorability holds given $\boldsymbol{X}$, then it also holds given only $p(\boldsymbol{X})$. Thus, adjusting for the one-dimensional propensity score is sufficient for unbiased estimates of causal effects.

For example, treatment could be defined as taking a prerequisite course high in retrieval practice ($T = 1$) versus a prerequisite course low in retrieval practice ($T = 0$). If a propensity score is computed based on all covariates that predict whether a student would take a high retrieval practice course, then students in the treatment condition can be compared to students in the control condition with similar propensity scores. In groups of subjects having similar propensity to receive treatment, differences between treated and control subjects are most plausibly attributable to the treatment; and within such groups average effect of treatment (e.g., mean performance in subsequent course) can be estimated. Several approaches to forming such groups exist, the simplest and coarsest of which is stratifying the distribution of propensity scores by quantile and estimating treatment effects only within groups formed by those quantiles. Partitioning the distribution of propensity scores by quintiles, for example, and then comparing treatment and control subjects within these five groups (and pooling the estimates) removes around 90% of the bias from observational studies (Cochran, 1968). Another simple approach is one-to-one

matching: yoking each subject in the treatment group to the subject in the control group with the most similar propensity score. In this situation, a *t* test can be used to test the null hypothesis that the causal effect is zero (Thoemmes & Kim, 2011). More sophisticated approaches such as one-to-many matching and inverse-propensity weighting achieve better results and are often used in practice.

The result of propensity score matching or weighting is to make the covariate distributions for the treatment and control groups resemble each other as closely as possible: the result is less model dependence, less potential for bias, and lower variance (Ho, Imai, King, & Stuart, 2007; 2011). There are many approaches to achieving covariate balance through propensity scores: I will use inverse-propensity-score weighting (IPW) together with regression analyses to perform this analysis (Lunceford & Davidian, 2004; Schafer & Kang, 2008). Inverse-propensity weighting—or weighting by the inverse probability of treatment—results in an adjusted population in which baseline covariates are independent of treatment condition (Austin & Stuart, 2015). Weighting allows for finer-grain adjustments than does stratification or one-to-one matching and it has the benefit of retaining observations that are in non-overlapping regions of both conditions' propensity-score distributions and assigning them a correspondingly low weight. However, extreme weights (in either direction) must be monitored because they can increase the variability of treatment-effect estimates (Kang & Schafer, 2008). Where appropriate, weights will be very liberally trimmed (to the 0.001st and 99.9th percentiles); trimming weights in this manner is common practice and is known to improve performance of inverse-propensity score weighting in cases where the propensity-score model produces extreme weights (e.g., Lee, Lessler, & Stuart).

For each analysis planned, I first generate propensity scores and check to ensure balance among the covariates (by comparing adjusted covariate distributions using

standardized mean differences, Kolmogorov-Smirnov statistics, and variance ratios). I then use inverse-propensity weighted regressions (weighted least-squares) to estimate the ATE. Combining a weighted regression model of the outcome with a propensity-score model of treatment exposure results in approximately unbiased effect estimation if either the exposure or the outcome model is correct, a beneficial property known as *double robustness* (Funk et al., 2011).

In inverse-propensity weighting (IPW), to estimate the ATE each subject is assigned a weight equal to the inverse of the probability of receiving the treatment that the subject actually received: subjects who were actually treated receive a weight of $1/p(X)$, while subjects who were actually not treated receive a weight of $1/(1 - p(T))$. A weighted regression minimizes the weighted sum of squares: each treatment observation counts as treatment to the degree to which it exhibited propensity *for not being treated* and each control observation counts as control to the degree to which it exhibited propensity *for being treated* (Imbens, 2000). By way of intuition, individuals with low propensity for treatment (based on background covariates) but who actually receive treatment are a rare and valuable source of counterfactual information: thus, they receive greater weight. Individuals with high propensity for treatment who nevertheless do not receive treatment are similarly valuable and also receive greater weight.

Though the techniques for doing so are relatively new, there is precedent for using propensity score matching and weighting methods with multilevel data (e.g., Arpino & Mealli, 2011; Li, Zaslavsky, & Landrum, 2013). Importantly, it has been shown that so long as the clustered nature of the data is accounted for in either the non-parametric propensity-score model or the parametric outcome model, the bias owing to such dependencies in the data can be effectively removed (e.g., Li, Zaslavsky, & Landrum, 2013).

EXPLORATORY LASSO REGULARIZED REGRESSION

In addition to estimating the causal effect of high retrieval practice in prerequisite courses on students' subsequent-course performance, it was also of interest to explore whether other features of prerequisite courses (specifically those associated with retrieval practice and in-class active learning) were predictive of successes in the subsequent course, and if so, which ones and to what extent. To achieve this, average subsequent-course grades were computed *for each prerequisite course*; these averages were then regressed on the full set of course-level predictor variables. Because there were over 40 predictor variables, lasso regression was used to improve model interpretability and prediction, but ordinary least-squares (OLS) results are provided for comparison.

The lasso was devised as a variable-selection and regularization technique in machine learning (Tibshirani, 1996; 2011). Regularization techniques are ways of constraining models by penalizing them as they become more complex (e.g., as the number of parameters grows) in order to produce more parsimonious solutions (i.e., one that retains only the most predictive variables), which is especially important as the number of predictors grows large. The lasso's objective function is the same as in OLS regression but with an added criterion: in addition to minimizing the sum of the squared deviations between observed and predicted values, the regression solution must also minimize the sum of the absolute values of the parameter estimates. Specifically, given a sample of N observations, a set of p predictor variables, and a single outcome measure y, the objective function of the lasso is given as

$$\min_{\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i - X_i'\beta)^2\right\} \; subject\ to \; \sum_{j=1}^{p}|\beta_j| \leq \lambda, \tag{7}$$

where λ is a regularization parameter that controls the amount of penalty and is usually chosen through cross-validation.

The more non-zero coefficients a model has, the more variance it can explain in the observed values (i.e., lower sum-of-squares), but the larger the sum of absolute coefficients, thus putting the two desiderata into tension. The optimal lasso solution will thus be a compromise, yielding accurate parameter estimates for a subset of highly predictive variables while shrinking smaller coefficients to zero (Helwig, 2017). Unlike its conceptual cousin, the stepwise or hierarchical regression, lasso regression adds regularization to combat overfitting (instead of optimizing R^2 or related fit criteria, a procedure which capitalizes on sampling error; Thompson, 1995), thus yielding models that tend to generalize better (i.e., perform well on new data; Yarkoni & Westfall, 2017).

Chapter Seven: Analysis

DATA

The dataset consists of 13,332 first-time-in-college students at the University of Texas at Austin (i.e., no students transferring from other colleges, no readmitted students) for whom I have demographic information who took at least one course sequence (i.e., a sequence of two courses for which the first course is a prerequisite for the second course) for which I have syllabus data. Syllabus data were collected for all high-enrollment undergraduate courses at UT Austin from Fall 2011 to Spring 2016 (see Part I). Because syllabi were coded on the basis of total enrollment, the dataset is limited to a subset of all possible prerequisite- and subsequent-course pairs for which I have syllabus information for the prerequisite course (for a total of twelve different prerequisite courses). All possible course sequences in the syllabus dataset are shown in Figure 38, along with the total number of unique course sections of each (given in parentheses).

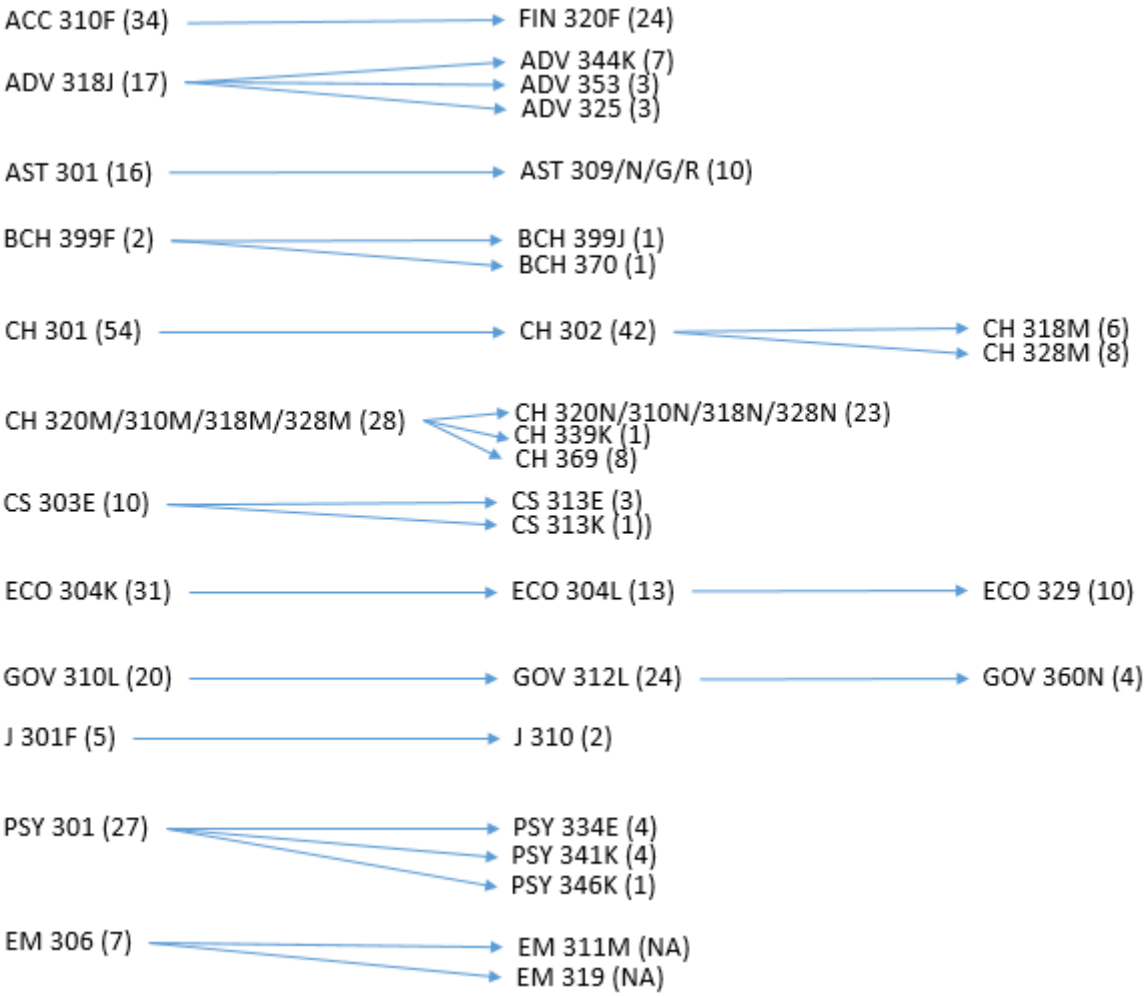

Figure 38 All prerequisite–subsequent course sequences in the syllabus dataset; credit for first course listed is prerequisite for the second in the official course catalog. The number of unique syllabi for each course is given in parentheses.

Across these twelve prerequisite courses, I have covariate and outcome data for the students described above—those who took at least one syllabus-coded prerequisite course and then continued on to complete a course sequence by taking one of the subsequent courses given in Figure 38. The total number per prerequisite course varies by prerequisite

course as a function of the number of prerequisite courses for which I have coded syllabus data. However, given that a student took one of these prerequisite courses, outcome data was used for any subsequent course the student took (i.e., not just those specific subsequent course sections for which I have syllabus data). See Results section for student totals by prerequisite course. Note that the majority of students in our dataset come from CH 301, ECO 304K, and GOV 310L.

Clearance to use student data

Student outcome data and covariate information were obtained from University registrar data. Use of such data for the purposes of this study was approved by the Internal Review Board at UT Austin (#2016070007), and the author of this document was a member of the approved research personnel covered thereby. Conditions of data use were agreed upon by the author and the Project 2021 Leadership team including Jane Huk, the leader of Research and Measurement, Executive Director Jamie Pennebaker, and COO Toni Wegner. Among the conditions agreed upon for clearance to use these data are that I keep specific courses and instructors anonymous and to present results in the aggregate, reporting across department, college, and course sequences that do not specifically identify specific prerequisite course sections.

Student background variables

The student data that were provided to me under terms described above included of measures of previous achievement (standardized grade in prerequisite course, SAT score equivalent, high school GPA rank, number of college credits earned in high school), demographic information (age, sex, race/ethnicity), socioeconomic status (mother's highest level of education), and several other relevant student variables (semester course

load, classification, college major). See Table 7 for variable scales, names, and descriptions. These data play a crucial role in the present study because they form the basis on which treatment and control groups are balanced. By achieving balance on measures of previous (high school) achievement, previous (high school) academic rigor, socio-economic status, demographic information, current (college) achievement, current (college) academic rigor (e.g., semester course load and current major), and year in college, I make the case that all plausible confounding factors affecting both the treatment-assignment mechanism (i.e., student selection of courses) and students' subsequent-course performance are as good as randomly assigned between treatment and control groups. See Results for formal tests of covariate balance for each of these variables before and after propensity-score weighting.

One caveat to be mentioned is that it is rarely advisable to condition on post-treatment variables (e.g., to control for them in a regression model or otherwise use them to equate groups) because they are definitionally measured after, and therefore potentially affected by, the treatment, causing treatment-effect estimates to be biased (Rosenbaum, 1984). In the present study, the treatment is an attribute of the prerequisite course, so any subsequent-course attributes are therefore post-treatment. However, several subsequent-course variables could be potential confounds, ranging from general course-specific effects to the number of credit hours undertaken during the subsequent-course semester. When the post-treatment variable is a surrogate for an unobserved pre-treatment variable, or when the treatment could not plausibly affect the post-treatment variable, the bias will be negligible (Rosenbaum, 1984). Proceeding with due caution, models will be fit and results will be reported with and without post-treatment covariates where appropriate.

Table 7 Variable names, scales, and descriptions

Variable	Name	Scale and description
Outcome measures		
Grade in subsequent course	*class_zscore.y*	Numeric: Standardized GPA (4.0 scale to z-score within each course) in subsequent course
Mean subsequence-course grade per prerequisite course	*meanz.y*	Numeric: Mean standardized GPA (same as above, averaged) in subsequent course
Previous achievement-related variables		
SAT score equivalent	*SAT_equivalent*	Numeric: SAT composite score (or institutionally equated ACT composite score): 800-1600
High school GPA percentile	*hspct2*	Numeric: High school GPA percentile in graduating class
Transfer credit from HS coursework	*transferredhours*	Numeric: Total number of credits transferred to UT from high school
Demographics		
Age	*age*	Numeric: Age in years at start of prerequisite semester
Gender	*sex*	Categorical: Indicator variable for female
Race or Ethnicity	*derivation*	Categorical: Indicator variables for Asian, Black, Hispanic (any), Hawaiian/Pacific Islander, Native American, Unknown, White, 2+ (excluding Black and Hispanic), 2+ (including 1+ Black)
University classification	*CLASSIFICATION*	Categorical: Indicator variables for Freshman, Sophomore, Junior, Senior, and Super-Senior
Socioeconomic status		
Mother's level of education	*motheredlevel*	Categorical: Indicator variables for No high school, Some high school, High school diploma or equivalent, Some college, Associate's degree, Bachelor's or four-year degree, Graduate or professional degree, and Unknown
Father's level of education	*fatheredlevel*	
Current achievement-related variables		
Grade in prerequisite course	*class_zscore.x*	Numeric: Standardized GPA in prerequisite course
Semester course load (credit hours taken)	*HRS_UNDERTAKEN.y*	Numeric: Total number of credit hours undertaken during subsequent-course semester
College major	*admitschool*	Categorical: Indicator variables for each college major (12 colleges)
Course/instructor effects		
Prerequisite course section	*unique_course.x*	Categorical: Indicators for each unique course section (209 sections)
Subsequent course section	*unique_course.y*	Categorical: Indicators for each subsequent course section (429 sections)
Instructor	*instructor_first*	Categorical: Indicators for each instructor (56 instructors)

Covariate balance assessment

To assess whether covariate balance between treatment and control groups has been achieved after propensity-score adjustment, distributions and summary statistics are computed for each covariate and compared between conditions. Graphical depictions of the distribution of each covariate in both conditions are provided both before and after adjustment. Thus, several lines of evidence will converge on a determination of covariate balance. The specific techniques to be used are discussed in more detail below.

Standardized mean differences

The difference in covariate means between treatment and control groups divided by the pooled standard deviation is a traditional indicator of balance appropriate when computing the ATE (Austin, 2009). Intuitively, there should be no mean differences between treatment and control conditions for any covariates when balanced. Commonly, a threshold of 0.1 is used for standardized mean differences in the absence of hypothesis tests. Each covariate will be examined with respect to this threshold.

Logistic regression of treatment on covariates

One possible way to formally assess covariate balance in a hypothesis-testing framework is to conduct a *t* test between treatment and control groups for each covariate. Another way to approach this question is to fit a logistic regression predicting treatment condition from all covariates. If covariates are balanced between conditions, then they should not be significantly predictive of treatment status. Accordingly, logistic regressions of treatment on each covariate are conducted before and after inverse propensity-score

weighting. Note that this procedure is very conservative using a nominal 0.05 significance level, but this is not necessarily problematic from the standpoint of assessing covariate balance.

Kolmogorov–Smirnov (K–S) test statistics

The K–S statistic for a continuous variable is a measure of the largest distance between the empirical cumulative distribution function for that variable between two groups. Effectively, it measures the similarity of two distributions, with a value of 0 representing complete overlap (i.e., identical distributions) and a value of 1 indicating no overlap. Thus, in the present study, K–S values close to zero are indicative of balanced covariate distributions between treatment and control.

Variance ratios

Recent research recommends that variance ratios be used to further examine distributional similarity between conditions for continuous variables (Austin, 2009; Imai, King, & Stuart, 2008). When variances are similar between conditions, the variance ratio will be close to 1. For balanced groups, a rule of thumb is that the variance ratio should be between 0.5 and 2 (e.g., Rubin, 2001; Stuart, 2010).

Treatment variables

The independent variable of principal interest in the present study is the quantity of graded retrieval-practice opportunities offered in the prerequisite course (total number of graded quizzes and exams including the final). In order to use propensity scores to achieve covariate balance, this course total was dichotomized to create a binary treatment variable. Specifically, treatment and control conditions were created using a median split of the

number of graded retrieval practice (as indicated in the prerequisite-course syllabus rubric), within each course sequence. That is, the median number of graded retrieval practice opportunities was determined for each prerequisite course: students taking prerequisite courses at the median or higher were assigned to treatment (high retrieval practice condition), while students taking courses below the median were assigned to control (low retrieval practice condition). This method of operationalizing treatment was decided on *a priori*, but a mean split was used to assess robustness. Note that these cutpoints are arbitrary and somewhat artificial; however, they are necessary when dichotomizing a continuous variable to ensure covariate balance. See Table 9 and Figures 39 and 44 for descriptive statistics and the distribution of graded retrieval practice opportunities by course dichotomized using the median and the mean, respectively. Notice that the mean split may support a more appropriate division between naturally occurring high and low retrieval-practice courses: the distributions produced by such a split are more distinctly separated.

Outcome variables

Two outcome measures of interest were examined in the present study. Of primary interest was the grade (4.0 scale, standardized in each course) earned in the subsequent course of a given course sequence. This was used as the dependent variable in the subsequent-course analysis to assess the impact of high retrieval practice. The second outcome measure was the average subsequent-course grade earned by students for each section of the prerequisite course. This represents the average performance of a prerequisite course section's students in their subsequent course and was used as the dependent variable in the exploratory lasso regression. Ideally a continuous measure of grades would be used, but the university does not keep such records. See Table 7 for all variables and descriptions.

Additional syllabus variables and outcome measures for lasso regression

Finally, all relevant prerequisite-course variables derived from the syllabus analysis were used in an exploratory lasso regression predicting mean subsequent-course performance in order to select the subset of these variables that are most associated with performance in the subsequent course. This data comprises all variables that appear in the correlation matrix presented in Appendix B and visualized in Figure 4, presented in Part I of the study. In addition, fixed effects of prerequisite-course instructor were included in both models to capture extraneous teacher effects beyond retrieval practice.

MODELING

Primary outcome analyses

The three modeling approaches outlined above (fixed effects, random effects, and OLS with cluster-robust standard errors) are estimated and compared for each of the treatment operationalizations (i.e., median-split and mean-split). Models are fit both before and after inverse propensity-score weighting. All significance tests reported are adjusted for unequal variances using the Satterthwaite degrees of freedom corrections (Satterthwaite, 1946).

All models are fit using R (R Core Team, 2018). Figures were made using the *ggplot2* package (Wickham, 2016a) or were manually created. Data manipulation was carried out using base R functions along with helper functions from the *dplyr* (Wickham, Fancois, Henry, & Mueller, 2017) and *tidyr* (Wickham, 2016b) packages. Standard regression models are fit using R base functions. Mixed models are fit using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014). Lasso regression models are fit using the *glmnet* package (Friedman, Hastie, & Tibshirani, 2010). Cluster-robust standard errors

are computed using the *clubSandwich* package (Pustejovsky, 2017). Heteroskedasticity-robust standard errors are computed using the *sandwich* package (Zeileis, 2004). Balance assessment diagnostics are provided in part by functions from the *cobalt* package (Griefer, 2017).

Fixed-effects model

In the first model, each unique subsequent-course section is included as a fixed effect. As discussed above, one cannot include fixed effects for prerequisite courses because the treatment variable of interest occurs at this level. It is possible, however, to control for the specific prerequisite course instructor to capture variability due to otherwise unmeasured prerequisite-course attributes. The fixed-effects model is given below

$$\begin{aligned} postGPA_{ip} = \beta_0 + \beta_1 Treatment_{ip} + \beta_2 SAT_{ip} + \beta_3 hsRank_{ip} + \beta_4 hsCredits_{ip} \\ + \beta_5 Age_{ip} + \beta_6 Sex_{ip} + \beta_7 Ethnicity_{ip} + \beta_8 SES_{ip} + \beta_9 Classification_{ip} \\ + \beta_{10} Major_{ip} + \beta_{11} preGPA_{ip} + \beta_{12} subLoad_{ip} + \tau_p + e_{ip} \end{aligned} \quad (8)$$

Where the τ_p represents individual prerequisite-instructor fixed-effects. The model is presented below in R syntax:

$$\begin{aligned} lm(postGPA \sim\ & Treatment + SAT + hsRank + hsCredits + Age + Sex \\ & + Ethnicity + SES + Classification + Major + preGPA \\ & + subseqLoad + prereqInstructor + subseqSection, weights \\ & = IPW\_ATE) \end{aligned}$$

Note that treatment, sex, ethnicity, SES, classification, major, prerequisite department, prerequisite instructor, and subsequent course section are treated as categorical variables, while the remaining variables are treated as numeric. See Table 7 for variable scales and descriptions.

Mixed-effects model

In the second modeling approach, random effects are included for each prerequisite and subsequent course section. This model is identical to the fixed-effects model shown above in every respect except that it does not include indicators for prerequisite instructor; instead, random effects are included for both prerequisite and subsequent course sections. Aside from this change, all covariates are the same for both models. Note that this is a crossed, rather than a nested, multilevel model. Specifically, for subject i in prerequisite course p and subsequent course s,

$$\begin{aligned} postGPA_{ips} = {} & \gamma_{000} + \gamma_{010} Treatment_p + \gamma_{100} SAT_{ips} + \gamma_{200} hsRank_{ips} \\ & + \gamma_{300} hsCredits_{ips} + \gamma_{400} Age_{ips} + \gamma_{500} Sex_{ips} + \gamma_{600} Ethnicity_{ips} \\ & + \gamma_{700} SES_{ips} + \gamma_{800} Classification_{ips} + \gamma_{900} Major_{ips} + \gamma_{A00} preGPA_{ips} \\ & + \gamma_{B00} subLoad_{ips} + u_{0p0} + u_{00s} + e_{ips} \end{aligned} \quad (9)$$

Where u_{0p0} and u_{00s} are individual prerequisite- and subsequent-course random effects (i.e., modeled as normally distributed with a mean of zero and variance $\sigma^2_{u_{0p0}}$ and $\sigma^2_{u_{00s}}$, respectively). We obtain estimates of the variation between prerequisite courses, $\sigma^2_{u_{0p0}}$, and between subsequent courses, $\sigma^2_{u_{00s}}$, and well as residual variation σ^2_e. The proportion of observed variation attributable to specific prerequisite or subsequent courses can then be computed using the intraclass correlation coefficient (ICC). For example, the proportion of variation in the outcome attributable to specific prerequisite courses is calculated as follows:

$$ICC_p = \frac{\sigma^2_{u_{0p0}}}{\sigma^2_{u_{0p0}} + \sigma^2_{u_{00s}} + \sigma^2_e} \quad (10)$$

This can also be interpreted as the correlation of student outcomes for any prerequisite course, regardless of the subsequent course. Similarly, the proportion of variation attributable to specific subsequent courses is calculated as follows:

$$ICC_s = \frac{\sigma^2_{u_{00s}}}{\sigma^2_{u_{0p0}} + \sigma^2_{u_{00s}} + \sigma^2_e} \quad (11)$$

Likewise, this can be interpreted as the correlation of student outcomes for any subsequent course, regardless of the prerequisite course. Finally, the proportion of variation in the outcome due to both the prerequisite and the subsequent course is calculated as follows:

$$ICC_{ps} = \frac{\sigma^2_{u_{0p0}} + \sigma^2_{u_{00s}}}{\sigma^2_{u_{0p0}} + \sigma^2_{u_{00s}} + \sigma^2_e} \quad (12)$$

This can be interpreted as the correlation of student outcomes for students in the same prerequisite and subsequent course.

This mixed-effects model with random effects for prerequisite and subsequent course section and inverse-propensity weights is presented below in R syntax appropriate for the lmer() function in the lme4 package, with weights (*IPW_ATE*, below) calculated as specified below in the section on propensity score modeling. Note that while this function uses restricted maximum likelihood estimation by default, restricted maximum likelihood and maximum likelihood estimation produce the same estimates for fixed effects, which are of primary interest, and given large samples like those in the present study, differences between random-effects estimates are negligible (e.g., Snijders & Bosker, 1999).

$$\begin{aligned} lmer(postGPA \sim\ & Treatment + SAT + hsRank + hsCredits + Age + Sex \\ & + Ethnicity + SES + Classification + Major + preGPA \\ & + subLoad + (1 \mid preSection) + (1 \mid subSection), \\ & weights = IPW\_ATE) \end{aligned}$$

Regression with cluster-robust standard-errors

Unlike the fixed- and random-effects models above, the third model does not explicitly estimate parameters for the unique effects of specific prerequisite- and subsequent-course sections. Instead, each unique prerequisite–subsequent course-section combination is treated as a unique cluster, and standard errors are computed that are robust to independence violations that arise from the clustered nature of these observations (Cameron & Miller, 2015). There are 1864 unique course-sequence clusters in the dataset with an average cluster size of 7.29 ($SD = 21.67, Min = 1, Max = 412$). Aside from the way unique prerequisite and subsequent courses are treated, all covariates (those shown in Table 7) remain the same across models. The model is identical to the fixed-effects model presented above in Equation 9 except that it lacks instructor fixed-effects. The syntax is presented below with a wrapper function from the *clubSandwich* package that calculates the cluster-robust standard errors. As recommended, the CR2 variance estimator is used with Satterthwaite degrees of freedom (Pustejovsky & Tipton, 2014).

$$coef\_test(\ lm(postGPA \sim Treatment + SAT + hsRank + hsCredits + Age + Sex + Ethnicity + SES + Classification + Major + preGPA + subseqLoad,\quad weights = IPW\_ATE),\quad vcov = ``CR2'', cluster = interaction(prereqSection, subseqSection)$$

Propensity-score model

The propensity-score model is a logistic regression model in which all pre-treatment covariates are used to predict treatment receipt as described above. The model is presented below in R syntax

$$glm(Treatment \sim SAT + hsRank + hsCredits + Sex + Ethnicity + SES + Classification + Major, family = ``binomial'')$$

All covariates are pre-treatment variables. Propensity scores are estimated from the fitted model by inputting specific covariate values for each student and getting their predicted probability as output. Propensity scores $p(X)$ are then used to generate the weights IPW_{ATE} as follows:

$$IPW_{ATE} = T\frac{1}{p(X)} + (1-T)\frac{1}{1-p(X)} \quad (13)$$

Weights are then normalized so that they sum to one within each condition (i.e., sum of weights for treated subjects is equal to one and the sum of weights for control subjects is equal to one).

Secondary models

For each of the three course sequences with the greatest number of students in the dataset, covariate balance is assessed, weights are calculated, models are fit, and treatment effects are estimated as described above within each course sequence. The three course sequences are CH 301–CH 302, ECO 304K–ECO 304L, and GOV 310L–GOV 312L. In addition to being large samples, these course sequences are extremely representative because many of them are required courses for popular majors at UT Austin. Perhaps most importantly, they allow

Lasso regression for variable selection

Lasso regression was performed by regressing the mean subsequent-course grade for each prerequisite course section on all 38 syllabus-derived variables related to course structure, requirements, and teaching practices as described above in order to find the subset of prerequisite course variables most predictive of subsequent-course success. A 10-fold cross-validation was used to choose the regularization parameter λ, which governs the

amount of shrinkage, and the value of λ that minimized the MSE was chosen. However, there were slight variations in the optimal value of lambda upon repeatedly fitting the same model, so I repeated entire the process 1000 times. Specifically, 10-fold cross-validation was used to select the value of λ that resulted in the lowest MSE and parameter estimates were obtained. This process was repeated, each time generating new parameter estimates, resulting in a distribution of estimates for each non-zero parameter. Distributions of each parameter are reported, along with a single set of such parameter estimates. These results are compared to the OLS solution obtained with the same data.

Chapter Eight: Results

RESEARCH QUESTION 1

Overall Causal Effect Estimates of High Graded Retrieval Practice (Median Split)

The creation of a dichotomous treatment variable using a median split of total graded retrieval practice elements per course resulted in a sample of 8,517 students in the high retrieval-practice (treatment) condition and 4,815 students in the low retrieval-practice (control) condition (values greater than the median were assigned to treatment). Weighting the sample for propensity-score adjustment means that certain individuals are overrepresented (or underrepresented) by design, causing the effective sample size to decrease as weights get more extreme. The effective sample size is calculated as the ratio of the squared sum of the weights to the sum of the squared weights; in the present case, this was 8,273 students in the treatment condition and 4,018 in the control condition, for a total of 12,291. See Table 8 for treatment and control sample sizes resulting from median splits and mean splits used to operationalize treatment. See Table 9 for mean, median, and standard deviation of retrieval practice elements overall and for each prerequisite course. See Figure 39 for histograms of number of courses by number of retrieval practice opportunities for each prerequisite course, with colors indicating median-split treatment assignment.

Table 8 Sample size of high and low retrieval practice (RP) conditions by prerequisite course under different treatment operationalizations

Prerequisite course	Median split		Mean split		Total
	Low RP	High RP	Low RP	High RP	
ADV 318J	162	0	65	97	162
AST 301	340	186	452	74	526
CH 301	3636	2615	2345	3906	6251
CH 302	40	18	40	18	58
CH 318M	149	0	11	138	149
CH 320M	196	0	196	0	196
CH 328M	68	0	68	0	68
CS 303E	365	0	19	346	365
ECO 304K	1469	1297	1352	1414	2766
GOV 310L	1840	523	824	1539	2363
GOV 312L	36	25	36	25	61
PSY 301	216	151	216	151	367
Total	8517	4815	5624	7708	13332

Table 9 Descriptive statistics for number of graded retrieval practice elements by prerequisite course

Course	Median	*M*	*SD*
ADV 318J	5	4.57	0.56
AST 301	4	5.5	5.17
CH 301	33	26.2	12.40
CH 302	13	15.2	12.10
CH 318M	4	3.93	0.26
CH 320M	4	4	0
CH 328M	4	4	0
CS 303E	33	32.1	4
ECO 304K	9	8.51	5.15
GOV 310L	10	8.22	3.77
GOV 312L	3	4.93	2.65
PSY 301	5	12	9.38
Overall	11	17.2	13.2

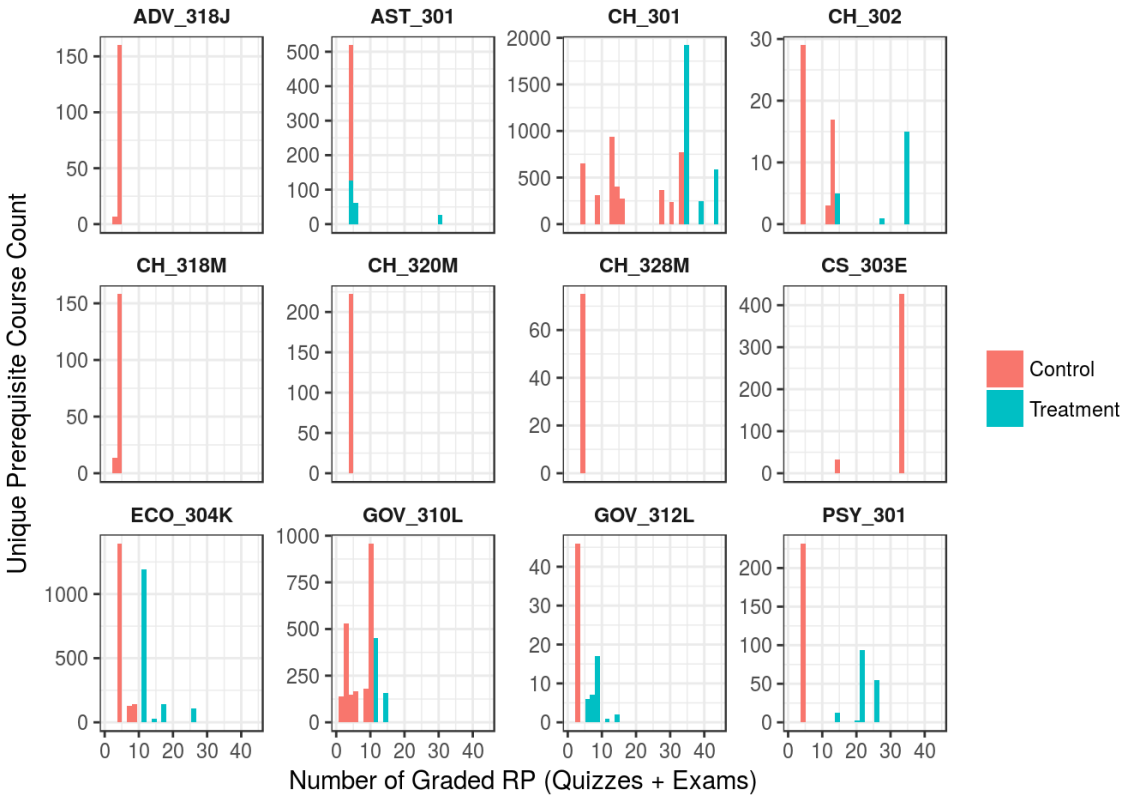

Figure 39 Distribution of graded retrieval practice opportunities by prerequisite course. Color indicates median-split treatment assignment. Note that vertical axis scales differ.

Covariate balance assessment

Prior to adjustment via inverse-propensity weighting, several covariates were unbalanced between treatment and control conditions. Figure 40 depicts standardized mean differences for each variable (or each level for categorical variables) both before (red) and after (blue) adjusting with inverse propensity-score weights. Before weighting, age and first-year classification had a standardized mean difference in excess of the 0.1 threshold

and several more were very near the threshold. For mean differences, variance ratios, and K–S statistics before and after adjustment, see Appendix C.

Given covariate balance, the covariates in question should not predict treatment status. As an additional check, a logistic regression of treatment on covariates is performed before and after weighting (Table 10). It can be seen that, prior to adjustment, treatment and control conditions differ with respect to high school rank, age, classification, and certain levels of ethnicity and major. After adjustment, however, no systematic differences remain between conditions.

To check that the propensity-score weighting is working as intended, distributions of the propensity score (or the logit propensity score) for treatment and control conditions are compared before and after weighting (Figure 41), showing the expected overlap after adjustment. In a similar fashion: distributions of each covariate are shown for each condition before and after adjustment. Densities are shown for continuous variables (Figure 42) and histograms are shown for categorical variables (Figure 43). Altogether, there is ample evidence that covariate balance has been achieved.

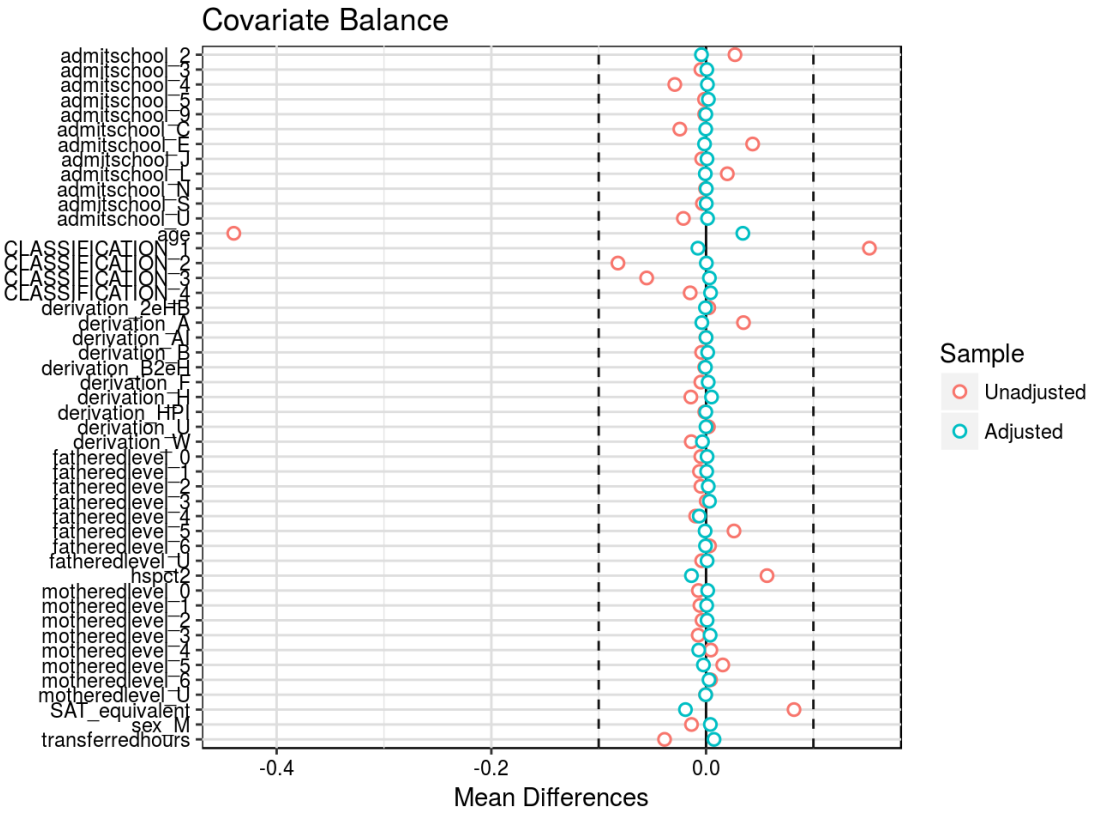

Figure 40 Love plot depicting standardized mean differences (treatment minus control) before and after propensity score adjustment.

Table 10 Logistic regression coefficients predicting treatment status before and after propensity score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	1.804	0.157	11.520	<.001	***	0.492	0.168	2.937	0.003	**
SAT_equivalent	0.000	0.000	1.161	0.246		0.000	0.000	-0.494	0.621	
hspct2	0.088	0.040	2.187	0.029	*	-0.032	0.042	-0.758	0.448	
transferredhours	0.000	0.000	0.177	0.859		0.000	0.000	0.140	0.889	
age	-0.082	0.007	-12.252	<.001	***	0.003	0.007	0.450	0.653	
sexW	-0.012	0.008	-1.377	0.169		-0.003	0.009	-0.335	0.737	
derivationAI	-0.049	0.090	-0.546	0.585		0.005	0.096	0.050	0.960	
derivationA	-0.007	0.024	-0.290	0.772		0.001	0.026	0.042	0.967	
derivationB2eH	-0.069	0.059	-1.180	0.238		-0.026	0.064	-0.409	0.683	
derivationB	-0.039	0.031	-1.235	0.217		0.009	0.033	0.274	0.784	
derivationF	-0.050	0.037	-1.339	0.181		0.027	0.039	0.673	0.501	
derivationHPI	-0.308	0.119	-2.580	0.010	**	-0.043	0.134	-0.318	0.751	
derivationH	-0.017	0.025	-0.670	0.503		0.007	0.026	0.262	0.793	
derivationU	0.089	0.073	1.208	0.227		0.007	0.078	0.087	0.931	
derivationW	-0.021	0.023	-0.895	0.371		0.003	0.025	0.119	0.905	
majorschool3	-0.021	0.034	-0.628	0.530		0.009	0.036	0.245	0.807	
majorschool4	-0.065	0.016	-4.207	0.000	***	0.007	0.017	0.426	0.670	
majorschool5	0.048	0.036	1.315	0.189		0.027	0.038	0.706	0.480	
majorschool9	-0.106	0.115	-0.922	0.356		-0.041	0.127	-0.323	0.747	
majorschoolC	-0.096	0.024	-4.084	0.000	***	0.000	0.025	-0.008	0.994	
majorschoolE	-0.012	0.013	-0.924	0.355		0.006	0.014	0.421	0.674	
majorschoolJ	-0.107	0.040	-2.704	0.007	**	0.023	0.042	0.559	0.576	
majorschoolL	0.028	0.016	1.734	0.083	.	0.001	0.017	0.043	0.966	
majorschoolN	-0.031	0.055	-0.558	0.577		0.013	0.058	0.231	0.818	
majorschoolS	-0.171	0.061	-2.799	0.005	**	0.008	0.065	0.122	0.903	
majorschoolU	-0.030	0.016	-1.922	0.055	.	0.003	0.017	0.164	0.870	
motheredlevel1	0.008	0.035	0.227	0.820		-0.006	0.037	-0.154	0.877	
motheredlevel2	0.032	0.031	1.002	0.316		-0.006	0.033	-0.189	0.850	
motheredlevel3	0.028	0.032	0.863	0.388		-0.001	0.034	-0.031	0.975	
motheredlevel4	0.043	0.031	1.372	0.170		-0.009	0.033	-0.286	0.775	
motheredlevel5	0.045	0.032	1.414	0.157		-0.008	0.034	-0.250	0.803	
motheredlevel6	0.048	0.034	1.402	0.161		0.004	0.036	0.124	0.902	
motheredlevelU	0.073	0.046	1.598	0.110		-0.025	0.048	-0.514	0.607	
fatheredlevel1	-0.026	0.036	-0.721	0.471		0.002	0.038	0.060	0.952	
fatheredlevel2	-0.002	0.032	-0.070	0.944		0.005	0.034	0.137	0.891	
fatheredlevel3	0.001	0.033	0.040	0.968		0.008	0.035	0.240	0.810	
fatheredlevel4	-0.012	0.032	-0.364	0.716		-0.002	0.034	-0.055	0.956	
fatheredlevel5	0.004	0.032	0.139	0.889		0.003	0.034	0.102	0.918	
fatheredlevel6	0.020	0.037	0.544	0.586		-0.005	0.039	-0.122	0.903	
fatheredlevelU	-0.038	0.044	-0.873	0.383		0.015	0.046	0.323	0.746	
CLASSIFICATION2	-0.066	0.011	-5.920	0.000	***	0.000	0.012	-0.036	0.971	
CLASSIFICATION3	-0.127	0.020	-6.356	0.000	***	0.007	0.021	0.333	0.739	
CLASSIFICATION4	-0.047	0.036	-1.299	0.194		0.042	0.038	1.121	0.262	

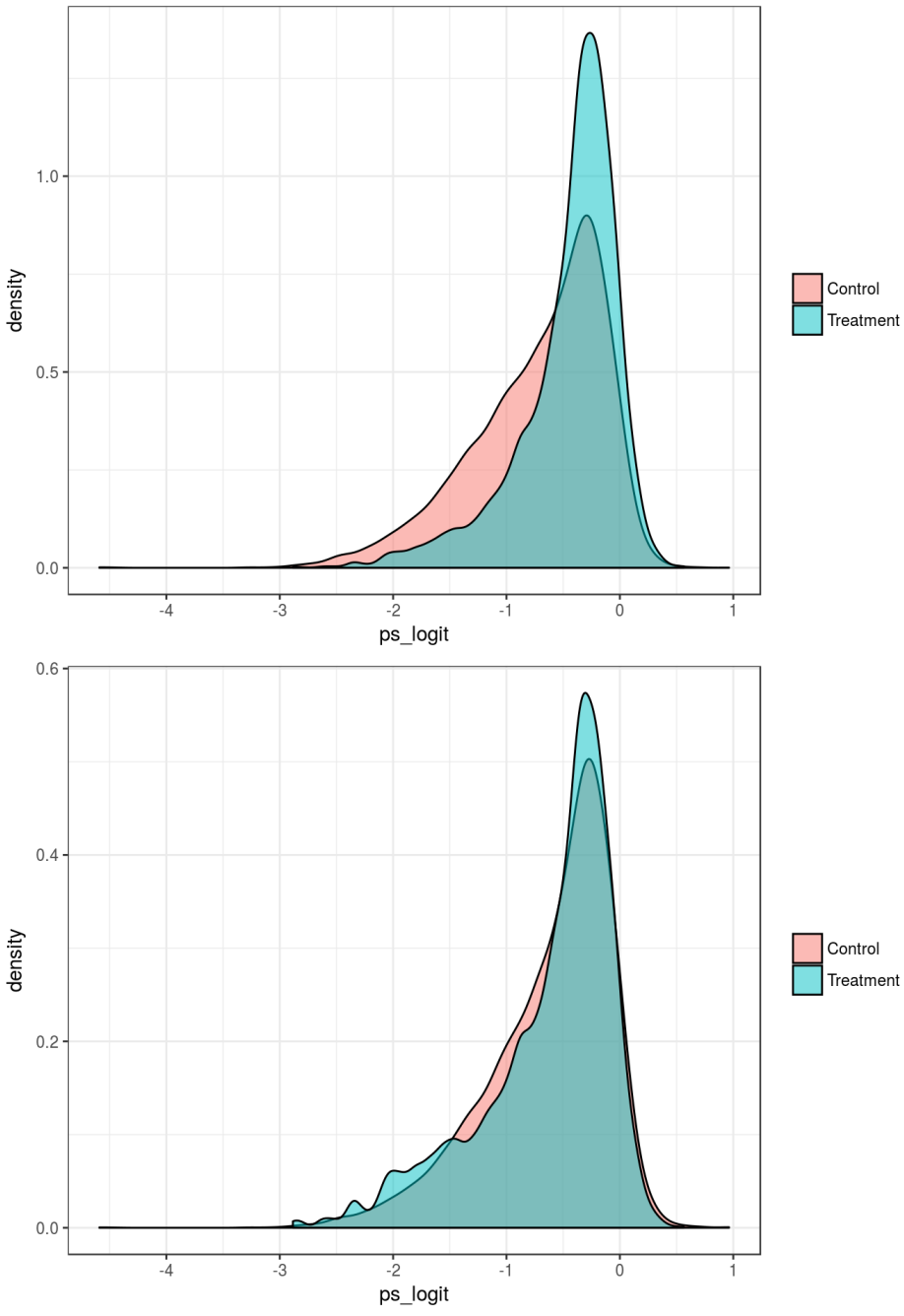

Figure 41 Distribution of propensity scores (logit scale) both before balancing (top panel) and after balancing (bottom panel)

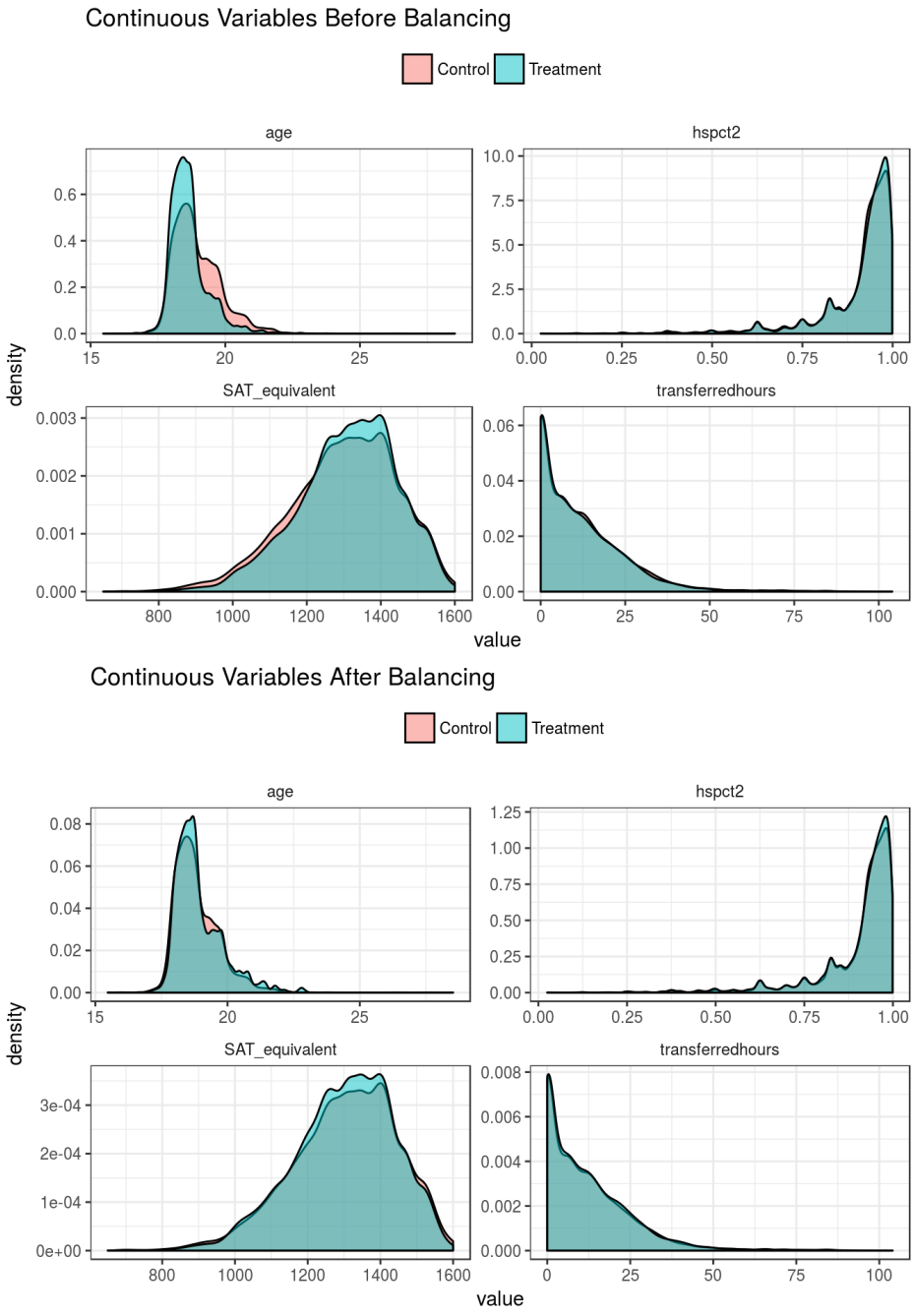

Figure 42 Distributions of continuous covariates before balancing (top panel) and after balancing (bottom panel)

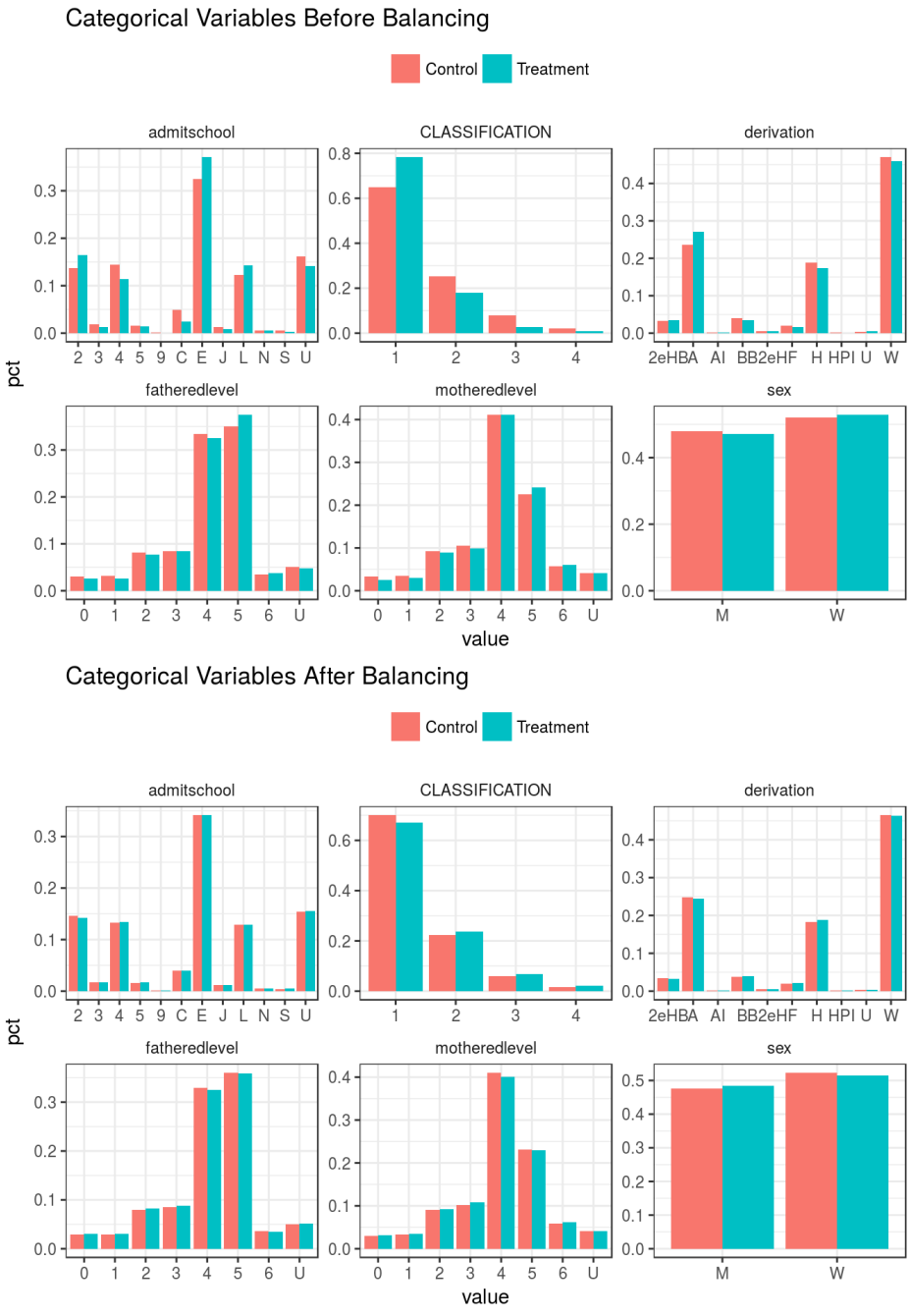

Figure 43 Distributions of categorical covariates before balancing (top panel) and after balancing (bottom panel)

Fixed-effects model

Results are reported for both the unweighted full fixed-effects model (i.e., including main effects of all covariates) and the inverse-propensity weighted full fixed-effects model as described in the Analysis section above (see Table 11 for a summary of treatment-effect estimates across all models). The unweighted average treatment effect estimate for this model was positive and significant, $ATE = 0.0687, SE = 0.0252, t = 2.732, p = 0.006$ The inverse-propensity weighted average treatment effect estimate was positive, similar in magnitude, and significant, $ATE = 0.0603, SE = 0.0246, t = 2.448, p = .014$. Thus, in this model the advantage for high retrieval-practice remained significant after covariate balance was achieved. Students in high retrieval-practice prerequisite courses were found to perform 0.06 standard deviations better in their subsequent course, all else being equal. Full regression output for this model is contained in Appendix D.

Random-effects model

Results are reported for both the unweighted full random-effects model (i.e., including main effects of all covariates) and the inverse-propensity weighted full random-effects model as described in the Analysis section above. The unweighted average treatment effect for this model was positive and significant, $ATE = 0.0569, SE = 0.0319, t = 1.786, p = .045$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, larger in magnitude, and significant, $ATE = 0.0665, SE = 0.0296, t = 2.247, p = .027$. Thus, modeling the relationship with random effects for prerequisite and subsequent courses produced ATE estimates that were similar

in magnitude and provided similar evidence of a treatment effect. Full regression output for this model is contained in Appendix D.

Cluster-robust standard errors model

Results are reported for both the unweighted full cluster-robust standard errors model and the inverse-propensity weighted full cluster-robust standard errors model as described in the Analysis section above. The unweighted average treatment effect for this model was positive but only marginally significant, $ATE = 0.0349, SE = 0.0179, p = 0.056$. The average treatment-effect estimate for the inverse-propensity weighted model was positive and similar in magnitude but only marginally significant, $ATE = 0.0346, SE = 0.0184, p = 0.062$. Notably, models using cluster-robust standard errors but not explicitly estimating individual course effects produced smaller ATE estimates than either of the other modeling approaches. Full regression output for these models is contained in Appendix D.

Table 11 Summary of average treatment effects estimates across all models and treatment operationalizations both before (left) and after (right) propensity-score adjustment

	Treatment	Model	Unadjusted			Adjusted		
			b (ATE)		*SE*	*b* (ATE)		*SE*
Overall								
	Median	Fixed effects	0.069	**	0.025	0.060	*	0.025
		Random effects	0.057	*	0.032	0.067	*	0.030
		Cluster-robust SE	0.036	.	0.018	0.035	.	0.019
	Mean	Fixed effects	0.069	*	0.029	0.067	*	0.029
		Random effects	0.056	.	0.029	0.057	.	0.029
		Cluster-robust SE	0.019		0.017	0.018		0.017
Chemistry								
	Median	Fixed effects	0.091	**	0.031	0.071	*	0.031
		Random effects	0.156	.	0.103	0.151		0.104
		Cluster-robust SE	0.062	*	0.028	0.063	*	0.028
	Mean	Fixed effects	0.123	***	0.035	0.100	**	0.034
		Random effects	0.263	**	0.091	0.241	**	0.086
		Cluster-robust SE	0.057	*	0.028	0.055	.	0.029
Economics								
	Median	Fixed effects	0.074		0.064	0.086		0.064
		Random effects	0.083	.	0.045	0.082	.	0.044
		Cluster-robust SE	0.048		0.037	0.047		0.037
	Mean	Fixed effects	0.159	.	0.094	0.159	.	0.093
		Random effects	0.101	*	0.043	0.103	*	0.045
		Cluster-robust SE	0.070	.	0.038	0.067	.	0.038
Government								
	Median	Fixed effects	0.044		0.051	0.025		0.046
		Random effects	-0.004		0.059	-0.013		0.053
		Cluster-robust SE	-0.055		0.037	-0.072		0.043
	Mean	Fixed effects	-0.044		0.065	-0.026		0.061
		Random effects	-0.057		0.047	-0.051		0.047
		Cluster-robust SE	-0.023		0.044	-0.034		0.047

Overall Causal Effect Estimates of High Graded Retrieval Practice (Mean Split)

Creation of a dichotomous treatment variable using a mean split of total graded retrieval practice elements per course resulted in a sample of 7708 students in the high retrieval-practice (treatment) condition and 5624 students in the low retrieval-practice (control) condition (values greater than the mean were labeled treatment). The effective sample size, after adjusting for covariates, was 7596 students in the treatment condition and 5489 in the control condition. See Table 8 for treatment and control sample sizes resulting from median splits and mean splits used to operationalize treatment. See Table 9 for mean, median, and standard deviation of retrieval practice elements overall and for each prerequisite course. See Figure 44 for histograms of number of courses by number of retrieval practice opportunities for each prerequisite course, with colors indicating mean-split treatment assignment. Compared to the median split, the mean-split treatment assignment appears to do a better job of capturing naturally occurring clusters of high and low retrieval practice courses. Notice, for example, the distinct separation of treatment and control course distributions with respect to treatment assignment for prerequisite courses CH 301 and GOV 310L.

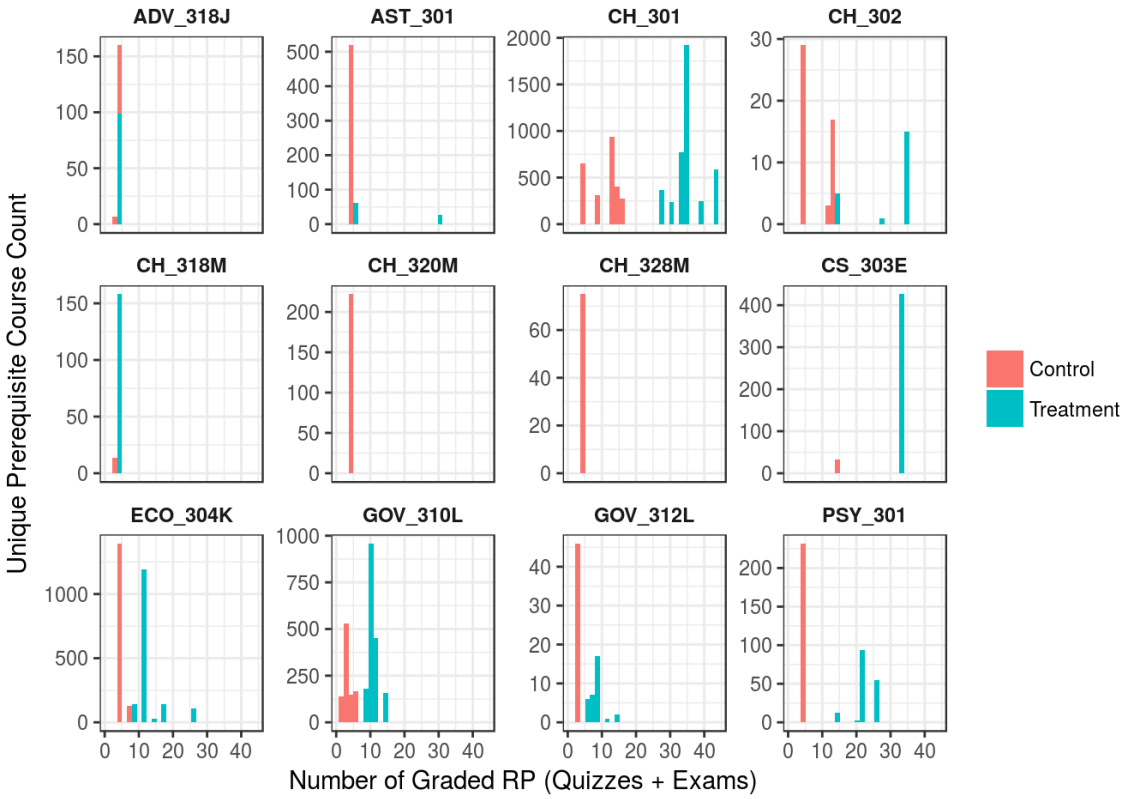

Figure 44 Distribution of graded retrieval practice opportunities by prerequisite course. Color indicates mean-split treatment assignment. Note that vertical axis scales differ.

Covariate balance assessment

Prior to adjustment with inverse-propensity weighting, several covariates were unbalanced between treatment and control conditions. Figure 45 depicts standardized mean differences for each variable (or each level for categorical variables) both before (red) and after (blue) adjusting with inverse propensity-score weights. Before weighting, age had a standardized mean difference in excess of the 0.1 threshold, and variables for SAT, high school rank, high school transfer credits, and certain indicators for ethnicity and major were

in excess of 0.05. Furthermore, adjusted variance ratios are all very close to one and adjusted K–S statistics are all close to zero. For mean differences, variance ratios, and K–S statistics before and after adjustment, see Appendix C.

Given covariate balance, the covariates in question should not predict treatment status. As an additional check, a logistic regression of treatment on covariates is performed before and after weighting (Table 12). It can be seen that, prior to adjustment, treatment and control conditions differ with respect to high school transfer credits, age, classification, mother's education level, and certain levels of ethnicity and major. After adjustment, however, no systematic differences remain between conditions.

To check that the propensity-score weighting has worked as intended, distributions of the propensity score (or the logit propensity score) for treatment and control conditions are compared before and after weighting (Figure 46), showing the expected overlap after adjustment. In a similar fashion: distributions of each covariate are shown for each condition before and after adjustment. Densities are shown for continuous variables (Figure 47) and histograms are shown for categorical variables (Figure 48). Altogether, there is ample evidence that covariate balance has been achieved.

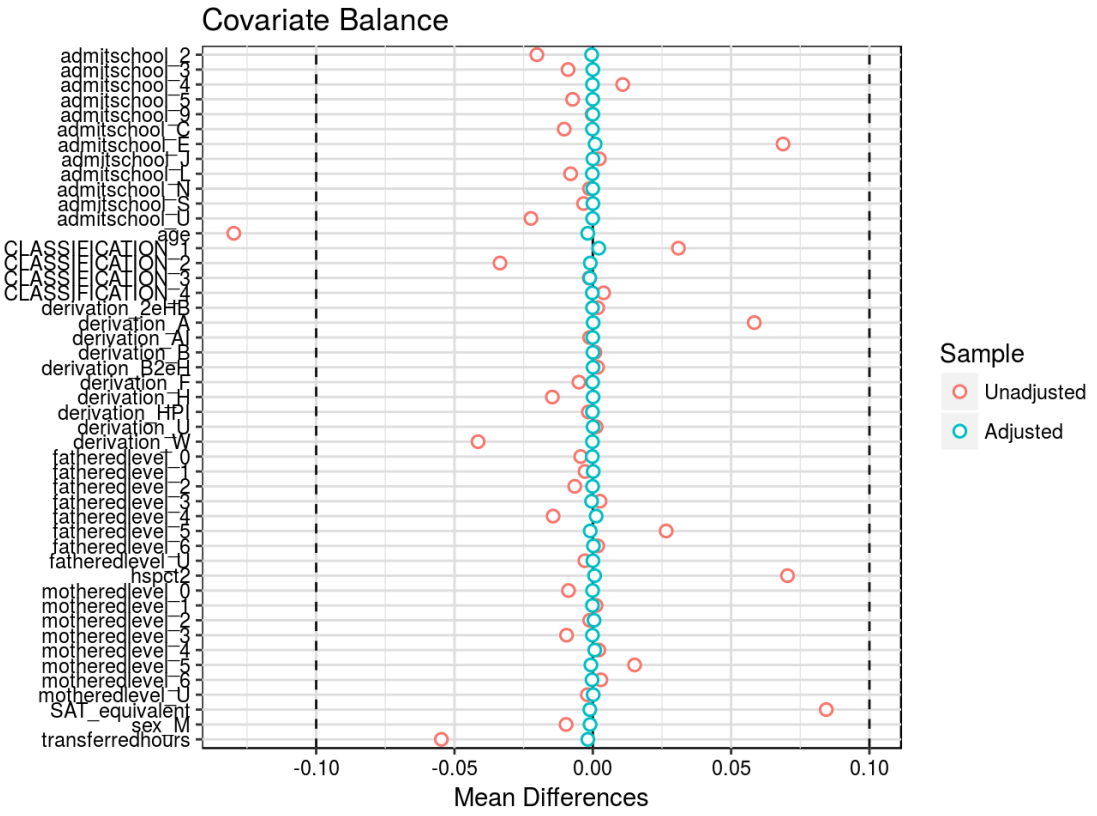

Figure 45 Love plot depicting standardized mean differences (treatment minus control) before and after propensity score adjustment

Table 12 Logistic regression coefficients predicting treatment status before and after propensity score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	1.127	0.164	6.881	0.000	***	0.485	0.166	2.928	0.003	**
SAT_equivalent	0.000	0.000	0.351	0.725		0.000	0.000	-0.025	0.980	
hspct2	0.051	0.042	1.205	0.228		0.001	0.043	0.034	0.973	
transferredhours	-0.001	0.000	-2.984	0.003	**	0.000	0.000	-0.054	0.957	
age	-0.037	0.007	-5.340	0.000	***	0.001	0.007	0.094	0.925	
sexW	0.004	0.009	0.465	0.642		0.001	0.009	0.096	0.923	
derivationAI	-0.133	0.094	-1.413	0.158		-0.002	0.097	-0.024	0.981	
derivationA	0.029	0.025	1.143	0.253		0.001	0.026	0.039	0.969	
derivationB2eH	0.070	0.061	1.135	0.256		0.005	0.063	0.078	0.938	
derivationB	0.005	0.033	0.143	0.886		0.001	0.033	0.028	0.977	
derivationF	-0.050	0.039	-1.276	0.202		0.000	0.040	-0.010	0.992	
derivationHPI	-0.366	0.125	-2.925	0.003	**	-0.029	0.132	-0.221	0.825	
derivationH	-0.012	0.026	-0.445	0.656		0.001	0.026	0.037	0.970	
derivationU	0.058	0.077	0.756	0.450		0.007	0.079	0.085	0.932	
derivationW	-0.025	0.024	-1.030	0.303		0.001	0.025	0.022	0.982	
majorschool3	-0.060	0.036	-1.685	0.092	.	0.001	0.036	0.022	0.982	
majorschool4	0.054	0.016	3.311	0.001	***	0.001	0.017	0.057	0.954	
majorschool5	-0.050	0.038	-1.309	0.191		0.001	0.039	0.017	0.986	
majorschool9	0.017	0.120	0.138	0.890		0.004	0.121	0.030	0.976	
majorschoolC	0.007	0.025	0.291	0.771		0.000	0.025	-0.002	0.998	
majorschoolE	0.077	0.013	5.761	0.000	***	0.001	0.014	0.089	0.929	
majorschoolJ	0.097	0.042	2.325	0.020	*	0.001	0.042	0.034	0.973	
majorschoolL	0.034	0.017	2.049	0.040	*	0.000	0.017	0.021	0.984	
majorschoolN	0.005	0.058	0.080	0.936		0.000	0.059	-0.006	0.996	
majorschoolS	-0.114	0.064	-1.789	0.074	.	0.002	0.065	0.033	0.974	
majorschoolU	0.025	0.016	1.521	0.128		0.000	0.017	0.021	0.984	
motheredlevel1	0.071	0.036	1.959	0.050	.	-0.001	0.037	-0.037	0.970	
motheredlevel2	0.069	0.033	2.098	0.036	*	0.001	0.034	0.029	0.977	
motheredlevel3	0.050	0.034	1.476	0.140		0.000	0.034	-0.012	0.990	
motheredlevel4	0.079	0.033	2.405	0.016	*	0.000	0.033	0.003	0.997	
motheredlevel5	0.079	0.033	2.359	0.018	*	-0.001	0.034	-0.022	0.983	
motheredlevel6	0.076	0.036	2.128	0.033	*	-0.002	0.037	-0.044	0.965	
motheredlevelU	0.065	0.048	1.356	0.175		0.002	0.049	0.046	0.963	
fatheredlevel1	-0.024	0.038	-0.634	0.526		0.003	0.039	0.079	0.937	
fatheredlevel2	-0.013	0.034	-0.395	0.693		0.001	0.034	0.035	0.972	
fatheredlevel3	0.008	0.035	0.231	0.817		0.000	0.035	0.013	0.990	
fatheredlevel4	-0.015	0.033	-0.458	0.647		0.003	0.034	0.084	0.933	
fatheredlevel5	0.001	0.034	0.036	0.971		0.002	0.034	0.046	0.963	
fatheredlevel6	0.005	0.039	0.140	0.889		0.003	0.039	0.087	0.930	
fatheredlevelU	-0.012	0.046	-0.267	0.789		0.000	0.047	0.000	1.000	
CLASSIFICATION2	-0.003	0.012	-0.218	0.828		-0.002	0.012	-0.179	0.858	
CLASSIFICATION3	0.043	0.021	2.039	0.041	*	-0.006	0.021	-0.268	0.789	
CLASSIFICATION4	0.140	0.038	3.713	0.000	***	-0.004	0.038	-0.113	0.910	

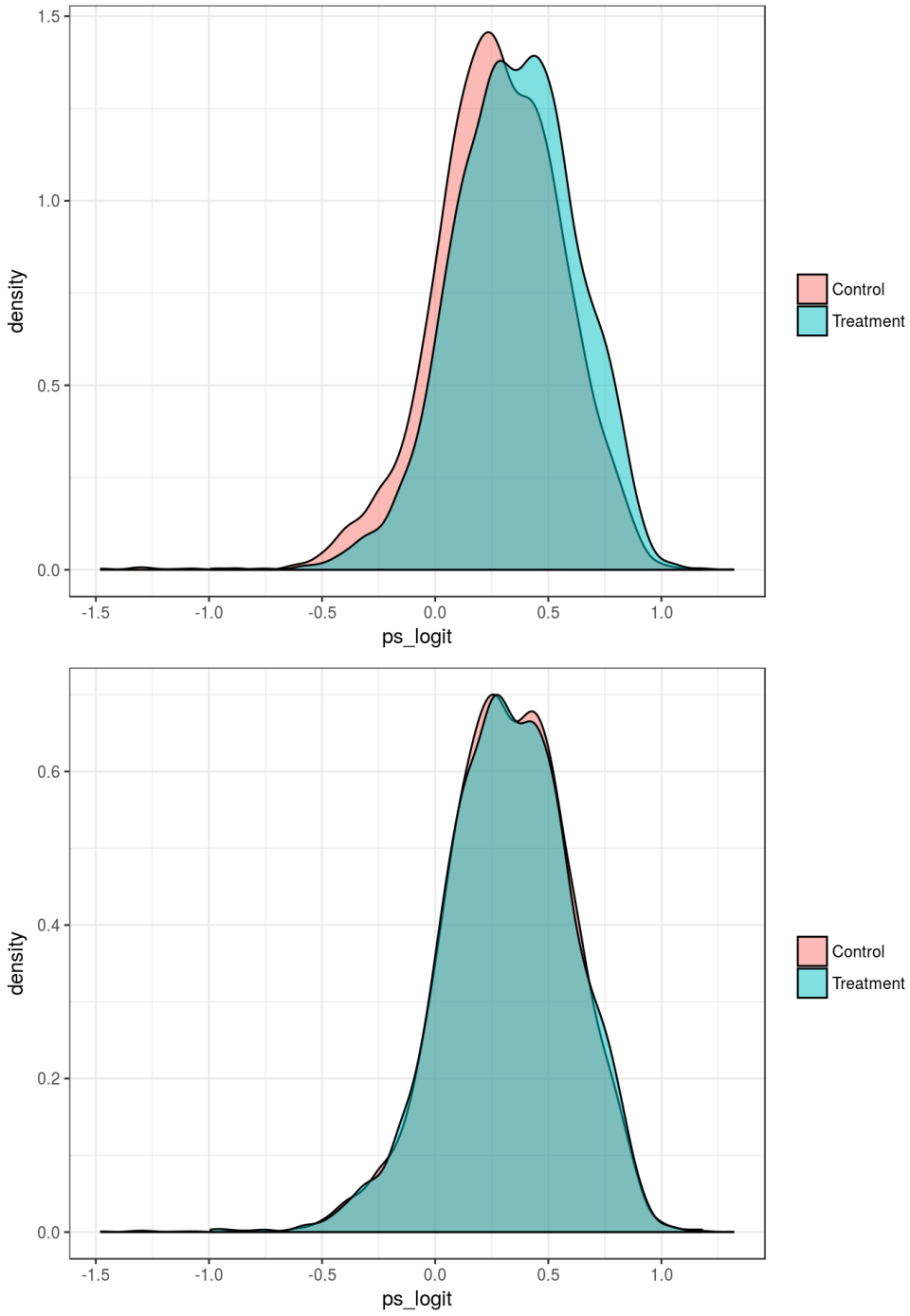

Figure 46 Distribution of propensity scores (logit scale) both before balancing (top panel) and after balancing (bottom panel)

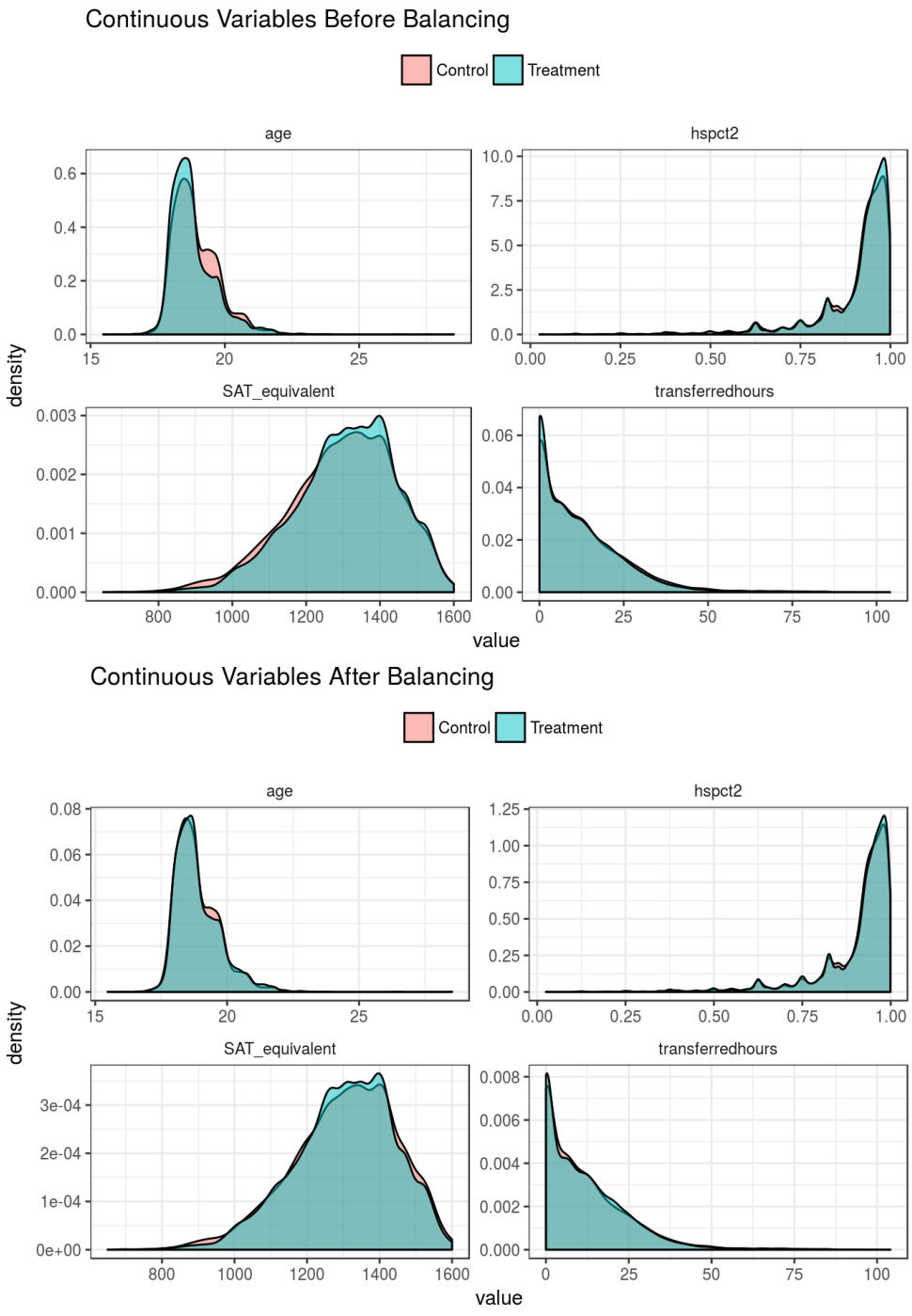

Figure 47 Distributions of continuous covariates before balancing (top panel) and after balancing (bottom panel)

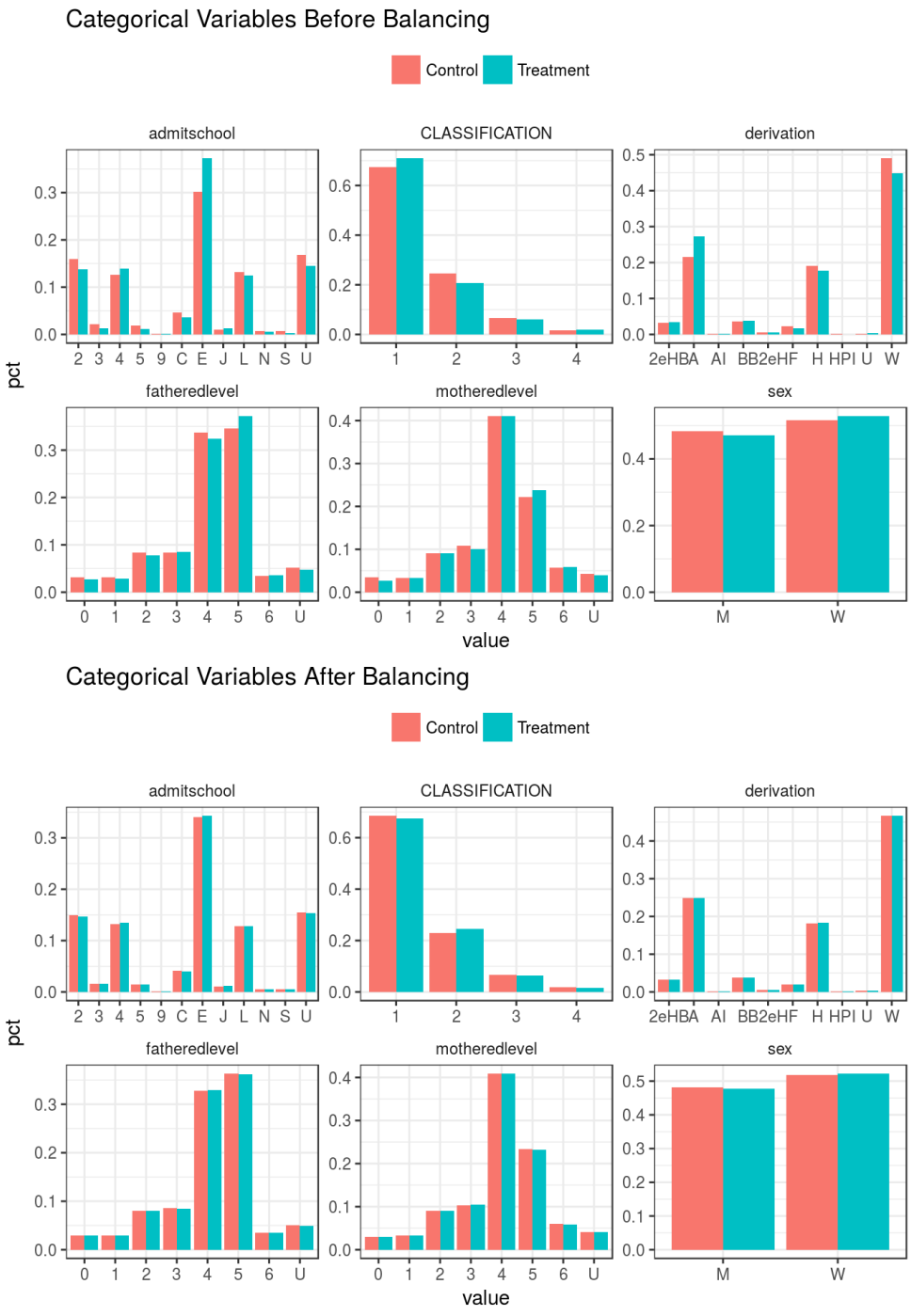

Figure 48 Distributions of categorical covariates before balancing (top panel) and after balancing (bottom panel).

Fixed-effects model

The unweighted average treatment effect estimate for this model was positive and significant, $ATE = 0.0690, SE = 0.0288, t = 2.393, p = .0167$. The inverse-propensity weighted average treatment effect estimate was positive, similar in magnitude, and significant, $ATE = 0.0673, SE = 0.0286, t = 2.354,\ p = .0187$. Thus, in this model the advantage for high retrieval-practice remained significant after covariate balance was achieved. Students in high retrieval-practice prerequisite courses were found to perform approximately 0.07 standard deviations better in their subsequent course, all else being equal. Full regression output for these models is contained in Appendix D. Notably, the size of the estimated effect was approximately the same as when operationalizing treatment using a median split.

Random-effects model

The unweighted average treatment effect for this model was positive but only marginally significant, $ATE = 0.0558, SE = 0.0286, t = 1.947, p = .0541$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, somewhat larger in magnitude, and still only marginally, $ATE = 0.0570, SE = 0.0293, t = 1.943, p = .0545$. Thus, modeling the relationship with random effects for prerequisite and subsequent courses produced somewhat smaller ATE estimates and provided somewhat less evidence for a treatment effect. Full regression output for these models is contained in Appendix D.

Cluster-robust standard errors model

The unweighted average treatment effect for this model was positive but not significant, $ATE = 0.0187, SE = 0.0171, p = 0.2732$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, similar in magnitude, but not significant, $ATE = 0.0178, SE = 0.0172, p = 0.3016$. Again, models using cluster-robust standard errors but not explicitly estimating individual course effects produced smaller ATE estimates than either of the other modeling approaches. Full regression output for these models is contained in Appendix D.

Causal Effect Estimates of High Graded Retrieval Practice (Median Split) For Chemistry

It is clear from Tables 7 and 8 (Figures 39 and 44) that different prerequisite courses differ greatly in the number of graded retrieval practice (RP) opportunities they tend to incorporate. For example, in terms of central tendency the median number of graded RP opportunities in CH 301 is 33 ($M = 26.20, SD = 12.4$), while the median in ECO 304K is 9 ($M = 8.51, SD = 3.77$) and the median in AST 301 is 4 ($M = 5.5, SD = 5.17$). Furthermore, the course sequences under consideration span a range of subjects, the nature of which may be differentially amenable to graded RP opportunities. For example, the second semester of an introductory chemistry course may make more direct use of material learned in the first semester, allowing additional retrieval practice in the first-semester course to have a greater impact on second-semester outcomes. To explore this hypothesis, the three prerequisite courses with the greatest number of students are selected—Principles of Chemistry I (CH 301; $n = 6251$), Introduction to Microeconomics (ECO 304K; $n = 2766$), and American Government (GOV 310L; $n = 2363$)—and the analyses are

repeated for each in turn: propensity scores weights are recalculated and renormalized within each condition, covariate balance assessment is performed over again, and the same models are fit and interpreted.

The creation of a dichotomous treatment variable using a median split of total graded retrieval practice elements ($Med = 33$) resulted in a sample of 2615 students in the high retrieval-practice (treatment) condition and 3636 students in the low retrieval-practice (control) condition (values greater than the median were assigned to treatment). The effective sample size, after adjusting for covariates, was 2474 students in the treatment condition and 3545 in the control condition. See Table 9 for mean, median, and standard deviation of retrieval practice elements.

Covariate balance assessment

As before, several covariates were unbalanced between treatment and control conditions prior to adjustment via inverse-propensity weighting. Figure 49 depicts standardized mean differences for each variable (or each level for categorical variables) both before and after adjustment using inverse propensity-score weights. Before weighting, age, transferred hours, and SAT score had a standardized mean difference in excess of the 0.1 threshold and several more were very near the threshold. For mean differences, variance ratios, and K–S statistics before and after adjustment, see Appendix C. Notice that in each case, variance ratios are closer to unity and K–S statistics are closer to zero after adjustment.

A logistic regression of treatment on covariates was performed before and after weighting (Table 13). Prior to adjustment, treatment and control conditions differed with respect to SAT scores, high school rank, transferred hours, age, classification, and certain

levels of ethnicity, SES, and major. After adjustment, however, no systematic differences remain between conditions.

To check that the propensity-score weighting is working as intended, distributions of the propensity score (or the logit propensity score) for treatment and control conditions are visually compared before and after weighting (Figure 50), showing the expected overlap after adjustment. In a similar fashion, distributions of each covariate are shown for each condition before and after adjustment. Densities are shown for continuous variables (Figure 51) and histograms are shown for categorical variables (Figure 52). Altogether, there is ample evidence that covariate balance has been achieved.

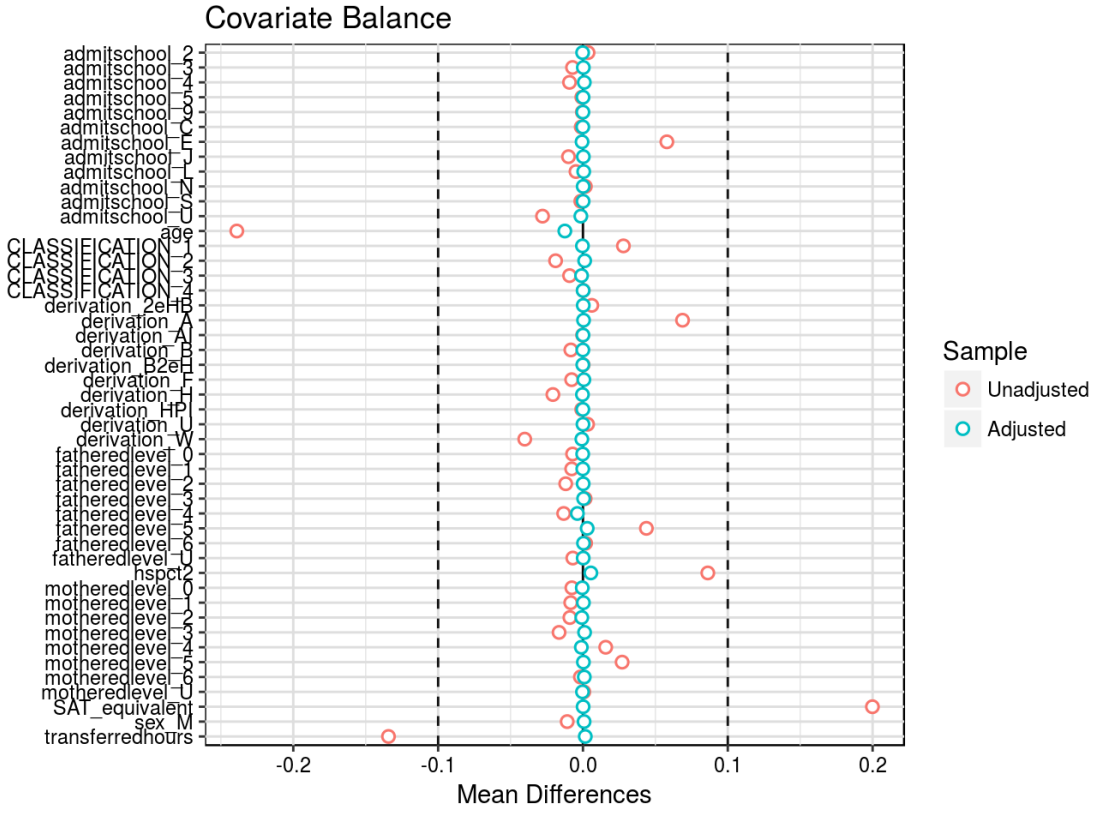

Figure 49 Love plot depicting standardized mean differences (treatment minus control) before and after propensity score adjustment

Table 13 Logistic regression coefficients predicting treatment status before and after propensity score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	1.306	0.268	4.869	0.000	***	0.608	0.281	2.169	0.030	*
SAT_equivalent	0.000	0.000	4.294	0.000	***	0.000	0.000	-0.090	0.929	
hspct2	0.291	0.084	3.456	0.001	***	0.017	0.087	0.199	0.843	
transferredhours	-0.002	0.001	-3.476	0.001	***	0.000	0.001	0.022	0.982	
age	-0.079	0.012	-6.872	0.000	***	-0.007	0.012	-0.544	0.586	
sexW	0.002	0.013	0.184	0.854		-0.002	0.013	-0.132	0.895	
derivationAI	-0.081	0.175	-0.460	0.646		0.002	0.180	0.013	0.989	
derivationA	-0.002	0.035	-0.044	0.965		-0.002	0.035	-0.058	0.954	
derivationB2eH	-0.017	0.091	-0.189	0.850		-0.007	0.095	-0.076	0.940	
derivationB	-0.072	0.047	-1.546	0.122		-0.004	0.048	-0.082	0.935	
derivationF	-0.115	0.057	-2.021	0.043	*	0.008	0.058	0.144	0.885	
derivationHPI	-0.254	0.175	-1.448	0.148		-0.015	0.183	-0.079	0.937	
derivationH	-0.029	0.036	-0.789	0.430		-0.002	0.037	-0.061	0.952	
derivationU	0.107	0.096	1.107	0.268		0.000	0.100	0.004	0.997	
derivationW	-0.060	0.034	-1.765	0.078	.	-0.002	0.035	-0.049	0.961	
majorschool3	0.005	0.059	0.076	0.939		0.010	0.061	0.157	0.876	
majorschool4	-0.039	0.040	-0.968	0.333		0.003	0.042	0.080	0.936	
majorschool5	0.035	0.092	0.385	0.700		0.010	0.095	0.107	0.915	
majorschool9	-0.055	0.249	-0.223	0.823		0.010	0.255	0.039	0.969	
majorschoolC	0.029	0.082	0.354	0.723		0.004	0.085	0.047	0.962	
majorschoolE	0.008	0.038	0.203	0.839		0.002	0.040	0.043	0.966	
majorschoolJ	-0.140	0.060	-2.322	0.020	*	0.007	0.062	0.111	0.912	
majorschoolL	-0.004	0.044	-0.082	0.935		0.005	0.046	0.112	0.911	
majorschoolN	0.089	0.089	1.006	0.315		0.010	0.091	0.115	0.909	
majorschoolS	-0.068	0.132	-0.517	0.606		0.018	0.134	0.137	0.891	
majorschoolU	-0.003	0.042	-0.073	0.942		0.001	0.043	0.013	0.990	
motheredlevel1	-0.009	0.051	-0.180	0.857		0.006	0.053	0.122	0.903	
motheredlevel2	0.019	0.047	0.397	0.692		0.002	0.049	0.046	0.964	
motheredlevel3	0.001	0.048	0.029	0.977		0.007	0.049	0.144	0.886	
motheredlevel4	0.049	0.046	1.049	0.294		0.004	0.048	0.091	0.928	
motheredlevel5	0.050	0.047	1.046	0.296		0.004	0.049	0.088	0.930	
motheredlevel6	0.029	0.051	0.574	0.566		0.008	0.052	0.157	0.875	
motheredlevelU	0.136	0.068	1.994	0.046	*	-0.004	0.070	-0.060	0.952	
fatheredlevel1	-0.018	0.054	-0.336	0.737		-0.002	0.056	-0.032	0.974	
fatheredlevel2	0.000	0.049	0.006	0.995		0.000	0.050	0.008	0.994	
fatheredlevel3	0.023	0.050	0.471	0.638		0.000	0.051	0.002	0.999	
fatheredlevel4	-0.008	0.048	-0.174	0.861		-0.003	0.049	-0.069	0.945	
fatheredlevel5	0.009	0.048	0.189	0.850		0.002	0.050	0.038	0.970	
fatheredlevel6	0.020	0.054	0.359	0.720		0.002	0.056	0.029	0.977	
fatheredlevelU	-0.098	0.066	-1.474	0.140		0.007	0.068	0.099	0.921	
CLASSIFICATION2	0.019	0.019	0.998	0.318		0.005	0.020	0.242	0.809	
CLASSIFICATION3	0.039	0.047	0.840	0.401		-0.005	0.049	-0.101	0.919	
CLASSIFICATION4	0.268	0.127	2.102	0.036	*	0.028	0.127	0.221	0.825	

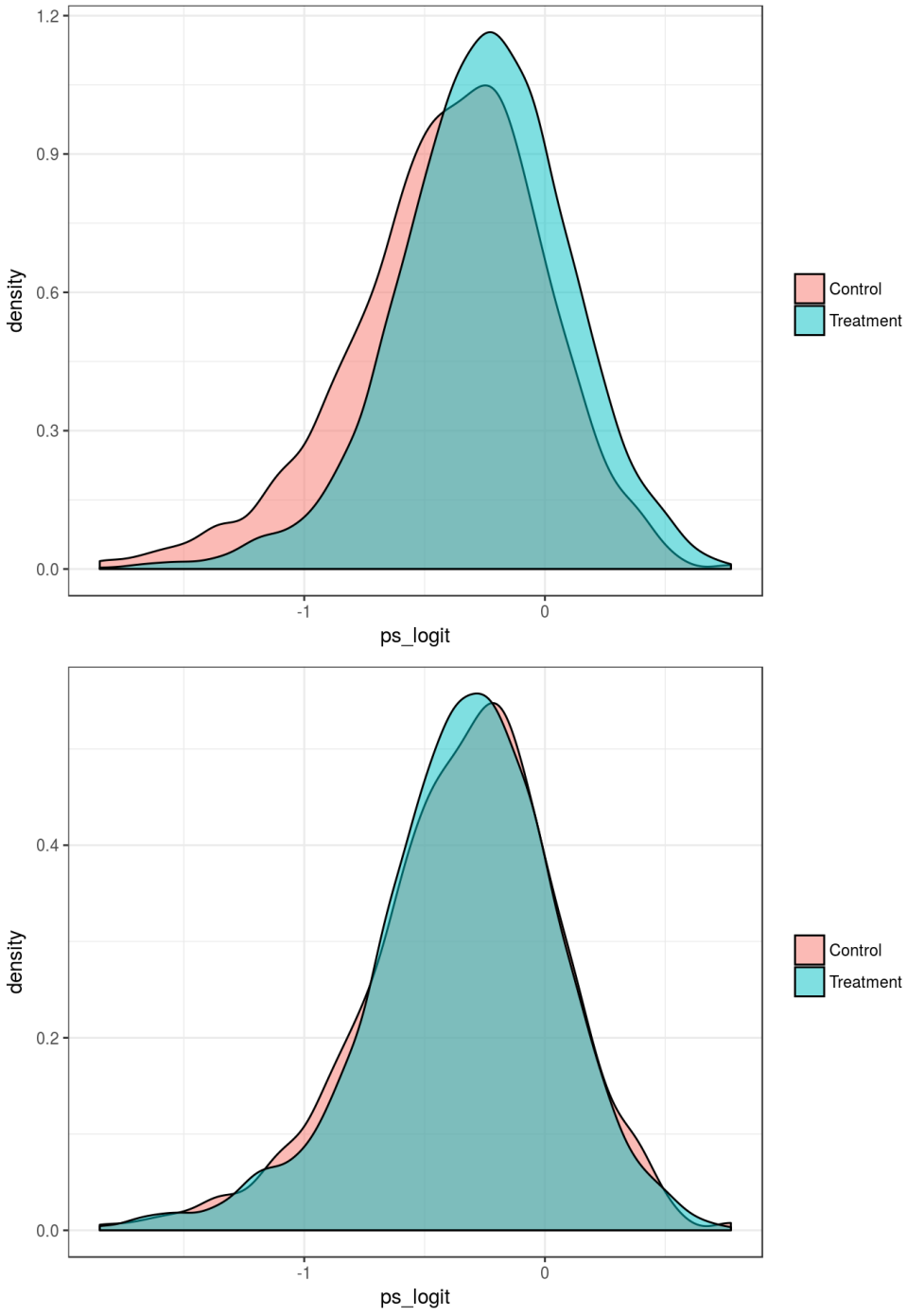

Figure 50 Distribution of propensity scores (logit scale) both before balancing (top panel) and after balancing (bottom panel)

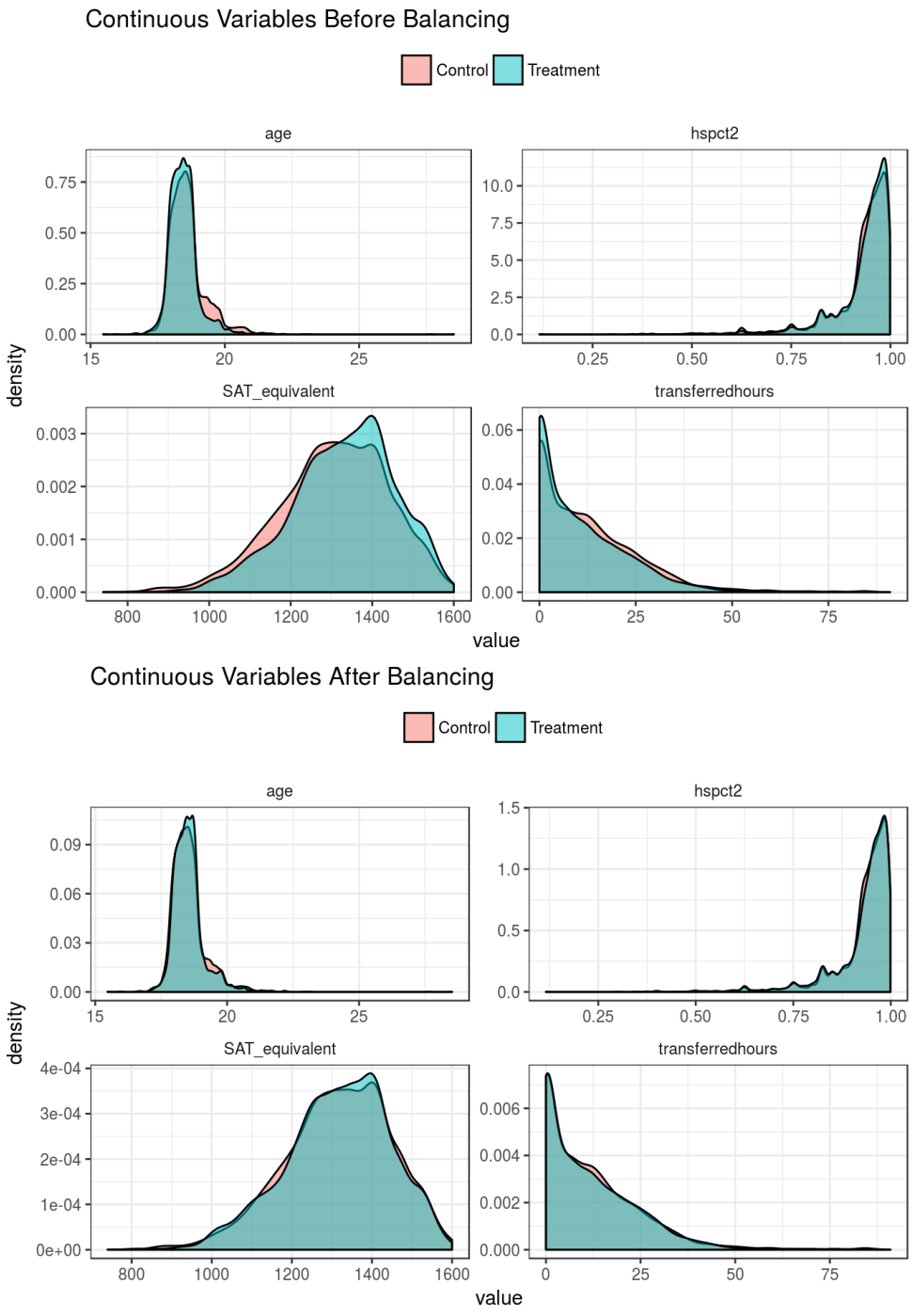

Figure 51 Distributions of continuous covariates before balancing (top panel) and after balancing (bottom panel)

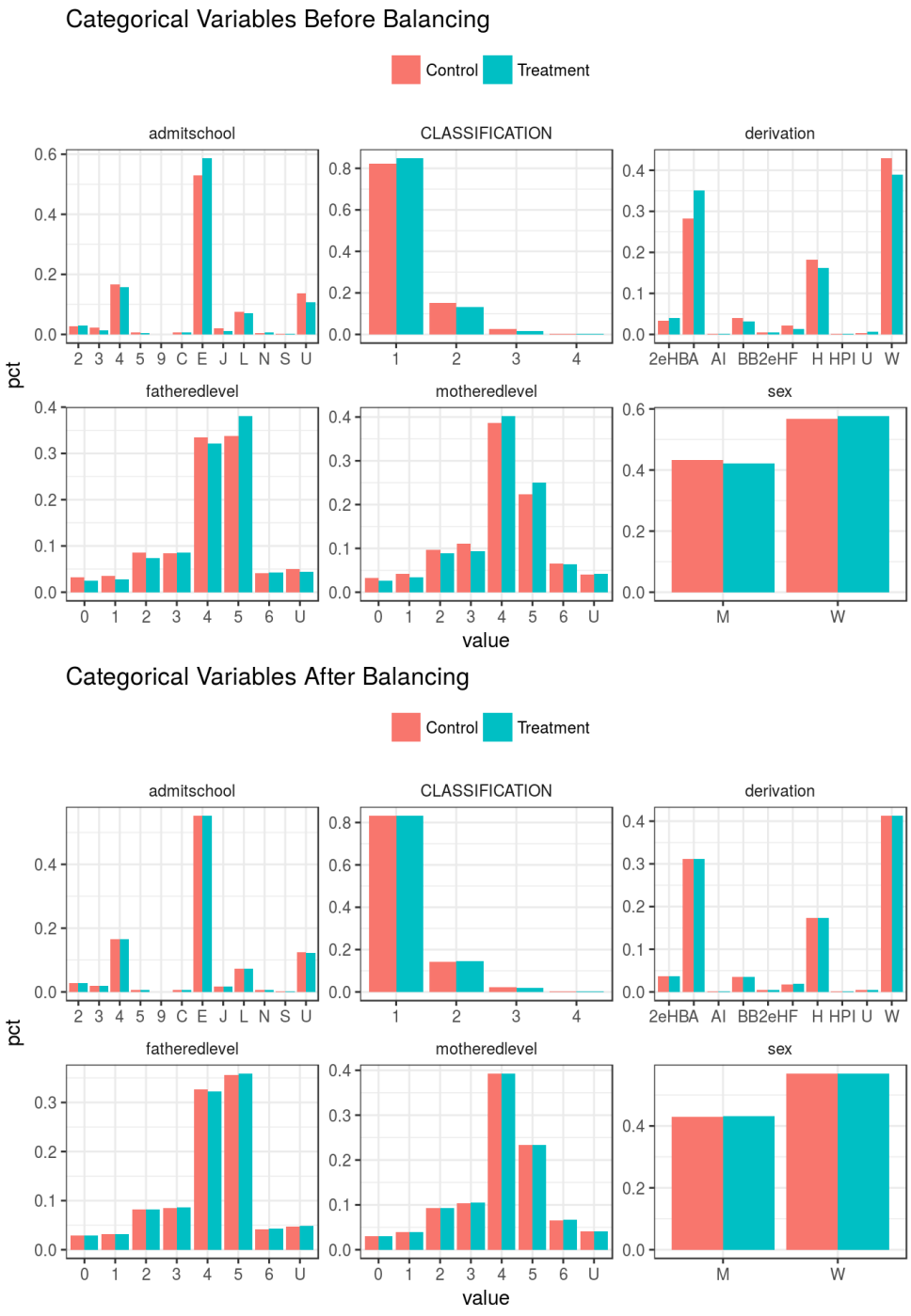

Figure 52 Distributions of categorical covariates before balancing (top panel) and after balancing (bottom panel).

Fixed-effects model

The unweighted average treatment effect estimate for this model was positive and significant, $ATE = 0.091, SE = 0.031, t = 2.914, p = .004$. The inverse-propensity weighted average treatment effect estimate was positive, similar in magnitude, and significant, $ATE = 0.071, SE = 0.031, t = 2.320,\ p = .020$. Thus, in this model the advantage for high retrieval-practice remained significant after covariate balance was achieved. Students in high retrieval-practice prerequisite courses were found to perform approximately 0.07 standard deviations better in their subsequent course, all else being equal. Full regression output for these models is contained in Appendix D. Notably, the size of the estimated effect was approximately the same as when operationalizing treatment using a median split.

Random-effects model

The unweighted average treatment effect for this model was positive and considerably larger than before, but only marginally significant, $ATE = 0.156, SE = 0.103, t = 1.507, p = .077$. The average treatment-effect estimate for the inverse-propensity weighted model was positive but not significant, $ATE = 0.151, SE = 0.104, t = 1.450, p = .159$. Thus, modeling the relationship with random effects for prerequisite and subsequent courses produced larger ATE estimates but provided somewhat less evidence for a treatment effect. Full regression output for these models is contained in Appendix D.

Cluster-robust standard errors model

The unweighted average treatment effect for this model was positive and significant, $ATE = 0.062, SE = 0.028, p = 0.032$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, similar in magnitude, and also significant, $ATE = 0.063, SE = 0.028, p = 0.027$. Here, models using cluster-robust standard errors but not explicitly estimating individual course effects produced smaller ATE estimates than either of the other modeling approaches. Full regression output for these models is contained in Appendix D.

Causal Effect Estimates of High Graded Retrieval Practice (Mean Split) For Chemistry

The creation of a dichotomous treatment variable using a mean split of total graded retrieval practice elements ($M = 26.2$) resulted in a sample of 3906 students in the high retrieval-practice (treatment) condition and 2345 students in the low retrieval-practice (control) condition (values greater than the mean were assigned to treatment). The effective sample size, after adjusting for covariates, was 3706 students in the treatment condition and 2130 in the control condition. See Table 9 for mean, median, and standard deviation of retrieval practice elements.

Covariate balance assessment

Several covariates were unbalanced between treatment and control conditions prior to adjustment via inverse-propensity weighting. Figure 53 depicts standardized mean differences for each variable (or each level for categorical variables) both before and after adjustment using inverse propensity-score weights. Before weighting, age, high school rank, transferred hours, and SAT score had a standardized mean difference in excess of the

0.1 threshold, with several other variables near the threshold. For mean differences, variance ratios, and K–S statistics before and after adjustment, see Appendix C. Notice that again, in each case, variance ratios are closer to unity and K–S statistics are closer to zero after adjustment.

A logistic regression of treatment on covariates was performed before and after weighting (Table 14). Prior to adjustment, treatment and control conditions differed with respect to SAT scores, high school rank, transferred hours, age, and certain levels of ethnicity and classification. After adjustment, however, no systematic differences remain between conditions.

To check that the propensity-score weighting is working as intended, distributions of the propensity score (or the logit propensity score) for treatment and control conditions are visually compared before and after weighting (Figure 54), showing the expected overlap after adjustment. In a similar fashion, distributions of each covariate are shown for each condition before and after adjustment. Densities are shown for continuous variables (Figure 55) and histograms are shown for categorical variables (Figure 56). Altogether, there is ample evidence that covariate balance has been achieved.

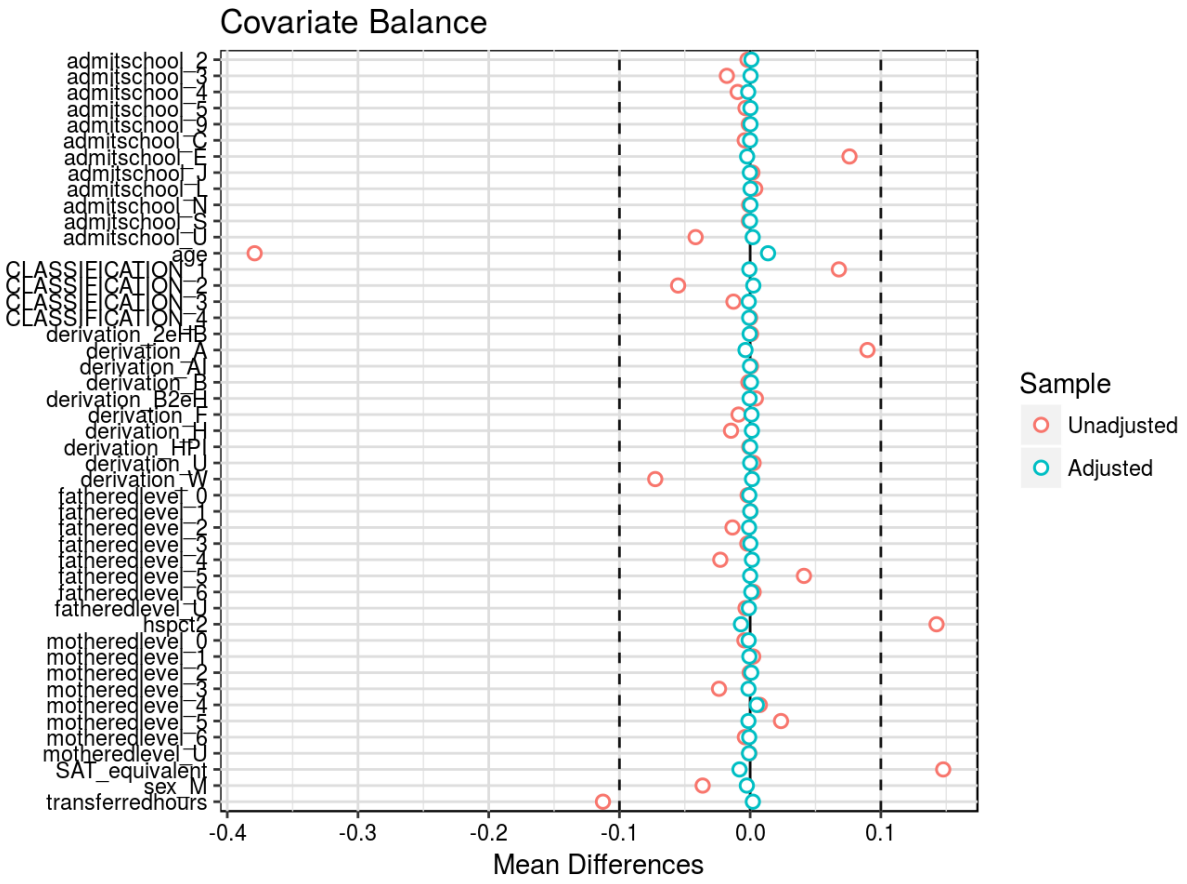

Figure 53 Love plot depicting standardized mean differences (treatment minus control) before and after propensity score adjustment.

Table 14 Logistic regression coefficients predicting treatment status before and after propensity score adjustment

Variable	Unadjusted					Adjusted			
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value
Intercept	2.317	0.260	8.904	<.001	***	0.401	0.266	1.510	0.131
SAT_equivalent	0.000	0.000	2.731	0.006	**	0.000	0.000	-0.194	0.846
hspct2	0.375	0.082	4.588	0.000	***	-0.016	0.087	-0.188	0.851
transferredhours	-0.001	0.001	-2.486	0.013	*	0.000	0.001	0.082	0.935
age	-0.123	0.011	-10.957	<.001	***	0.007	0.011	0.581	0.561
sexW	0.017	0.013	1.316	0.188		0.003	0.013	0.233	0.816
derivationAI	0.132	0.170	0.778	0.437		-0.018	0.177	-0.103	0.918
derivationA	0.035	0.034	1.050	0.294		0.002	0.035	0.047	0.963
derivationB2eH	0.202	0.088	2.290	0.022	*	-0.015	0.092	-0.164	0.870
derivationB	-0.005	0.045	-0.111	0.911		0.007	0.048	0.139	0.890
derivationF	-0.090	0.055	-1.631	0.103		0.017	0.057	0.297	0.766
derivationHPI	-0.198	0.170	-1.165	0.244		0.006	0.184	0.030	0.976
derivationH	0.005	0.035	0.135	0.893		0.006	0.037	0.152	0.879
derivationU	0.105	0.093	1.123	0.261		0.006	0.100	0.058	0.954
derivationW	-0.038	0.033	-1.166	0.244		0.003	0.035	0.077	0.938
majorschool3	-0.058	0.057	-1.017	0.309		-0.013	0.062	-0.208	0.835
majorschool4	0.001	0.039	0.023	0.981		-0.012	0.042	-0.283	0.777
majorschool5	0.004	0.089	0.049	0.961		-0.009	0.096	-0.089	0.929
majorschool9	-0.144	0.241	-0.598	0.550		0.069	0.232	0.297	0.766
majorschoolC	0.024	0.080	0.297	0.767		-0.019	0.085	-0.223	0.823
majorschoolE	0.040	0.037	1.080	0.280		-0.011	0.040	-0.271	0.787
majorschoolJ	0.064	0.058	1.101	0.271		-0.014	0.062	-0.226	0.821
majorschoolL	0.068	0.043	1.604	0.109		-0.011	0.046	-0.236	0.813
majorschoolN	0.006	0.086	0.068	0.945		-0.004	0.092	-0.045	0.964
majorschoolS	0.045	0.128	0.348	0.728		-0.025	0.138	-0.184	0.854
majorschoolU	0.005	0.041	0.125	0.900		-0.009	0.044	-0.214	0.830
motheredlevel1	0.025	0.049	0.498	0.618		0.005	0.052	0.100	0.921
motheredlevel2	0.033	0.046	0.732	0.464		0.013	0.048	0.277	0.782
motheredlevel3	-0.016	0.046	-0.353	0.724		0.008	0.048	0.156	0.876
motheredlevel4	0.045	0.045	1.003	0.316		0.014	0.047	0.296	0.768
motheredlevel5	0.043	0.046	0.935	0.350		0.010	0.048	0.212	0.832
motheredlevel6	0.013	0.049	0.272	0.786		0.008	0.051	0.149	0.882
motheredlevelU	0.064	0.066	0.971	0.332		0.007	0.069	0.100	0.920
fatheredlevel1	-0.005	0.052	-0.102	0.919		0.006	0.055	0.107	0.915
fatheredlevel2	-0.032	0.047	-0.677	0.498		-0.002	0.050	-0.045	0.964
fatheredlevel3	-0.017	0.048	-0.360	0.719		0.003	0.051	0.052	0.958
fatheredlevel4	-0.033	0.046	-0.725	0.468		0.002	0.049	0.051	0.960
fatheredlevel5	-0.010	0.047	-0.208	0.835		0.002	0.049	0.044	0.965
fatheredlevel6	-0.001	0.053	-0.016	0.987		0.007	0.055	0.129	0.897
fatheredlevelU	-0.053	0.064	-0.818	0.414		-0.001	0.067	-0.022	0.983
CLASSIFICATION2	-0.025	0.019	-1.302	0.193		0.000	0.020	-0.010	0.992
CLASSIFICATION3	0.046	0.045	1.013	0.311		-0.022	0.048	-0.463	0.643
CLASSIFICATION4	0.335	0.124	2.705	0.007	**	-0.080	0.119	-0.668	0.504

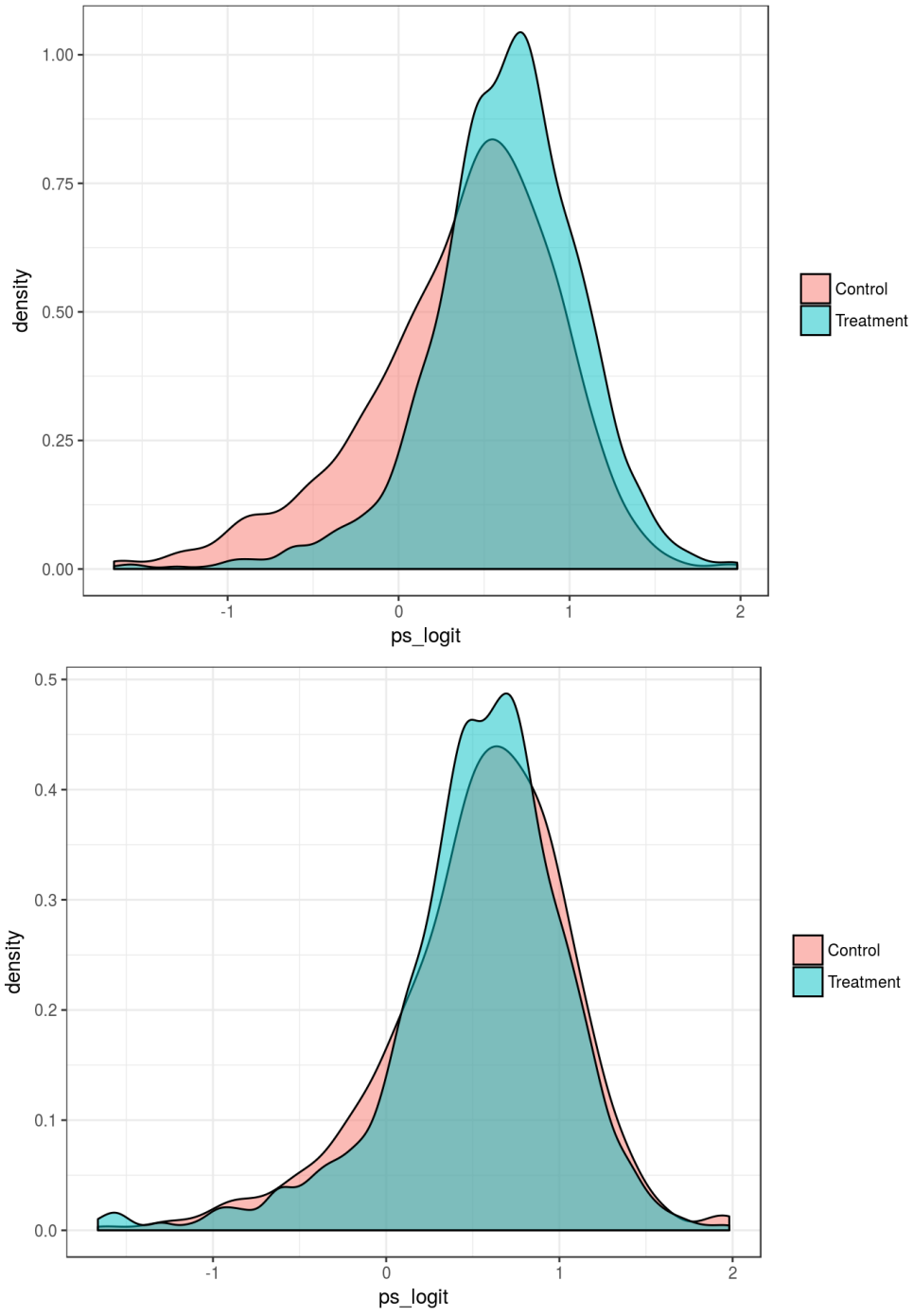

Figure 54 Distribution of propensity scores (logit scale) both before balancing (top panel) and after balancing (bottom panel).

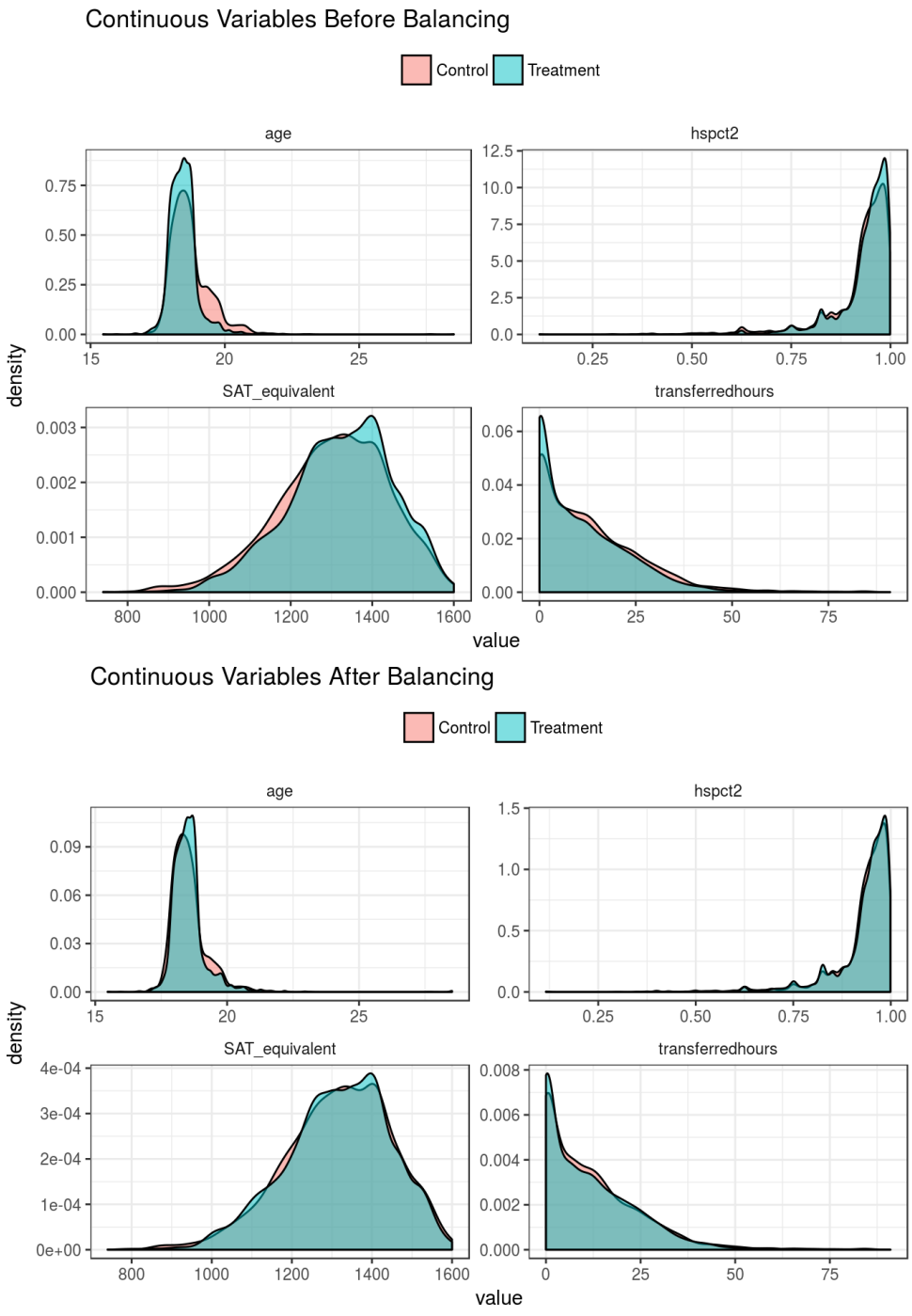

Figure 55 Distributions of continuous covariates before balancing (top panel) and after balancing (bottom panel).

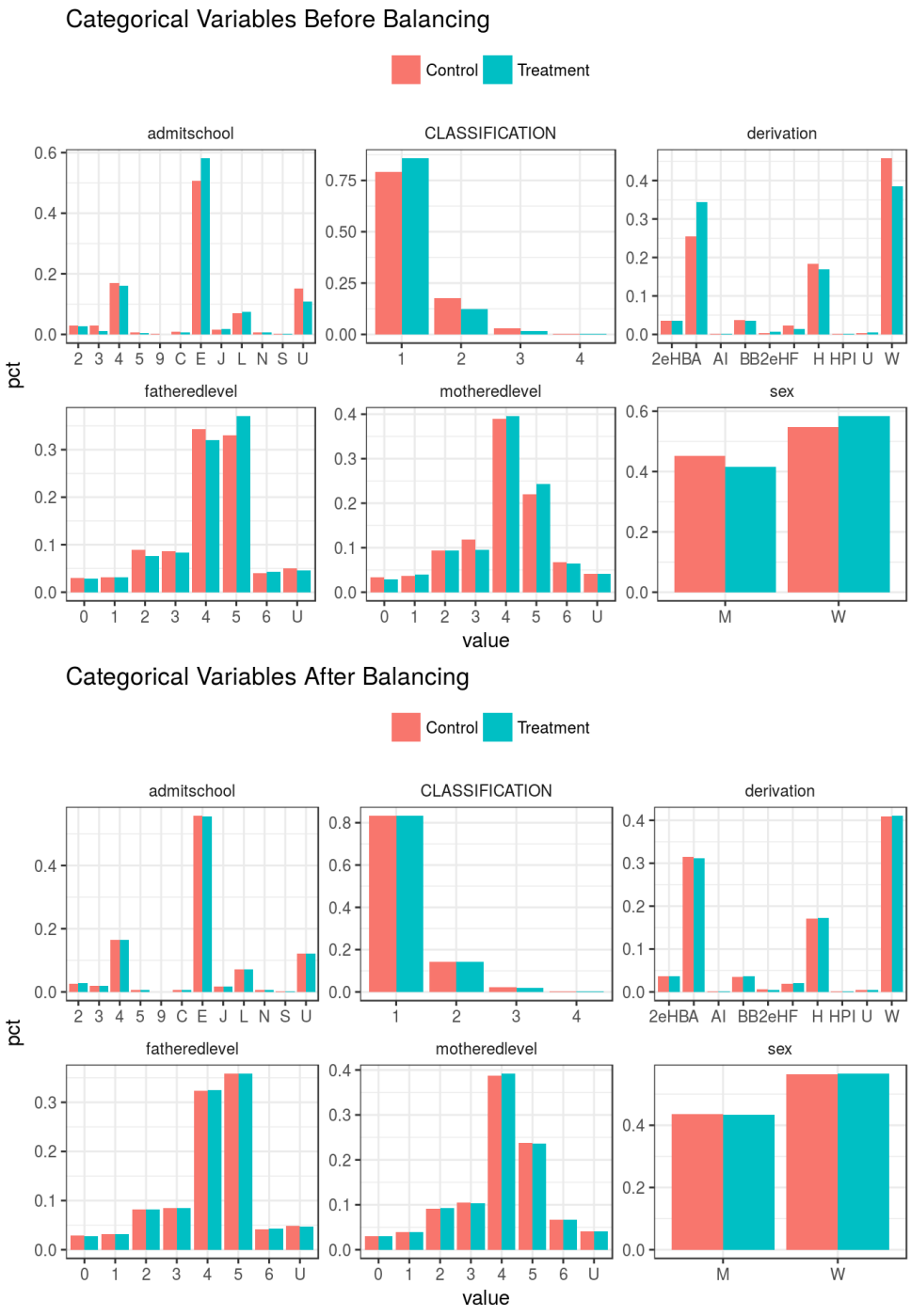

Figure 56 Distributions of categorical covariates before balancing (top panel) and after balancing (bottom panel).

Fixed-effects model

The unweighted average treatment effect estimate for the fixed-effects model was positive and significant, $ATE = 0.123, SE = 0.035, t = 3.561, p < .001$. The inverse-propensity weighted average treatment effect estimate was positive, slightly smaller in magnitude, and significant, $ATE = 0.100, SE = 0.034, t = 2.924,\ p = .003$. Thus, in this model the advantage for high retrieval-practice remained significant after covariate balance was achieved. Students in high retrieval-practice prerequisite courses were found to perform approximately 0.1 standard deviations better in their subsequent chemistry course, all else being equal. Full regression output for these models is contained in Appendix D. Notably, the size of the estimated effect was slightly larger here than when operationalizing treatment using a median split.

Random-effects model

The unweighted average treatment effect for the random-effects model was positive, considerably larger than before, and significant, $ATE = 0.263, SE = 0.091, t = 2.890, p = .007$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, slightly smaller in magnitude, and significant, $ATE = 0.241, SE = 0.086, t = 2.810, p = .009$. Thus, modeling the relationship with random effects for prerequisite and subsequent courses produced larger estimates of the ATE. Full regression output for these models is contained in Appendix D.

Cluster-robust standard errors model

The unweighted average treatment effect for the model with cluster-robust standard errors was positive and significant, $ATE = 0.057, SE = 0.028, p = 0.047$. The average

treatment-effect estimate for the inverse-propensity weighted model was positive, similar in magnitude, but only marginally significant, $ATE = 0.055, SE = 0.029, p = 0.066$. Here, models using cluster-robust standard errors but not explicitly estimating individual course effects produced smaller ATE estimates than either of the other modeling approaches. Full regression output for these models is contained in Appendix D.

Causal Effect Estimates of High Graded Retrieval Practice (Median Split) For Economics

The creation of a dichotomous treatment variable using a median split of total graded retrieval practice elements ($Med = 9$) resulted in a sample of 1297 students in the high retrieval-practice (treatment) condition and 1469 students in the low retrieval-practice (control) condition (values greater than the mean were assigned to treatment). The effective sample size, after adjusting for covariates, was 1260 students in the treatment condition and 1443 in the control condition. See Table 9 for mean, median, and standard deviation of retrieval practice elements.

Covariate balance assessment

Several covariates were unbalanced between treatment and control conditions prior to adjustment via inverse-propensity weighting. Figure 57 depicts standardized mean differences for each variable (or each level for categorical variables) both before and after adjustment using inverse propensity-score weights. Specifically, before weighting, high school transferred hours had a standardized mean difference in excess of the 0.1 threshold, with several other variables near the threshold. For mean differences, variance ratios, and K–S statistics before and after adjustment, see Appendix C. Notice that again, in each case, variance ratios are closer to unity and K–S statistics are closer to zero after adjustment.

A logistic regression of treatment on covariates was performed before and after weighting (Table 15). Prior to adjustment, treatment and control conditions differed with respect to SAT scores and transferred hours. After adjustment, however, no systematic differences remain between conditions.

To check that the propensity-score weighting is working as intended, distributions of the propensity score (or the logit propensity score) for treatment and control conditions are visually compared before and after weighting (Figure 58), showing the expected overlap after adjustment. In a similar fashion, distributions of each covariate are shown for each condition before and after adjustment. Densities are shown for continuous variables (Figure 59) and histograms are shown for categorical variables (Figure 60). Altogether, there is ample evidence that covariate balance has been achieved.

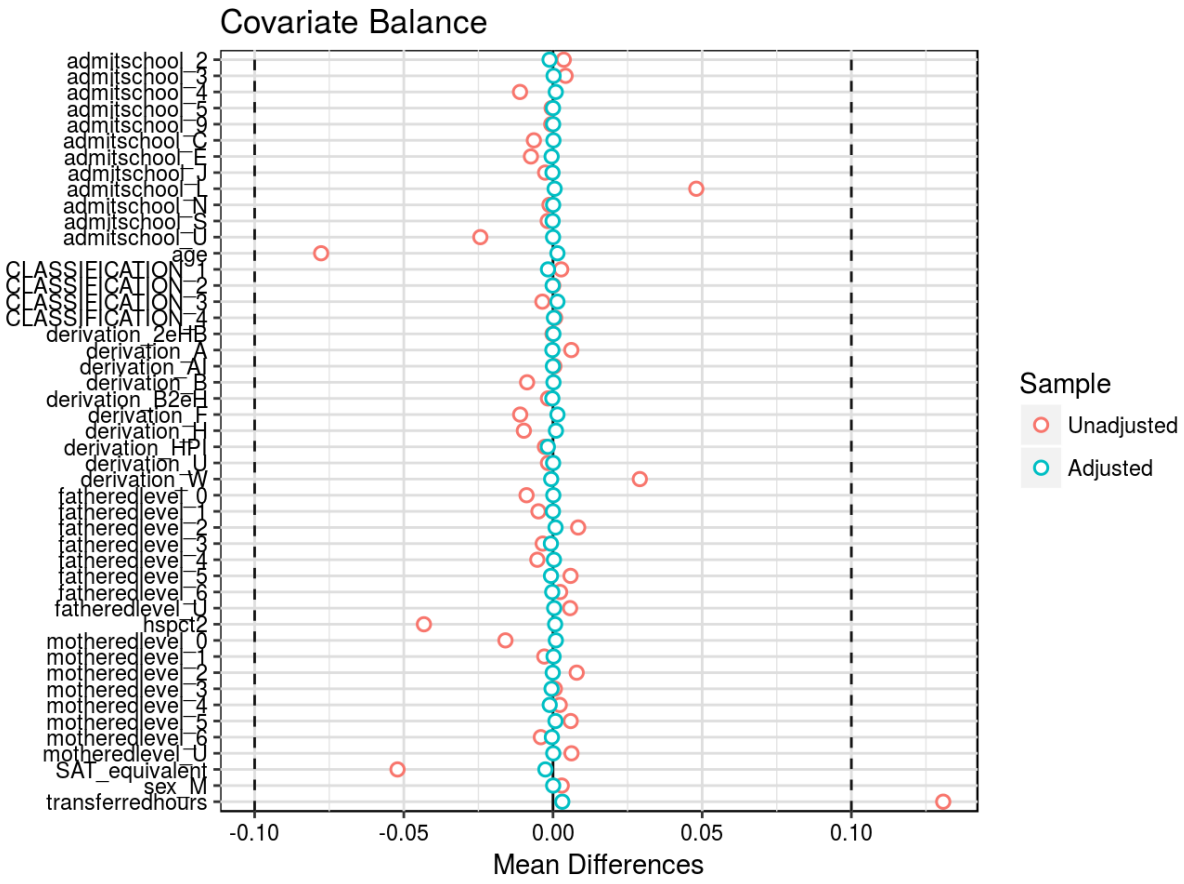

Figure 57 Love plot depicting standardized mean differences (treatment minus control) before and after propensity score adjustment

Table 15 Logistic regression coefficients predicting treatment status before and after propensity score adjustment

Variable	Unadjusted					Adjusted			
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value
Intercept	1.136	0.404	2.812	0.005	**	0.529	0.411	1.288	0.198
SAT_equivalent	0.000	0.000	-2.077	0.038	*	0.000	0.000	-0.045	0.964
hspct2	-0.083	0.083	-0.992	0.321		0.000	0.084	0.000	1.000
transferredhours	0.002	0.001	2.863	0.004	**	0.000	0.001	0.005	0.996
age	-0.026	0.018	-1.410	0.159		-0.001	0.019	-0.062	0.950
sexW	-0.006	0.020	-0.291	0.771		0.000	0.020	-0.009	0.993
derivationAI	0.052	0.169	0.306	0.760		-0.002	0.173	-0.011	0.991
derivationA	0.015	0.061	0.245	0.806		-0.001	0.062	-0.018	0.986
derivationB2eH	-0.134	0.162	-0.827	0.409		-0.017	0.167	-0.102	0.919
derivationB	-0.099	0.080	-1.227	0.220		0.000	0.081	-0.001	0.999
derivationF	-0.103	0.084	-1.233	0.218		0.013	0.084	0.152	0.879
derivationHPI	-0.442	0.256	-1.726	0.084	.	-0.501	0.332	-1.509	0.131
derivationH	-0.032	0.063	-0.508	0.611		0.000	0.064	-0.002	0.998
derivationU	-0.110	0.161	-0.683	0.495		0.001	0.162	0.009	0.993
derivationW	0.011	0.058	0.190	0.850		-0.001	0.059	-0.014	0.989
majorschool3	0.192	0.140	1.366	0.172		0.011	0.143	0.076	0.940
majorschool4	-0.053	0.051	-1.034	0.301		0.003	0.051	0.050	0.960
majorschool5	-0.014	0.109	-0.132	0.895		-0.002	0.110	-0.016	0.987
majorschool9	-0.132	0.289	-0.457	0.648		0.009	0.288	0.031	0.975
majorschoolC	-0.086	0.063	-1.367	0.172		0.001	0.063	0.012	0.990
majorschoolE	-0.036	0.039	-0.908	0.364		-0.003	0.040	-0.071	0.943
majorschoolJ	-0.317	0.208	-1.523	0.128		-0.028	0.217	-0.131	0.896
majorschoolL	0.052	0.028	1.817	0.069	.	0.000	0.029	0.015	0.988
majorschoolN	-0.145	0.208	-0.697	0.486		0.004	0.207	0.017	0.986
majorschoolS	-0.158	0.168	-0.937	0.349		-0.012	0.174	-0.071	0.944
majorschoolU	-0.054	0.028	-1.892	0.059	.	0.000	0.029	-0.017	0.987
motheredlevel1	0.103	0.093	1.109	0.267		-0.005	0.092	-0.059	0.953
motheredlevel2	0.142	0.077	1.855	0.064	.	-0.008	0.076	-0.102	0.919
motheredlevel3	0.124	0.080	1.558	0.119		-0.009	0.079	-0.110	0.912
motheredlevel4	0.121	0.077	1.579	0.114		-0.009	0.076	-0.113	0.910
motheredlevel5	0.122	0.079	1.555	0.120		-0.007	0.078	-0.092	0.927
motheredlevel6	0.096	0.085	1.130	0.259		-0.008	0.084	-0.091	0.927
motheredlevelU	0.159	0.118	1.354	0.176		-0.016	0.118	-0.138	0.890
fatheredlevel1	-0.015	0.092	-0.160	0.873		0.004	0.094	0.039	0.969
fatheredlevel2	0.042	0.079	0.529	0.597		0.010	0.079	0.133	0.894
fatheredlevel3	-0.003	0.083	-0.038	0.970		0.006	0.083	0.073	0.942
fatheredlevel4	0.003	0.079	0.033	0.974		0.008	0.078	0.096	0.923
fatheredlevel5	0.011	0.080	0.142	0.887		0.007	0.079	0.087	0.931
fatheredlevel6	0.038	0.098	0.385	0.701		0.005	0.098	0.053	0.958
fatheredlevelU	0.038	0.114	0.332	0.740		0.013	0.115	0.113	0.910
CLASSIFICATION2	0.024	0.030	0.779	0.436		0.001	0.031	0.021	0.984
CLASSIFICATION3	0.022	0.065	0.341	0.733		0.014	0.065	0.216	0.829
CLASSIFICATION4	0.077	0.134	0.577	0.564		0.017	0.137	0.121	0.904

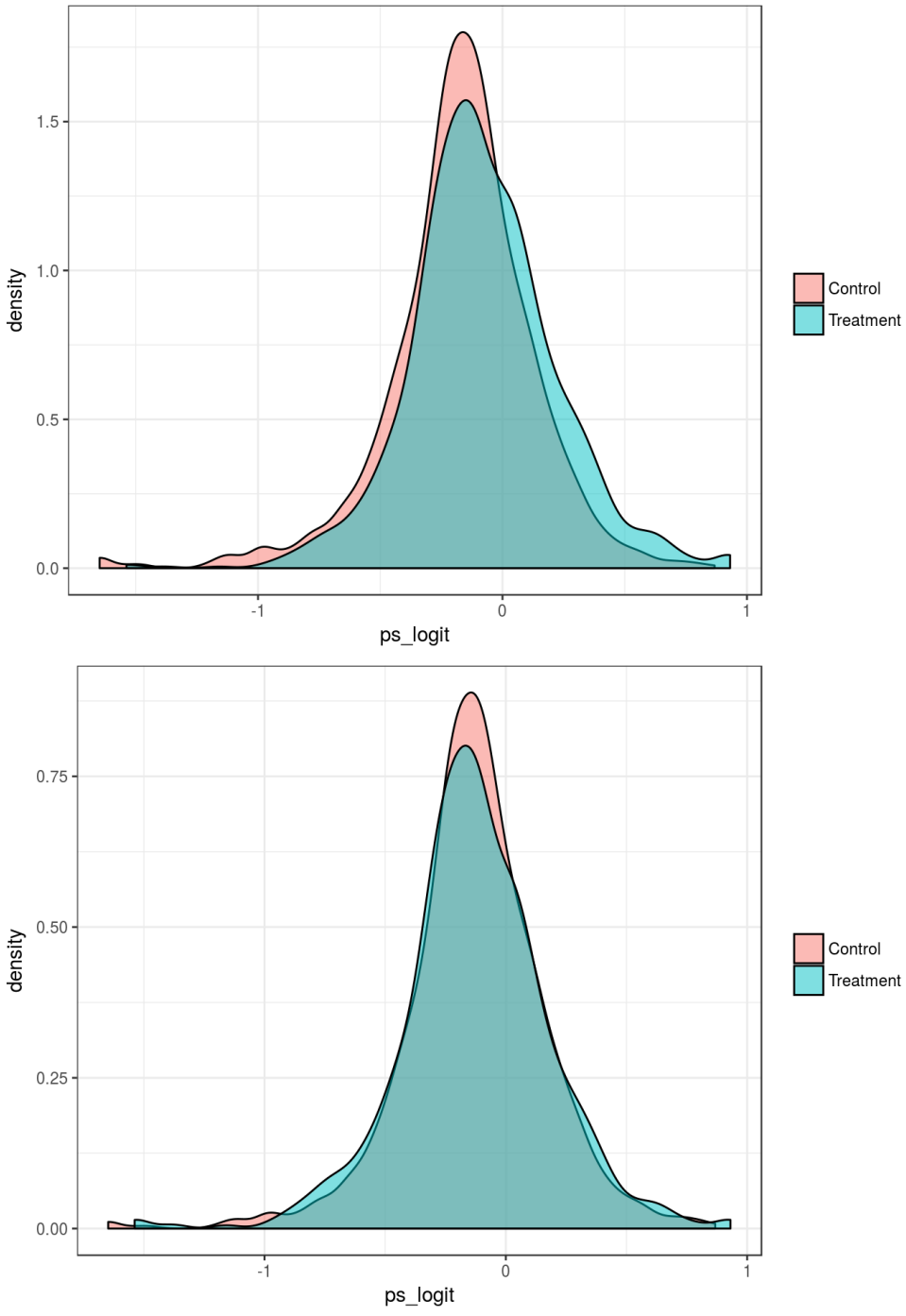

Figure 58 Distribution of propensity scores (logit scale) both before balancing (top panel) and after balancing (bottom panel)

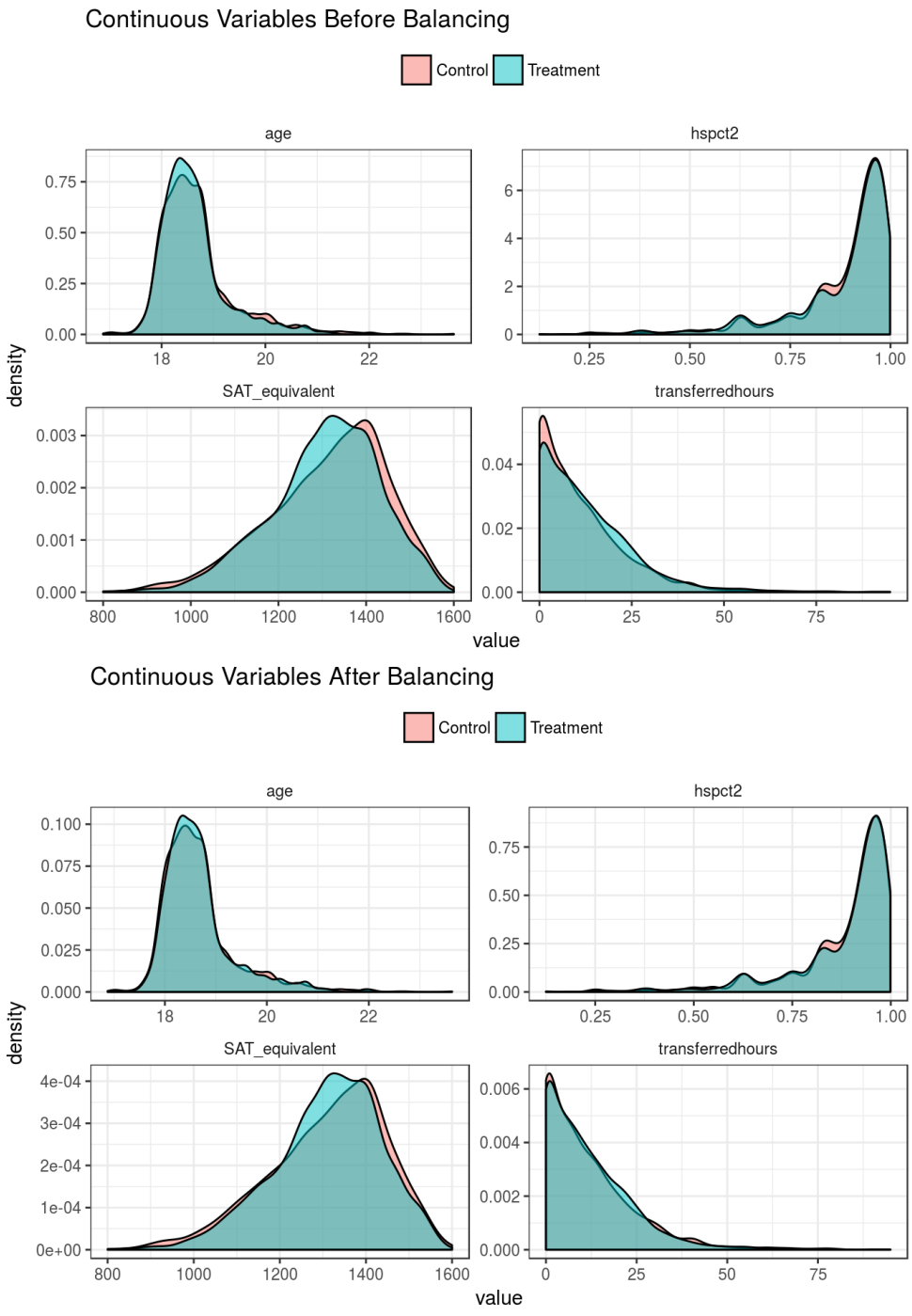

Figure 59 Distributions of continuous covariates before balancing (top panel) and after balancing (bottom panel)

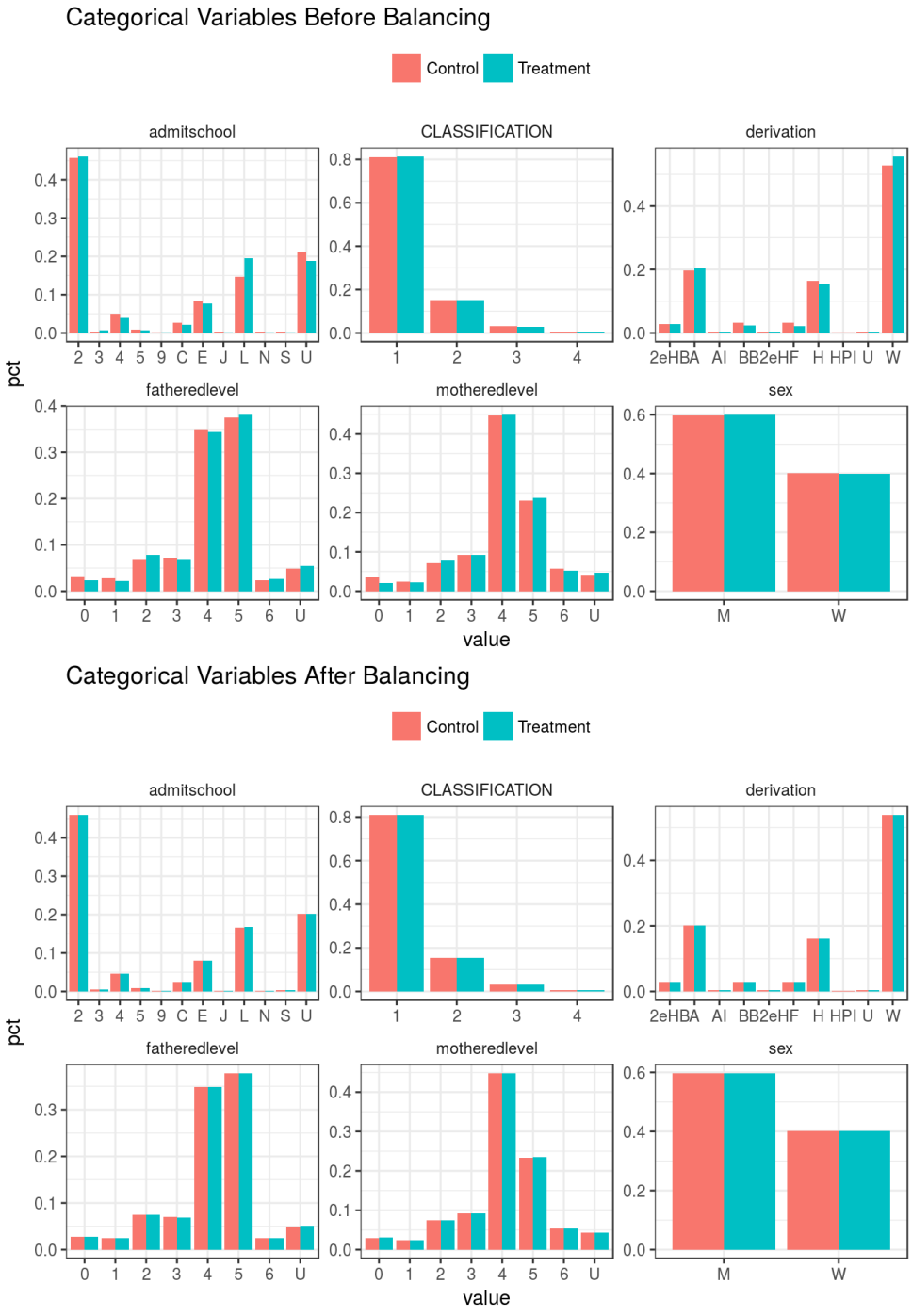

Figure 60 Distributions of categorical covariates before balancing (top panel) and after balancing (bottom panel).

Fixed-effects model

The unweighted average treatment effect estimate for the fixed-effects model was positive but not significant, $ATE = 0.074, SE = 0.064, t = 1.167, p = .243$. The inverse-propensity weighted average treatment effect estimate was positive and larger in magnitude, but not significant, $ATE = 0.086, SE = 0.064, t = 1.350,\ p = .177$. Thus, for the economics course sequence the advantage for high retrieval-practice was not significant after covariate balance was achieved. Full regression output for these models is contained in Appendix D.

Random-effects model

The unweighted average treatment effect for the random-effects model was positive but only marginally significant, $ATE = 0.083, SE = 0.045, t = 1.857, p = .077$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, similar in magnitude, but only marginally significant, $ATE = 0.082, SE = 0.044, t = 1.840, p = .081$. Again, for the economics course sequence, the effect of high retrieval-practice in the prerequisite course was not significant. Full regression output for these models is contained in Appendix D.

Cluster-robust standard errors model

The unweighted average treatment effect for the model with cluster-robust standard errors was positive but not significant, $ATE = 0.048, SE = 0.037, p = 0.198$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, similar in magnitude, but not significant, $ATE = 0.047, SE = 0.037, p = 0.214$. Here again, models using cluster-robust standard errors but not explicitly estimating individual

course effects produced smaller ATE estimates than either of the other modeling approaches. Full regression output for these models is contained in Appendix D.

Causal Effect Estimates of High Graded Retrieval Practice (Mean Split) For Economics

The creation of a dichotomous treatment variable using a mean split of total graded retrieval practice elements in the prerequisite course ECO 304K ($M = 8.51$) resulted in a sample of 1414 students in the high retrieval-practice (treatment) condition and 1352 students in the low retrieval-practice (control) condition (values greater than the mean were assigned to treatment). The effective sample size, after adjusting for covariates, was 1376 students in the treatment condition and 1327 in the control condition. See Table 9 for mean, median, and standard deviation of retrieval practice elements.

Covariate balance assessment

Only one covariate appeared to be unbalanced between treatment and control conditions prior to adjustment via inverse-propensity weighting. Figure 61 depicts standardized mean differences for each variable (or each level for categorical variables) both before and after adjustment using inverse propensity-score weights. Specifically, before weighting, high school transferred hours had a standardized mean difference in excess of the 0.1 threshold. For mean differences, variance ratios, and K–S statistics before and after adjustment, see Appendix C. Notice that again, in each case, variance ratios are closer to unity and K–S statistics are closer to zero after adjustment.

A logistic regression of treatment on covariates was performed before and after weighting (Table 16). Prior to adjustment, treatment and control conditions differed only

with respect to transferred hours. After adjustment, however, no systematic differences remained between conditions.

To check that the propensity-score weighting worked as intended, distributions of the propensity score (or the logit propensity score) for treatment and control conditions are visually compared before and after weighting (Figure 62), showing the expected overlap after adjustment. In a similar fashion, distributions of each covariate are shown for each condition before and after adjustment. Densities are shown for continuous variables (Figure 63) and histograms are shown for categorical variables (Figure 64). Altogether, there is ample evidence that covariate balance has been achieved.

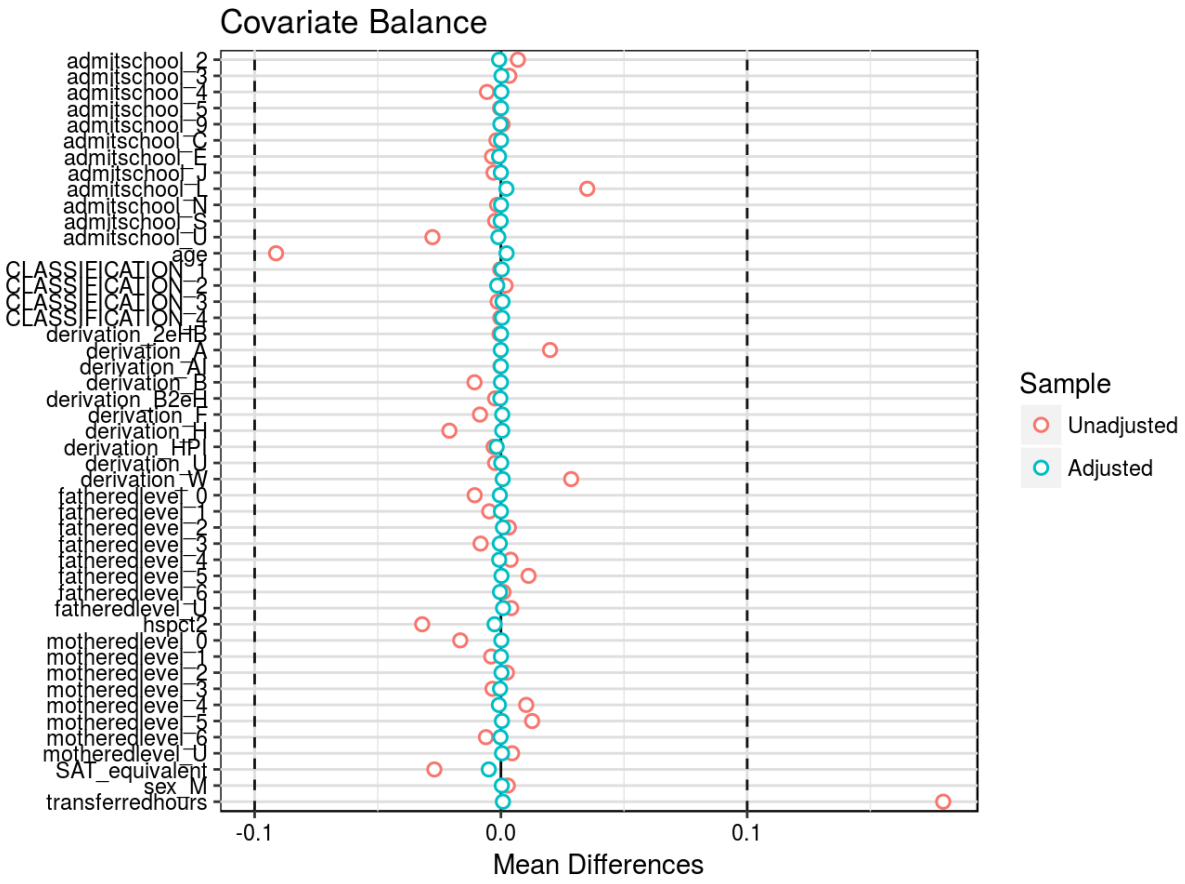

Figure 61 Love plot depicting standardized mean differences (treatment minus control) before and after propensity score adjustment

Table 16 Logistic regression coefficients predicting treatment status before and after propensity score adjustment

Variable	Unadjusted					Adjusted			
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value
Intercept	1.217	0.404	3.014	0.003	**	0.505	0.409	1.233	0.218
SAT_equivalent	0.000	0.000	-1.726	0.084	.	0.000	0.000	-0.143	0.887
hspct2	-0.060	0.083	-0.719	0.472		-0.004	0.084	-0.051	0.959
transferredhours	0.004	0.001	4.390	0.000	***	0.000	0.001	-0.028	0.977
age	-0.033	0.018	-1.795	0.073	.	0.001	0.019	0.040	0.968
sexW	-0.009	0.020	-0.470	0.638		0.000	0.020	-0.009	0.992
derivationAI	0.020	0.169	0.117	0.907		-0.001	0.174	-0.006	0.996
derivationA	0.035	0.061	0.579	0.563		0.000	0.062	0.005	0.996
derivationB2eH	-0.164	0.162	-1.016	0.310		-0.018	0.167	-0.107	0.915
derivationB	-0.105	0.080	-1.306	0.192		-0.002	0.081	-0.023	0.982
derivationF	-0.065	0.084	-0.778	0.437		0.004	0.085	0.044	0.965
derivationHPI	-0.461	0.256	-1.800	0.072	.	-0.502	0.333	-1.506	0.132
derivationH	-0.041	0.063	-0.643	0.521		-0.001	0.064	-0.009	0.993
derivationU	-0.146	0.161	-0.901	0.367		0.008	0.161	0.050	0.960
derivationW	0.012	0.058	0.204	0.839		0.000	0.059	0.007	0.995
majorschool3	0.154	0.140	1.096	0.273		0.012	0.144	0.081	0.936
majorschool4	-0.016	0.051	-0.309	0.758		-0.001	0.051	-0.025	0.980
majorschool5	0.002	0.109	0.022	0.982		-0.002	0.109	-0.019	0.985
majorschool9	0.162	0.289	0.561	0.575		-0.037	0.274	-0.134	0.893
majorschoolC	-0.038	0.063	-0.607	0.544		-0.001	0.063	-0.014	0.989
majorschoolE	-0.018	0.039	-0.447	0.655		-0.005	0.040	-0.121	0.903
majorschoolJ	-0.342	0.208	-1.646	0.100	.	-0.012	0.213	-0.058	0.954
majorschoolL	0.037	0.028	1.315	0.189		0.002	0.029	0.062	0.951
majorschoolN	-0.173	0.208	-0.835	0.404		0.002	0.207	0.010	0.992
majorschoolS	-0.191	0.168	-1.136	0.256		-0.012	0.173	-0.071	0.943
majorschoolU	-0.051	0.028	-1.782	0.075	.	-0.003	0.029	-0.111	0.911
motheredlevel1	0.088	0.093	0.947	0.344		-0.003	0.092	-0.032	0.975
motheredlevel2	0.119	0.077	1.546	0.122		-0.002	0.076	-0.022	0.982
motheredlevel3	0.107	0.080	1.339	0.181		-0.003	0.079	-0.043	0.966
motheredlevel4	0.117	0.077	1.519	0.129		-0.004	0.076	-0.049	0.961
motheredlevel5	0.119	0.079	1.508	0.132		-0.003	0.078	-0.038	0.970
motheredlevel6	0.075	0.085	0.889	0.374		-0.001	0.084	-0.017	0.986
motheredlevelU	0.124	0.118	1.054	0.292		-0.011	0.118	-0.089	0.929
fatheredlevel1	0.008	0.092	0.091	0.928		0.007	0.093	0.072	0.943
fatheredlevel2	0.051	0.079	0.647	0.518		0.011	0.079	0.142	0.887
fatheredlevel3	0.004	0.083	0.046	0.963		0.007	0.082	0.090	0.928
fatheredlevel4	0.027	0.079	0.348	0.728		0.008	0.078	0.096	0.924
fatheredlevel5	0.032	0.080	0.404	0.686		0.008	0.079	0.105	0.916
fatheredlevel6	0.049	0.098	0.507	0.612		0.004	0.098	0.040	0.968
fatheredlevelU	0.068	0.114	0.600	0.549		0.017	0.115	0.146	0.884
CLASSIFICATION2	0.017	0.030	0.549	0.583		-0.002	0.031	-0.076	0.939
CLASSIFICATION3	0.021	0.065	0.317	0.751		0.006	0.065	0.087	0.930
CLASSIFICATION4	0.004	0.134	0.031	0.975		0.020	0.135	0.152	0.880

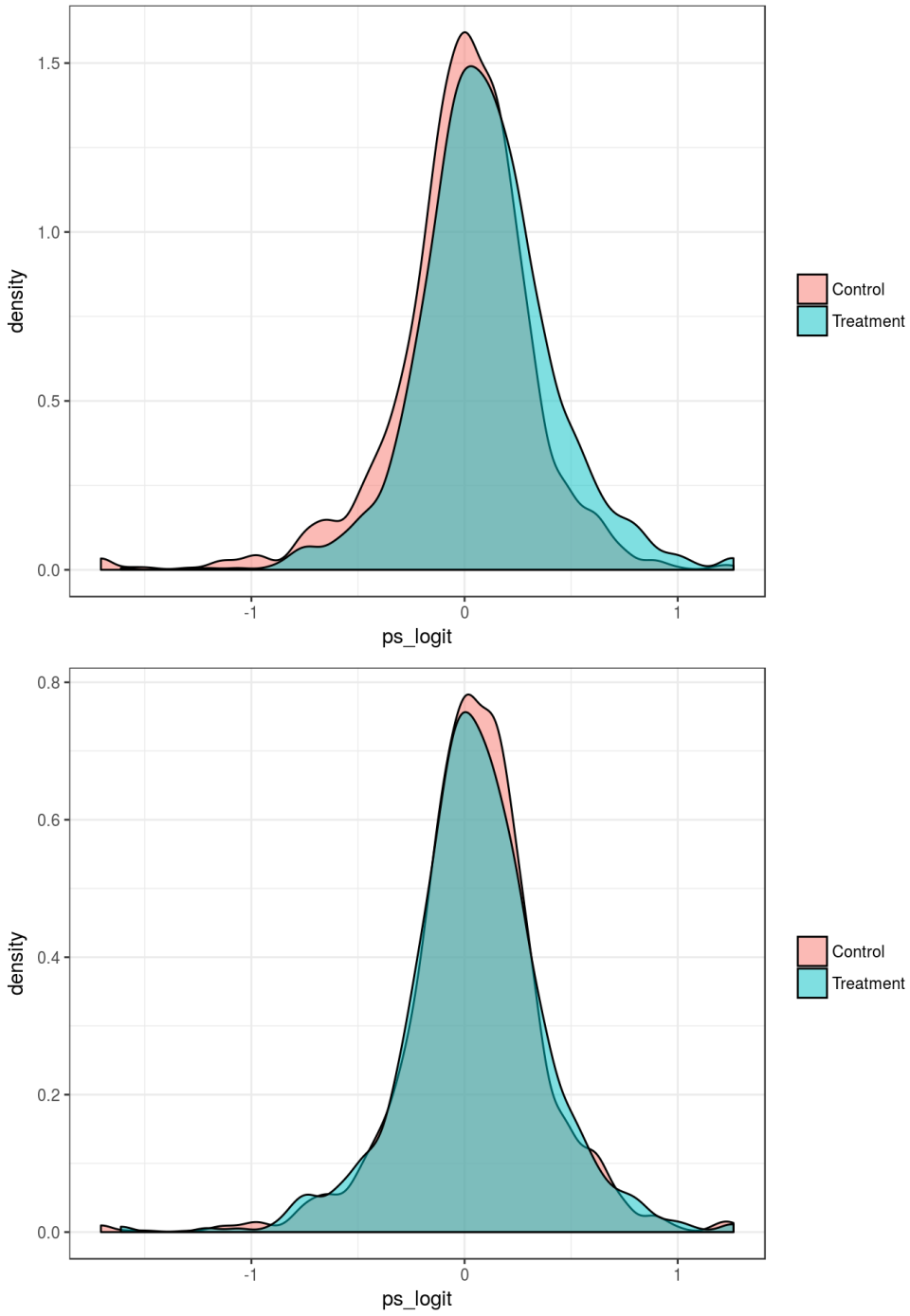

Figure 62 Distribution of propensity scores (logit scale) both before balancing (top panel) and after balancing (bottom panel)

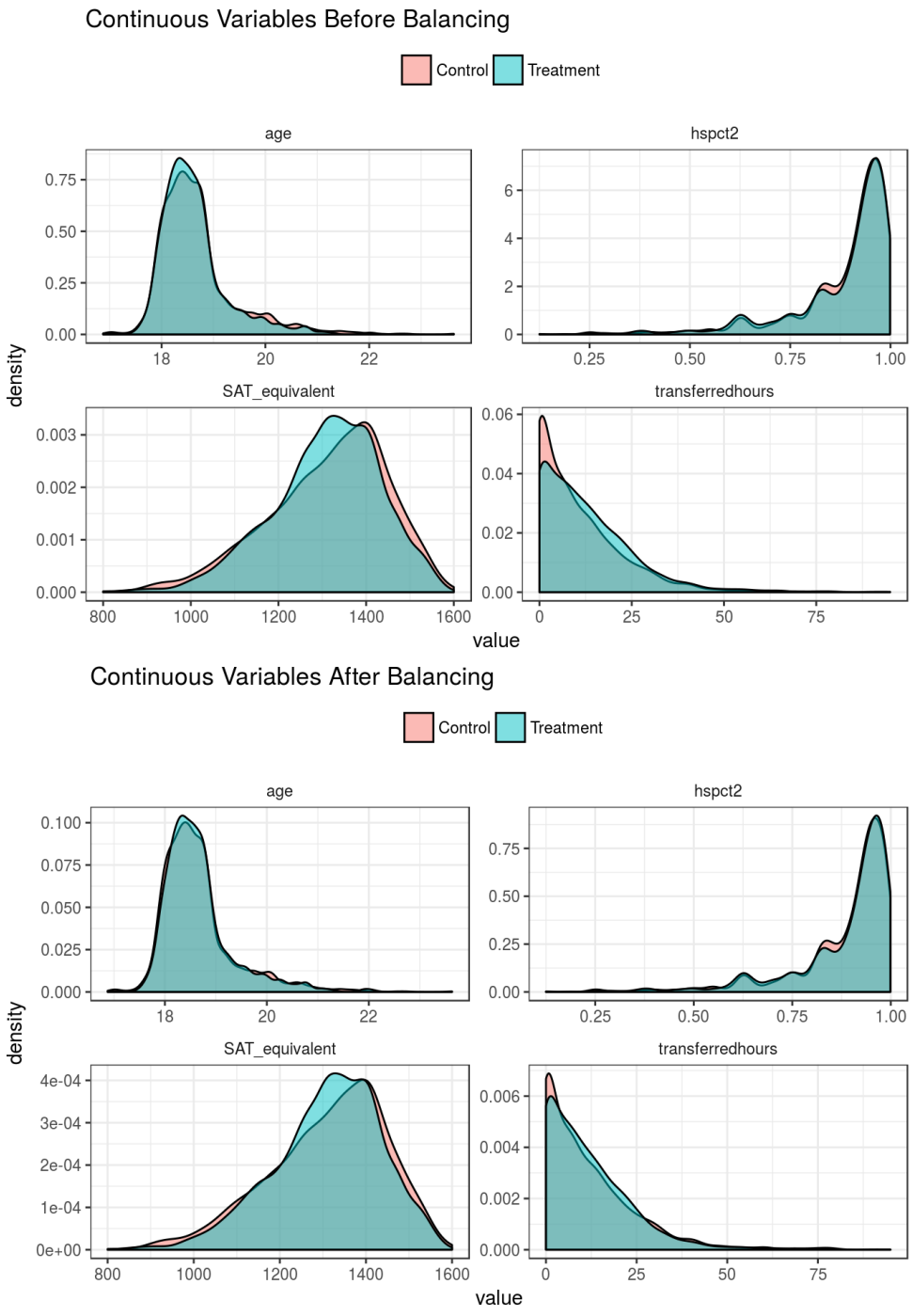

Figure 63 Distributions of continuous covariates before balancing (top panel) and after balancing (bottom panel)

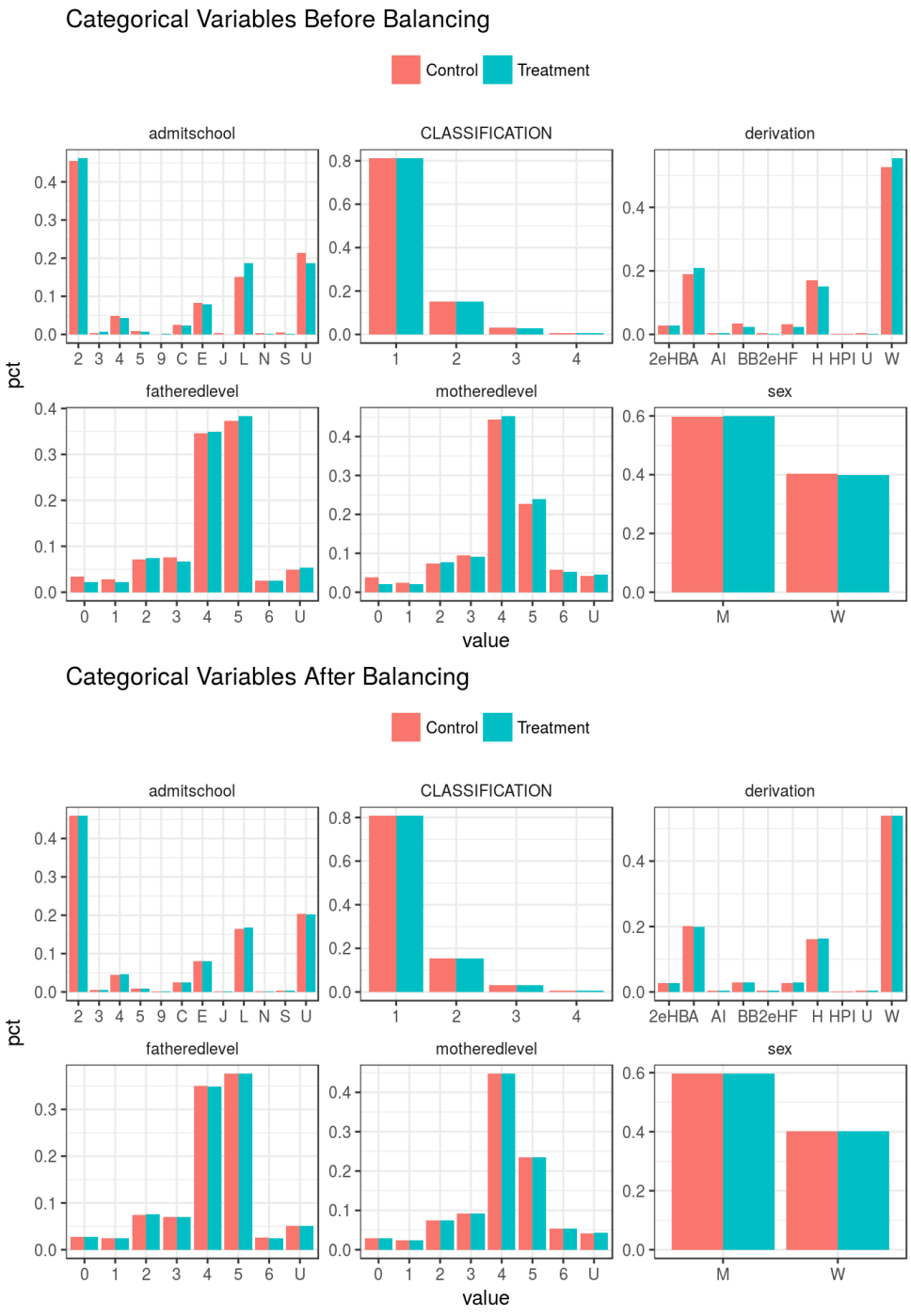

Figure 64 Distributions of categorical covariates before balancing (top panel) and after balancing (bottom panel).

Fixed-effects model

The unweighted average treatment effect estimate for the fixed-effects model was positive but only marginally significant, $ATE = 0.159, SE = 0.094, t = 1.706, p = .089$. The inverse-propensity weighted average treatment effect estimate was positive and similar in magnitude, but still only marginally significant, $ATE = 0.159, SE = 0.093, t = 1.706, p = .088$. Thus, for the economics course sequence the advantage for high retrieval-practice was not significant after covariate balance was achieved. Full regression output for these models is contained in Appendix D.

Random-effects model

The unweighted average treatment effect for the random-effects model was positive and significant, $ATE = 0.101, SE = 0.043, t = 2.356, p = .028$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, similar in magnitude, and significant, $ATE = 0.103, SE = 0.045, t = 2.309, p = .031$. Here, for the economics course sequence, the effect of high retrieval-practice in the prerequisite course was significant after covariate balance was achieved. Full regression output for these models is contained in Appendix D.

Cluster-robust standard errors model

The unweighted average treatment effect for the model with cluster-robust standard errors was positive but only marginally significant, $ATE = 0.070, SE = 0.038, p = 0.069$. The average treatment-effect estimate for the inverse-propensity weighted model was positive, similar in magnitude, and again only marginally significant, $ATE = 0.067, SE = 0.038, p = 0.082$. Here again, models using cluster-robust standard

errors but not explicitly estimating individual course effects produced smaller ATE estimates than either of the other modeling approaches. Full regression output for these models is contained in Appendix D.

Causal Effect Estimates of High Graded Retrieval Practice (Median Split) For Government

The creation of a dichotomous treatment variable using a median split of total graded retrieval practice elements ($Med = 10$) resulted in a sample of 523 students in the high retrieval-practice (treatment) condition and 1840 students in the low retrieval-practice (control) condition (values greater than the median were assigned to treatment). The effective sample size, after adjusting for covariates, was 445 students in the treatment condition and 1820 in the control condition. See Table 9 for mean, median, and standard deviation of retrieval practice elements.

Covariate balance assessment

Several covariates were unbalanced between treatment and control conditions prior to adjustment via inverse-propensity weighting. Figure 65 depicts standardized mean differences for each variable (or each level for categorical variables) both before and after adjustment using inverse propensity-score weights. Specifically, before weighting, age had a standardized mean difference in excess of the 0.1 threshold (indeed, exceeding 2.0), with several other variables near the threshold. For mean differences, variance ratios, and K–S statistics before and after adjustment, see Appendix C. Notice that again, in each case, variance ratios are closer to unity and K–S statistics are closer to zero after adjustment.

A logistic regression of treatment on covariates was performed before and after weighting (Table 17). Prior to adjustment, treatment and control conditions differed with

respect to high school rank, age, and certain levels of major. After adjustment, however, no systematic differences remain between conditions.

To check that the propensity-score weighting is working as intended, distributions of the propensity score (or the logit propensity score) for treatment and control conditions are visually compared before and after weighting (Figure 66), showing the expected overlap after adjustment. In a similar fashion, distributions of each covariate are shown for each condition before and after adjustment. Densities are shown for continuous variables (Figure 67) and histograms are shown for categorical variables (Figure 68). Altogether, there is ample evidence that covariate balance has been achieved.

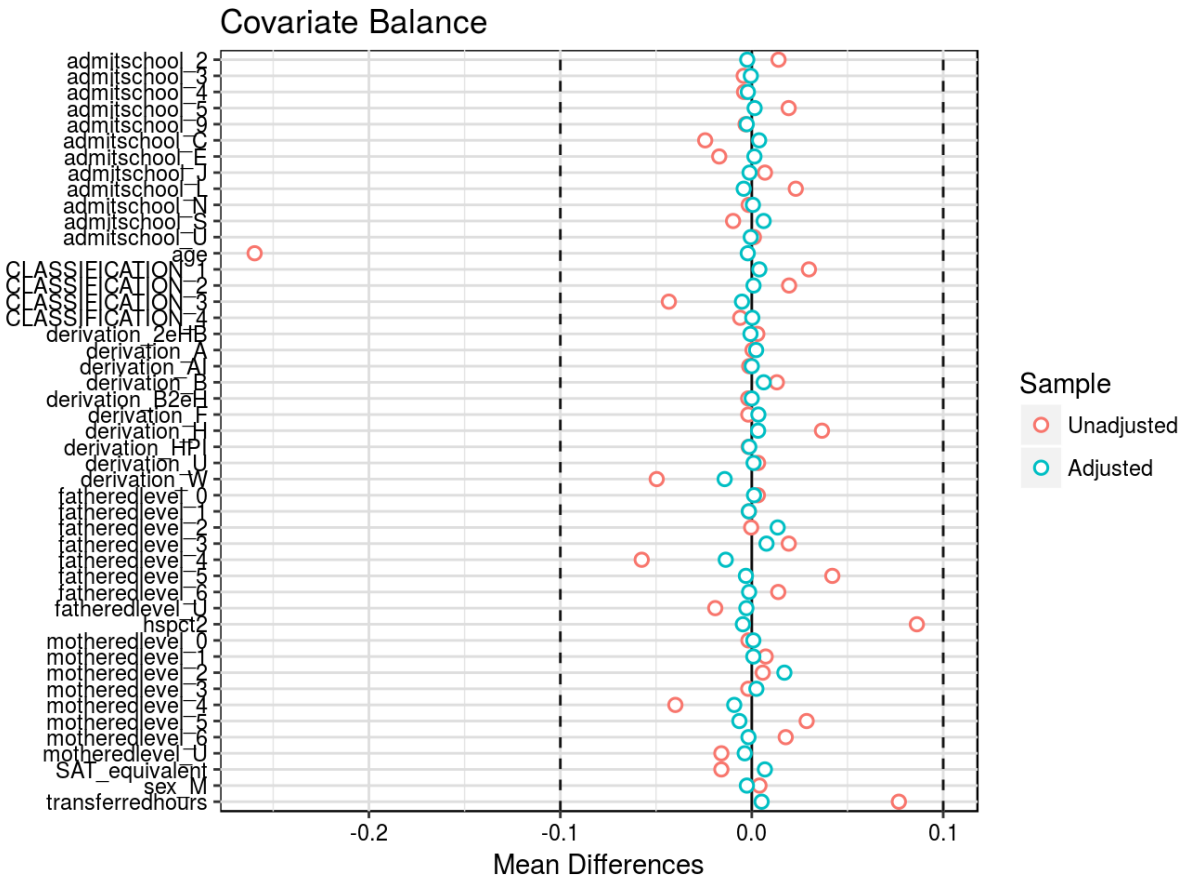

Figure 65 Love plot depicting standardized mean differences (treatment minus control) before and after propensity score adjustment

Table 17 Logistic regression coefficients predicting treatment status before and after propensity score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	1.309	0.314	4.169	0.000	***	0.370	0.377	0.980	0.327	
SAT_equivalent	0.000	0.000	0.031	0.976		0.000	0.000	0.769	0.442	
hspct2	0.142	0.066	2.155	0.031	*	0.012	0.082	0.143	0.886	
transferredhours	0.001	0.001	1.310	0.190		0.000	0.001	0.383	0.701	
age	-0.062	0.014	-4.530	0.000	***	0.002	0.017	0.093	0.926	
sexW	-0.017	0.018	-0.912	0.362		0.002	0.022	0.068	0.946	
derivationAI	-0.101	0.164	-0.615	0.539		0.006	0.201	0.029	0.977	
derivationA	-0.037	0.054	-0.692	0.489		0.012	0.066	0.174	0.862	
derivationB2eH	-0.096	0.112	-0.858	0.391		-0.012	0.138	-0.086	0.931	
derivationB	0.046	0.063	0.732	0.464		0.031	0.076	0.403	0.687	
derivationF	-0.009	0.081	-0.106	0.916		0.058	0.098	0.593	0.553	
derivationHPI	-0.284	0.246	-1.156	0.248		-0.508	0.411	-1.238	0.216	
derivationH	0.005	0.052	0.096	0.924		0.011	0.064	0.171	0.864	
derivationU	0.392	0.243	1.613	0.107		0.207	0.296	0.701	0.484	
derivationW	-0.034	0.049	-0.705	0.481		0.002	0.060	0.032	0.974	
majorschool3	-0.030	0.059	-0.506	0.613		0.001	0.073	0.014	0.989	
majorschool4	0.002	0.036	0.063	0.950		-0.001	0.045	-0.014	0.989	
majorschool5	0.129	0.054	2.374	0.018	*	0.025	0.066	0.371	0.711	
majorschool9	-0.165	0.172	-0.957	0.339		-0.501	0.289	-1.731	0.084	.
majorschoolC	-0.047	0.038	-1.231	0.219		0.017	0.047	0.372	0.710	
majorschoolE	-0.019	0.032	-0.579	0.563		0.007	0.040	0.186	0.852	
majorschoolJ	0.123	0.082	1.487	0.137		-0.010	0.104	-0.092	0.927	
majorschoolL	0.018	0.033	0.529	0.597		0.002	0.041	0.053	0.958	
majorschoolN	-0.062	0.096	-0.644	0.519		0.034	0.114	0.294	0.769	
majorschoolS	-0.183	0.084	-2.166	0.030	*	0.086	0.097	0.894	0.372	
majorschoolU	-0.006	0.033	-0.182	0.855		0.007	0.040	0.177	0.860	
motheredlevel1	0.060	0.074	0.816	0.415		0.001	0.093	0.014	0.989	
motheredlevel2	0.034	0.068	0.503	0.615		0.022	0.085	0.253	0.801	
motheredlevel3	0.016	0.070	0.225	0.822		-0.002	0.087	-0.018	0.985	
motheredlevel4	0.026	0.069	0.369	0.712		-0.008	0.087	-0.094	0.926	
motheredlevel5	0.048	0.071	0.671	0.503		-0.012	0.089	-0.137	0.891	
motheredlevel6	0.093	0.076	1.233	0.218		-0.015	0.095	-0.155	0.877	
motheredlevelU	0.003	0.094	0.033	0.974		-0.039	0.117	-0.332	0.740	
fatheredlevel1	-0.053	0.077	-0.693	0.488		-0.029	0.096	-0.304	0.761	
fatheredlevel2	-0.037	0.069	-0.542	0.588		0.012	0.085	0.137	0.891	
fatheredlevel3	0.008	0.070	0.112	0.911		0.010	0.087	0.118	0.906	
fatheredlevel4	-0.046	0.069	-0.670	0.503		-0.015	0.086	-0.176	0.860	
fatheredlevel5	0.006	0.070	0.088	0.930		-0.006	0.087	-0.064	0.949	
fatheredlevel6	0.018	0.079	0.233	0.816		-0.021	0.098	-0.216	0.829	
fatheredlevelU	-0.072	0.088	-0.823	0.410		0.000	0.108	-0.002	0.998	
CLASSIFICATION2	0.014	0.020	0.698	0.485		-0.005	0.024	-0.194	0.846	
CLASSIFICATION3	-0.018	0.034	-0.527	0.598		-0.013	0.042	-0.313	0.754	
CLASSIFICATION4	0.073	0.053	1.370	0.171		-0.012	0.066	-0.185	0.853	

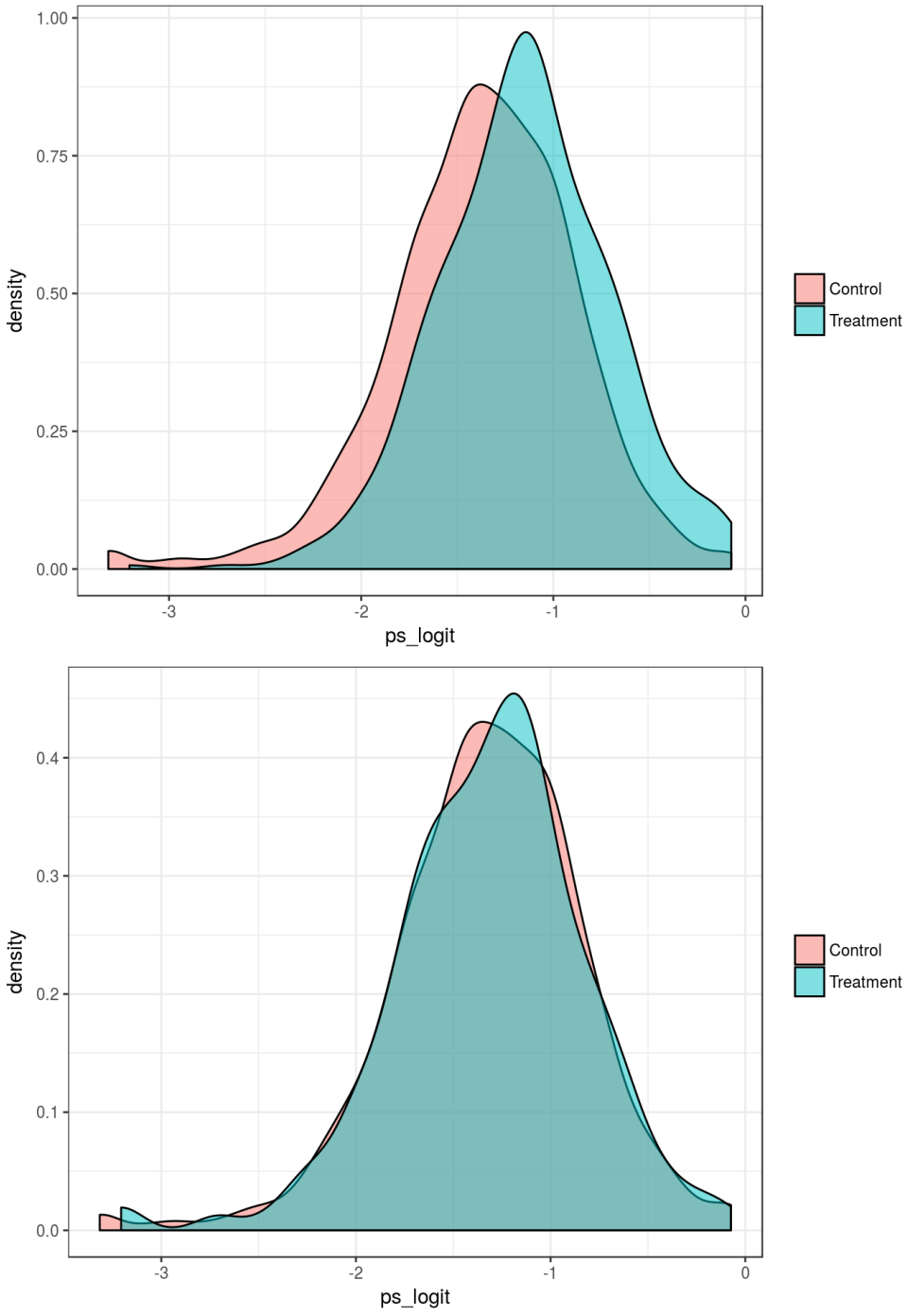

Figure 66 Distribution of propensity scores (logit scale) both before balancing (top panel) and after balancing (bottom panel)

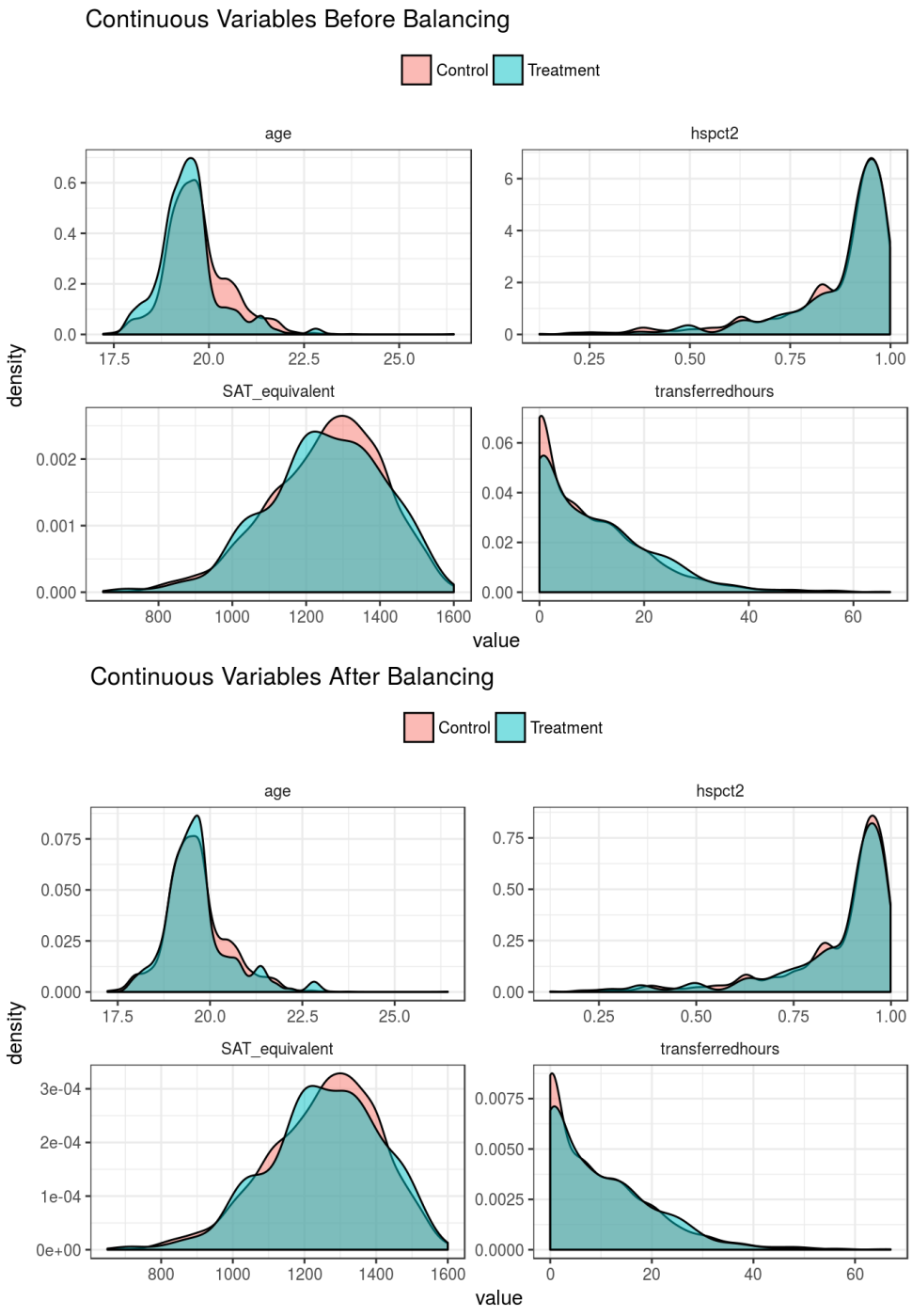

Figure 67 Distributions of continuous covariates before balancing (top panel) and after balancing (bottom panel)

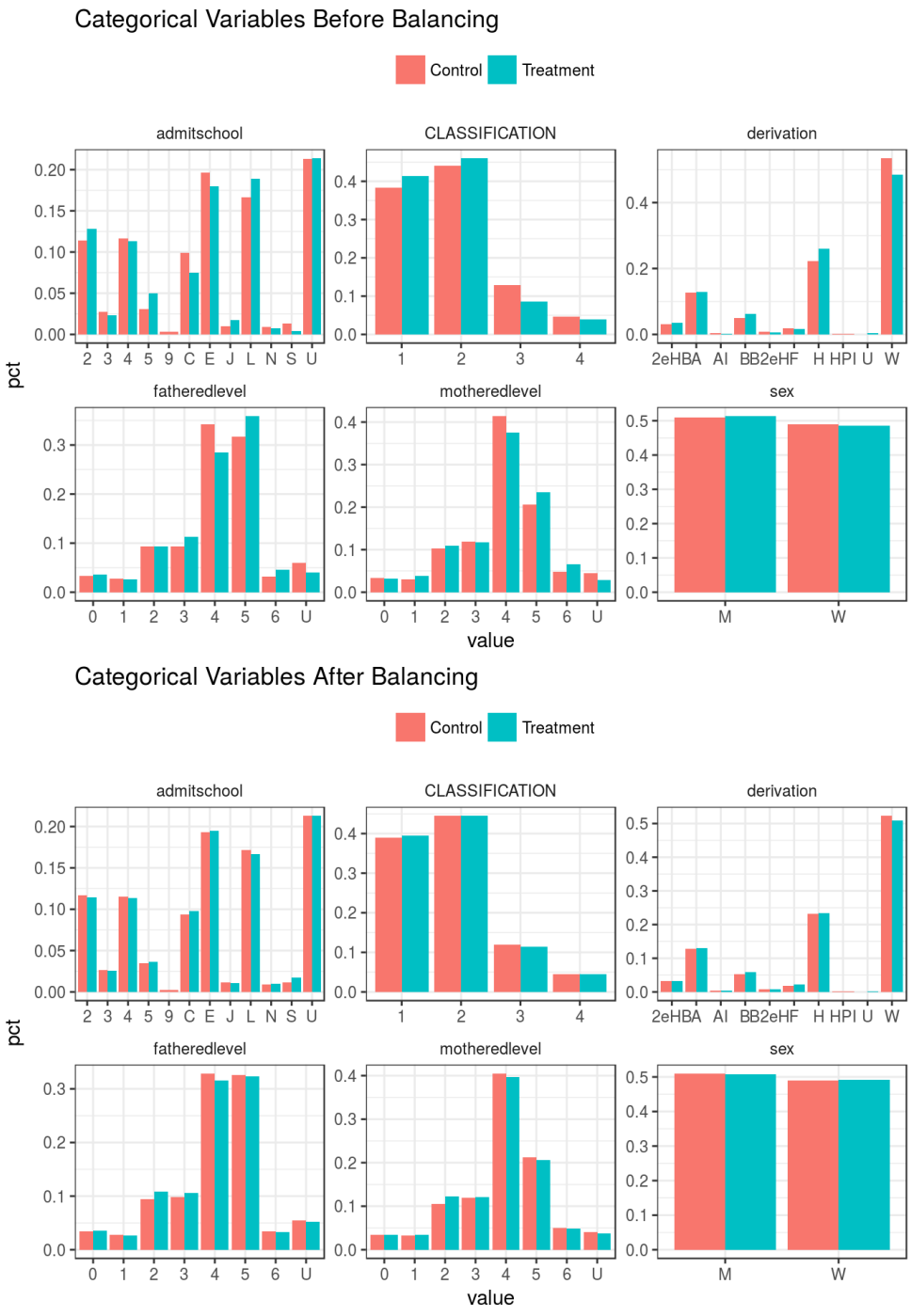

Figure 68 Distributions of categorical covariates before balancing (top panel) and after balancing (bottom panel).

Fixed-effects model

The unweighted average treatment effect estimate for the fixed-effects model was positive but not significant, $ATE = 0.044, SE = 0.051, t = 0.848, p = .397$. The inverse-propensity weighted average treatment effect estimate was positive, much smaller in magnitude, and not significant, $ATE = 0.025, SE = 0.046, t = 0.540,\ p = .589$. Thus, for the government course sequence the advantage for high retrieval-practice (i.e., above the median) was not significant after covariate balance was achieved. Full regression output for these models is contained in Appendix D.

Random-effects model

The unweighted average treatment effect for the random-effects model was negative, very small, and not significant, $ATE = -0.004, SE = 0.059, t = -0.073, p = .943$. The average treatment-effect estimate for the inverse-propensity weighted model was negative, larger in magnitude, and not significant, $ATE = -0.013, SE = 0.053, t = 1.840, p = .814$. Again, for the government course sequence, the effect of high retrieval-practice in the prerequisite course was not significant. Full regression output for these models is contained in Appendix D.

Cluster-robust standard errors model

The unweighted average treatment effect for the model with cluster-robust standard errors was negative and large in magnitude but not significant, $ATE = -0.055, SE = 0.037, p = 0.152$. The average treatment-effect estimate for the inverse-propensity weighted model was negative, larger in magnitude, but not significant, $ATE =$

$-0.072, SE = 0.043, p = 0.121$. Full regression output for these models is contained in Appendix D.

Causal Effect Estimates of High Graded Retrieval Practice (Mean Split) For Government

The creation of a dichotomous treatment variable using a mean split of total graded retrieval practice elements ($M = 8.22$) resulted in a sample of 1539 students in the high retrieval-practice (treatment) condition and 824 students in the low retrieval-practice (control) condition (values greater than the median were assigned to treatment). The effective sample size, after adjusting for covariates, was 1499 students in the treatment condition and 775 in the control condition. See Table 9 for mean, median, and standard deviation of retrieval practice elements.

Covariate balance assessment

Several covariates were unbalanced between treatment and control conditions prior to adjustment via inverse-propensity weighting. Figure 69 depicts standardized mean differences for each variable (or each level for categorical variables) both before and after adjustment using inverse propensity-score weights. Specifically, before weighting, SAT and transferred hours had standardized mean differences in excess of the 0.1 threshold. For mean differences, variance ratios, and K–S statistics before and after adjustment, see Appendix C. Notice that again, in each case, variance ratios are closer to unity and K–S statistics are closer to zero after adjustment.

A logistic regression of treatment on covariates was performed before and after weighting (Table 18). Prior to adjustment, treatment and control conditions differed with

respect to certain levels of ethnicity, major, SES, and classification. After adjustment, however, no systematic differences remained between conditions.

To confirm that the propensity-score weighting has worked as intended, distributions of the propensity score (or the logit propensity score) for treatment and control conditions are visually compared before and after weighting (Figure 70), showing the expected overlap after adjustment. In a similar fashion, distributions of each covariate are shown for each condition before and after adjustment. Densities are shown for continuous variables (Figure 71) and histograms are shown for categorical variables (Figure 72). Altogether, there is ample evidence that covariate balance has been achieved.

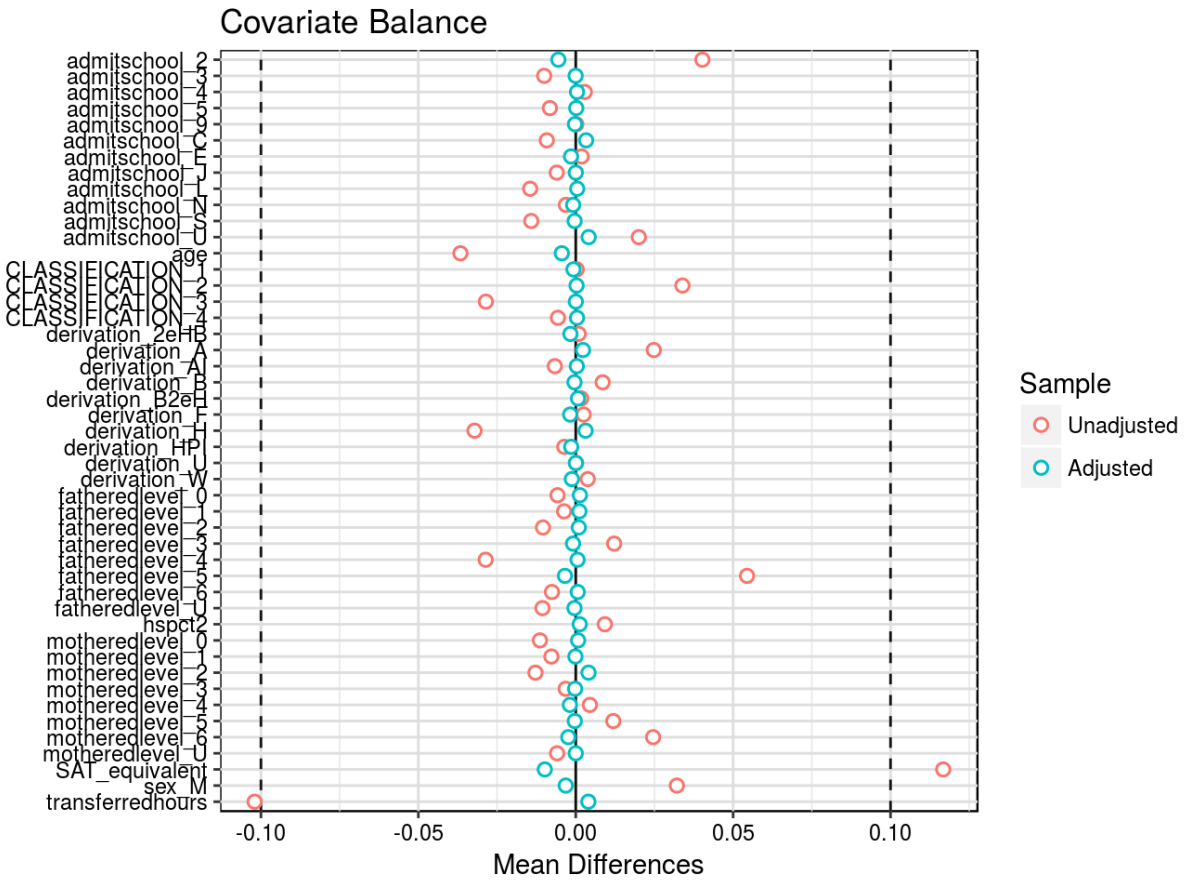

Figure 69 Love plot depicting standardized mean differences (treatment minus control) before and after propensity score adjustment

Table 18 Logistic regression coefficients predicting treatment status before and after propensity score adjustment

Variable	Unadjusted					Adjusted			
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value
Intercept	0.338	0.361	0.937	0.349		0.567	0.386	1.468	0.142
SAT_equivalent	0.000	0.000	1.510	0.131		0.000	0.000	-0.154	0.877
hspct2	0.030	0.076	0.393	0.694		0.003	0.081	0.040	0.968
transferredhours	-0.001	0.001	-1.466	0.143		0.000	0.001	-0.053	0.958
age	0.008	0.016	0.502	0.616		-0.003	0.017	-0.209	0.835
sexW	-0.007	0.021	-0.312	0.755		0.001	0.022	0.060	0.952
derivationAI	-0.499	0.189	-2.647	0.008	**	0.030	0.197	0.155	0.877
derivationA	0.042	0.062	0.675	0.500		0.016	0.065	0.246	0.806
derivationB2eH	0.047	0.128	0.365	0.715		0.038	0.141	0.273	0.785
derivationB	0.087	0.072	1.208	0.227		0.005	0.076	0.072	0.943
derivationF	0.076	0.093	0.821	0.412		-0.010	0.096	-0.108	0.914
derivationHPI	-0.666	0.283	-2.356	0.019	*	-0.487	0.392	-1.240	0.215
derivationH	0.007	0.060	0.111	0.912		0.011	0.063	0.178	0.859
derivationU	-0.053	0.280	-0.189	0.850		0.022	0.303	0.073	0.942
derivationW	0.009	0.056	0.162	0.871		0.011	0.059	0.193	0.847
majorschool3	-0.136	0.068	-1.997	0.046	*	0.009	0.072	0.128	0.898
majorschool4	-0.061	0.041	-1.465	0.143		0.011	0.044	0.260	0.795
majorschool5	-0.097	0.062	-1.557	0.120		0.012	0.067	0.173	0.863
majorschool9	0.006	0.198	0.029	0.977		-0.004	0.204	-0.021	0.983
majorschoolC	-0.083	0.044	-1.899	0.058	.	0.019	0.047	0.400	0.689
majorschoolE	-0.057	0.037	-1.535	0.125		0.009	0.039	0.217	0.828
majorschoolJ	-0.179	0.095	-1.887	0.059	.	0.013	0.101	0.126	0.900
majorschoolL	-0.079	0.038	-2.069	0.039	*	0.010	0.041	0.258	0.797
majorschoolN	-0.081	0.110	-0.735	0.462		-0.001	0.119	-0.010	0.992
majorschoolS	-0.352	0.097	-3.625	0.000	***	0.000	0.105	-0.001	0.999
majorschoolU	-0.034	0.038	-0.909	0.364		0.015	0.040	0.368	0.713
motheredlevel1	0.040	0.085	0.473	0.636		-0.004	0.091	-0.041	0.967
motheredlevel2	0.103	0.078	1.311	0.190		0.012	0.086	0.145	0.884
motheredlevel3	0.108	0.080	1.349	0.178		0.001	0.087	0.015	0.988
motheredlevel4	0.120	0.079	1.515	0.130		0.003	0.086	0.033	0.974
motheredlevel5	0.112	0.081	1.369	0.171		0.006	0.089	0.065	0.948
motheredlevel6	0.239	0.087	2.754	0.006	**	-0.007	0.094	-0.075	0.940
motheredlevelU	0.124	0.109	1.145	0.252		0.009	0.117	0.074	0.941
fatheredlevel1	-0.050	0.089	-0.563	0.574		-0.002	0.096	-0.024	0.981
fatheredlevel2	-0.085	0.079	-1.076	0.282		-0.014	0.087	-0.157	0.875
fatheredlevel3	-0.036	0.080	-0.450	0.653		-0.015	0.088	-0.168	0.867
fatheredlevel4	-0.096	0.079	-1.217	0.224		-0.011	0.087	-0.126	0.900
fatheredlevel5	-0.036	0.080	-0.448	0.654		-0.014	0.088	-0.160	0.873
fatheredlevel6	-0.115	0.090	-1.274	0.203		-0.009	0.099	-0.087	0.931
fatheredlevelU	-0.123	0.101	-1.218	0.223		-0.019	0.109	-0.173	0.862
CLASSIFICATION2	0.001	0.023	0.045	0.964		0.002	0.025	0.098	0.922
CLASSIFICATION3	-0.088	0.039	-2.273	0.023	*	0.006	0.041	0.134	0.893
CLASSIFICATION4	-0.083	0.061	-1.356	0.175		0.011	0.066	0.168	0.867

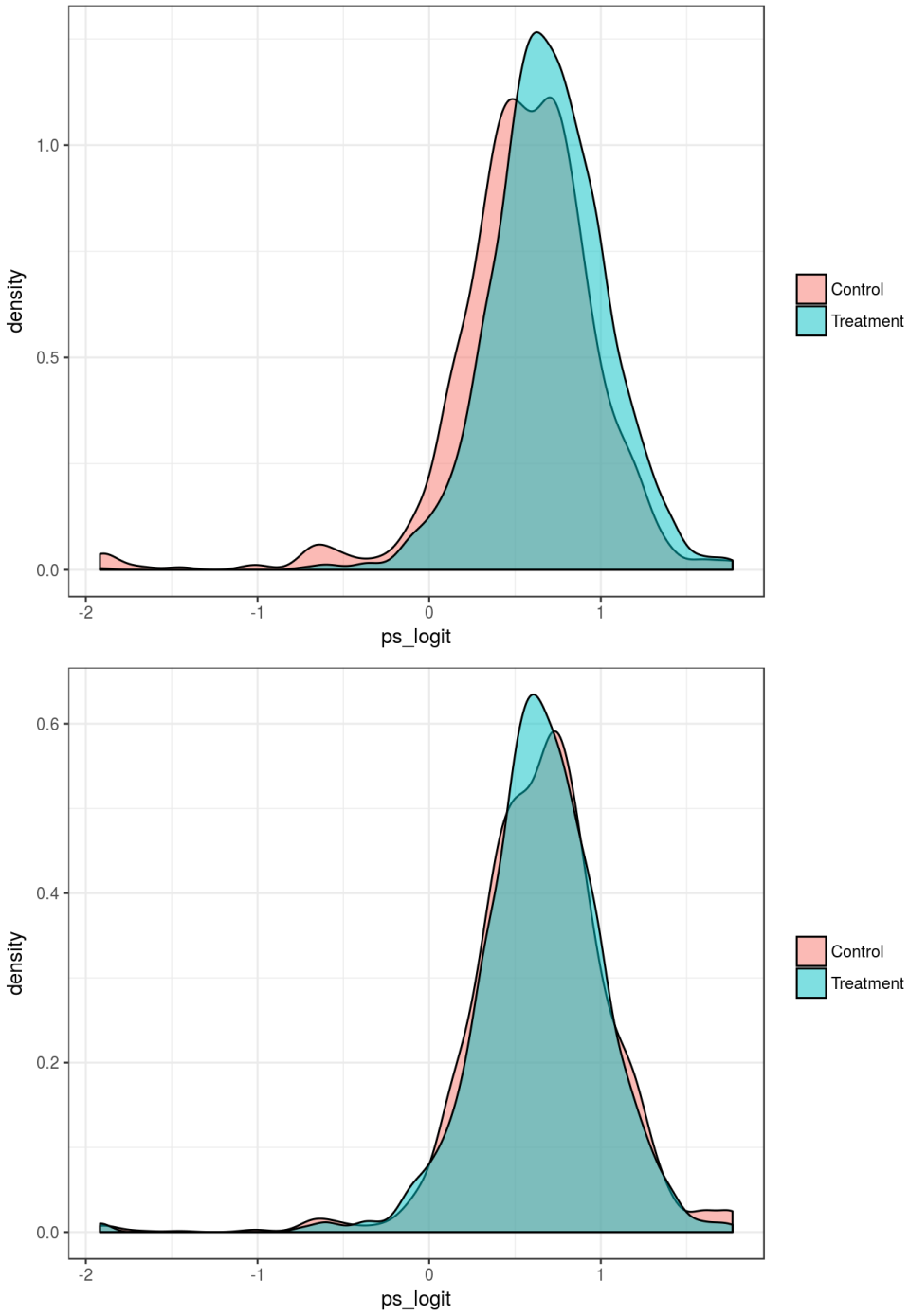

Figure 70 Distribution of propensity scores (logit scale) both before balancing (top panel) and after balancing (bottom panel)

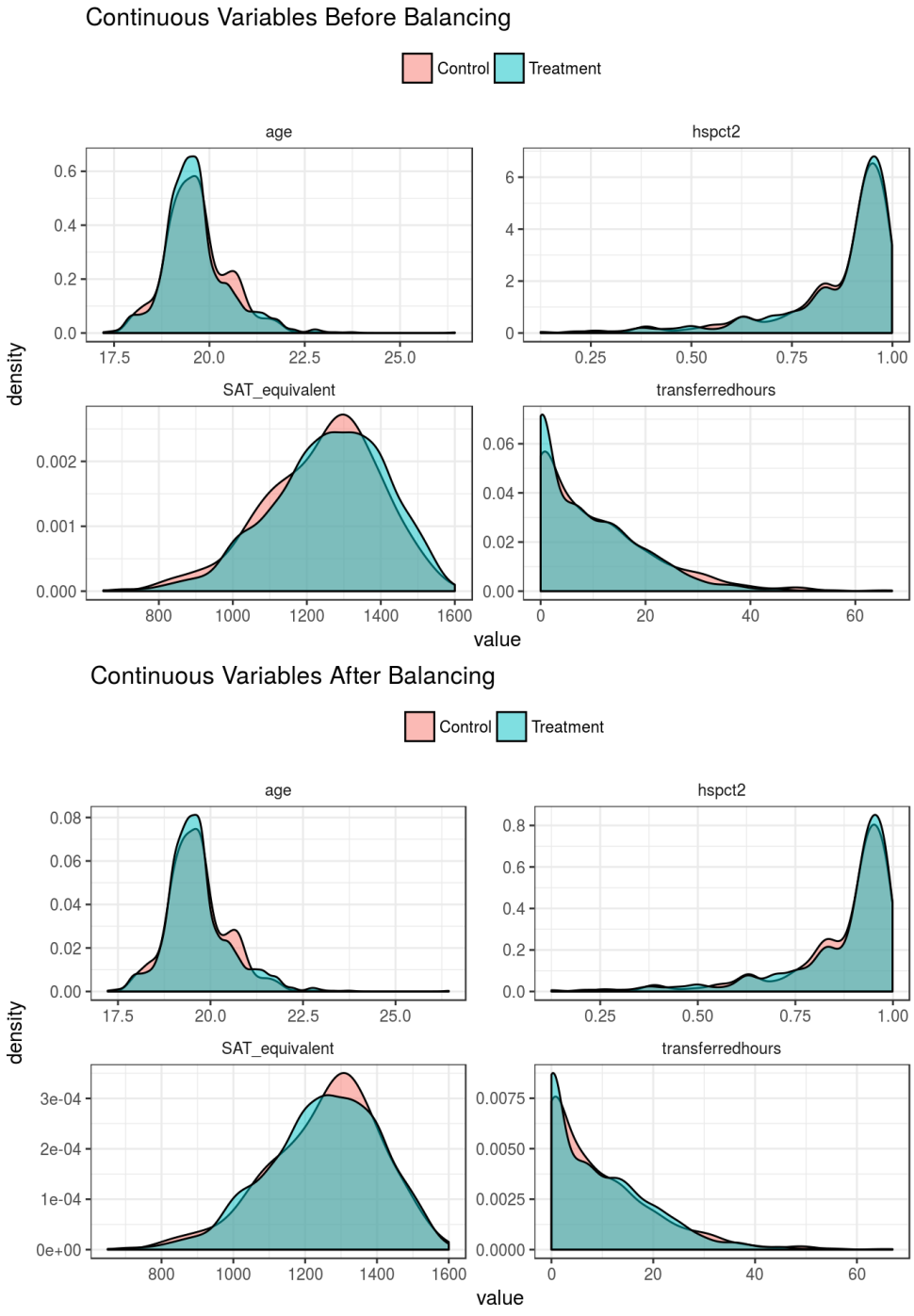

Figure 71 Distributions of continuous covariates before balancing (top panel) and after balancing (bottom panel)

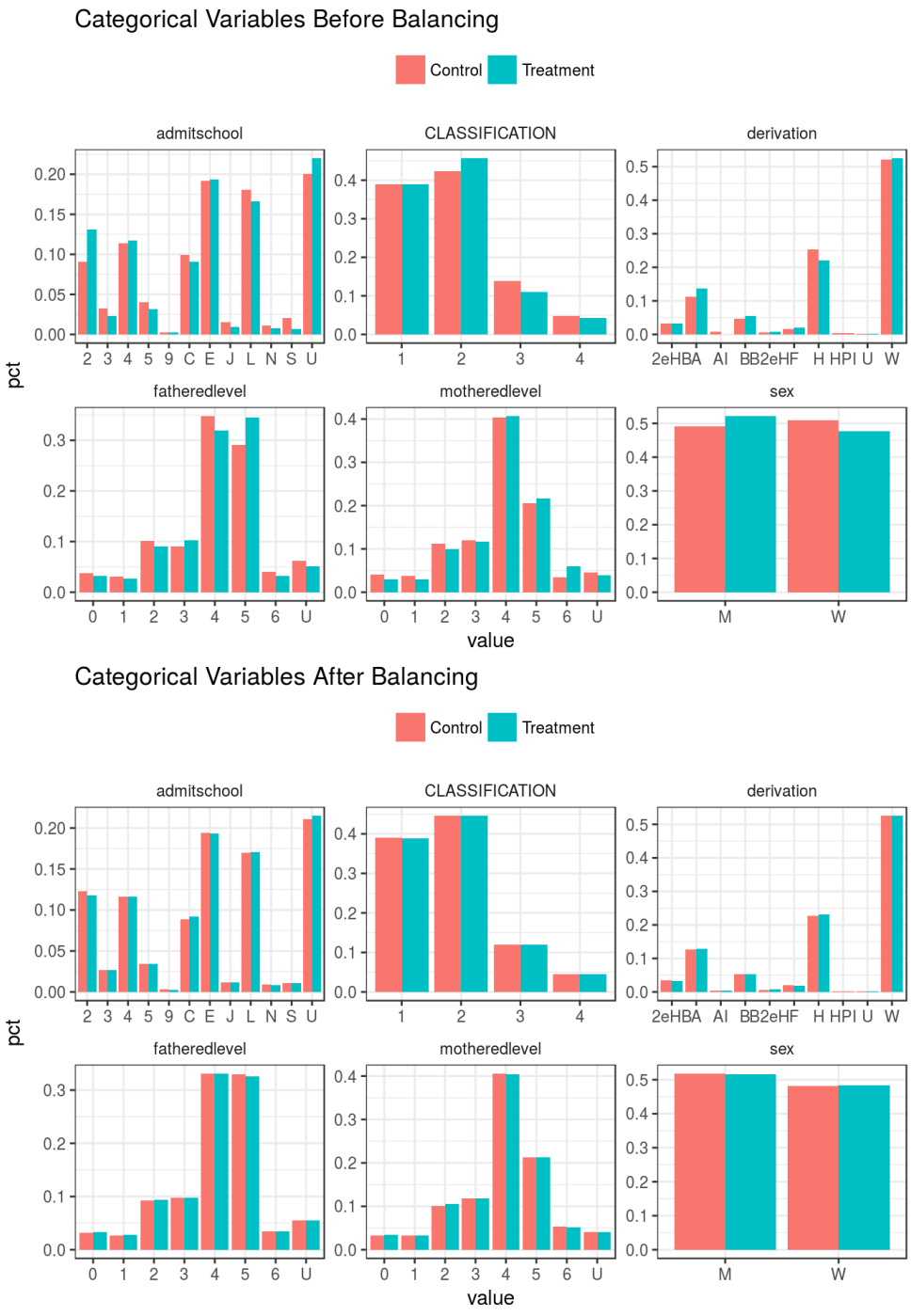

Figure 72 Distributions of categorical covariates before balancing (top panel) and after balancing (bottom panel).

Fixed-effects model

The unweighted average treatment effect estimate for the fixed-effects model was negative but not significant, $ATE = -0.044, SE = 0.065, t = -0.681, p = .496$. The inverse-propensity weighted average treatment effect estimate was positive, much smaller in magnitude, and not significant, $ATE = -0.026, SE = 0.061, t = -0.42,\ p = .674$. Thus, for the government course sequence the advantage for high retrieval-practice (i.e., above the mean) was not significant after covariate balance was achieved. Full regression output for these models is contained in Appendix D.

Random-effects model

The unweighted average treatment effect for the random-effects model was negative and not significant, $ATE = -0.057, SE = 0.047, t = -1.222, p = .243$. The average treatment-effect estimate for the inverse-propensity weighted model was negative and similarly large in magnitude, but not significant, $ATE = -0.051, SE = 0.047, t = 1.099, p = .296$. Again, for the government course sequence, the effect of high retrieval-practice in the prerequisite course was not significant. Full regression output for these models is contained in Appendix D.

Cluster-robust standard errors model

The unweighted average treatment effect for the model with cluster-robust standard errors was negative and smaller in magnitude but not significant, $ATE = -0.23, SE = 0.044, p = 0.608$. The average treatment-effect estimate for the inverse-propensity weighted model was negative, larger in magnitude, but not significant, $ATE =$

$-0.034, SE = 0.047, p = 0.121$. Full regression output for these models is contained in Appendix D.

RESEARCH QUESTION 2

Pre-requisite course variables predicting subsequent-course success

Lasso and ordinary least squares (OLS) regressions were performed to assess the extent to which prerequisite course features predict the average grade each course's students earned in a subsequent course. All lasso coefficient estimates except fixed effects of instructor are included in Table 19 along with associated OLS regression output. The initial lasso solution identified a 5-variable solution for predicting subsequent course grade: social media ($\beta = 0.042$), cumulative exams ($\beta = 0.084$), exam dates ($\beta = 0.063$), flipped classroom ($\beta = 0.019$), and number of quizzes ($\beta = 0.002$) in the prerequisite course all significantly and positively predicted the average grade students from those courses went on to earn in their subsequent course. Two additional variables were selected that were not pedagogically relevant: year of course ($\beta = -0.138$) and whether the course was a core course ($\beta = 0.015$). Note that lasso is not typically used for inference, and thus no hypothesis tests are conducted here.

Due to slight variations in the optimal regularization parameter selected using cross-validation, the entire process was repeated 1000 times, generating distributions for each non-zero parameter estimate. These are shown in Figure 73, along with the mean and standard deviation of the estimates. The mean of the distribution of each parameter is close to the lasso parameter estimates reported above: social media had a mean β of 0.045 ($SD = 0.005$), cumulative exams had a mean β of 0.096 ($SD = 0.010$), exam dates had a mean

β of 0.075 ($SD = 0.007$), flipped classroom had a mean β of 0.026 ($SD = 0.006$), number of quizzes had a mean β of 0.002 ($SD = 0.0003$), and group activities had a mean β of -0.007, ($SD = 0.006$). Note that group activities was estimated to be zero in the lasso estimates reported above, but upon repeated simulation it has a non-zero (indeed, a negative) mean estimate. However, the standard deviation for these estimates is quite large, and it can be seen in Figure 73 that the distribution is nontrivially overlapping zero.

The OLS solution identified only three variables that significantly predictive of average subsequent course performance: cumulative final exam ($\beta = 0.739, p = .001$), extra credit ($\beta = -0.308, p = .029$), and number of quizzes ($\beta = 0.017, p = .03$). Notably, the lasso and OLS regressions agreed on only a single predictor: the number of quizzes in the prerequisite course was significantly and positively related to subsequent-course performance in both procedures.

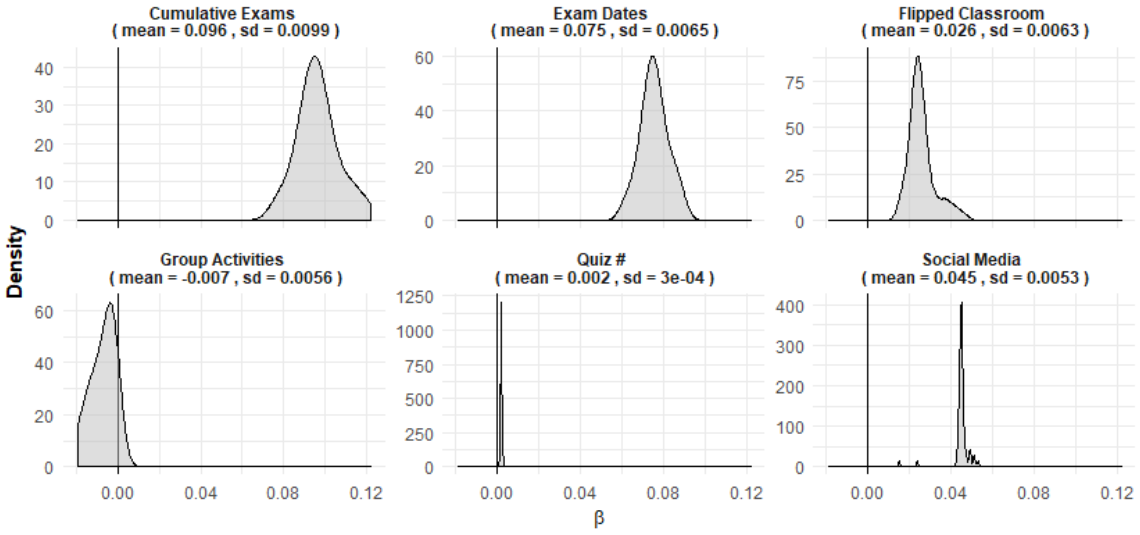

Figure 73 Distributions of lasso regression effect estimates after 1000 replications of 10-fold cross-validation were performed to select the regularization parameter. Vertical line indicates zero effect.

Table 19 Lasso and OLS regression coefficient estimates

	Lasso estimate	OLS estimate		*SE*	*t*	p-value
Instructor effects	-	-		-	-	-
Year	0.015	-0.015		0.036	-0.417	0.677
Online/SMOC	-	0.195		0.223	0.875	0.383
Core Course	-0.138	0.066		0.252	0.263	0.793
Flag Course	-	0.152		0.117	1.299	0.197
Course Level	-	0.281		0.211	1.334	0.185
Office Hours	-	0.083		0.103	0.807	0.421
Reading Acts	-	0.163		0.181	0.899	0.371
Watching Acts	-	0.338		0.204	1.656	0.100
Doing Acts	-	-0.077		0.277	-0.280	0.780
Social Media	0.042	0.045		0.213	0.211	0.833
Community Learning Ops	-	-0.120		0.118	-1.023	0.308
SLO-Knowledge	-	-0.314		0.331	-0.949	0.345
SLO-Skills	-	0.115		0.189	0.609	0.544
SLO-Social/Emotional	-	-0.005		0.282	-0.018	0.985
Course Topics	-	-0.566	.	0.305	-1.855	0.066
Dates for Topics	-	0.057		0.270	0.213	0.832
Total Enrollment	-	0.000		0.000	0.611	0.543
Cumulative Exams	0.084	0.308		0.221	1.392	0.167
Grade Choice	-	-0.327	.	0.173	-1.884	0.062
Cumulative Final	-	0.739	***	0.214	3.461	0.001
Exam Dates	0.063	0.322		0.269	1.198	0.233
Assignment Dates	-	-0.136		0.129	-1.052	0.295
Projects/Presentations	-	0.245		0.216	1.135	0.259
Participation %	-	0.002		0.017	0.098	0.922
Attendance Enforced	-	0.332	.	0.188	1.762	0.081
Flipped Classroom	0.019	0.383	.	0.220	1.742	0.084
Extra Credit	-	-0.308	*	0.139	-2.212	0.029
In-Class Active	-	0.001		0.121	0.004	0.997
Group Activities	-	-0.371	.	0.193	-1.917	0.058
Informal RP	-	0.067		0.133	0.502	0.616
Exam #	-	-0.203		0.183	-1.108	0.270
Exam %	-	0.008		0.011	0.751	0.454
Quiz #	0.002	0.017	*	0.008	2.198	0.030
Quiz %	-	-0.003		0.016	-0.166	0.869
In-Class Assignment #	-	-0.009		0.024	-0.349	0.728
In-Class Assignment %	-	0.034		0.030	1.143	0.255
Homework #	-	-0.010		0.008	-1.171	0.244
Homework %	-	0.002		0.007	0.333	0.740

Chapter Nine: Discussion Part II

In Part II of this study, a novel approach—subsequent-course analysis—was developed in an effort to connect course-level variables to students' subsequent learning outcomes in related college coursework. In general, this approach uses information from course syllabi to identify specific learning activities and teaching practices; it uses institutional records to track students' learning outcomes across specific course sequences; it uses propensity-score methods and the potential outcomes framework to ask what students' outcomes would have been if their course sequences had been different; and it uses techniques for modeling the correlated nature of observations inherent in the course-sequence paradigm.

Based on the overwhelming academic consensus that retrieving information from memory benefits long-term retention and transfer of learning, subsequent-course analysis was used to formally test the hypothesis that receiving extra retrieval practice in an introductory course can improve students' outcomes in subsequent courses in the same discipline, which are presumed to continue to build upon this foundational material. Converging lines of evidence—including the treatment effect estimates from the *a priori* model specification—support the conclusion that taking a prerequisite course with many opportunities for retrieval practice can improve transfer of learning to subsequent, related coursework relative to taking a prerequisite course with few opportunities for retrieval practice. These effects were obtained in an observational study using the highest standards of statistical control, ensuring that treatment individuals were indistinguishable from control individuals in every meaningful way except for their treatment status (i.e., how many graded retrieval practice opportunities were offered in their prerequisite course).

However, the estimated effects tended to be small, and despite having a large sample, statistical significance was sensitive to both model specification and treatment operationalization. The discipline of the course sequence appeared to make an important difference as well, with the effect appearing strong in the Chemistry sequence, weak in the Economics sequence, but nonexistent in the Government sequence. These findings and their qualifications are discussed in more detail below.

It was also hypothesized that of all prerequisite course attributes derived from syllabi, those related to retrieval practice and active learning during class-time (e.g., number of in-class assignments, description of a flipped classroom) would be most predictive of better performance in the subsequent course. This hypothesis received support from the foregoing investigation as well: a widely used variable-selection technique from machine learning identified a subset of six course features (out of 39) that were most predictive of average performance in the next course in the sequence even after controlling for individual instructor effects, and three of those—cumulative exams, flipped classroom, and number of quizzes—were distinctly related to active learning and spaced retrieval practice. Indeed, the strongest effect was that of cumulative exams, one of the few course variables that unambiguously incorporates spacing of content. Unexpectedly, having a social media page for the course was strongly associated with subsequent course success as well, and conscientious instructors (those who provided dates for all exams in their syllabi) had students who went on to do better in their subsequent course. At this point, these findings are merely associational: an important direction for future research would be to subject each of these putative effects to a rigorous subsequent-course analysis to discover whether they can be given a causal interpretation.

Propensity-score adjustment and covariate balance

The propensity score models of treatment assignment were successful in achieving covariate balance in every case. No adjustments needed to be made to the original additive, linear model specification. Specifically, no higher-order terms or interactions needed to be added to the propensity-score model to achieve balance.

In one sense, the propensity score adjustment was crucial: regardless of how the treatment effect was operationalized, there were always covariates that were markedly unbalanced between the treatment and control groups (something that, with a sample this large, we would not expect to see if students were randomly assigned to courses). In every case, students who received treatment (i.e., took a high retrieval practice course) had a higher predicted probability of doing so based on their background covariates than did students who did not receive treatment (i.e., took a low retrieval practice course). This can be seen by comparing the distribution of logit propensity scores for treatment and control groups (e.g., Figure 46, top panel). If background covariates were indeed independent of the type of prerequisite course that students took (as in a randomized experiment), then these distributions would be overlapping almost perfectly (i.e., showing no systematic deviations in center or spread). Inverse propensity weighting provides a good approximation to this ideal situation (Figure 46, bottom panel).

It is interesting to examine which covariates showed the greatest extent of imbalance before propensity-score adjustment. In general, the control group tended to be older than the treatment group, to have lower SAT scores, to have had lower high school GPAs than their peers, and to be taking the prerequisite course later in college (i.e., less likely to be taking it as a Freshman). This suggests that students with higher previous academic achievement may be seeking out courses that end up being high in retrieval

practice (perhaps due to the professor having a reputation for good teaching), while students with lower previous achievement may be choosing courses that happen to be lower in retrieval practice (perhaps due to the professor having a reputation for a lighter workload). Regardless of these speculations, it is clear that failure to adjust for covariate imbalance biases the treatment group in favor of students with higher previous achievement. As previous achievement is certainly related to subsequent course performance, this represents a serious confound that was obviated by the use of propensity-score adjustment.

By subject, the Chemistry course sequence showed greater covariate imbalance than either the Economics or the Government course sequence, and the nature of the imbalance (e.g., the specific covariates and the direction of the imbalance) often differed as well. For example, in the Chemistry sequence, treatment students had more high school transfer credits than control students on average, while in the Economics sequence treatment students had fewer than control students. As I discuss below, retrieval practice may be more important in certain course sequences than in others; to the extent that this is true, we would expect to see more imbalance in places where it matters more to course difficulty. Recall that the median number of graded retrieval practice elements was 33 in the Chemistry prerequisite course CH 301 (M = 26.2, SD = 12.4), but only 9 in the Economics prerequisite ECO 304K (M = 8.51, SD = 5.15) and 10 in the Government prerequisite course GOV 310L (M = 8.22, SD = 3.77). Note the greater variability in the Chemistry course as well.

The effect of retrieval practice on subsequent course performance

Overall average treatment effect estimates from inverse propensity-score weighted fixed- and random-effects models ranged from 0.057 to 0.067 and were also very comparable in their standard errors (see Table 11); using a median split, both estimates (fixed, *ATE* = 0.060; random, *ATE* = 0.067) were statistically significant. Here, taking a course that is above the median in retrieval practice opportunities is estimated to raise students' subsequent-course grades by 0.060 to 0.067 standard deviations, a small effect that could move a student from the 50th percentile to the 52nd. However, when a mean split was used to assess robustness, the effect was only significant for the fixed-effects model. Estimates from the cluster-robust standard error models ranged from 0.018 to 0.035 and had correspondingly smaller standard errors; they were not significant using either the mean or median split, though they approached significance in the latter. The median-split operationalization produced slightly larger ATE estimates than did the mean split: estimates from the fixed- and random-effects models were both significant using a median split, while only the fixed-effect model estimate was significant using the mean split (indeed, this was found to be the case across the board). Because the effects were relatively small and sensitive to modeling decisions, one should interpret these findings as tentative and as requiring further investigation.

The estimated effects in Chemistry tended to be larger than the overall estimates; using a median split, estimates from the fixed- and random effects models were 0.071 and 0.150, respectively, though only the former reached significance. The estimate from the cluster-robust standard errors model was 0.063 and was also significant. The estimates obtained using a mean split were larger still: the estimate from the fixed-effects model was 0.10 and the estimate from the random-effects model was 0.24, and both were significant.

These estimates suggest that the effect of retrieval practice on subsequent-course performance is more pronounced in the Chemistry course sequence than in those of other disciplines.

This speculation receives support from the less compelling treatment effects observed in the Economics and Government course sequences. In the Economics sequence, treatment effect estimates were almost as large as those in Chemistry, but only a single estimate was statistically significant. However, in the Government sequence, treatment effect estimates were much closer to zero (indeed, some were negative) and none of them reached significance.

The Chemistry course sequence consists of prerequisite course CH 301 (Principles of Chemistry I) and subsequent course CH 302 (Principles of Chemistry II): the first course covers topics such as atomic theory, bonding, and intermolecular forces, and the second course builds directly upon them with topics such as thermodynamics, chemical equilibria, and reaction kinetics. Because an understanding of the prerequisite course topics is a requirement for understanding topics covered in the subsequent course, the effects of retrieval practice on retention and transfer may take on greater importance.

The Economics course sequence consists of ECO 304K (Microeconomics) and ECO 304L (Macroeconomics). Though credit for the former course is a university prerequisite for enrollment in the latter, the subject matter may not build in the same way as it does in the Chemistry sequence. However, certain foundational concepts in the prerequisite Microeconomics course (e.g., supply and demand curves) certainly come up again in the Macroeconomics course (e.g., aggregate supply and demand). Indeed, because

the effect estimates were comparable to those in Chemistry, perhaps failure to achieve significance was due to a relatively smaller sample of Economics courses.

The Government course sequence consists of GOV 310L (American Government) and GOV12L (Issues and Policies in American Government). In contrast to the previous two sequences, the Government sequence is required for all undergraduate students at UT Austin. While the description in the official course headnote states that GOV 312L "assumes basic knowledge of government from GOV 310L," the description of the prerequisite course focuses more on issues related to Texas state and local government. Perhaps the failure to observe any significant effect of retrieval practice in this sequence is on account of the material in each course being relatively more independent.

To speculate further, perhaps the variability in effect estimates observed between disciplines is the result of differences in dosage rather than differences in subject matter. Recall that the first course in the Chemistry sequence gave a total of 26 quizzes and exams on average, over three times as many as were given in the average Government prerequisite course. Furthermore, in the Chemistry prerequisite course, the standard deviation was 12.4, relative to only 3.77 in the Government prerequisite course. Perhaps an effect would emerge in the Government sequence if there was greater variability in the number of retrieval practice opportunities offered. As it stands though, it is not possible to disentangle whether the differential effectiveness is attributable to differences in the course content, to differences in the treatment dosage, or to something else entirely. Future studies may examine the effect of such course-level retrieval practice variables in domains that more explicitly build on each other, such as mathematics and languages.

Neither is it possible, with the data at hand, to tease apart the motivational effects of frequent quizzing and testing from the strictly cognitive mnemonic effects. Perhaps the benefits of retrieval practice are mediated by student motivation: a course in which students have to study repeatedly for quizzes and tests may cause those students to be more conscientious about their coursework in general (e.g., "I'm already at the library with my bookbag open, so I guess I'll go ahead and study for my other courses too"). Putting this question to the test would require additional data beyond those I was given access too and represents an important direction for future study. Other questions that warrant future study include whether or not the effects observed and reported herein generalize to smaller courses, upper division courses, or courses offered at different institutions. Furthermore, it would be interesting to examine the effect of certain course-level variables on student evaluations of teaching. It is my intention to address these questions in future research.

Overall, this study has contributed much-needed information about course-level variables in high-enrollment colleges courses at a large public university, shedding light on the extent to which effective learning practices are being used and what other course variables make the difference when it comes to preparing students for success in their subsequent coursework. The findings presented herein have the potential to directly improve teaching at UT Austin by spurring the development of new resources for faculty that support the incorporation of spaced retrieval practice into their courses. One idea is to develop a syllabus template for faculty members to use, with presets that nudge them toward best practices in subtle ways. For example, the template could include more fields for exams and quizzes by default, or it could make certain best-practices (e.g., cumulative assignments, in-class activities, class social media involvement) required rather than

optional. Regardless of the specific features, the goal remains the same: to encourage instructors to reflect more deeply on their course design and to adopt practices that will improve learning outcomes for their students long after their final grades are submitted.

Appendix A

Table A1 Complete syllabus codebook

Variable	Definition	Code	Format
Course	Name of course as it appears in catalogs, course schedules, and student records	Abbreviation and Course Number (e.g., GOV 312L)	Entry
Department	Department offering credit for course	Department name (e.g., The Government course sequence)	Entry
Semester	Semester course is offered	Year, Semester (e.g., 2015, Fall)	Entry
Unique Number	Unique number of course as it appears in course schedule and student records	5 digit number (e.g., 37715)	Entry
Course Format	Indication that the course is face to face, hybrid, or online course	F2F/Hybrid/Online/SMOC	Forced choice
Room Number	Location of face to face course meetings	Building and Room Number (e.g., BUR 106)	Entry
Multiple Sections	Indication that course meets simultaneously with multiple sections/unique numbers	Yes/No	Forced choice
Multiple Sections - other unique numbers	List of additional unique numbers associated with course	5 digit number, 5 digit number, …	Entry
Class Meetings - Days	Days that course meets	MWF, TTh, MTWThF, etc.	Entry
Class Meetings - Times	Time of day that course meets	8:00am-9:30am, 12:30pm-2:30pm, etc.	Entry

Cultural Diversity in the United States Flag	Indication that course carries the cultural diversity in the United States flag	Yes/No	Forced choice
Ethics and Leadership Flag	Indication that course carries the ethics and leadership flag	Yes/No	Forced choice
Global Cultures Flag	Indication that course carries the global cultures flag	Yes/No	Forced choice
Independent Inquiry Flag	Indication that course carries the independent inquiry flag	Yes/No	Forced choice
Quantitative Reasoning Flag	Indication that course carries the quantitative reasoning flag	Yes/No	Forced choice
Writing Flag	Indication that course carries the writing flag	Yes/No	Forced choice
Core Course	Indication that course satisfies a core curriculum requirement	Yes/No	Forced choice
Team-taught	Indication that more than one instructor is involved in teaching the course	Yes/No	Forced choice
Instructor	Instructor of record teaching the course	Last Name, First Name (e.g., Pennebaker, James)	Entry
Co-instructor	Additional instructor of record teaching the course; if no co-instructor this field will be left blank	Last Name, First Name (e.g., Pennebaker, James) (Leave blank if no co-instructor)	Entry
Instructor Office Hours	Instructor time devoted to office hours	Total number of hours (enter as a number)	Quantitative entry
# of TAs	Number of TAs assigned to support the course	Total number of TAs; if zero, enter "0"	Quantitative entry
TA Office Hours	TA time devoted to office hours	Total number of hours (enter as a number)	Quantitative entry

Course Resources: Reading	Indication that course has required reading materials outside of class time	Yes/No	Forced choice
Reading Materials	List of the required reading materials in the course (e.g., textbook, readings posted in LMS, etc.)	List of materials	Open-ended entry
Course Resources: Watching	Indication that course has required watching activities outside of class time	Yes/No	Forced choice
Watching Activities	List of required watching activities in the course (e.g., recorded lectures, YouTube videos, TED talks, etc.)	List of activities	Open-ended entry
Course Resources: Doing	Indication that course has required practice activities outside of class time	Yes/No	Forced choice
Doing Activities	List of required practice activities in the course (e.g., Quest, textbook website, Canvas/Blackboard, iClicker, etc.)	List of activities	Open-ended entry
Social Media	Course social media resources are listed in syllabus	Yes/No	Forced choice
Learning Management System	Course learning management system (LMS) such as Canvas, Blackboard, Moodle, etc., is listed in the syllabus	Yes/No	Forced choice
Community Learning Opportunities	Community learning opportunities (e.g., TA-led sessions, exam-review sessions, study groups, Sanger Learning Center resources, etc.) are listed in the syllabus	Yes/No	Forced choice

Stated Learning Objectives - Knowledge	Knowledge-level learning objectives are clearly listed in the syllabus (e.g., topics to be learned within the course; knowledge to be gained as a result of taking the course, etc.)	Yes/No	Forced choice
Suggested Learning Objectives - Knowledge	Knowledge-level learning objectives are NOT clearly listed in the syllabus, but language appears in the syllabus that suggest knowledge-level learning objectives are associated with the course	Yes/No; N/A if previous code is "Yes"	Forced choice
Stated Learning Objectives - Skills	Skill-level learning objectives are clearly listed in the syllabus (e.g., quantitative reasoning skills, critical thinking skills, procedural skills associated with discipline, etc.)	Yes/No	Forced choice
Suggested Learning Objectives - Skills	Skill-level learning objectives are NOT clearly listed in the syllabus, but language appears in the syllabus that suggest skill-level learning objectives are associated with the course	Yes/No; N/A if previous code is "Yes"	Forced choice
Stated Learning Objectives - Socio-emotional	Socio-emotional learning objectives are clearly listed in the syllabus (e.g., teamwork/collaborative learning skills, self-awareness, self-management, social awareness, responsible decision-making, etc.)	Yes/No	Forced choice
Suggested Learning Objectives - Socio-emotional	Socio-emotional learning objectives are NOT clearly listed in the syllabus, but language appears in the syllabus that suggest socio-emotional learning objectives	Yes/No; N/A if previous code is "Yes"	Forced choice

	are associated with the course		
List of Course Topics	Course topics are listed in the syllabus	Yes/No	Forced choice
Dates for Course Topics	Dates for covering course topics are listed in the syllabus	Yes/No	Forced choice
Number of Exams	Number of exams/tests given in the course (excluding final exam and quizzes)	Total number of exams/tests	Quantitative entry
Exams Grade %	The percentage of final course grade that is accounted for by performance on exams/tests	Percentage of grade	Quantitative entry
Multiple Choice Exam Items	Exams contain multiple choice or matching items	Yes/No/Unclear	Forced choice
Short Answer Exam Items	Exams contain open-ended short answer items	Yes/No/Unclear	Forced choice
Essay Exam Items	Exams contain essay questions	Yes/No/Unclear	Forced choice
Cumulative Exams	Exams/tests in the course are described as cumulative in the syllabus	Yes/No/Unclear	Forced choice
Drop Lowest Exam Score	Students can drop their lowest exam/text score (e.g., lowest exam score will not be counted towards final course grade)	Yes/No/Unclear	Forced choice
Re-test Opportunity	Students have the opportunity to re-take exams to improve their score	Yes/No/Unclear	Forced choice
Final Exam Grade %	The percentage of final course grade that is accounted for by	Percentage of grade (if no final exam, enter 0%)	Quantitative entry

	performance on the final exam		
Multiple Choice Final Exam Items	Final exam contain multiple choice or matching items	Yes/No/Unclear	Forced choice
Short Answer Final Exam Items	Final exam contain open-ended short answer items	Yes/No/Unclear	Forced choice
Essay Final Exam Items	Final exam contain essay questions	Yes/No/Unclear	Forced choice
Cumulative Final Exam	The final exam is described as cumulative in the syllabus	Yes/No/Unclear	Forced choice
Alternative Assessment Option Weighting	Students have options in how their final grade is calculated (e.g., optional final exam, lowest test score counts for a less %, exams are worth increasing % of final grade further into the course, etc.)	Yes/No/Unclear	Forced choice
Calendar of Exam Dates	Indication that the syllabus has a calendar that includes all exam dates	Yes/No	Forced choice
Calendar of All Assessments/Assignments with Due Dates	Indication that the syllabus has a calendar that includes all assessments/assignments	Yes/No	Forced choice
Number of Quizzes	Number of quizzes (e.g., short graded assessments)	Total number of quizzes	Quantitative entry
Quiz Grade %	The percentage of final course grade that is accounted for by performance on quizzes; if not final exam exists in the course record 0%	Percentage	Quantitative entry
Multiple Choice Quiz Items	Quizzes contain multiple choice or matching items	Yes/No/Unclear	Forced choice

Short Answer Quiz Items	Quizzes contain open-ended short answer items	Yes/No/Unclear	Forced choice
Quiz Delivery (Online or Paper)	Format for administering quizzes	Online/Paper/Both/Unclear	Forced choice
Number of In-Class Assignments	Number of in-class assignments (e.g., work completed during class)	Total number of assignments	Quantitative entry
In-Class Assignment Grade %	The percentage of final course grade that is accounted for by performance on in-class assignments	Percentage	Quantitative entry
Types of In-Class Assignments	List of the in-class assignments (e.g., problem-solving activities, writing activities, etc.)	List of assignments	Open-ended entry
Group Assignments	Assignments completed in pairs or in groups are counted toward course grade	Yes/No/Unclear	Forced choice
In-class Active Learning	Evidence of active learning and student engagement is listed in the syllabus (e.g., group discussions, iClicker questioning, group or individual problem-solving, student-led activities, etc.)	Yes/No/Unclear	Forced choice
Types of In-class Active Learning	List of the types of in-class active learning mentioned in the syllabus	List of types of in-class active learning	Open-ended entry
Retrieval Practice	Evidence of retrieval practice opportunities for students is listed in the syllabus (e.g., practice/ungraded quizzes, iClicker questions during class, pop quizzes, practice tests, copies of old exams etc.)	Yes/No	Forced choice

Types of Retrieval Practices	List of the types of retrieval practices mentioned in the syllabus	List of retrieval practices	Open-ended entry
Projects or Presentations	Assignments in the form of projects or presentations exist in the course	Yes/No	Forced choice
Lab or Breakout Session	There is a required lab or TA session associated with the course (e.g., discussion section that meets for 1 hour a week)	Yes/No	Forced choice
Participation % of grade	The percentage of final course grade that is accounted for by participation during in-class activities	Percentage	Quantitative entry
Attendance Requirement Enforced	Attendance requirement is enforced in the classroom (e.g., "TAs will take attendance," "iClicker responses will mark attendance and count towards participation," etc.)	Yes/No	Forced choice
# HW Assignments	Number of homework assignments (graded work completed outside of class) in the course	Total number of HW assignments	Quantitative entry
HW Assignments Grade %	The percentage of final course grade that is accounted for by performance on homework assignments	Percentage	Quantitative entry
Types of HW Assignments	List of the types of homework assignments (e.g., problems to solve, writing assignment, presentation, discussion post, etc.)	List of assignments	Open-ended entry

Flipped Classroom	The course is described as a "flipped classroom" in the course syllabus	Yes/No	Forced choice
Extra Credit	Extra credit opportunities are listed in the course sylalbus	Yes/No	Forced choice
Extra Credit Points	Number of extra credit points available to earn in the course	Number of points	Quantitative entry
Syllabus Pages	Number of pages of the syllabus	Number of pages	Quantitative entry
Syllabus Word Count	Number of words in the syllabus	Number of words	Quantitative entry
Notes	Anything interesting or confusing about the course to make note of	Enter any notes regarding syllabus data that may not have been captured by the coding scheme	Open-end entry
Learning Objectives	Cut and paste the learning objectives/course goals/outcomes listed in the syllabus here	List of learning objectives	Text entry

Appendix B

Table B1 Correlation coefficients for all course variables.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	**Year**	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	**Online/SMOC**	0.36	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	**Core Course**	0.05	0.01	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	**Flag Course**	0.21	-0.13	0.50	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	**Office Hours**	-0.03	-0.03	0.26	-0.05	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	**Reading Acts**	-0.03	0.53	-0.34	-0.44	-0.04	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	**Watching Acts**	0.08	0.12	0.19	0.04	0.03	-0.03	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-
8	**Doing Acts**	0.16	0.14	-0.05	0.15	-0.19	-0.10	0.20	1.00	-	-	-	-	-	-	-	-	-	-	-	-
9	**Social Media**	0.26	0.26	-0.24	-0.16	0.07	-0.14	0.07	0.23	1.00	-	-	-	-	-	-	-	-	-	-	-
10	**Community Learn Ops**	0.26	0.11	0.16	0.17	-0.02	-0.33	-0.02	0.28	0.55	1.00	-	-	-	-	-	-	-	-	-	-
11	**SLO-Knowledge**	0.04	0.03	-0.20	-0.19	0.14	0.26	0.27	0.14	0.18	0.04	1.00	-	-	-	-	-	-	-	-	-
12	**SLO-Skills**	0.09	-0.09	0.10	0.10	0.01	0.02	0.04	0.22	-0.05	0.11	0.68	1.00	-	-	-	-	-	-	-	-
13	**SLO-Social/Emotional**	0.14	0.00	0.31	0.10	-0.02	0.21	0.30	0.27	-0.25	0.00	0.52	0.71	1.00	-	-	-	-	-	-	-
14	**Course Topics**	-0.02	-0.09	-0.10	0.02	0.02	0.12	0.20	-0.04	-0.04	-0.09	0.24	0.21	0.17	1.00	-	-	-	-	-	-
15	**Dates for Topics**	-0.12	0.00	-0.02	0.01	0.07	0.25	0.25	-0.04	-0.18	-0.23	0.24	0.02	0.17	0.88	1.00	-	-	-	-	-
16	**Total Enrollment**	0.07	0.39	0.22	0.08	-0.03	0.00	0.16	0.05	0.20	0.07	-0.05	-0.05	-0.02	0.08	0.12	1.00	-	-	-	-
17	**Cumulative Exams**	-0.07	-0.41	-0.05	-0.14	0.12	-0.15	-0.25	-0.12	0.08	0.14	-0.19	0.01	0.00	-0.08	-0.22	-0.14	1.00	-	-	-
18	**Grade Choice**	-0.06	-0.25	0.10	-0.11	0.30	-0.49	-0.30	-0.01	0.07	0.25	-0.12	0.22	-0.10	-0.19	-0.34	0.05	0.34	1.00	-	-
19	**Cumulative Final**	0.00	-0.27	0.05	0.12	0.12	-0.42	-0.21	0.11	0.03	0.18	-0.18	0.26	-0.11	-0.13	-0.42	-0.06	0.18	0.56	1.00	-
20	**Exam Dates**	0.04	0.05	0.02	0.00	0.18	-0.10	-0.04	0.07	0.24	0.24	0.06	0.17	0.10	0.62	0.54	0.07	0.29	0.25	-0.05	1.00
21	**Assignment Dates**	0.00	0.23	-0.03	-0.18	0.27	0.33	0.16	-0.16	-0.03	-0.25	0.11	-0.14	0.00	0.56	0.70	0.14	-0.19	-0.08	-0.26	0.53
22	**Projects/Presentations**	-0.13	-0.26	-0.04	0.01	-0.05	0.18	0.19	0.36	-0.21	-0.27	0.05	0.09	0.17	0.13	0.21	-0.16	-0.12	-0.23	0.02	-0.26
23	**Participation %**	0.07	0.01	0.00	0.17	-0.35	0.07	0.01	0.11	-0.18	0.06	-0.01	-0.01	0.16	0.05	0.11	-0.07	-0.20	-0.27	-0.16	-0.07
24	**Attendance Enforced**	0.07	-0.01	-0.03	0.01	-0.18	0.34	-0.24	-0.06	-0.16	0.01	0.24	0.10	0.19	0.09	0.21	-0.10	-0.14	-0.31	-0.15	-0.12
25	**Flipped Classroom**	0.07	-0.57	0.07	-0.02	0.01	-0.48	0.31	0.30	0.14	0.31	0.21	0.22	0.31	-0.06	-0.15	-0.09	0.40	0.34	0.22	0.00
26	**Extra Credit**	-0.04	-0.26	-0.23	-0.14	-0.07	-0.20	0.06	0.06	0.28	0.10	0.21	0.07	0.13	-0.05	-0.06	0.02	0.04	0.22	0.07	0.04
27	**In-Class Active**	0.11	-0.42	0.15	0.18	-0.16	-0.34	-0.15	0.14	-0.15	0.35	0.06	0.17	0.05	-0.03	-0.10	-0.22	0.07	0.18	0.20	-0.15
28	**Group Activities**	0.07	-0.26	0.05	0.19	-0.14	0.09	0.12	0.18	0.05	0.18	0.25	0.11	0.21	0.19	0.25	-0.17	-0.08	-0.40	-0.12	-0.19
29	**Informal RP**	0.10	-0.30	0.01	0.03	-0.07	-0.37	-0.18	0.14	0.08	0.40	-0.04	0.18	-0.04	-0.05	-0.30	-0.13	0.33	0.40	0.35	0.06
30	**Credit Hours**	0.01	0.05	0.30	0.09	0.03	0.02	0.07	-0.03	0.05	0.04	-0.10	-0.09	-0.03	0.02	-0.01	0.10	0.14	0.03	0.07	-0.06
31	**Course Level**	0.02	-0.10	-0.67	-0.48	0.05	0.23	-0.07	0.04	0.21	-0.11	0.05	-0.07	-0.03	0.03	-0.12	-0.15	0.22	0.14	0.18	-0.06
32	**Exam #**	-0.02	-0.14	0.04	-0.10	0.29	-0.30	-0.32	-0.13	-0.01	0.11	-0.13	0.06	-0.21	-0.11	-0.20	-0.08	0.54	0.76	0.31	0.31
33	**Exam %**	-0.09	-0.39	-0.06	-0.10	0.16	-0.31	-0.36	-0.50	0.02	0.02	-0.21	-0.01	-0.23	0.07	-0.14	-0.10	0.66	0.54	0.26	0.30
34	**Quiz #**	0.12	0.20	0.12	0.02	0.00	-0.14	0.12	0.15	0.15	0.18	0.10	0.13	0.07	-0.08	-0.04	0.15	-0.17	0.07	0.11	-0.02
35	**Quiz %**	0.06	0.32	0.08	0.00	-0.12	0.09	0.34	0.12	0.13	-0.02	0.16	0.03	0.12	0.04	0.11	0.27	-0.44	-0.33	-0.11	-0.13
36	**In-Class Assignment #**	0.03	-0.03	0.06	-0.09	0.16	0.01	0.02	-0.07	-0.04	0.11	0.12	0.08	0.08	-0.01	0.02	-0.03	0.09	0.15	0.05	0.08
37	**In-Class Assignment %**	0.06	0.04	0.02	-0.03	0.06	0.12	-0.01	-0.07	0.05	0.11	0.16	0.04	-0.04	-0.02	0.05	-0.05	-0.01	-0.03	-0.04	0.01
38	**Homework #**	0.12	0.16	-0.02	0.16	-0.03	-0.20	-0.15	0.74	0.17	0.20	0.02	0.07	0.09	-0.04	-0.05	0.02	-0.18	0.15	0.21	0.13
39	**Homework %**	0.02	0.25	0.01	0.09	0.00	0.27	0.18	0.64	-0.11	-0.08	0.11	-0.02	0.17	-0.15	0.08	-0.02	-0.49	-0.38	-0.20	-0.26

		21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
1	**Year**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	**Online/SMOC**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	**Core Course**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	**Flag Course**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	**Office Hours**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	**Reading Acts**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	**Watching Acts**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	**Doing Acts**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	**Social Media**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	**Community Learn Ops**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	**SLO-Knowledge**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	**SLO-Skills**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13	**SLO-Social/Emotional**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14	**Course Topics**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15	**Dates for Topics**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16	**Total Enrollment**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	**Cumulative Exams**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	**Grade Choice**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19	**Cumulative Final**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
20	**Exam Dates**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	**Assignment Dates**	1.00		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22	**Projects/Presentations**	0.03	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
23	**Participation %**	0.06	0.22	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
24	**Attendance Enforced**	0.25	0.25	0.61	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
25	**Flipped Classroom**	-0.17	0.14	0.04	0.03	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-
26	**Extra Credit**	-0.06	-0.16	-0.12	-0.07	0.38	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-
27	**In-Class Active**	-0.25	0.14	0.20	0.24	0.48	0.11	1.00	-	-	-	-	-	-	-	-	-	-	-	-
28	**Group Activities**	0.00	0.57	0.28	0.42	0.26	-0.03	0.50	1.00	-	-	-	-	-	-	-	-	-	-	-
29	**Informal RP**	-0.29	-0.13	0.02	0.01	0.49	0.21	0.73	-0.04	1.00	-	-	-	-	-	-	-	-	-	-
30	**Credit Hours**	-0.04	-0.03	-0.22	0.04	0.06	0.18	0.12	-0.01	0.21	1.00	-	-	-	-	-	-	-	-	-
31	**Course Level**	0.02	-0.05	-0.28	-0.18	0.22	0.50	-0.15	-0.14	0.12	0.00	1.00	-	-	-	-	-	-	-	-
32	**Exam #**	-0.04	-0.29	-0.27	-0.29	0.13	0.05	0.08	-0.35	0.24	0.16	0.04	1.00	-	-	-	-	-	-	-
33	**Exam %**	-0.10	-0.39	-0.36	-0.33	0.07	0.14	-0.04	-0.31	0.19	0.13	0.20	0.67	1.00	-	-	-	-	-	-
34	**Quiz #**	-0.12	-0.01	-0.04	0.10	0.12	-0.03	0.17	0.00	0.13	0.02	-0.13	-0.06	-0.26	1.00	-	-	-	-	-
35	**Quiz %**	0.01	-0.03	-0.04	-0.04	-0.07	0.01	-0.08	-0.04	-0.12	-0.02	-0.06	-0.46	-0.58	0.48	1.00	-	-	-	-
36	**In-Class Assignment #**	0.12	-0.03	-0.05	0.15	0.20	0.02	0.19	0.07	0.16	0.04	-0.02	0.08	-0.04	0.06	-0.08	1.00	-	-	-
37	**In-Class Assignment %**	0.04	0.04	-0.05	0.23	0.08	-0.11	0.19	0.21	0.04	0.06	-0.21	-0.07	-0.16	0.00	-0.07	0.49	1.00	-	-
38	**Homework #**	-0.02	-0.04	0.01	0.01	0.15	0.06	0.10	-0.05	0.12	-0.02	0.10	0.02	-0.26	0.15	0.06	0.03	-0.03	1.00	-
39	**Homework %**	0.07	0.47	0.20	0.18	-0.10	-0.12	-0.02	0.25	-0.22	-0.11	-0.09	-0.45	-0.77	0.00	0.08	-0.03	-0.04	0.38	1.00

Appendix C

Table C1 Overall mean differences and standardized mean differences for all covariates before and after propensity score adjustment (median split)

Variable	Unadjusted		Adjusted	
	M_{diff}	M_{diff} (standardized)	M_{diff}	M_{diff} (standardized)
SAT_equivalent	11.3758	0.0819	-2.6617	-0.0192
hspct2	0.0063	0.0568	-0.0015	-0.0136
transferredhours	-0.4833	-0.0387	0.0927	0.0074
age	-0.3535	-0.4399	0.0276	0.0343
sex_M	-0.0135	-0.0136	0.0040	0.0040
derivation_2eHB	0.0026	0.0025	-0.0005	-0.0005
derivation_AI	-0.0001	-0.0002	0.0000	0.0000
derivation_A	0.0349	0.0349	-0.0040	-0.0039
derivation_B2eH	-0.0013	-0.0013	-0.0006	-0.0006
derivation_B	-0.0041	-0.0041	0.0015	0.0015
derivation_F	-0.0049	-0.0048	0.0020	0.0020
derivation_HPI	-0.0012	-0.0012	-0.0002	-0.0002
derivation_H	-0.0142	-0.0141	0.0052	0.0051
derivation_U	0.0022	0.0022	0.0000	0.0000
derivation_W	-0.0138	-0.0138	-0.0034	-0.0034
majorschool_2	0.0269	0.0270	-0.0042	-0.0043
majorschool_3	-0.0048	-0.0048	0.0008	0.0007
majorschool_4	-0.0291	-0.0291	0.0013	0.0013
majorschool_5	-0.0014	-0.0014	0.0021	0.0022
majorschool_9	-0.0014	-0.0013	-0.0002	-0.0002
majorschool_C	-0.0244	-0.0244	-0.0003	-0.0003
majorschool_E	0.0434	0.0434	-0.0014	-0.0014
majorschool_J	-0.0041	-0.0041	0.0009	0.0009
majorschool_L	0.0198	0.0198	-0.0007	-0.0007
majorschool_N	-0.0005	-0.0005	0.0002	0.0002
majorschool_S	-0.0034	-0.0034	0.0002	0.0002
majorschool_U	-0.0211	-0.0211	0.0014	0.0014
motheredlevel_0	-0.0071	-0.0071	0.0015	0.0015
motheredlevel_1	-0.0057	-0.0057	0.0006	0.0006
motheredlevel_2	-0.0037	-0.0037	0.0010	0.0010
motheredlevel_3	-0.0072	-0.0073	0.0038	0.0038
motheredlevel_4	0.0045	0.0045	-0.0068	-0.0068
motheredlevel_5	0.0156	0.0156	-0.0025	-0.0024
motheredlevel_6	0.0042	0.0042	0.0028	0.0028
motheredlevel_U	-0.0004	-0.0004	-0.0003	-0.0003
fatheredlevel_0	-0.0048	-0.0049	0.0010	0.0010
fatheredlevel_1	-0.0061	-0.0061	0.0005	0.0006
fatheredlevel_2	-0.0048	-0.0048	0.0020	0.0020
fatheredlevel_3	0.0002	0.0002	0.0032	0.0032
fatheredlevel_4	-0.0097	-0.0097	-0.0063	-0.0064
fatheredlevel_5	0.0260	0.0260	-0.0009	-0.0009
fatheredlevel_6	0.0033	0.0033	-0.0005	-0.0005
fatheredlevel_U	-0.0041	-0.0040	0.0010	0.0010
CLASSIFICATION_1	0.1522	0.1522	-0.0076	-0.0076

CLASSIFICATION_2	-0.0821	-0.0821	0.0003	0.0003
CLASSIFICATION_3	-0.0553	-0.0553	0.0031	0.0031
CLASSIFICATION_4	-0.0148	-0.0148	0.0042	0.0042

Table C2 Variance ratios and K–S statistics for continuous covariates (median split)

Variable	Unadjusted		Adjusted	
	Variance ratio	K–S statistic	Variance ratio	K–S statistic
SAT_equivalent	1.1771	0.0456	1.0677	0.0177
hspct2	1.154	0.0362	1.075	0.0156
transferredhours	1.0643	0.0168	1.0481	0.0072
age	1.6593	0.2169	1.1245	0.0340

Table C3 Overall mean differences and standardized mean differences for all covariates before and after propensity score adjustment (mean split)

Variable	Unadjusted		Adjusted	
	M_{diff}	M_{diff} (standardized)	M_{diff}	M_{diff} (standardized)
SAT_equivalent	11.8997	0.0844	-0.1513	-0.0011
hspct2	0.0079	0.0704	0.0001	0.0006
transferredhours	-0.6893	-0.0547	-0.0226	-0.0018
age	-0.1097	-0.1298	-0.0016	-0.0018
sex_M	-0.0097	-0.0097	-0.0009	-0.0010
derivation_2eHB	0.0018	0.0018	-0.0001	-0.0001
derivation_AI	-0.0011	-0.0012	-0.0001	0.0000
derivation_A	0.0584	0.0584	0.0001	0.0001
derivation_B2eH	0.0018	0.0017	0.0001	0.0001
derivation_B	0.0007	0.0008	0.0000	0.0001
derivation_F	-0.0050	-0.0050	-0.0001	-0.0001
derivation_HPI	-0.0016	-0.0016	-0.0001	-0.0001
derivation_H	-0.0146	-0.0146	0.0001	0.0001
derivation_U	0.0012	0.0012	0.0000	0.0001
derivation_W	-0.0415	-0.0415	-0.0002	-0.0001
majorschool_2	-0.0202	-0.0202	-0.0005	-0.0004
majorschool_3	-0.0089	-0.0089	0.0000	0.0000
majorschool_4	0.0109	0.0108	-0.0001	-0.0001
majorschool_5	-0.0073	-0.0073	0.0000	0.0000
majorschool_9	-0.0002	-0.0003	0.0000	0.0000
majorschool_C	-0.0104	-0.0103	-0.0001	-0.0001
majorschool_E	0.0688	0.0688	0.0008	0.0008
majorschool_J	0.0023	0.0024	0.0000	0.0000
majorschool_L	-0.0080	-0.0080	-0.0001	-0.0002
majorschool_N	-0.0012	-0.0012	0.0000	0.0000
majorschool_S	-0.0034	-0.0033	0.0000	0.0000
majorschool_U	-0.0224	-0.0224	0.0000	-0.0001
motheredlevel_0	-0.0088	-0.0088	-0.0001	-0.0001
motheredlevel_1	0.0012	0.0012	-0.0001	-0.0001
motheredlevel_2	-0.0011	-0.0011	0.0005	0.0004
motheredlevel_3	-0.0095	-0.0095	-0.0001	-0.0001
motheredlevel_4	0.0021	0.0022	0.0007	0.0007
motheredlevel_5	0.0151	0.0151	-0.0007	-0.0007
motheredlevel_6	0.0029	0.0029	-0.0003	-0.0003
motheredlevel_U	-0.0020	-0.0020	0.0002	0.0001
fatheredlevel_0	-0.0044	-0.0044	-0.0002	-0.0002
fatheredlevel_1	-0.0028	-0.0028	0.0002	0.0001
fatheredlevel_2	-0.0065	-0.0065	0.0000	-0.0001
fatheredlevel_3	0.0026	0.0026	-0.0005	-0.0004
fatheredlevel_4	-0.0144	-0.0144	0.0012	0.0012
fatheredlevel_5	0.0265	0.0265	-0.0009	-0.0009
fatheredlevel_6	0.0018	0.0018	0.0002	0.0002
fatheredlevel_U	-0.0029	-0.0029	0.0000	0.0000
CLASSIFICATION_1	0.0309	0.0310	0.0022	0.0021
CLASSIFICATION_2	-0.0335	-0.0335	-0.0009	-0.0009
CLASSIFICATION_3	-0.0013	-0.0013	-0.0011	-0.0011
CLASSIFICATION_4	0.0039	0.0039	-0.0001	-0.0002

Table C4 Variance ratios and K–S statistics for continuous covariates (mean split)

Variable	Unadjusted		Adjusted	
	Variance ratio	K–S statistic	Variance ratio	K–S statistic
SAT_equivalent	1.1166	0.0402	1.0838	0.0202
hspct2	1.1634	0.0425	1.0244	0.0209
transferredhours	1.0818	0.0219	1.0561	0.0136
age	1.0272	0.0867	1.1199	0.0258

Table C5 Overall mean differences and standardized mean differences for all covariates in the Chemistry course sequence before and after propensity score adjustment (median split)

Variable	Unadjusted		Adjusted	
	M_{diff}	M_{diff} (standardized)	M_{diff}	M_{diff} (standardized)
SAT_equivalent	25.9553	0.1999	0.0155	0.0001
hspct2	0.0067	0.0862	0.0004	0.0055
transferredhours	-1.6689	-0.1343	0.0200	0.0016
age	-0.1472	-0.2390	-0.0077	-0.0125
sex_M	-0.0108	-0.0109	0.0008	0.0008
derivation_2eHB	0.0061	0.0061	0.0002	0.0002
derivation_AI	-0.0003	-0.0002	0.0000	0.0000
derivation_A	0.0688	0.0688	0.0004	0.0004
derivation_B2eH	0.0002	0.0001	-0.0001	-0.0001
derivation_B	-0.0083	-0.0084	0.0000	-0.0001
derivation_F	-0.0078	-0.0078	0.0006	0.0006
derivation_HPI	-0.0009	-0.0009	-0.0001	-0.0001
derivation_H	-0.0208	-0.0208	-0.0004	-0.0003
derivation_U	0.0032	0.0032	0.0000	0.0000
derivation_W	-0.0402	-0.0402	-0.0008	-0.0008
majorschool_2	0.0035	0.0035	-0.0002	-0.0002
majorschool_3	-0.0072	-0.0072	0.0003	0.0003
majorschool_4	-0.0093	-0.0093	0.0009	0.0009
majorschool_5	-0.0008	-0.0008	0.0001	0.0001
majorschool_9	-0.0004	-0.0004	0.0000	0.0000
majorschool_C	-0.0012	-0.0012	-0.0001	-0.0001
majorschool_E	0.0580	0.0580	-0.0007	-0.0006
majorschool_J	-0.0100	-0.0100	0.0003	0.0003
majorschool_L	-0.0048	-0.0048	0.0005	0.0005
majorschool_N	0.0017	0.0017	0.0001	0.0002
majorschool_S	-0.0015	-0.0015	0.0001	0.0001
majorschool_U	-0.0279	-0.0279	-0.0014	-0.0014
motheredlevel_0	-0.0077	-0.0077	-0.0005	-0.0005
motheredlevel_1	-0.0084	-0.0084	0.0003	0.0003
motheredlevel_2	-0.0091	-0.0090	-0.0008	-0.0008
motheredlevel_3	-0.0165	-0.0165	0.0011	0.0011
motheredlevel_4	0.0157	0.0157	-0.0011	-0.0011
motheredlevel_5	0.0271	0.0270	0.0003	0.0003
motheredlevel_6	-0.0018	-0.0018	0.0009	0.0009
motheredlevel_U	0.0006	0.0006	-0.0004	-0.0003
fatheredlevel_0	-0.0071	-0.0072	-0.0001	-0.0002
fatheredlevel_1	-0.0077	-0.0078	-0.0001	-0.0001
fatheredlevel_2	-0.0119	-0.0119	0.0002	0.0001
fatheredlevel_3	0.0015	0.0015	0.0005	0.0005
fatheredlevel_4	-0.0134	-0.0133	-0.0039	-0.0038
fatheredlevel_5	0.0438	0.0438	0.0029	0.0029
fatheredlevel_6	0.0018	0.0019	0.0003	0.0004

fatheredlevel_U	-0.0070	-0.0070	0.0002	0.0002
CLASSIFICATION_1	0.0279	0.0279	-0.0004	-0.0004
CLASSIFICATION_2	-0.0189	-0.0189	0.0012	0.0012
CLASSIFICATION_3	-0.0092	-0.0092	-0.0009	-0.0010
CLASSIFICATION_4	0.0002	0.0002	0.0001	0.0001

Table C6 Variance ratios and K–S statistics for continuous covariates in the Chemistry course sequence (median split)

Variable	Unadjusted		Adjusted	
	Variance ratio	K–S statistic	Variance ratio	K–S statistic
SAT_equivalent	1.1458	0.0836	1.0480	0.0124
hspct2	1.3138	0.0491	1.0224	0.0251
transferredhours	1.0489	0.0781	1.1020	0.0162
age	1.7223	0.0957	1.2500	0.0330

Table C7 Overall mean differences and standardized mean differences for all covariates in the Chemistry course sequence before and after propensity score adjustment (mean split)

Variable	Unadjusted		Adjusted	
	M_{diff}	M_{diff} (standardized)	M_{diff}	M_{diff} (standardized)
SAT_equivalent	19.4772	0.1477	-1.0735	-0.0081
hspct2	0.0114	0.1426	-0.0006	-0.0070
transferredhours	-1.4033	-0.1126	0.0256	0.0021
age	-0.2447	-0.3791	0.0088	0.0137
sex_M	-0.0363	-0.0363	-0.0025	-0.0025
derivation_2eHB	0.0007	0.0007	-0.0003	-0.0004
derivation_AI	0.0006	0.0007	-0.0001	-0.0001
derivation_A	0.0897	0.0898	-0.0036	-0.0036
derivation_B2eH	0.0043	0.0044	-0.0004	-0.0004
derivation_B	-0.0014	-0.0013	0.0007	0.0006
derivation_F	-0.0089	-0.0089	0.0010	0.0010
derivation_HPI	-0.0007	-0.0007	0.0000	0.0000
derivation_H	-0.0147	-0.0147	0.0014	0.0013
derivation_U	0.0026	0.0026	0.0001	0.0000
derivation_W	-0.0726	-0.0726	0.0014	0.0014
majorschool_2	-0.0020	-0.0020	0.0010	0.0010
majorschool_3	-0.0179	-0.0178	0.0003	0.0003
majorschool_4	-0.0095	-0.0095	-0.0015	-0.0015
majorschool_5	-0.0036	-0.0036	0.0002	0.0002
majorschool_9	-0.0010	-0.0010	0.0002	0.0002
majorschool_C	-0.0042	-0.0042	-0.0001	-0.0001
majorschool_E	0.0760	0.0760	-0.0023	-0.0023
majorschool_J	0.0017	0.0017	-0.0002	-0.0002
majorschool_L	0.0038	0.0039	0.0003	0.0003
majorschool_N	-0.0008	-0.0008	0.0002	0.0002
majorschool_S	-0.0010	-0.0009	-0.0001	-0.0001
majorschool_U	-0.0417	-0.0417	0.0019	0.0019
motheredlevel_0	-0.0044	-0.0044	-0.0011	-0.0011
motheredlevel_1	0.0023	0.0023	-0.0006	-0.0006
motheredlevel_2	-0.0004	-0.0004	0.0008	0.0008
motheredlevel_3	-0.0239	-0.0238	-0.0013	-0.0013
motheredlevel_4	0.0074	0.0074	0.0051	0.0051
motheredlevel_5	0.0236	0.0236	-0.0013	-0.0013
motheredlevel_6	-0.0041	-0.0041	-0.0007	-0.0008
motheredlevel_U	-0.0007	-0.0007	-0.0009	-0.0008
fatheredlevel_0	-0.0020	-0.0019	-0.0006	-0.0006
fatheredlevel_1	0.0001	0.0002	0.0001	0.0001
fatheredlevel_2	-0.0136	-0.0135	-0.0008	-0.0008
fatheredlevel_3	-0.0021	-0.0022	0.0002	0.0001
fatheredlevel_4	-0.0229	-0.0228	0.0014	0.0014
fatheredlevel_5	0.0411	0.0412	0.0000	0.0000
fatheredlevel_6	0.0027	0.0027	0.0009	0.0009

fatheredlevel_U	-0.0036	-0.0036	-0.0010	-0.0010
CLASSIFICATION_1	0.0679	0.0679	-0.0006	-0.0006
CLASSIFICATION_2	-0.0551	-0.0551	0.0024	0.0024
CLASSIFICATION_3	-0.0128	-0.0128	-0.0010	-0.0011
CLASSIFICATION_4	0.0000	0.0000	-0.0007	-0.0007

Table C8 Variance ratios and K–S statistics for continuous covariates in the Chemistry course sequence (mean split)

Variable	Unadjusted		Adjusted	
	Variance ratio	K–S statistic	Variance ratio	K–S statistic
SAT_equivalent	1.1300	0.06	1.0458	0.0176
hspct2	1.4082	0.0733	1.1243	0.0277
transferredhours	1.0011	0.065	1.1087	0.0203
age	1.7799	0.1698	1.0962	0.0509

Table C9 Overall mean differences and standardized mean differences for all covariates in the Economics course sequence before and after propensity score adjustment (median split)

Variable	Unadjusted		Adjusted	
	M_{diff}	M_{diff} (standardized)	M_{diff}	M_{diff} (standardized)
SAT_equivalent	-6.7560	-0.0521	-0.3312	-0.0026
hspct2	-0.0055	-0.0433	0.0001	0.0007
transferredhours	1.5896	0.1308	0.0379	0.0031
age	-0.0537	-0.0777	0.0010	0.0015
sex_M	0.0029	0.0029	0.0001	0.0001
derivation_2eHB	-0.0001	-0.0002	0.0002	0.0001
derivation_AI	0.0005	0.0005	0.0000	0.0000
derivation_A	0.0061	0.0061	-0.0002	-0.0002
derivation_B2eH	-0.0017	-0.0017	-0.0002	-0.0002
derivation_B	-0.0087	-0.0087	0.0002	0.0002
derivation_F	-0.0110	-0.0110	0.0015	0.0015
derivation_HPI	-0.0027	-0.0027	-0.0017	-0.0017
derivation_H	-0.0097	-0.0098	0.0010	0.0010
derivation_U	-0.0017	-0.0017	0.0001	0.0000
derivation_W	0.0291	0.0291	-0.0006	-0.0006
majorschool_2	0.0036	0.0036	-0.0012	-0.0011
majorschool_3	0.0042	0.0042	0.0002	0.0002
majorschool_4	-0.0111	-0.0111	0.0009	0.0009
majorschool_5	-0.0005	-0.0005	0.0000	0.0000
majorschool_9	-0.0006	-0.0006	0.0000	0.0000

majorschool_C	-0.0064	-0.0064	0.0002	0.0001
majorschool_E	-0.0074	-0.0074	-0.0004	-0.0005
majorschool_J	-0.0026	-0.0026	-0.0002	-0.0002
majorschool_L	0.0481	0.0480	0.0005	0.0005
majorschool_N	-0.0012	-0.0012	0.0001	0.0000
majorschool_S	-0.0018	-0.0018	-0.0001	-0.0001
majorschool_U	-0.0243	-0.0244	0.0000	0.0000
motheredlevel_0	-0.0160	-0.0159	0.0010	0.0010
motheredlevel_1	-0.0029	-0.0029	0.0002	0.0002
motheredlevel_2	0.0079	0.0079	0.0000	-0.0001
motheredlevel_3	0.0006	0.0006	-0.0005	-0.0005
motheredlevel_4	0.0023	0.0023	-0.0011	-0.0011
motheredlevel_5	0.0059	0.0059	0.0008	0.0008
motheredlevel_6	-0.0041	-0.0041	-0.0004	-0.0004
motheredlevel_U	0.0062	0.0062	0.0001	0.0001
fatheredlevel_0	-0.0089	-0.0089	0.0001	0.0001
fatheredlevel_1	-0.0048	-0.0049	-0.0001	0.0000
fatheredlevel_2	0.0085	0.0084	0.0008	0.0009
fatheredlevel_3	-0.0034	-0.0034	-0.0008	-0.0007
fatheredlevel_4	-0.0053	-0.0053	0.0003	0.0003
fatheredlevel_5	0.0058	0.0059	-0.0007	-0.0007
fatheredlevel_6	0.0024	0.0024	-0.0002	-0.0003
fatheredlevel_U	0.0057	0.0057	0.0004	0.0004
CLASSIFICATION_1	0.0028	0.0028	-0.0017	-0.0017
CLASSIFICATION_2	0.0001	0.0001	-0.0001	-0.0001
CLASSIFICATION_3	-0.0035	-0.0036	0.0014	0.0015
CLASSIFICATION_4	0.0008	0.0007	0.0003	0.0003

Table C10 Variance ratios and K–S statistics for continuous covariates in the Economics course sequence (median split)

Variable	Unadjusted		Adjusted	
	Variance ratio	K–S statistic	Variance ratio	K–S statistic
SAT_equivalent	1.2041	0.0678	1.1998	0.0473
hspct2	1.1288	0.0341	1.0134	0.0293
transferredhours	1.1000	0.0722	1.1232	0.0285
age	1.2692	0.0448	1.1024	0.0329

Table C11 Overall mean differences and standardized mean differences for all covariates in the Economics course sequence before and after propensity score adjustment (mean split)

Variable	Unadjusted		Adjusted	
	M_{diff}	M_{diff} (standardized)	M_{diff}	M_{diff} (standardized)
SAT_equivalent	-3.5187	-0.0270	-0.6407	-0.0049
hspct2	-0.0041	-0.0320	-0.0003	-0.0026
transferredhours	2.1742	0.1796	0.0099	0.0008
age	-0.0635	-0.0913	0.0016	0.0023
sex_M	0.0028	0.0028	0.0004	0.0004
derivation_2eHB	-0.0005	-0.0005	0.0000	0.0001
derivation_AI	-0.0002	-0.0002	-0.0001	0.0000
derivation_A	0.0199	0.0200	0.0000	-0.0001
derivation_B2eH	-0.0024	-0.0023	-0.0002	-0.0002
derivation_B	-0.0108	-0.0107	0.0001	0.0000
derivation_F	-0.0085	-0.0085	0.0006	0.0005
derivation_HPI	-0.0030	-0.0030	-0.0017	-0.0017
derivation_H	-0.0210	-0.0209	0.0005	0.0005
derivation_U	-0.0024	-0.0023	0.0001	0.0001
derivation_W	0.0286	0.0285	0.0007	0.0007
majorschool_2	0.0069	0.0069	-0.0007	-0.0007
majorschool_3	0.0034	0.0034	0.0003	0.0002
majorschool_4	-0.0057	-0.0056	0.0002	0.0002
majorschool_5	-0.0003	-0.0004	0.0000	0.0001
majorschool_9	0.0007	0.0007	-0.0002	-0.0002
majorschool_C	-0.0018	-0.0018	0.0000	0.0000
majorschool_E	-0.0036	-0.0036	-0.0007	-0.0007
majorschool_J	-0.0030	-0.0030	0.0000	0.0000
majorschool_L	0.0351	0.0351	0.0023	0.0022
majorschool_N	-0.0016	-0.0015	0.0001	0.0000
majorschool_S	-0.0023	-0.0023	-0.0001	-0.0001
majorschool_U	-0.0278	-0.0278	-0.0010	-0.0010
motheredlevel_0	-0.0165	-0.0165	0.0002	0.0002
motheredlevel_1	-0.0039	-0.0039	0.0000	0.0000
motheredlevel_2	0.0024	0.0024	0.0002	0.0003
motheredlevel_3	-0.0034	-0.0034	-0.0004	-0.0003
motheredlevel_4	0.0103	0.0103	-0.0008	-0.0008
motheredlevel_5	0.0126	0.0127	0.0004	0.0004
motheredlevel_6	-0.0061	-0.0061	-0.0002	-0.0002
motheredlevel_U	0.0046	0.0045	0.0005	0.0005
fatheredlevel_0	-0.0107	-0.0107	-0.0004	-0.0004
fatheredlevel_1	-0.0048	-0.0047	0.0000	0.0000
fatheredlevel_2	0.0033	0.0032	0.0008	0.0008
fatheredlevel_3	-0.0082	-0.0083	-0.0005	-0.0004
fatheredlevel_4	0.0040	0.0039	-0.0007	-0.0007
fatheredlevel_5	0.0112	0.0112	0.0003	0.0002
fatheredlevel_6	0.0011	0.0011	-0.0004	-0.0004

fatheredlevel_U	0.0041	0.0042	0.0009	0.0009
CLASSIFICATION_1	-0.0003	-0.0003	0.0003	0.0003
CLASSIFICATION_2	0.0019	0.0019	-0.0014	-0.0015
CLASSIFICATION_3	-0.0013	-0.0013	0.0006	0.0007
CLASSIFICATION_4	-0.0002	-0.0003	0.0005	0.0005

Table C12 Variance ratios and K–S statistics for continuous covariates in the Economics course sequence (mean split)

Variable	Unadjusted		Adjusted	
	Variance ratio	K–S statistic	Variance ratio	K–S statistic
SAT_equivalent	1.2271	0.0579	1.2013	0.0451
hspct2	1.0804	0.0304	1.0152	0.0282
transferredhours	1.1129	0.0982	1.2181	0.0462
age	1.3671	0.0362	1.1512	0.0365

Table C13 Overall mean differences and standardized mean differences for all covariates in the Government course sequence before and after propensity score adjustment (median split)

Variable	Unadjusted		Adjusted	
	M_{diff}	M_{diff} (standardized)	M_{diff}	M_{diff} (standardized)
SAT_equivalent	-2.4817	-0.0159	1.0689	0.0068
hspct2	0.0121	0.0862	-0.0006	-0.0047
transferredhours	0.8010	0.0768	0.0535	0.0051
age	-0.2121	-0.2597	-0.0017	-0.0021
sex_M	0.0040	0.0040	-0.0026	-0.0026
derivation_2eHB	0.0029	0.0029	-0.0006	-0.0007
derivation_AI	-0.0014	-0.0013	0.0000	0.0000
derivation_A	0.0004	0.0004	0.0022	0.0023
derivation_B2eH	-0.0019	-0.0019	0.0000	0.0000
derivation_B	0.0131	0.0131	0.0062	0.0062
derivation_F	-0.0018	-0.0018	0.0034	0.0034
derivation_HPI	-0.0016	-0.0016	-0.0013	-0.0013
derivation_H	0.0366	0.0367	0.0033	0.0033
derivation_U	0.0033	0.0033	0.0010	0.0009
derivation_W	-0.0496	-0.0497	-0.0142	-0.0142
majorschool_2	0.0140	0.0140	-0.0024	-0.0024
majorschool_3	-0.0043	-0.0042	-0.0005	-0.0005
majorschool_4	-0.0040	-0.0040	-0.0021	-0.0020
majorschool_5	0.0193	0.0193	0.0015	0.0015
majorschool_9	-0.0033	-0.0033	-0.0026	-0.0026
majorschool_C	-0.0243	-0.0243	0.0039	0.0038
majorschool_E	-0.0170	-0.0170	0.0014	0.0014
majorschool_J	0.0069	0.0069	-0.0011	-0.0012
majorschool_L	0.0230	0.0230	-0.0042	-0.0042
majorschool_N	-0.0016	-0.0016	0.0006	0.0006
majorschool_S	-0.0098	-0.0098	0.0062	0.0062
majorschool_U	0.0011	0.0011	-0.0006	-0.0005
motheredlevel_0	-0.0017	-0.0017	0.0008	0.0008
motheredlevel_1	0.0072	0.0073	0.0008	0.0008
motheredlevel_2	0.0057	0.0057	0.0170	0.0170
motheredlevel_3	-0.0019	-0.0018	0.0025	0.0024
motheredlevel_4	-0.0399	-0.0399	-0.0092	-0.0092
motheredlevel_5	0.0287	0.0287	-0.0065	-0.0065
motheredlevel_6	0.0177	0.0177	-0.0017	-0.0017
motheredlevel_U	-0.0159	-0.0159	-0.0037	-0.0037
fatheredlevel_0	0.0031	0.0032	0.0012	0.0012
fatheredlevel_1	-0.0015	-0.0015	-0.0015	-0.0015
fatheredlevel_2	-0.0003	-0.0003	0.0136	0.0135
fatheredlevel_3	0.0193	0.0193	0.0077	0.0077
fatheredlevel_4	-0.0575	-0.0575	-0.0136	-0.0136
fatheredlevel_5	0.0421	0.0421	-0.0031	-0.0031
fatheredlevel_6	0.0138	0.0138	-0.0014	-0.0014

fatheredlevel_U	-0.0190	-0.0191	-0.0028	-0.0028
CLASSIFICATION_1	0.0298	0.0298	0.0039	0.0039
CLASSIFICATION_2	0.0195	0.0195	0.0009	0.0009
CLASSIFICATION_3	-0.0433	-0.0433	-0.0052	-0.0051
CLASSIFICATION_4	-0.0060	-0.0060	0.0003	0.0003

Table C14 Variance ratios and K–S statistics for continuous covariates in the Government course sequence (median split)

Variable	Unadjusted		Adjusted	
	Variance ratio	K–S statistic	Variance ratio	K–S statistic
SAT_equivalent	1.0419	0.0378	1.0144	0.0308
hspct2	1.2653	0.0482	1.0007	0.0326
transferredhours	1.073	0.0429	1.065	0.0179
age	1.1217	0.1463	1.0856	0.0595

Table C15 Overall mean differences and standardized mean differences for all covariates in the Government course sequence before and after propensity score adjustment (mean split)

Variable	Unadjusted		Adjusted	
	M_{diff}	M_{diff} (standardized)	M_{diff}	M_{diff} (standardized)
SAT_equivalent	18.1128	0.1167	-1.5251	-0.0098
hspct2	0.0014	0.0092	0.0002	0.0013
transferredhours	-1.0701	-0.1020	0.0420	0.0040
age	-0.0305	-0.0366	-0.0037	-0.0044
sex_M	0.0321	0.0321	-0.0032	-0.0032
derivation_2eHB	0.0009	0.0009	-0.0016	-0.0017
derivation_AI	-0.0067	-0.0066	0.0003	0.0003
derivation_A	0.0248	0.0248	0.0023	0.0023
derivation_B2eH	0.0017	0.0017	0.0007	0.0007
derivation_B	0.0086	0.0086	-0.0004	-0.0004
derivation_F	0.0025	0.0025	-0.0018	-0.0018
derivation_HPI	-0.0036	-0.0036	-0.0015	-0.0015
derivation_H	-0.0321	-0.0322	0.0031	0.0031
derivation_U	0.0001	0.0001	0.0001	0.0001
derivation_W	0.0039	0.0038	-0.0012	-0.0012
majorschool_2	0.0403	0.0402	-0.0056	-0.0055
majorschool_3	-0.0101	-0.0100	0.0000	0.0000
majorschool_4	0.0029	0.0029	0.0004	0.0004
majorschool_5	-0.0082	-0.0082	0.0002	0.0002
majorschool_9	0.0002	0.0002	-0.0003	-0.0002
majorschool_C	-0.0092	-0.0092	0.0033	0.0033
majorschool_E	0.0019	0.0019	-0.0015	-0.0015
majorschool_J	-0.0061	-0.0060	0.0000	0.0000
majorschool_L	-0.0145	-0.0145	0.0004	0.0005
majorschool_N	-0.0031	-0.0031	-0.0009	-0.0008
majorschool_S	-0.0141	-0.0141	-0.0003	-0.0004
majorschool_U	0.0201	0.0200	0.0041	0.0041
motheredlevel_0	-0.0114	-0.0114	0.0008	0.0008
motheredlevel_1	-0.0077	-0.0077	-0.0001	-0.0001
motheredlevel_2	-0.0128	-0.0128	0.0041	0.0041
motheredlevel_3	-0.0031	-0.0032	-0.0002	-0.0002
motheredlevel_4	0.0045	0.0045	-0.0018	-0.0019
motheredlevel_5	0.0119	0.0119	-0.0003	-0.0003
motheredlevel_6	0.0246	0.0246	-0.0024	-0.0024
motheredlevel_U	-0.0059	-0.0059	0.0001	0.0000
fatheredlevel_0	-0.0058	-0.0058	0.0013	0.0013
fatheredlevel_1	-0.0037	-0.0037	0.0012	0.0012
fatheredlevel_2	-0.0104	-0.0104	0.0010	0.0010
fatheredlevel_3	0.0122	0.0122	-0.0009	-0.0009
fatheredlevel_4	-0.0286	-0.0286	0.0006	0.0006
fatheredlevel_5	0.0544	0.0544	-0.0034	-0.0034
fatheredlevel_6	-0.0075	-0.0076	0.0006	0.0006

fatheredlevel_U	-0.0106	-0.0106	-0.0004	-0.0004
CLASSIFICATION_1	0.0003	0.0003	-0.0007	-0.0007
CLASSIFICATION_2	0.0339	0.0339	0.0002	0.0003
CLASSIFICATION_3	-0.0285	-0.0285	0.0000	0.0000
CLASSIFICATION_4	-0.0056	-0.0057	0.0004	0.0004

Table C16 Variance ratios and K–S statistics for continuous covariates in the Government course sequence (mean split)

Variable	Unadjusted		Adjusted	
	Variance ratio	K–S statistic	Variance ratio	K–S statistic
SAT_equivalent	1.0092	0.0578	1.0018	0.0354
hspct2	1.0287	0.0208	1.0056	0.0251
transferredhours	1.2474	0.0358	1.0907	0.0361
age	1.0628	0.0549	1.1227	0.0372

Appendix D

Table D1 Full regression output for fixed-effects models (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.312	0.759	-1.728	0.084	.	-1.446	0.946	-1.528	0.127	
High RP	0.069	0.025	2.732	0.006	**	0.060	0.025	2.448	0.014	*
SAT equivalent	0.001	0.000	8.843	<.001	***	0.001	0.000	9.193	<.001	***
derivationAI	0.093	0.139	0.672	0.501		0.075	0.138	0.543	0.587	
derivationA	-0.050	0.037	-1.349	0.177		-0.010	0.037	-0.278	0.781	
derivationB2eH	-0.002	0.090	-0.022	0.982		-0.007	0.092	-0.075	0.940	
derivationB	-0.083	0.048	-1.715	0.086	.	-0.015	0.048	-0.321	0.748	
derivationF	0.067	0.058	1.159	0.247		0.109	0.057	1.931	0.053	.
derivationHPI	0.152	0.182	0.836	0.403		0.283	0.189	1.495	0.135	
derivationH	-0.083	0.038	-2.169	0.030	*	-0.051	0.038	-1.350	0.177	
derivationU	-0.263	0.112	-2.350	0.019	*	-0.255	0.110	-2.312	0.021	*
derivationW	-0.008	0.036	-0.232	0.816		0.027	0.036	0.747	0.455	
motheredlevel1	0.013	0.054	0.244	0.807		0.021	0.053	0.395	0.693	
motheredlevel2	-0.056	0.048	-1.165	0.244		-0.034	0.048	-0.717	0.474	
motheredlevel3	-0.008	0.050	-0.153	0.878		0.024	0.049	0.489	0.625	
motheredlevel4	0.009	0.048	0.186	0.853		0.020	0.047	0.416	0.677	
motheredlevel5	0.024	0.049	0.483	0.629		0.023	0.048	0.466	0.641	
motheredlevel6	0.030	0.053	0.574	0.566		0.048	0.052	0.929	0.353	
motheredlevelU	0.067	0.070	0.957	0.339		0.092	0.069	1.338	0.181	
fatheredlevel1	-0.002	0.056	-0.030	0.976		-0.090	0.055	-1.619	0.106	
fatheredlevel2	0.037	0.050	0.744	0.457		0.022	0.049	0.446	0.655	
fatheredlevel3	0.008	0.051	0.152	0.879		-0.021	0.050	-0.408	0.683	
fatheredlevel4	0.104	0.049	2.121	0.034	*	0.093	0.049	1.911	0.056	.
fatheredlevel5	0.115	0.050	2.303	0.021	*	0.100	0.049	2.020	0.043	*
fatheredlevel6	0.100	0.057	1.756	0.079	.	0.079	0.057	1.401	0.161	
fatheredlevelU	0.017	0.068	0.252	0.801		-0.020	0.066	-0.301	0.764	
age	-0.046	0.012	-3.975	<.001	***	-0.044	0.012	-3.791	0.000	***
class_zscore.x	0.726	0.009	79.509	<.001	***	0.728	0.009	80.350	<.001	***
CLASSIFICATION2	0.021	0.018	1.131	0.258		0.028	0.018	1.534	0.125	
CLASSIFICATION3	0.015	0.034	0.433	0.665		-0.046	0.033	-1.394	0.163	
CLASSIFICATION4	0.027	0.059	0.466	0.641		0.096	0.057	1.664	0.096	.
hspct2	0.441	0.064	6.884	0.000	***	0.474	0.063	7.514	0.000	***
sexW	0.017	0.013	1.254	0.210		0.007	0.013	0.515	0.607	
transferredhours	0.002	0.001	3.248	0.001	**	0.001	0.001	2.506	0.012	*
majorschool3	0.012	0.054	0.214	0.830		0.057	0.053	1.077	0.281	
majorschool4	-0.020	0.028	-0.718	0.473		-0.018	0.028	-0.639	0.523	
majorschool5	-0.099	0.058	-1.714	0.087	.	-0.046	0.056	-0.827	0.408	
majorschool9	0.109	0.182	0.599	0.549		0.087	0.186	0.468	0.640	
majorschoolC	0.010	0.039	0.252	0.801		0.033	0.039	0.839	0.402	
majorschoolE	0.052	0.025	2.093	0.036	*	0.050	0.025	2.010	0.044	*
majorschoolJ	0.013	0.063	0.212	0.832		-0.060	0.062	-0.971	0.332	
majorschoolL	0.055	0.026	2.092	0.036		0.054	0.026	2.070	0.038	*
majorschoolN	-0.048	0.086	-0.553	0.580		-0.090	0.085	-1.066	0.287	
majorschoolS	-0.022	0.095	-0.230	0.818		-0.062	0.094	-0.658	0.511	
majorschoolU	0.040	0.026	1.567	0.117		0.044	0.026	1.724	0.085	.
HRS_UNDERTAKEN.y	0.033	0.003	10.656	<.001	***	0.030	0.003	9.804	<.001	

Note. *N*=13332*; RMSE* = 0.7077; *df* = 12809, R^2 = .4635*; *p < .05; **p < .01; ***p < .001*

Table D2 Full regression output for random-effects models (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.093	0.265	-4.131	<.001	***	-1.085	0.265	-4.103	<.001	***
High RP	0.057	0.032	1.786	0.045	*	0.067	0.030	2.247	0.027	
SAT equivalent	0.001	0.000	8.407	<.001	***	0.001	0.000	8.376	<.001	***
derivationAI	0.098	0.137	0.714	0.479		0.092	0.136	0.678	0.498	
derivationA	-0.052	0.036	-1.433	0.147		-0.064	0.036	-1.762	0.078	
derivationB2eH	-0.004	0.089	-0.045	0.948		-0.023	0.090	-0.261	0.794	
derivationB	-0.072	0.048	-1.515	0.129		-0.079	0.048	-1.652	0.099	
derivationF	0.058	0.057	1.024	0.303		0.058	0.057	1.023	0.307	.
derivationHPI	0.152	0.181	0.841	0.404		0.078	0.184	0.426	0.670	
derivationH	-0.080	0.038	-2.118	0.034	*	-0.087	0.038	-2.315	0.021	
derivationU	-0.278	0.111	-2.494	0.013	*	-0.279	0.112	-2.496	0.013	*
derivationW	-0.007	0.035	-0.211	0.833		-0.012	0.035	-0.344	0.731	
motheredlevel1	-0.013	0.053	-0.237	0.819		-0.036	0.011	-3.152	0.002	**
motheredlevel2	-0.074	0.048	-1.541	0.125		0.002	0.052	0.034	0.973	
motheredlevel3	-0.025	0.049	-0.512	0.618		-0.078	0.047	-1.636	0.102	
motheredlevel4	-0.008	0.047	-0.161	0.881		-0.032	0.049	-0.650	0.516	
motheredlevel5	0.003	0.049	0.059	0.945		-0.016	0.047	-0.336	0.737	
motheredlevel6	0.019	0.052	0.369	0.699		-0.002	0.048	-0.036	0.971	
motheredlevelU	0.060	0.070	0.858	0.389		0.019	0.052	0.368	0.713	
fatheredlevel1	0.021	0.055	0.386	0.705		0.062	0.069	0.902	0.367	
fatheredlevel2	0.054	0.049	1.100	0.273		0.024	0.055	0.433	0.665	
fatheredlevel3	0.033	0.050	0.650	0.522		0.061	0.049	1.234	0.217	
fatheredlevel4	0.123	0.048	2.541	0.011	*	0.044	0.050	0.881	0.378	
fatheredlevel5	0.132	0.049	2.701	0.007	**	0.132	0.048	2.724	0.006	*
fatheredlevel6	0.123	0.056	2.185	0.030	*	0.138	0.049	2.822	0.005	*
fatheredlevelU	0.019	0.067	0.287	0.775		0.115	0.056	2.037	0.042	.
age	-0.036	0.011	-3.143	0.001	**	0.015	0.067	0.232	0.817	
class_zscore.x	0.721	0.009	80.194	<.001	***	0.722	0.009	80.068	<.001	***
CLASSIFICATION2	0.018	0.018	1.033	0.384		0.007	0.018	0.395	0.693	
CLASSIFICATION3	0.003	0.032	0.085	0.926		0.003	0.032	0.099	0.921	.
CLASSIFICATION4	-0.006	0.057	-0.099	0.851		0.012	0.057	0.219	0.827	
hspct2	0.438	0.063	6.964	<.001	***	0.407	0.063	6.465	0.000	***
sexW	0.017	0.013	1.297	0.202		0.020	0.013	1.545	0.122	
transferredhours	0.002	0.001	3.106	0.002	**	0.002	0.001	3.074	0.002	*
majorschool3	0.009	0.053	0.164	0.851		0.009	0.053	0.164	0.870	
majorschool4	-0.029	0.027	-1.044	0.287		-0.030	0.027	-1.097	0.272	
majorschool5	-0.083	0.056	-1.466	0.144		-0.083	0.056	-1.463	0.144	
majorschool9	0.185	0.175	1.058	0.284		0.182	0.173	1.052	0.293	
majorschoolC	0.002	0.038	0.047	0.996		0.002	0.037	0.066	0.947	
majorschoolE	0.046	0.024	1.908	0.058	.	0.047	0.024	1.934	0.053	.
majorschoolJ	-0.013	0.063	-0.201	0.851		-0.038	0.062	-0.605	0.545	
majorschoolL	0.054	0.026	2.090	0.036	*	0.048	0.026	1.864	0.062	*
majorschoolN	-0.018	0.085	-0.216	0.819		-0.007	0.085	-0.084	0.933	
majorschoolS	-0.047	0.094	-0.498	0.619		-0.038	0.093	-0.408	0.683	
majorschoolU	0.040	0.025	1.590	0.113		0.040	0.025	1.572	0.116	.
HRS_UNDERTAKEN.y	0.031	0.003	10.471	<.001	***	0.032	0.003	10.715	<.001	***

Note. *MSE = .49803;* $s_p^2 = 0.01963$; $s_s^2 = 0.0162$; *$p < .05$; **$p < .01$; ***$p < .001$

Table D3 Full regression output for cluster-robust standard errors models (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*d.f.*	p-value		Estimate	*SE*	*d.f.*	p-value	
Intercept	-1.163	0.270	196.9	< 0.001	***	-1.265	0.320	90.94	<0.001	***
High RP	0.036	0.018	119.4	0.056	.	0.035	0.019	126.76	0.061	.
SAT equivalent	0.001	0.000	196.8	< 0.001	***	0.001	0.000	103.64	<0.001	***
derivationAI	0.150	0.099	31.9	0.138		0.105	0.129	16.11	0.427	
derivationA	-0.040	0.031	118.5	0.200		-0.003	0.035	93.71	0.942	
derivationB2eH	0.000	0.103	61.4	0.998		-0.002	0.095	33.20	0.980	
derivationB	-0.062	0.047	147.7	0.189		0.003	0.062	75.22	0.961	
derivationF	0.077	0.058	141.6	0.184		0.120	0.062	57.17	0.058	.
derivationHPI	0.152	0.136	16.1	0.279		0.273	0.138	2.20	0.174	
derivationH	-0.068	0.033	129.7	0.044	*	-0.041	0.038	97.96	0.273	
derivationU	-0.267	0.110	31.6	0.021	*	-0.256	0.111	31.21	0.028	*
derivationW	0.009	0.031	123.8	0.775		0.037	0.035	94.88	0.291	
motheredlevel1	-0.031	0.011	193.5	0.007	**	-0.028	0.014	85.07	0.048	*
motheredlevel2	-0.025	0.058	129.8	0.670		-0.001	0.068	64.45	0.984	
motheredlevel3	-0.076	0.055	130.0	0.169		-0.043	0.069	48.64	0.535	
motheredlevel4	-0.023	0.056	130.9	0.682		0.013	0.068	46.77	0.843	
motheredlevel5	-0.004	0.054	122.6	0.940		0.018	0.064	43.74	0.776	
motheredlevel6	0.004	0.054	125.6	0.934		0.018	0.065	44.28	0.781	
motheredlevelU	0.018	0.057	132.8	0.749		0.048	0.068	51.33	0.480	
fatheredlevel1	0.060	0.081	103.7	0.462		0.103	0.087	30.22	0.242	
fatheredlevel2	0.031	0.067	133.7	0.648		-0.058	0.080	69.36	0.466	
fatheredlevel3	0.056	0.060	142.6	0.351		0.036	0.074	61.24	0.634	
fatheredlevel4	0.041	0.062	144.2	0.514		0.003	0.076	67.84	0.966	
fatheredlevel5	0.129	0.059	133.7	0.031	*	0.111	0.072	61.32	0.128	
fatheredlevel6	0.137	0.060	135.0	0.024	*	0.114	0.072	62.97	0.121	
fatheredlevelU	0.126	0.068	139.3	0.067	.	0.100	0.080	68.69	0.216	
age	0.030	0.083	112.1	0.720		-0.019	0.093	28.18	0.839	
class_zscore.x	0.705	0.017	170.7	< 0.001	***	0.707	0.018	79.60	<0.001	***
CLASSIFICATION2	0.016	0.017	263.6	0.362		0.023	0.020	139.99	0.254	
CLASSIFICATION3	-0.002	0.031	233.4	0.950		-0.055	0.043	67.41	0.208	
CLASSIFICATION4	-0.029	0.060	119.7	0.625		0.024	0.091	13.89	0.793	
hspct2	0.421	0.076	277.5	< 0.001	***	0.455	0.102	86.06	<0.001	***
sexW	0.013	0.014	192.5	0.330		0.000	0.016	136.50	0.993	
transferredhours	0.002	0.001	191.2	0.006	**	0.001	0.001	113.35	0.083	.
majorschool3	-0.026	0.054	123.0	0.627		0.027	0.060	45.95	0.660	
majorschool4	-0.035	0.026	257.9	0.177		-0.029	0.029	139.84	0.322	
majorschool5	-0.077	0.065	173.3	0.240		-0.019	0.075	46.00	0.798	
majorschool9	0.195	0.136	13.2	0.174		0.185	0.086	1.90	0.170	
majorschoolC	0.007	0.041	242.3	0.871		0.008	0.051	74.68	0.870	
majorschoolE	0.027	0.024	132.9	0.259		0.033	0.024	111.09	0.175	
majorschoolJ	-0.042	0.066	49.0	0.524		-0.126	0.108	10.63	0.271	
majorschoolL	0.048	0.027	218.6	0.078	.	0.050	0.029	165.89	0.081	.
majorschoolN	0.012	0.113	69.1	0.915		-0.029	0.116	10.81	0.809	
majorschoolS	-0.015	0.089	51.1	0.866		-0.060	0.106	8.88	0.584	
majorschoolU	0.034	0.025	213.3	0.180		0.043	0.027	138.07	0.117	
C_HRS_UNDERTAKEN.y	0.031	0.003	228.6	< 0.001	***	0.028	0.004	128.12	<0.001	***

Note. *p < .05; **p < .01; ***p < .001

Table D4 Full regression output for fixed-effects models (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.297	0.759	-1.708	0.088	.	-1.214	0.789	-1.538	0.124	
High RP	0.069	0.029	2.393	0.017	*	0.067	0.029	2.352	0.019	*
SAT equivalent	0.001	0.000	8.831	<.001	***	0.001	0.000	8.890	<.001	***
derivationAI	0.097	0.139	0.698	0.485		0.085	0.139	0.608	0.543	
derivationA	-0.049	0.037	-1.340	0.180		-0.059	0.037	-1.612	0.107	
derivationB2eH	-0.004	0.090	-0.047	0.962		-0.016	0.091	-0.171	0.864	
derivationB	-0.083	0.048	-1.713	0.087	.	-0.088	0.048	-1.833	0.067	.
derivationF	0.067	0.058	1.158	0.247		0.054	0.057	0.942	0.346	
derivationHPI	0.151	0.182	0.829	0.407		0.057	0.188	0.304	0.761	
derivationH	-0.083	0.038	-2.180	0.029	*	-0.093	0.038	-2.447	0.014	*
derivationU	-0.264	0.112	-2.363	0.018	*	-0.218	0.113	-1.928	0.054	*
derivationW	-0.008	0.036	-0.225	0.822		-0.015	0.036	-0.432	0.666	
motheredlevel1	0.012	0.054	0.224	0.823		0.023	0.054	0.437	0.662	
motheredlevel2	-0.058	0.048	-1.194	0.232		-0.063	0.048	-1.306	0.192	
motheredlevel3	-0.009	0.050	-0.187	0.851		-0.011	0.050	-0.223	0.823	
motheredlevel4	0.007	0.048	0.149	0.881		0.003	0.048	0.054	0.957	
motheredlevel5	0.022	0.049	0.443	0.658		0.016	0.049	0.324	0.746	
motheredlevel6	0.028	0.053	0.537	0.591		0.030	0.053	0.575	0.565	
motheredlevelU	0.066	0.070	0.931	0.352		0.052	0.070	0.744	0.457	
fatheredlevel1	-0.002	0.056	-0.042	0.966		0.001	0.056	0.020	0.984	
fatheredlevel2	0.037	0.050	0.733	0.464		0.042	0.050	0.838	0.402	
fatheredlevel3	0.008	0.051	0.156	0.876		0.015	0.051	0.292	0.770	
fatheredlevel4	0.104	0.049	2.118	0.034	*	0.109	0.049	2.218	0.027	*
fatheredlevel5	0.115	0.050	2.298	0.022	*	0.121	0.050	2.419	0.016	*
fatheredlevel6	0.101	0.057	1.774	0.076	.	0.103	0.057	1.806	0.071	.
fatheredlevelU	0.018	0.068	0.258	0.796		0.029	0.068	0.434	0.664	
age	-0.047	0.012	-4.022	<.001	***	-0.052	0.012	-4.464	<.001	***
class_zscore.x	0.727	0.009	79.542	<.001	***	0.722	0.009	78.529	<.001	***
CLASSIFICATION2	0.021	0.018	1.134	0.257		0.017	0.018	0.915	0.360	
CLASSIFICATION3	0.015	0.034	0.448	0.654		0.026	0.034	0.787	0.431	
CLASSIFICATION4	0.028	0.059	0.478	0.632		0.047	0.058	0.802	0.422	
hspct2	0.441	0.064	6.875	<.001	***	0.434	0.064	6.776	<.001	***
sexW	0.017	0.013	1.275	0.202		0.016	0.013	1.218	0.223	
transferredhours	0.002	0.001	3.267	0.001	**	0.002	0.001	3.541	<.001	***
majorschool3	0.011	0.054	0.210	0.833		0.013	0.054	0.233	0.816	
majorschool4	-0.020	0.028	-0.704	0.482		-0.018	0.029	-0.623	0.534	
majorschool5	-0.095	0.058	-1.637	0.102	.	-0.099	0.058	-1.705	0.088	.
majorschool9	0.106	0.182	0.580	0.562		0.108	0.181	0.595	0.552	
majorschoolC	0.009	0.039	0.242	0.809		0.007	0.039	0.175	0.861	
majorschoolE	0.053	0.025	2.122	0.034	*	0.054	0.025	2.176	0.030	*
majorschoolJ	0.015	0.063	0.231	0.818		0.037	0.063	0.585	0.559	
majorschoolL	0.055	0.026	2.102	0.036		0.056	0.026	2.133	0.033	*
majorschoolN	-0.044	0.086	-0.510	0.610		-0.033	0.086	-0.383	0.701	
majorschoolS	-0.019	0.095	-0.205	0.838		-0.010	0.095	-0.110	0.912	
majorschoolU	0.040	0.026	1.564	0.118		0.037	0.026	1.456	0.145	
HRS_UNDERTAKEN.y	0.032	0.003	10.626	<.001	***	0.033	0.003	10.906	<.001	***

Note. N=13332; $RMSE$ = 0.7078; df = 12809, R^2 = .4634; *p < .05; **p < .01; ***p < .001

Table D5 Full regression output for random-effects models (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.088	0.264	-4.118	0.000	***	-1.025	0.262	-3.911	0.000	***
High RP	0.056	0.029	1.947	0.054	.	0.057	0.029	1.943	0.055	.
SAT equivalent	0.001	0.000	8.398	<.001	***	0.001	0.000	8.402	<.001	***
derivationAI	0.099	0.137	0.722	0.471		0.077	0.138	0.561	0.575	
derivationA	-0.053	0.036	-1.441	0.150		-0.062	0.036	-1.708	0.088	.
derivationB2eH	-0.006	0.089	-0.064	0.949		-0.014	0.090	-0.155	0.877	
derivationB	-0.072	0.048	-1.521	0.128		-0.077	0.048	-1.611	0.107	
derivationF	0.058	0.057	1.026	0.305		0.048	0.057	0.856	0.392	
derivationHPI	0.153	0.181	0.846	0.398		0.060	0.187	0.323	0.747	
derivationH	-0.080	0.038	-2.118	0.034	*	-0.089	0.038	-2.369	0.018	*
derivationU	-0.278	0.111	-2.500	0.012	*	-0.231	0.113	-2.053	0.040	*
derivationW	-0.007	0.035	-0.203	0.839		-0.015	0.035	-0.414	0.679	
motheredlevel1	-0.013	0.053	-0.254	0.800		-0.040	0.011	-3.531	<.001	***
motheredlevel2	-0.074	0.048	-1.553	0.120		0.001	0.053	0.015	0.988	
motheredlevel3	-0.026	0.049	-0.522	0.602		-0.079	0.048	-1.646	0.100	.
motheredlevel4	-0.008	0.047	-0.175	0.861		-0.027	0.049	-0.545	0.586	
motheredlevel5	0.002	0.049	0.045	0.964		-0.012	0.047	-0.249	0.803	
motheredlevel6	0.019	0.052	0.363	0.717		-0.003	0.049	-0.057	0.955	
motheredlevelU	0.060	0.070	0.854	0.393		0.022	0.052	0.424	0.672	
fatheredlevel1	0.021	0.055	0.384	0.701		0.046	0.070	0.660	0.509	
fatheredlevel2	0.054	0.049	1.097	0.273		0.023	0.055	0.426	0.670	
fatheredlevel3	0.032	0.050	0.640	0.522		0.058	0.049	1.193	0.233	
fatheredlevel4	0.123	0.048	2.537	0.011	*	0.038	0.050	0.762	0.446	
fatheredlevel5	0.132	0.049	2.696	0.007	**	0.127	0.048	2.645	0.008	**
fatheredlevel6	0.123	0.056	2.184	0.029	*	0.139	0.049	2.841	0.005	**
fatheredlevelU	0.019	0.067	0.286	0.775		0.124	0.056	2.211	0.027	*
age	-0.037	0.011	-3.199	0.001	**	0.032	0.067	0.484	0.629	
class_zscore.x	0.721	0.009	80.210	<.001	***	0.717	0.009	79.339	<.001	***
CLASSIFICATION2	0.017	0.018	0.971	0.331		0.015	0.018	0.823	0.411	
CLASSIFICATION3	0.000	0.032	-0.005	0.996		0.013	0.032	0.392	0.695	
CLASSIFICATION4	-0.009	0.057	-0.157	0.875		0.015	0.057	0.270	0.787	
hspct2	0.439	0.063	6.968	<.001	***	0.430	0.063	6.833	<.001	***
sexW	0.017	0.013	1.302	0.193		0.017	0.013	1.318	0.187	
transferredhours	0.002	0.001	3.129	0.002	**	0.002	0.001	3.350	0.001	***
majorschool3	0.009	0.053	0.166	0.868		0.015	0.053	0.278	0.781	
majorschool4	-0.030	0.027	-1.097	0.273		-0.029	0.027	-1.052	0.293	
majorschool5	-0.081	0.056	-1.428	0.153		-0.081	0.057	-1.431	0.152	
majorschool9	0.184	0.175	1.054	0.292		0.191	0.174	1.103	0.270	
majorschoolC	0.000	0.038	0.006	0.995		0.000	0.038	-0.008	0.993	
majorschoolE	0.046	0.024	1.894	0.058	.	0.047	0.024	1.958	0.050	.
majorschoolJ	-0.015	0.063	-0.232	0.817		0.010	0.063	0.156	0.876	
majorschoolL	0.054	0.026	2.102	0.036	*	0.056	0.026	2.180	0.029	*
majorschoolN	-0.019	0.085	-0.224	0.823		-0.003	0.085	-0.029	0.976	
majorschoolS	-0.047	0.094	-0.503	0.615		-0.044	0.094	-0.466	0.641	
majorschoolU	0.039	0.025	1.561	0.119		0.036	0.025	1.432	0.152	
HRS_UNDERTAKEN.y	0.031	0.003	10.452	<.001	*	0.032	0.003	10.685	<.001	***

Note. *MSE = .49795;* $s_p^2 = 0.01934$; $s_s^2 = 0.01664$; **p < .05; **p < .01; ***p < .001*

Table D6 Full regression output for cluster-robust standard errors models (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*d.f.*	p-value		Estimate	*SE*	*d.f.*	p-value	
Intercept	-1.117	0.270	201.8	< 0.001	***	-1.063	0.282	142.21	< 0.001	***
High RP	0.019	0.017	230.4	0.273		0.018	0.017	210.91	0.302	
SAT equivalent	0.001	0.000	196.1	< 0.001	***	0.001	0.000	233.92	< 0.001	***
derivationAI	0.151	0.098	31.9	0.135		0.125	0.107	29.87	0.252	
derivationA	-0.041	0.031	118.5	0.191		-0.052	0.032	104.99	0.111	
derivationB2eH	-0.004	0.103	61.4	0.969		-0.005	0.102	52.93	0.958	
derivationB	-0.064	0.047	147.8	0.180		-0.069	0.048	157.49	0.150	
derivationF	0.076	0.058	141.7	0.188		0.065	0.060	132.65	0.279	
derivationHPI	0.148	0.138	16.1	0.300		0.067	0.186	7.16	0.729	
derivationH	-0.068	0.033	129.7	0.042	*	-0.078	0.034	118.61	0.023	*
derivationU	-0.265	0.110	31.6	0.022	*	-0.217	0.125	19.30	0.100	.
derivationW	0.009	0.031	123.7	0.782		0.002	0.032	107.23	0.946	
motheredlevel1	-0.033	0.011	197.6	0.004	**	-0.036	0.012	117.09	0.003	**
motheredlevel2	-0.026	0.059	129.7	0.657		-0.013	0.061	144.31	0.830	
motheredlevel3	-0.076	0.055	130.0	0.167		-0.083	0.057	134.82	0.147	
motheredlevel4	-0.023	0.056	130.8	0.683		-0.025	0.058	133.47	0.662	
motheredlevel5	-0.004	0.054	122.5	0.941		-0.010	0.055	125.72	0.860	
motheredlevel6	0.005	0.054	125.4	0.932		-0.004	0.056	128.92	0.943	
motheredlevelU	0.019	0.057	132.5	0.745		0.019	0.060	136.89	0.749	
fatheredlevel1	0.061	0.081	103.6	0.452		0.048	0.079	105.46	0.546	
fatheredlevel2	0.030	0.067	133.7	0.653		0.035	0.069	156.52	0.615	
fatheredlevel3	0.056	0.060	142.6	0.350		0.062	0.063	157.70	0.328	
fatheredlevel4	0.041	0.062	144.2	0.516		0.049	0.065	158.85	0.453	
fatheredlevel5	0.129	0.059	133.7	0.032	*	0.137	0.061	149.28	0.027	*
fatheredlevel6	0.137	0.060	135.0	0.024	*	0.147	0.063	151.88	0.020	*
fatheredlevelU	0.127	0.068	139.3	0.065	.	0.133	0.069	162.54	0.058	.
age	0.029	0.083	112.1	0.730		0.044	0.083	120.06	0.596	
class_zscore.x	0.705	0.016	170.8	< 0.001	***	0.700	0.015	217.42	< 0.001	***
CLASSIFICATION2	0.014	0.017	265.1	0.433		0.012	0.018	281.70	0.517	
CLASSIFICATION3	-0.007	0.031	231.6	0.816		0.009	0.032	202.77	0.787	
CLASSIFICATION4	-0.033	0.060	120.3	0.576		-0.004	0.059	117.67	0.943	
hspct2	0.423	0.075	277.9	< 0.001	***	0.415	0.074	270.52	< 0.001	***
sexW	0.013	0.014	192.8	0.346		0.014	0.014	217.24	0.306	
transferredhours	0.002	0.001	190.9	0.005	**	0.002	0.001	221.47	0.003	**
majorschool3	-0.026	0.054	123.1	0.633		-0.020	0.055	122.05	0.709	
majorschool4	-0.038	0.026	257.5	0.137		-0.043	0.026	253.86	0.104	
majorschool5	-0.074	0.065	173.4	0.258		-0.077	0.066	166.74	0.249	
majorschool9	0.191	0.136	13.2	0.184		0.195	0.139	12.91	0.184	
majorschoolC	0.003	0.041	241.8	0.938		0.003	0.041	239.16	0.936	
majorschoolE	0.025	0.024	136.5	0.298		0.027	0.024	139.68	0.266	
majorschoolJ	-0.048	0.066	49.0	0.472		-0.020	0.064	49.25	0.758	
majorschoolL	0.049	0.027	219.1	0.075	.	0.049	0.028	219.64	0.079	.
majorschoolN	0.011	0.113	69.1	0.923		0.032	0.118	65.07	0.784	
majorschoolS	-0.019	0.089	51.1	0.831		-0.016	0.088	46.25	0.856	
majorschoolU	0.033	0.025	213.4	0.197		0.027	0.025	210.54	0.282	
HRS_UNDERTAKEN.y	0.031	0.003	228.7	< 0.001	***	0.032	0.003	274.93	< 0.001	***

Note. *p < .05; **p < .01; ***p < .001

Table D7 Full regression output for fixed-effects models in the Chemistry course sequence (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.154	0.405	-2.850	0.004	**	-1.393	0.403	-3.459	0.001	
High RP	0.091	0.031	2.914	0.004	**	0.071	0.031	2.320	0.020	*
SAT equivalent	0.000	0.000	3.978	0.000	***	0.000	0.000	3.954	0.000	***
derivationAI	0.181	0.240	0.756	0.450		0.282	0.237	1.188	0.235	
derivationA	-0.042	0.047	-0.897	0.370		-0.032	0.047	-0.680	0.497	
derivationB2eH	-0.018	0.125	-0.142	0.887		-0.019	0.126	-0.154	0.878	
derivationB	-0.065	0.064	-1.011	0.312	.	-0.038	0.064	-0.604	0.546	
derivationF	0.048	0.078	0.616	0.538		0.163	0.077	2.119	0.034	.
derivationHPI	0.014	0.240	0.057	0.954		0.165	0.242	0.682	0.495	
derivationH	-0.047	0.050	-0.944	0.345		-0.047	0.049	-0.946	0.344	
derivationU	-0.272	0.132	-2.064	0.039	*	-0.230	0.132	-1.734	0.083	*
derivationW	0.001	0.047	0.026	0.980		0.004	0.046	0.077	0.939	
motheredlevel1	-0.048	0.070	-0.686	0.493		-0.051	0.070	-0.724	0.469	
motheredlevel2	-0.092	0.065	-1.427	0.154		-0.046	0.065	-0.709	0.478	
motheredlevel3	-0.049	0.066	-0.740	0.460		-0.009	0.066	-0.130	0.896	
motheredlevel4	-0.064	0.064	-1.014	0.310		-0.029	0.064	-0.448	0.654	
motheredlevel5	-0.032	0.065	-0.497	0.619		0.002	0.065	0.038	0.970	
motheredlevel6	-0.045	0.069	-0.651	0.515		-0.009	0.069	-0.130	0.896	
motheredlevelU	-0.069	0.094	-0.736	0.462		-0.040	0.093	-0.429	0.668	
fatheredlevel1	-0.028	0.074	-0.377	0.706		-0.053	0.075	-0.709	0.478	
fatheredlevel2	-0.007	0.067	-0.103	0.918		-0.033	0.067	-0.484	0.628	
fatheredlevel3	-0.018	0.069	-0.260	0.795		-0.055	0.068	-0.796	0.426	
fatheredlevel4	0.104	0.066	1.578	0.115		0.071	0.066	1.078	0.281	.
fatheredlevel5	0.120	0.067	1.800	0.072	.	0.083	0.067	1.249	0.212	*
fatheredlevel6	0.167	0.075	2.225	0.026	*	0.112	0.075	1.501	0.133	
fatheredlevelU	0.108	0.091	1.185	0.236		0.066	0.091	0.731	0.465	
age	-0.051	0.017	-2.985	0.003	**	-0.041	0.017	-2.353	0.019	***
class_zscore.x	0.847	0.013	64.959	< 2e-16	***	0.848	0.013	65.642	< 2e-16	***
CLASSIFICATION2	0.018	0.027	0.659	0.510		0.025	0.027	0.940	0.347	
CLASSIFICATION3	-0.050	0.065	-0.766	0.444		-0.101	0.065	-1.555	0.120	
CLASSIFICATION4	0.218	0.175	1.242	0.214		0.175	0.168	1.038	0.299	.
hspct2	0.503	0.118	4.256	0.000	***	0.587	0.118	4.963	0.000	***
sexW	0.036	0.018	1.975	0.048	*	0.039	0.018	2.175	0.030	
transferredhours	0.002	0.001	2.851	0.004	**	0.002	0.001	2.433	0.015	*
majorschool3	0.061	0.082	0.747	0.455		0.071	0.081	0.869	0.385	
majorschool4	-0.025	0.056	-0.441	0.660		-0.026	0.056	-0.456	0.649	
majorschool5	-0.179	0.127	-1.413	0.158		-0.225	0.127	-1.774	0.076	
majorschool9	0.290	0.342	0.849	0.396		0.241	0.339	0.710	0.478	
majorschoolC	0.113	0.113	1.003	0.316		0.109	0.113	0.964	0.335	
majorschoolE	0.071	0.053	1.324	0.185		0.064	0.054	1.191	0.234	*
majorschoolJ	0.082	0.083	0.980	0.327		0.060	0.083	0.722	0.470	
majorschoolL	0.022	0.061	0.360	0.719		0.013	0.061	0.221	0.825	*
majorschoolN	0.031	0.122	0.257	0.797		0.049	0.121	0.406	0.685	
majorschoolS	0.308	0.181	1.699	0.089	.	0.245	0.178	1.373	0.170	
majorschoolU	0.054	0.058	0.939	0.348		0.059	0.058	1.021	0.307	.
C_HRS_UNDERTAKEN.y	0.036	0.004	8.288	< 2e-16	***	0.036	0.004	8.280	< 2e-16	

Note. $n = 6251$; *RMSE* = 0.6625; $df = 6117$, $R^2 = .5140$; *$p < .05$; **$p < .01$; ***$p < .001$

Table D8 Full regression output for random-effects models in the Chemistry course sequence (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.171	0.399	-2.936	<.001	***	-1.442	0.400	-3.611	0.000	***
High RP	0.156	0.103	1.507	0.077	.	0.151	0.104	1.450	0.159	
SAT equivalent	0.000	0.000	3.348	<.001	***	0.000	0.000	3.349	0.001	***
derivationAI	0.174	0.237	0.734	0.476		0.262	0.234	1.119	0.263	
derivationA	-0.050	0.047	-1.073	0.152		-0.046	0.046	-0.986	0.324	
derivationB2eH	-0.026	0.123	-0.215	0.964		-0.029	0.124	-0.230	0.818	
derivationB	-0.067	0.063	-1.059	0.130		-0.046	0.063	-0.733	0.464	
derivationF	0.031	0.077	0.405	0.306		0.136	0.076	1.783	0.075	.
derivationHPI	0.005	0.236	0.020	0.401		0.159	0.238	0.666	0.505	
derivationH	-0.044	0.049	-0.897	0.034	*	-0.050	0.049	-1.035	0.301	
derivationU	-0.303	0.130	-2.328	0.013	*	-0.253	0.131	-1.931	0.054	.
derivationW	-0.004	0.046	-0.084	0.833		-0.008	0.046	-0.166	0.868	
motheredlevel1	-0.059	0.069	-0.857	0.812		-0.014	0.017	-0.841	0.400	
motheredlevel2	-0.096	0.064	-1.504	0.123		-0.060	0.069	-0.872	0.383	
motheredlevel3	-0.061	0.065	-0.950	0.608		-0.047	0.064	-0.737	0.461	
motheredlevel4	-0.067	0.062	-1.077	0.872		-0.018	0.065	-0.279	0.780	
motheredlevel5	-0.035	0.064	-0.551	0.953		-0.028	0.063	-0.450	0.653	
motheredlevel6	-0.057	0.068	-0.830	0.712		0.002	0.064	0.025	0.980	
motheredlevelU	-0.065	0.092	-0.701	0.391		-0.019	0.068	-0.274	0.784	
fatheredlevel1	-0.011	0.073	-0.154	0.699		-0.027	0.092	-0.292	0.771	
fatheredlevel2	0.013	0.066	0.201	0.271		-0.029	0.073	-0.394	0.694	
fatheredlevel3	0.002	0.067	0.024	0.516		-0.012	0.066	-0.175	0.861	
fatheredlevel4	0.112	0.065	1.734	0.011	*	-0.035	0.067	-0.520	0.603	
fatheredlevel5	0.129	0.065	1.976	0.007	**	0.077	0.065	1.198	0.231	
fatheredlevel6	0.179	0.074	2.432	0.029	*	0.094	0.065	1.439	0.150	
fatheredlevelU	0.116	0.090	1.289	0.774		0.124	0.073	1.682	0.093	.
age	-0.025	0.017	-1.474	0.002	**	0.068	0.089	0.767	0.443	
class_zscore.x	0.858	0.013	66.479	<.001	***	0.857	0.013	67.127	<.001	***
CLASSIFICATION2	0.024	0.026	0.894	0.302		0.031	0.026	1.172	0.241	
CLASSIFICATION3	-0.055	0.064	-0.854	0.932		-0.104	0.064	-1.622	0.105	
CLASSIFICATION4	0.204	0.173	1.180	0.921		0.164	0.166	0.988	0.323	
hspct2	0.486	0.117	4.172	0.000	***	0.566	0.117	4.847	0.000	***
sexW	0.040	0.018	2.258	0.195		0.043	0.018	2.455	0.014	*
transferredhours	0.002	0.001	2.643	0.002	**	0.002	0.001	2.205	0.027	*
majorschool3	0.092	0.081	1.136	0.870		0.094	0.080	1.170	0.242	
majorschool4	-0.022	0.055	-0.399	0.297		-0.023	0.055	-0.422	0.673	
majorschool5	-0.098	0.126	-0.782	0.143		-0.138	0.126	-1.101	0.271	
majorschool9	0.401	0.338	1.186	0.290		0.393	0.335	1.174	0.240	
majorschoolC	0.136	0.111	1.224	0.963		0.123	0.111	1.105	0.269	
majorschoolE	0.063	0.053	1.205	0.056	.	0.056	0.053	1.069	0.285	
majorschoolJ	0.072	0.083	0.863	0.840		0.044	0.083	0.532	0.595	
majorschoolL	0.028	0.060	0.467	0.037	*	0.016	0.060	0.259	0.796	
majorschoolN	0.036	0.121	0.299	0.829		0.058	0.119	0.489	0.625	
majorschoolS	0.278	0.179	1.554	0.618		0.203	0.176	1.153	0.249	
majorschoolU	0.063	0.057	1.104	0.112		0.066	0.057	1.154	0.249	
C_HRS_UNDERTAKEN.y	0.034	0.004	7.949	<.001	***	0.034	0.004	7.989	0.000	***

Note. *MSE = .42613;* $s_p^2 = 0.08561$; s_s^2*=0.05498; *p < .05; **p < .01; ***p < .001*

Table D9 Full regression output for cluster-robust standard errors models in the Chemistry course sequence (median split) before and after adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*d.f.*	p-value		Estimate	*SE*	*d.f.*	p-value	
Intercept	-0.758	0.354	47.7	0.038	*	-1.117	0.386	44.69	0.006	**
High RP	0.062	0.028	41.5	0.032	*	0.063	0.028	43.98	0.027	*
SAT equivalent	0.000	0.000	54.3	0.001	**	0.000	0.000	48.45	0.003	**
derivationAI	0.219	0.162	7.5	0.217		0.291	0.215	5.29	0.232	
derivationA	-0.034	0.036	38.9	0.352		-0.024	0.037	37.55	0.527	
derivationB2eH	0.030	0.130	18.7	0.818		0.027	0.119	19.07	0.823	
derivationB	-0.042	0.059	45.2	0.480		-0.023	0.063	39.88	0.719	
derivationF	0.062	0.080	46.5	0.438		0.181	0.089	28.91	0.052	.
derivationHPI	0.041	0.169	7.5	0.815		0.186	0.200	2.73	0.426	
derivationH	-0.024	0.038	41.4	0.527		-0.025	0.038	38.74	0.514	
derivationU	-0.278	0.136	17.0	0.056	.	-0.228	0.138	16.25	0.118	
derivationW	0.012	0.037	40.3	0.750		0.013	0.038	38.80	0.740	
motheredlevel1	-0.053	0.016	44.1	0.002	**	-0.037	0.019	39.67	0.057	.
motheredlevel2	-0.083	0.074	39.4	0.268		-0.082	0.078	36.18	0.296	
motheredlevel3	-0.106	0.069	39.4	0.133		-0.056	0.075	36.06	0.462	
motheredlevel4	-0.059	0.077	40.2	0.442		-0.021	0.084	36.38	0.807	
motheredlevel5	-0.069	0.070	37.6	0.329		-0.035	0.077	34.48	0.651	
motheredlevel6	-0.041	0.072	38.4	0.570		-0.009	0.077	35.65	0.908	
motheredlevelU	-0.059	0.073	40.9	0.420		-0.022	0.078	36.43	0.781	
fatheredlevel1	-0.077	0.102	30.7	0.456		-0.038	0.100	26.23	0.708	
fatheredlevel2	-0.012	0.092	39.9	0.893		-0.037	0.098	37.73	0.709	
fatheredlevel3	0.015	0.086	44.3	0.863		-0.007	0.089	40.80	0.938	
fatheredlevel4	0.010	0.091	44.1	0.914		-0.025	0.095	40.20	0.791	
fatheredlevel5	0.122	0.086	41.6	0.161		0.089	0.090	38.79	0.330	
fatheredlevel6	0.136	0.087	41.6	0.124		0.102	0.089	38.77	0.258	
fatheredlevelU	0.175	0.099	44.8	0.083	.	0.121	0.098	41.00	0.226	
age	0.130	0.117	33.4	0.274		0.076	0.114	26.28	0.508	
class_zscore.x	0.820	0.018	39.9	<.001	***	0.821	0.019	36.34	<.001	***
CLASSIFICATION2	0.015	0.023	64.4	0.519		0.022	0.024	51.11	0.355	
CLASSIFICATION3	-0.036	0.074	43.0	0.626		-0.102	0.086	29.95	0.244	
CLASSIFICATION4	0.228	0.134	13.4	0.112		0.178	0.131	6.66	0.220	
hspct2	0.497	0.132	63.3	<.001	***	0.571	0.135	49.17	<.001	***
sexW	0.036	0.019	53.6	0.069	.	0.039	0.020	48.37	0.059	.
transferredhours	0.002	0.001	52.2	0.025	*	0.001	0.001	44.21	0.081	.
majorschool3	0.057	0.080	58.2	0.478		0.065	0.086	34.69	0.453	
majorschool4	0.008	0.054	53.6	0.887		0.004	0.054	47.65	0.948	
majorschool5	-0.160	0.177	27.6	0.373		-0.182	0.166	17.62	0.287	
majorschool9	0.432	0.076	2.0	0.030	*	0.382	0.091	1.33	0.100	.
majorschoolC	0.110	0.111	37.6	0.331		0.099	0.106	22.68	0.359	
majorschoolE	0.071	0.053	46.0	0.189		0.064	0.052	41.28	0.229	
majorschoolJ	0.076	0.091	45.0	0.410		0.025	0.092	35.33	0.791	
majorschoolL	0.018	0.061	56.7	0.770		0.010	0.058	47.79	0.871	
majorschoolN	0.098	0.153	39.8	0.524		0.103	0.147	37.48	0.488	
majorschoolS	0.248	0.099	10.6	0.030	*	0.170	0.125	4.60	0.238	
majorschoolU	0.067	0.056	55.2	0.233		0.073	0.058	48.50	0.208	
C_HRS_UNDERTAKEN.y	0.034	0.005	58.6	<.001	***	0.034	0.005	52.10	<.001	***

Note. *p < .05; **p < .01; ***p < .001

Table D10 Full regression output for fixed-effects models in the Chemistry course sequence (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.194	0.405	-2.946	0.003	***	-1.360	0.387	-3.512	<.001	***
High RP	0.123	0.035	3.561	0.000	**	0.100	0.034	2.924	0.003	**
SAT equivalent	0.000	0.000	4.005	0.000	***	0.000	0.000	4.015	0.000	***
derivationAI	0.189	0.240	0.790	0.429		0.106	0.236	0.448	0.654	
derivationA	-0.042	0.047	-0.881	0.378		-0.054	0.047	-1.134	0.257	
derivationB2eH	-0.025	0.125	-0.203	0.839		-0.165	0.124	-1.334	0.182	
derivationB	-0.067	0.064	-1.041	0.298		-0.050	0.064	-0.779	0.436	
derivationF	0.050	0.078	0.642	0.521		0.008	0.076	0.108	0.914	
derivationHPI	0.010	0.239	0.042	0.967		0.005	0.244	0.019	0.985	
derivationH	-0.048	0.050	-0.963	0.336		-0.059	0.050	-1.179	0.238	
derivationU	-0.277	0.132	-2.107	0.035	*	-0.174	0.133	-1.306	0.192	
derivationW	0.002	0.047	0.033	0.974		-0.001	0.047	-0.028	0.977	
motheredlevel1	-0.051	0.070	-0.732	0.464		-0.025	0.070	-0.357	0.721	
motheredlevel2	-0.096	0.065	-1.481	0.139		-0.068	0.064	-1.054	0.292	
motheredlevel3	-0.052	0.066	-0.790	0.430		-0.018	0.065	-0.274	0.784	
motheredlevel4	-0.069	0.064	-1.083	0.279		-0.048	0.063	-0.772	0.440	
motheredlevel5	-0.037	0.065	-0.575	0.565		-0.012	0.064	-0.180	0.857	
motheredlevel6	-0.050	0.069	-0.723	0.470		-0.024	0.069	-0.347	0.729	
motheredlevelU	-0.075	0.094	-0.801	0.423		-0.081	0.092	-0.877	0.381	
fatheredlevel1	-0.028	0.074	-0.377	0.706		-0.050	0.075	-0.675	0.500	
fatheredlevel2	-0.006	0.067	-0.091	0.928		-0.025	0.067	-0.371	0.710	
fatheredlevel3	-0.018	0.069	-0.259	0.796		-0.035	0.068	-0.510	0.610	
fatheredlevel4	0.105	0.066	1.601	0.109		0.071	0.066	1.084	0.278	
fatheredlevel5	0.120	0.067	1.805	0.071	.	0.092	0.067	1.381	0.167	
fatheredlevel6	0.167	0.075	2.235	0.025	*	0.127	0.075	1.697	0.090	.
fatheredlevelU	0.110	0.091	1.210	0.226		0.111	0.090	1.227	0.220	
age	-0.051	0.017	-2.977	0.003	**	-0.049	0.016	-3.064	0.002	**
class_zscore.x	0.848	0.013	65.014	<.001	***	0.841	0.013	63.485	<.001	***
CLASSIFICATION2	0.019	0.027	0.719	0.472		0.009	0.027	0.318	0.750	
CLASSIFICATION3	-0.049	0.065	-0.761	0.447		-0.041	0.065	-0.634	0.526	
CLASSIFICATION4	0.212	0.175	1.213	0.225		-0.058	0.160	-0.363	0.717	
hspct2	0.499	0.118	4.224	<.001	***	0.541	0.118	4.600	<.001	***
sexW	0.036	0.018	2.011	0.044	*	0.033	0.018	1.838	0.066	.
transferredhours	0.002	0.001	2.844	0.004	**	0.003	0.001	3.554	<.001	***
majorschool3	0.059	0.082	0.729	0.466		0.058	0.082	0.704	0.482	
majorschool4	-0.026	0.056	-0.470	0.639		-0.018	0.057	-0.323	0.747	
majorschool5	-0.174	0.127	-1.373	0.170		-0.170	0.128	-1.329	0.184	
majorschool9	0.293	0.341	0.859	0.390		0.215	0.310	0.693	0.488	
majorschoolC	0.111	0.113	0.987	0.324		0.080	0.113	0.713	0.476	
majorschoolE	0.070	0.053	1.305	0.192		0.076	0.055	1.392	0.164	
majorschoolJ	0.077	0.083	0.928	0.353		0.090	0.084	1.074	0.283	
majorschoolL	0.016	0.061	0.269	0.788		-0.003	0.062	-0.049	0.961	
majorschoolN	0.029	0.122	0.235	0.814		0.094	0.123	0.765	0.444	
majorschoolS	0.307	0.181	1.694	0.090	.	0.266	0.184	1.448	0.148	
majorschoolU	0.052	0.058	0.898	0.369		0.046	0.059	0.775	0.438	
C_HRS_UNDERTAKEN.y	0.036	0.004	8.301	<.001	***	0.039	0.004	8.947	<.001	***

Note. $n = 6251$; *RMSE* = 0.6622; $df = 6117$, $R^2 = .5143$; *$p < .05$; **$p < .01$; ***$p < .001$

Table D11 Full regression output for random-effects models in the Chemistry course sequence (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.244	0.400	-3.113	0.002	**	-1.309	0.378	-3.466	<.001	***
High RP	0.263	0.091	2.890	0.007	**	0.241	0.086	2.810	0.009	**
SAT equivalent	0.000	0.000	3.333	<.001	***	0.000	0.000	3.441	<.001	***
derivationAI	0.172	0.237	0.728	0.467		0.121	0.233	0.521	0.602	
derivationA	-0.050	0.047	-1.073	0.283		-0.063	0.047	-1.342	0.180	
derivationB2eH	-0.029	0.123	-0.233	0.816		-0.173	0.122	-1.413	0.158	
derivationB	-0.066	0.063	-1.057	0.291		-0.048	0.063	-0.757	0.449	
derivationF	0.032	0.077	0.418	0.676		-0.009	0.075	-0.123	0.902	
derivationHPI	0.005	0.236	0.022	0.983		-0.007	0.241	-0.031	0.976	
derivationH	-0.044	0.049	-0.899	0.369		-0.053	0.049	-1.080	0.280	
derivationU	-0.304	0.130	-2.341	0.019	*	-0.173	0.132	-1.316	0.188	
derivationW	-0.004	0.046	-0.080	0.936		-0.008	0.046	-0.176	0.860	
motheredlevel1	-0.059	0.069	-0.858	0.391		-0.026	0.016	-1.621	0.105	
motheredlevel2	-0.096	0.064	-1.503	0.133		-0.032	0.068	-0.475	0.635	
motheredlevel3	-0.061	0.065	-0.946	0.344		-0.069	0.063	-1.103	0.270	
motheredlevel4	-0.067	0.062	-1.080	0.280		-0.032	0.064	-0.502	0.616	
motheredlevel5	-0.035	0.064	-0.553	0.580		-0.049	0.062	-0.789	0.430	
motheredlevel6	-0.056	0.068	-0.824	0.410		-0.012	0.063	-0.186	0.852	
motheredlevelU	-0.065	0.092	-0.706	0.480		-0.032	0.067	-0.475	0.635	
fatheredlevel1	-0.012	0.073	-0.165	0.869		-0.072	0.091	-0.791	0.429	
fatheredlevel2	0.012	0.066	0.188	0.851		-0.026	0.073	-0.359	0.720	
fatheredlevel3	0.001	0.067	0.008	0.993		0.000	0.066	0.004	0.997	
fatheredlevel4	0.111	0.065	1.722	0.085	.	-0.010	0.067	-0.153	0.878	
fatheredlevel5	0.128	0.065	1.961	0.050	*	0.083	0.064	1.292	0.196	
fatheredlevel6	0.178	0.074	2.419	0.016	*	0.109	0.065	1.681	0.093	.
fatheredlevelU	0.115	0.090	1.285	0.199		0.148	0.073	2.023	0.043	*
age	-0.025	0.017	-1.450	0.147		0.128	0.089	1.435	0.151	
class_zscore.x	0.858	0.013	66.512	<.001	***	0.853	0.013	65.096	<.001	***
CLASSIFICATION2	0.024	0.026	0.906	0.365		0.013	0.026	0.499	0.618	
CLASSIFICATION3	-0.055	0.064	-0.853	0.394		-0.044	0.064	-0.697	0.486	
CLASSIFICATION4	0.202	0.173	1.168	0.243		0.031	0.159	0.192	0.847	
hspct2	0.484	0.117	4.150	<.001	***	0.533	0.116	4.593	<.001	***
sexW	0.040	0.018	2.259	0.024	*	0.038	0.018	2.132	0.033	*
transferredhours	0.002	0.001	2.642	0.008	**	0.003	0.001	3.273	0.001	**
majorschool3	0.092	0.081	1.136	0.256		0.093	0.081	1.151	0.250	
majorschool4	-0.023	0.055	-0.410	0.682		-0.018	0.056	-0.328	0.743	
majorschool5	-0.097	0.126	-0.775	0.438		-0.082	0.127	-0.647	0.517	
majorschool9	0.399	0.338	1.181	0.238		0.408	0.307	1.328	0.184	
majorschoolC	0.136	0.111	1.223	0.222		0.107	0.111	0.966	0.334	
majorschoolE	0.064	0.053	1.211	0.226		0.070	0.054	1.305	0.192	
majorschoolJ	0.070	0.083	0.842	0.400		0.085	0.083	1.022	0.307	
majorschoolL	0.027	0.060	0.454	0.650		0.014	0.061	0.223	0.823	
majorschoolN	0.037	0.121	0.303	0.762		0.109	0.122	0.891	0.373	
majorschoolS	0.278	0.179	1.556	0.120		0.229	0.181	1.265	0.206	
majorschoolU	0.062	0.057	1.095	0.273		0.059	0.058	1.019	0.308	
HRS_UNDERTAKEN.y	0.034	0.004	7.947	<.001	***	0.036	0.004	8.542	<.001	***

Note. *MSE = .4260;* $s_p^2 = 0.0749$; $s_s^2 = 0.05260$; *$p < .05$; **$p < .01$; ***$p < .001$

Table D12 Full regression output for cluster-robust standard errors models in the Chemistry course sequence (mean split) before and after adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*d.f.*	p-value		Estimate	*SE*	*d.f.*	p-value	
Intercept	-0.801	0.358	48.2	0.030	*	-1.014	0.416	12.74	0.030	*
High RP	0.057	0.028	56.1	0.047	*	0.055	0.029	49.30	0.066	.
SAT equivalent	0.000	0.000	53.6	<0.001	***	0.000	0.000	55.73	<0.001	***
derivationAI	0.207	0.165	7.5	0.247		0.196	0.116	1.89	0.241	
derivationA	-0.036	0.036	38.9	0.322		-0.054	0.041	24.74	0.199	
derivationB2eH	0.017	0.128	18.7	0.893		-0.099	0.165	6.06	0.570	
derivationB	-0.046	0.060	45.3	0.445		-0.032	0.062	40.52	0.608	
derivationF	0.061	0.081	46.5	0.456		0.014	0.113	35.45	0.902	
derivationHPI	0.037	0.176	7.4	0.840		0.019	0.192	7.18	0.923	
derivationH	-0.026	0.038	41.5	0.492		-0.040	0.039	28.60	0.321	
derivationU	-0.278	0.136	16.9	0.056	.	-0.170	0.168	4.64	0.362	
derivationW	0.010	0.037	40.3	0.783		0.001	0.040	25.45	0.971	
motheredlevel1	-0.051	0.016	45.0	0.003	**	-0.044	0.018	7.98	0.035	*
motheredlevel2	-0.085	0.074	39.4	0.256		-0.067	0.083	36.81	0.426	
motheredlevel3	-0.107	0.069	39.4	0.128		-0.093	0.078	34.28	0.238	
motheredlevel4	-0.058	0.077	40.2	0.452		-0.040	0.082	32.60	0.623	
motheredlevel5	-0.069	0.070	37.5	0.334		-0.063	0.079	30.86	0.429	
motheredlevel6	-0.040	0.072	38.4	0.579		-0.031	0.081	31.54	0.701	
motheredlevelU	-0.058	0.072	40.9	0.429		-0.048	0.081	34.20	0.555	
fatheredlevel1	-0.072	0.105	30.7	0.493		-0.085	0.101	27.66	0.406	
fatheredlevel2	-0.013	0.092	39.9	0.885		-0.025	0.096	44.04	0.793	
fatheredlevel3	0.017	0.087	44.3	0.849		0.012	0.093	45.20	0.900	
fatheredlevel4	0.012	0.092	44.1	0.895		0.006	0.096	43.61	0.953	
fatheredlevel5	0.124	0.086	41.6	0.160		0.104	0.090	42.38	0.251	
fatheredlevel6	0.137	0.088	41.6	0.125		0.125	0.093	42.68	0.184	
fatheredlevelU	0.176	0.099	44.8	0.082	.	0.160	0.103	46.31	0.127	
age	0.127	0.117	33.4	0.286		0.138	0.115	34.66	0.238	
class_zscore.x	0.821	0.018	39.6	<0.001	***	0.814	0.018	48.26	<0.001	***
CLASSIFICATION2	0.018	0.024	64.4	0.452		0.008	0.027	59.57	0.754	
CLASSIFICATION3	-0.036	0.074	43.0	0.626		-0.025	0.080	23.51	0.760	
CLASSIFICATION4	0.226	0.132	13.4	0.110		-0.011	0.169	3.13	0.953	
hspct2	0.492	0.130	63.5	<0.001	***	0.545	0.127	38.99	<0.001	***
sexW	0.035	0.019	53.7	0.073	.	0.034	0.021	54.06	0.110	
transferredhours	0.002	0.001	51.7	0.025	*	0.002	0.001	59.08	0.010	*
majorschool3	0.061	0.080	58.2	0.451		0.070	0.089	49.11	0.435	
majorschool4	0.005	0.053	53.6	0.919		0.010	0.057	50.79	0.866	
majorschool5	-0.157	0.177	27.6	0.381		-0.120	0.161	20.02	0.463	
majorschool9	0.436	0.076	2.0	0.030	*	0.419	0.100	1.19	0.118	
majorschoolC	0.110	0.111	37.6	0.329		0.078	0.110	26.20	0.485	
majorschoolE	0.069	0.053	46.1	0.201		0.080	0.057	44.01	0.168	
majorschoolJ	0.064	0.091	44.9	0.491		0.089	0.088	39.06	0.316	
majorschoolL	0.014	0.061	56.8	0.820		-0.005	0.068	58.54	0.941	
majorschoolN	0.103	0.153	39.8	0.503		0.183	0.150	34.72	0.231	
majorschoolS	0.241	0.102	10.6	0.038	*	0.169	0.096	9.91	0.108	
majorschoolU	0.067	0.055	55.2	0.233		0.066	0.061	54.52	0.284	
C_HRS_UNDERTAKEN.y	0.034	0.005	58.8	<0.001	***	0.037	0.005	71.70	<0.001	***

Note. *p < .05; **p < .01; ***p < .001

Table D13 Full regression output for fixed-effects models in the Economics course sequence (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-0.568	0.944	-0.602	0.547		-0.634	0.959	-0.661	0.509	
High RP	0.074	0.064	1.167	0.243		0.086	0.064	1.350	0.177	
SAT equivalent	0.001	0.000	4.689	<.001	***	0.001	0.000	4.712	<.001	***
derivationAI	0.136	0.243	0.559	0.576		0.149	0.245	0.606	0.545	
derivationA	-0.100	0.088	-1.129	0.259		-0.082	0.088	-0.938	0.348	
derivationB2eH	-0.064	0.232	-0.277	0.781		0.049	0.237	0.206	0.837	
derivationB	-0.152	0.116	-1.308	0.191		-0.160	0.115	-1.385	0.166	
derivationF	0.074	0.121	0.614	0.539		0.043	0.120	0.356	0.722	
derivationHPI	0.133	0.369	0.361	0.718		0.153	0.471	0.325	0.746	
derivationH	-0.123	0.091	-1.353	0.176		-0.112	0.090	-1.234	0.217	
derivationU	-0.307	0.232	-1.322	0.186		-0.276	0.231	-1.199	0.231	
derivationW	-0.012	0.084	-0.147	0.883		-0.003	0.084	-0.041	0.968	
motheredlevel1	0.151	0.134	1.130	0.258		0.205	0.131	1.559	0.119	
motheredlevel2	0.020	0.111	0.184	0.854		0.065	0.108	0.602	0.547	
motheredlevel3	0.162	0.115	1.408	0.159		0.216	0.112	1.931	0.054	.
motheredlevel4	0.104	0.110	0.938	0.348		0.160	0.107	1.494	0.135	
motheredlevel5	0.150	0.113	1.324	0.186		0.207	0.110	1.878	0.060	.
motheredlevel6	0.171	0.122	1.405	0.160		0.222	0.119	1.863	0.063	.
motheredlevelU	0.300	0.169	1.770	0.077	.	0.405	0.168	2.408	0.016	*
fatheredlevel1	-0.117	0.134	-0.870	0.385		-0.198	0.134	-1.477	0.140	
fatheredlevel2	-0.069	0.114	-0.607	0.544		-0.138	0.112	-1.231	0.218	
fatheredlevel3	-0.101	0.119	-0.843	0.399		-0.156	0.118	-1.324	0.185	
fatheredlevel4	0.013	0.114	0.116	0.908		-0.037	0.112	-0.330	0.741	
fatheredlevel5	-0.013	0.115	-0.112	0.910		-0.064	0.113	-0.567	0.571	
fatheredlevel6	-0.206	0.141	-1.465	0.143		-0.261	0.139	-1.877	0.061	.
fatheredlevelU	-0.249	0.164	-1.519	0.129		-0.348	0.163	-2.131	0.033	*
age	-0.072	0.028	-2.578	0.010	**	-0.072	0.028	-2.568	0.010	*
class_zscore.x	0.671	0.020	34.102	<.001	***	0.668	0.020	33.885	<.001	***
CLASSIFICATION2	0.018	0.044	0.411	0.681		0.023	0.044	0.517	0.605	
CLASSIFICATION3	-0.055	0.096	-0.579	0.562		-0.062	0.094	-0.660	0.510	
CLASSIFICATION4	-0.042	0.197	-0.212	0.832		-0.034	0.198	-0.174	0.862	
hspct2	0.545	0.124	4.415	0.000	***	0.552	0.124	4.469	0.000	***
sexW	0.057	0.029	1.961	0.050	*	0.059	0.029	2.026	0.043	*
transferredhours	0.002	0.001	1.604	0.109		0.002	0.001	1.721	0.085	.
majorschool3	0.074	0.203	0.366	0.715		0.116	0.204	0.568	0.570	
majorschool4	0.008	0.074	0.110	0.913		-0.010	0.074	-0.132	0.895	
majorschool5	0.083	0.157	0.530	0.596		0.076	0.156	0.485	0.628	
majorschool9	0.138	0.423	0.327	0.744		0.078	0.413	0.189	0.850	
majorschoolC	0.074	0.091	0.806	0.420		0.106	0.091	1.167	0.243	
majorschoolE	0.148	0.058	2.547	0.011	*	0.153	0.058	2.642	0.008	**
majorschoolJ	-0.369	0.300	-1.230	0.219		-0.776	0.309	-2.514	0.012	*
majorschoolL	0.077	0.043	1.811	0.070	.	0.076	0.043	1.783	0.075	.
majorschoolN	-0.197	0.299	-0.659	0.510		-0.224	0.295	-0.759	0.448	
majorschoolS	-0.429	0.242	-1.770	0.077	.	-0.351	0.247	-1.421	0.155	
majorschoolU	0.049	0.042	1.180	0.238		0.061	0.041	1.458	0.145	
C_HRS_UNDERTAKEN.y	0.027	0.007	4.101	<.001	***	0.028	0.007	4.234	<.001	***

Note. $n = 2766$; *RMSE* = 0.7124; $df = 2688$, $R^2 = .4257$; *$p < .05$; **$p < .01$; ***$p < .001$

Table D14 Full regression output for random-effects models in the Economics course sequence (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.010	0.611	-1.653	0.099	.	-0.997	0.612	-1.629	0.104	
High RP	0.083	0.045	1.857	0.077	.	0.082	0.044	1.840	0.081	.
SAT equivalent	0.001	0.000	4.704	<.001	***	0.001	0.000	4.704	<.001	***
derivationAI	0.124	0.242	0.514	0.608		0.136	0.245	0.558	0.577	
derivationA	-0.092	0.088	-1.045	0.296		-0.075	0.087	-0.860	0.390	
derivationB2eH	-0.067	0.232	-0.290	0.772		0.040	0.236	0.168	0.866	
derivationB	-0.135	0.116	-1.168	0.243		-0.147	0.115	-1.276	0.202	
derivationF	0.062	0.120	0.519	0.604		0.037	0.119	0.307	0.759	
derivationHPI	0.139	0.367	0.378	0.706		0.156	0.470	0.332	0.740	
derivationH	-0.121	0.091	-1.334	0.182		-0.109	0.090	-1.212	0.226	
derivationU	-0.319	0.232	-1.377	0.169		-0.295	0.230	-1.283	0.200	
derivationW	-0.003	0.084	-0.042	0.967		0.005	0.083	0.062	0.951	
motheredlevel1	0.120	0.133	0.902	0.367		-0.057	0.027	-2.090	0.037	*
motheredlevel2	0.008	0.110	0.070	0.944		0.181	0.131	1.384	0.166	
motheredlevel3	0.149	0.114	1.302	0.193		0.058	0.108	0.540	0.589	
motheredlevel4	0.087	0.110	0.789	0.430		0.209	0.111	1.874	0.061	.
motheredlevel5	0.133	0.113	1.182	0.237		0.149	0.107	1.395	0.163	
motheredlevel6	0.163	0.121	1.346	0.178		0.196	0.110	1.785	0.074	.
motheredlevelU	0.289	0.169	1.715	0.086	.	0.220	0.119	1.850	0.064	.
fatheredlevel1	-0.096	0.133	-0.719	0.472		0.399	0.167	2.381	0.017	*
fatheredlevel2	-0.046	0.113	-0.402	0.688		-0.171	0.133	-1.285	0.199	
fatheredlevel3	-0.078	0.119	-0.655	0.513		-0.110	0.112	-0.980	0.327	
fatheredlevel4	0.037	0.113	0.329	0.742		-0.127	0.117	-1.087	0.277	
fatheredlevel5	0.011	0.114	0.100	0.920		-0.011	0.111	-0.097	0.922	
fatheredlevel6	-0.172	0.140	-1.230	0.219		-0.037	0.112	-0.327	0.744	
fatheredlevelU	-0.234	0.163	-1.434	0.152		-0.225	0.138	-1.628	0.104	
age	-0.054	0.027	-1.995	0.046	*	-0.331	0.162	-2.038	0.042	*
class_zscore.x	0.670	0.020	34.183	<.001	***	0.666	0.020	33.963	<.001	***
CLASSIFICATION2	0.036	0.044	0.814	0.415		0.040	0.044	0.921	0.357	
CLASSIFICATION3	-0.023	0.094	-0.249	0.804		-0.027	0.093	-0.296	0.767	
CLASSIFICATION4	-0.009	0.193	-0.046	0.964		0.008	0.194	0.042	0.966	
hspct2	0.561	0.123	4.577	<.001	***	0.568	0.123	4.628	<.001	***
sexW	0.054	0.029	1.864	0.062	.	0.056	0.029	1.925	0.054	.
transferredhours	0.002	0.001	1.714	0.087	.	0.002	0.001	1.828	0.068	.
majorschool3	0.083	0.202	0.412	0.680		0.130	0.204	0.639	0.523	
majorschool4	0.043	0.074	0.580	0.562		0.024	0.073	0.330	0.741	
majorschool5	0.108	0.156	0.694	0.488		0.101	0.155	0.651	0.515	
majorschool9	0.171	0.415	0.413	0.680		0.107	0.408	0.261	0.794	
majorschoolC	0.096	0.091	1.061	0.289		0.127	0.090	1.402	0.161	
majorschoolE	0.186	0.057	3.258	<.001	**	0.190	0.057	3.315	<.001	***
majorschoolJ	-0.370	0.299	-1.238	0.216		-0.807	0.308	-2.622	0.009	**
majorschoolL	0.097	0.042	2.341	0.019	*	0.095	0.042	2.287	0.022	*
majorschoolN	-0.148	0.298	-0.496	0.620		-0.153	0.294	-0.521	0.602	
majorschoolS	-0.436	0.242	-1.802	0.072	.	-0.346	0.247	-1.405	0.160	
majorschoolU	0.070	0.041	1.709	0.088	.	0.081	0.041	1.957	0.050	.
C_HRS_UNDERTAKEN.y	0.026	0.007	4.030	<.001	***	0.027	0.007	4.178	<.001	***

Note. *MSE = .50533;* $s_p^2 = 0.007757$; $s_s^2 = 0.001149$; *p < .05; **p < .01; ***p < .001

Table D15 Full regression output for cluster-robust standard errors models in the Economics course sequence (median split) before and after adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*d.f.*	p-value		Estimate	*SE*	*d.f.*	p-value	
Intercept	-1.049	0.639	81.0	0.104		-1.078	0.649	74.26	0.101	
High RP	0.048	0.037	63.6	0.198		0.047	0.037	63.22	0.214	
SAT equivalent	0.001	0.000	69.2	<.001	***	0.001	0.000	65.84	<.001	***
derivationAI	0.143	0.206	11.3	0.501		0.160	0.207	11.05	0.457	
derivationA	-0.086	0.087	41.3	0.330		-0.069	0.086	38.78	0.429	
derivationB2eH	-0.075	0.281	10.7	0.794		0.031	0.311	8.65	0.923	
derivationB	-0.155	0.125	46.1	0.222		-0.165	0.130	37.98	0.212	
derivationF	0.070	0.133	46.2	0.605		0.045	0.130	29.66	0.729	
derivationHPI	0.090	0.332	3.4	0.803		0.105	0.333	3.36	0.772	
derivationH	-0.122	0.089	43.0	0.175		-0.111	0.087	40.57	0.209	
derivationU	-0.321	0.222	10.7	0.176		-0.301	0.253	6.69	0.274	
derivationW	-0.004	0.087	38.0	0.967		0.005	0.085	35.77	0.950	
motheredlevel1	-0.055	0.027	79.4	0.048	*	-0.056	0.028	74.85	0.051	.
motheredlevel2	0.129	0.141	44.3	0.364		0.189	0.151	39.45	0.216	
motheredlevel3	0.005	0.110	40.3	0.963		0.056	0.118	31.27	0.637	
motheredlevel4	0.146	0.111	44.3	0.196		0.208	0.118	32.33	0.087	.
motheredlevel5	0.092	0.101	40.8	0.367		0.155	0.106	29.02	0.156	
motheredlevel6	0.135	0.104	42.5	0.204		0.198	0.110	30.14	0.080	.
motheredlevelU	0.164	0.115	47.3	0.161		0.223	0.120	35.80	0.071	.
fatheredlevel1	0.283	0.179	37.4	0.121		0.389	0.192	31.13	0.052	.
fatheredlevel2	-0.086	0.150	39.6	0.567		-0.161	0.155	38.09	0.305	
fatheredlevel3	-0.047	0.113	45.4	0.681		-0.111	0.118	36.48	0.352	
fatheredlevel4	-0.073	0.116	48.6	0.536		-0.124	0.125	39.73	0.326	
fatheredlevel5	0.041	0.112	44.3	0.717		-0.009	0.119	35.56	0.939	
fatheredlevel6	0.017	0.111	45.1	0.882		-0.033	0.118	36.61	0.781	
fatheredlevelU	-0.168	0.135	56.4	0.220		-0.222	0.137	46.88	0.111	
age	-0.210	0.176	37.6	0.241		-0.304	0.190	32.00	0.119	
class_zscore.x	0.666	0.024	71.2	<.001	***	0.662	0.023	67.99	<.001	***
CLASSIFICATION2	0.037	0.045	72.6	0.417		0.043	0.047	68.20	0.358	
CLASSIFICATION3	-0.007	0.096	60.3	0.941		-0.008	0.099	41.83	0.933	
CLASSIFICATION4	0.007	0.164	19.3	0.964		0.026	0.160	16.20	0.875	
hspct2	0.603	0.143	61.6	<.001	***	0.609	0.139	57.31	<.001	***
sexW	0.056	0.028	63.9	0.051	.	0.057	0.029	59.41	0.057	.
transferredhours	0.002	0.001	72.7	0.121		0.002	0.001	69.77	0.118	
majorschool3	0.074	0.167	7.6	0.670		0.119	0.215	6.86	0.596	
majorschool4	0.042	0.077	76.5	0.591		0.024	0.080	53.48	0.762	
majorschool5	0.133	0.128	17.6	0.313		0.126	0.126	15.48	0.333	
majorschool9	0.222	0.428	2.1	0.654		0.160	0.282	1.45	0.646	
majorschoolC	0.111	0.088	38.6	0.217		0.143	0.085	34.17	0.101	
majorschoolE	0.196	0.062	65.1	0.002	**	0.202	0.063	55.54	0.002	**
majorschoolJ	-0.366	0.352	5.5	0.342		-0.797	0.658	1.55	0.379	
majorschoolL	0.109	0.040	65.8	0.008	**	0.109	0.040	66.88	0.009	**
majorschoolN	-0.131	0.159	5.4	0.443		-0.132	0.236	3.31	0.612	
majorschoolS	-0.384	0.311	6.7	0.259		-0.293	0.440	5.34	0.533	
majorschoolU	0.080	0.045	65.0	0.079	.	0.091	0.044	58.84	0.044	*
C_HRS_UNDERTAKEN.y	0.026	0.007	75.3	<.001	***	0.027	0.007	70.34	<.001	***

Note. *p < .05; **p < .01; ***p < .001

Table D16 Full regression output for fixed-effects models in the Economics course sequence (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-0.650	0.946	-0.687	0.492		-0.638	0.933	-0.684	0.494	
High RP	0.159	0.094	1.701	0.089	.	0.159	0.093	1.706	0.088	.
SAT equivalent	0.001	0.000	4.710	<.001	***	0.001	0.000	4.706	<.001	***
derivationAI	0.133	0.243	0.549	0.583		0.144	0.248	0.583	0.560	
derivationA	-0.101	0.088	-1.147	0.251		-0.094	0.088	-1.070	0.285	
derivationB2eH	-0.066	0.232	-0.285	0.775		0.053	0.238	0.224	0.822	
derivationB	-0.151	0.116	-1.302	0.193		-0.175	0.115	-1.516	0.130	
derivationF	0.072	0.121	0.597	0.551		0.042	0.121	0.345	0.730	
derivationHPI	0.132	0.368	0.359	0.720		0.149	0.474	0.314	0.753	
derivationH	-0.124	0.091	-1.362	0.173		-0.120	0.091	-1.316	0.188	
derivationU	-0.300	0.232	-1.292	0.196		-0.286	0.230	-1.245	0.213	
derivationW	-0.013	0.084	-0.151	0.880		-0.010	0.084	-0.113	0.910	
motheredlevel1	0.150	0.133	1.122	0.262		0.200	0.131	1.526	0.127	
motheredlevel2	0.020	0.111	0.184	0.854		0.051	0.108	0.474	0.636	
motheredlevel3	0.162	0.115	1.416	0.157		0.204	0.112	1.817	0.069	.
motheredlevel4	0.103	0.110	0.932	0.352		0.143	0.108	1.330	0.183	
motheredlevel5	0.150	0.113	1.328	0.184		0.194	0.111	1.754	0.079	.
motheredlevel6	0.173	0.122	1.422	0.155		0.211	0.120	1.763	0.078	.
motheredlevelU	0.299	0.169	1.765	0.078	.	0.394	0.169	2.335	0.020	*
fatheredlevel1	-0.109	0.134	-0.812	0.417		-0.140	0.134	-1.047	0.295	
fatheredlevel2	-0.068	0.114	-0.593	0.554		-0.114	0.113	-1.012	0.312	
fatheredlevel3	-0.099	0.119	-0.828	0.408		-0.133	0.118	-1.128	0.260	
fatheredlevel4	0.016	0.114	0.141	0.888		-0.006	0.112	-0.053	0.958	
fatheredlevel5	-0.011	0.115	-0.097	0.922		-0.038	0.114	-0.331	0.741	
fatheredlevel6	-0.205	0.141	-1.454	0.146		-0.242	0.139	-1.738	0.082	.
fatheredlevelU	-0.243	0.164	-1.484	0.138		-0.322	0.164	-1.968	0.049	*
age	-0.072	0.028	-2.591	0.010	**	-0.074	0.028	-2.649	0.008	**
class_zscore.x	0.671	0.020	34.129	<.001	***	0.672	0.020	33.943	<.001	***
CLASSIFICATION2	0.019	0.044	0.429	0.668		0.017	0.044	0.389	0.698	
CLASSIFICATION3	-0.057	0.096	-0.598	0.550		-0.041	0.094	-0.431	0.667	
CLASSIFICATION4	-0.033	0.197	-0.166	0.868		-0.020	0.196	-0.101	0.920	
hspct2	0.542	0.123	4.392	<.001	***	0.539	0.124	4.342	<.001	***
sexW	0.058	0.029	1.990	0.047	*	0.059	0.029	2.028	0.043	*
transferredhours	0.002	0.001	1.578	0.115		0.002	0.001	1.801	0.072	.
majorschool3	0.072	0.203	0.353	0.724		0.113	0.206	0.546	0.585	
majorschool4	0.004	0.074	0.051	0.959		0.003	0.074	0.035	0.972	
majorschool5	0.080	0.157	0.513	0.608		0.060	0.155	0.385	0.700	
majorschool9	0.129	0.423	0.304	0.761		-0.054	0.401	-0.135	0.893	
majorschoolC	0.071	0.091	0.780	0.436		0.088	0.091	0.966	0.334	
majorschoolE	0.148	0.058	2.550	0.011	*	0.151	0.058	2.596	0.009	**
majorschoolJ	-0.369	0.300	-1.229	0.219		-0.805	0.305	-2.638	0.008	**
majorschoolL	0.075	0.043	1.753	0.080	.	0.071	0.043	1.660	0.097	.
majorschoolN	-0.198	0.299	-0.661	0.509		-0.249	0.297	-0.839	0.401	
majorschoolS	-0.425	0.242	-1.756	0.079	.	-0.380	0.247	-1.536	0.125	
majorschoolU	0.048	0.042	1.164	0.245		0.050	0.042	1.211	0.226	
C_HRS_UNDERTAKEN.y	0.027	0.007	4.086	<.001	***	0.027	0.007	4.044	<.001	***

Note. $n = 2766$; *RMSE* = 0.7122; $df = 2688$, $R^2 = .4261$; *$p < .05$; **$p < .01$; ***$p < .001$

Table D17 Full regression output for random-effects models in the Economics course sequence (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.039	0.610	-1.703	0.089	.	-0.970	0.614	-1.581	0.114	
High RP	0.101	0.043	2.356	0.028	*	0.103	0.045	2.309	0.031	*
SAT equivalent	0.001	0.000	4.716	0.000	***	0.001	0.000	4.657	<.001	***
derivationAI	0.126	0.242	0.521	0.603		0.135	0.247	0.545	0.586	
derivationA	-0.092	0.088	-1.051	0.293		-0.087	0.088	-0.994	0.321	
derivationB2eH	-0.065	0.232	-0.281	0.779		0.052	0.237	0.218	0.828	
derivationB	-0.135	0.116	-1.164	0.245		-0.159	0.115	-1.378	0.168	
derivationF	0.062	0.120	0.517	0.605		0.036	0.120	0.297	0.766	
derivationHPI	0.142	0.367	0.386	0.700		0.161	0.473	0.340	0.734	
derivationH	-0.120	0.091	-1.328	0.184		-0.116	0.090	-1.287	0.198	
derivationU	-0.317	0.232	-1.368	0.171		-0.311	0.229	-1.359	0.174	
derivationW	-0.004	0.084	-0.043	0.966		-0.001	0.084	-0.006	0.995	
motheredlevel1	0.120	0.133	0.904	0.366		-0.057	0.027	-2.077	0.038	*
motheredlevel2	0.007	0.110	0.067	0.947		0.172	0.131	1.316	0.188	
motheredlevel3	0.148	0.114	1.298	0.194		0.042	0.108	0.391	0.696	
motheredlevel4	0.086	0.110	0.783	0.434		0.192	0.112	1.720	0.085	.
motheredlevel5	0.132	0.113	1.175	0.240		0.130	0.107	1.207	0.227	
motheredlevel6	0.163	0.121	1.347	0.178		0.180	0.110	1.630	0.103	
motheredlevelU	0.289	0.169	1.714	0.087	.	0.204	0.119	1.710	0.087	.
fatheredlevel1	-0.096	0.133	-0.725	0.468		0.384	0.168	2.285	0.022	*
fatheredlevel2	-0.046	0.113	-0.409	0.683		-0.121	0.133	-0.914	0.361	
fatheredlevel3	-0.078	0.119	-0.656	0.512		-0.087	0.112	-0.774	0.439	
fatheredlevel4	0.036	0.113	0.321	0.748		-0.107	0.117	-0.912	0.362	
fatheredlevel5	0.011	0.114	0.094	0.925		0.017	0.111	0.149	0.881	
fatheredlevel6	-0.173	0.140	-1.233	0.218		-0.013	0.113	-0.112	0.911	
fatheredlevelU	-0.235	0.163	-1.439	0.150		-0.208	0.139	-1.496	0.135	
age	-0.053	0.027	-1.967	0.049	*	-0.310	0.163	-1.902	0.057	.
class_zscore.x	0.669	0.020	34.172	<.001	***	0.670	0.020	34.026	<.001	***
CLASSIFICATION2	0.036	0.044	0.820	0.413		0.034	0.044	0.773	0.440	
CLASSIFICATION3	-0.023	0.094	-0.243	0.808		-0.008	0.093	-0.082	0.935	
CLASSIFICATION4	-0.006	0.192	-0.031	0.975		0.012	0.193	0.061	0.951	
hspct2	0.563	0.123	4.590	<.001	***	0.551	0.123	4.468	<.001	***
sexW	0.054	0.029	1.875	0.061	.	0.056	0.029	1.927	0.054	.
transferredhours	0.002	0.001	1.660	0.097	.	0.002	0.001	1.852	0.064	.
majorschool3	0.083	0.202	0.409	0.682		0.130	0.206	0.632	0.527	
majorschool4	0.042	0.074	0.565	0.572		0.038	0.074	0.517	0.605	
majorschool5	0.109	0.156	0.696	0.487		0.083	0.154	0.539	0.590	
majorschool9	0.162	0.415	0.390	0.696		-0.019	0.391	-0.049	0.961	
majorschoolC	0.096	0.091	1.057	0.291		0.107	0.090	1.188	0.235	
majorschoolE	0.186	0.057	3.261	0.001	**	0.187	0.057	3.252	0.001	**
majorschoolJ	-0.366	0.299	-1.223	0.222		-0.841	0.304	-2.766	0.006	**
majorschoolL	0.098	0.041	2.355	0.019	*	0.091	0.042	2.180	0.029	*
majorschoolN	-0.145	0.298	-0.487	0.627		-0.172	0.295	-0.583	0.560	
majorschoolS	-0.431	0.242	-1.784	0.075	.	-0.380	0.247	-1.537	0.124	
majorschoolU	0.071	0.041	1.728	0.084	.	0.069	0.041	1.674	0.094	.
C_HRS_UNDERTAKEN.y	0.026	0.007	4.034	<.001	***	0.026	0.007	3.945	<.001	***

Note. *MSE = .50527;* $s_p^2 = 0.006912$; $s_s^2 = 0.001141$; *p < .05; **p < .01; ***p < .001

Table D18 Full regression output for cluster-robust standard errors models in the Economics course sequence (mean split) before and after adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*d.f.*	p-value		Estimate	*SE*	*d.f.*	p-value	
Intercept	-1.082	0.640	80.9	0.094	.	-1.054	0.648	79.09	0.108	
High RP	0.070	0.038	67.6	0.069	.	0.067	0.038	67.65	0.082	.
SAT equivalent	0.001	0.000	69.2	<.001	***	0.001	0.000	68.76	<.001	***
derivationAI	0.145	0.207	11.3	0.500		0.161	0.206	10.96	0.452	
derivationA	-0.088	0.088	41.3	0.323		-0.082	0.087	39.50	0.353	
derivationB2eH	-0.070	0.281	10.7	0.807		0.044	0.298	8.73	0.887	
derivationB	-0.152	0.126	46.1	0.231		-0.177	0.132	41.14	0.186	
derivationF	0.069	0.133	46.3	0.607		0.043	0.130	37.98	0.742	
derivationHPI	0.100	0.333	3.4	0.781		0.110	0.335	3.36	0.762	
derivationH	-0.121	0.089	43.0	0.181		-0.117	0.087	41.47	0.186	
derivationU	-0.317	0.223	10.7	0.184		-0.321	0.264	6.71	0.266	
derivationW	-0.004	0.088	38.0	0.964		0.001	0.086	36.53	0.989	
motheredlevel1	-0.054	0.027	79.4	0.052	.	-0.056	0.028	78.67	0.050	*
motheredlevel2	0.128	0.141	44.2	0.368		0.178	0.149	41.74	0.238	
motheredlevel3	0.004	0.110	40.3	0.974		0.038	0.118	35.50	0.747	
motheredlevel4	0.144	0.111	44.3	0.201		0.189	0.117	37.10	0.116	
motheredlevel5	0.090	0.101	40.8	0.380		0.134	0.106	33.88	0.215	
motheredlevel6	0.132	0.105	42.5	0.213		0.180	0.109	35.29	0.109	
motheredlevelU	0.164	0.115	47.3	0.162		0.204	0.120	41.11	0.096	.
fatheredlevel1	0.282	0.179	37.4	0.123		0.371	0.192	35.00	0.061	.
fatheredlevel2	-0.088	0.150	39.6	0.562		-0.106	0.160	38.92	0.512	
fatheredlevel3	-0.048	0.112	45.4	0.670		-0.089	0.120	40.64	0.464	
fatheredlevel4	-0.073	0.116	48.6	0.532		-0.100	0.125	43.30	0.427	
fatheredlevel5	0.039	0.111	44.3	0.728		0.020	0.118	40.05	0.864	
fatheredlevel6	0.015	0.110	45.1	0.894		-0.006	0.118	40.78	0.957	
fatheredlevelU	-0.169	0.135	56.4	0.215		-0.203	0.137	50.12	0.145	
age	-0.213	0.176	37.6	0.234		-0.278	0.193	35.00	0.158	
class_zscore.x	0.666	0.024	71.3	<.001	***	0.666	0.024	70.63	<.001	***
CLASSIFICATION2	0.037	0.045	72.6	0.417		0.036	0.047	74.34	0.455	
CLASSIFICATION3	-0.007	0.096	60.3	0.938		0.009	0.099	45.93	0.929	
CLASSIFICATION4	0.011	0.163	19.3	0.949		0.032	0.160	15.50	0.845	
hspct2	0.603	0.143	61.6	<.001	***	0.596	0.141	57.85	<.001	***
sexW	0.057	0.028	63.9	0.049	*	0.057	0.029	63.38	0.055	.
transferredhours	0.002	0.001	72.8	0.149		0.002	0.001	60.87	0.134	
majorschool3	0.073	0.167	7.6	0.677		0.123	0.212	6.82	0.582	
majorschool4	0.040	0.077	76.5	0.603		0.036	0.078	68.80	0.648	
majorschool5	0.132	0.127	17.6	0.311		0.111	0.128	16.27	0.401	
majorschool9	0.205	0.408	2.1	0.664		0.010	0.479	1.36	0.986	
majorschoolC	0.109	0.089	38.6	0.224		0.123	0.087	37.74	0.166	
majorschoolE	0.195	0.062	65.1	0.002	**	0.198	0.063	61.67	0.003	**
majorschoolJ	-0.357	0.354	5.5	0.356		-0.829	0.698	1.49	0.390	
majorschoolL	0.109	0.040	65.8	0.008	**	0.104	0.041	68.13	0.013	*
majorschoolN	-0.127	0.159	5.4	0.460		-0.148	0.247	3.37	0.586	
majorschoolS	-0.378	0.311	6.7	0.265		-0.327	0.453	5.22	0.501	
majorschoolU	0.081	0.045	65.0	0.075	.	0.081	0.045	63.11	0.076	.
C_HRS_UNDERTAKEN.y	0.026	0.007	75.4	<.001	***	0.026	0.007	75.08	<.001	***

Note. *p < .05; **p < .01; ***p < .001

Table D19 Full regression output for fixed-effects models in the Government course sequence (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.144	0.682	-1.679	0.093	.	-0.941	0.615	-1.530	0.126	
High RP	0.044	0.051	0.848	0.397		0.025	0.046	0.540	0.589	
SAT equivalent	0.001	0.000	7.670	<.001	***	0.001	0.000	6.829	<.001	***
derivationAI	0.115	0.303	0.379	0.705		-0.063	0.302	-0.210	0.834	
derivationA	0.022	0.098	0.228	0.820		0.046	0.096	0.477	0.634	
derivationB2eH	0.120	0.202	0.597	0.551		-0.096	0.199	-0.481	0.630	
derivationB	-0.085	0.114	-0.748	0.455		-0.066	0.111	-0.595	0.552	
derivationF	0.270	0.145	1.861	0.063	.	0.218	0.141	1.550	0.121	
derivationHPI	0.962	0.443	2.169	0.030	*	1.146	0.583	1.965	0.050	*
derivationH	-0.039	0.094	-0.411	0.681		-0.017	0.093	-0.185	0.853	
derivationU	-0.025	0.444	-0.056	0.955		-0.127	0.425	-0.298	0.766	
derivationW	0.010	0.089	0.114	0.909		0.023	0.088	0.267	0.790	
motheredlevel1	0.054	0.134	0.406	0.685		-0.014	0.132	-0.108	0.914	
motheredlevel2	0.064	0.123	0.520	0.603		-0.017	0.122	-0.136	0.892	
motheredlevel3	0.183	0.126	1.446	0.148		0.128	0.124	1.030	0.303	
motheredlevel4	0.241	0.125	1.932	0.054	.	0.164	0.124	1.322	0.186	
motheredlevel5	0.254	0.128	1.981	0.048	*	0.114	0.127	0.897	0.370	
motheredlevel6	0.307	0.136	2.251	0.024	*	0.226	0.135	1.672	0.095	.
motheredlevelU	0.326	0.171	1.910	0.056	.	0.128	0.168	0.762	0.446	
fatheredlevel1	-0.013	0.141	-0.089	0.929		-0.066	0.140	-0.469	0.639	
fatheredlevel2	0.003	0.124	0.021	0.983		-0.018	0.122	-0.150	0.880	
fatheredlevel3	-0.029	0.126	-0.231	0.817		-0.038	0.124	-0.303	0.762	
fatheredlevel4	-0.041	0.124	-0.330	0.742		-0.002	0.122	-0.015	0.988	
fatheredlevel5	-0.016	0.126	-0.130	0.896		0.053	0.124	0.428	0.669	
fatheredlevel6	-0.035	0.142	-0.248	0.804		-0.006	0.140	-0.040	0.968	
fatheredlevelU	-0.061	0.159	-0.385	0.701		-0.026	0.155	-0.171	0.864	
age	-0.038	0.026	-1.456	0.145		-0.033	0.025	-1.325	0.185	
class_zscore.x	0.562	0.021	26.926	<.001	***	0.586	0.021	28.147	<.001	***
CLASSIFICATION2	0.072	0.039	1.859	0.063	.	0.088	0.038	2.288	0.022	*
CLASSIFICATION3	0.076	0.065	1.160	0.246		0.059	0.064	0.922	0.356	
CLASSIFICATION4	0.042	0.103	0.410	0.682		0.165	0.102	1.613	0.107	
hspct2	0.138	0.122	1.134	0.257		0.141	0.120	1.179	0.239	
sexW	0.005	0.033	0.140	0.889		-0.033	0.033	-1.005	0.315	
transferredhours	0.000	0.002	0.025	0.980		-0.001	0.002	-0.412	0.680	
majorschool3	-0.100	0.108	-0.929	0.353		-0.060	0.105	-0.570	0.568	
majorschool4	-0.041	0.066	-0.621	0.535		-0.060	0.065	-0.926	0.354	
majorschool5	-0.115	0.100	-1.148	0.251		-0.095	0.097	-0.978	0.328	
majorschool9	-0.250	0.314	-0.797	0.426		-0.212	0.413	-0.513	0.608	
majorschoolC	0.010	0.069	0.146	0.884		-0.052	0.068	-0.762	0.446	
majorschoolE	0.005	0.059	0.084	0.933		-0.064	0.058	-1.101	0.271	
majorschoolJ	0.077	0.150	0.515	0.606		0.094	0.150	0.629	0.529	
majorschoolL	0.125	0.061	2.057	0.040	*	0.088	0.059	1.471	0.141	
majorschoolN	-0.229	0.174	-1.320	0.187		-0.408	0.163	-2.498	0.013	*
majorschoolS	0.083	0.154	0.541	0.589		0.034	0.143	0.239	0.811	
majorschoolU	0.031	0.060	0.522	0.602		0.040	0.058	0.690	0.490	
C_HRS_UNDERTAKEN.y	0.017	0.007	2.259	0.024	*	0.013	0.007	1.767	0.077	.

Note. $n = 2363$; *RMSE* = 0.7293; $df = 2217$, $R^2 = .4491$; *$p < .05$; **$p < .01$; ***$p < .001$

Table D20 Full regression output for random-effects models in the Government course sequence (median split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.442	0.596	-2.418	0.016	*	-1.107	0.567	-1.953	0.051	.
High RP	-0.004	0.059	-0.073	0.943		-0.013	0.053	-0.247	0.814	
SAT equivalent	0.001	0.000	7.699	<.001	***	0.001	0.000	6.739	<.001	***
derivationAI	0.168	0.294	0.570	0.569		-0.163	0.289	-0.565	0.572	
derivationA	0.011	0.096	0.118	0.906		0.025	0.094	0.263	0.793	
derivationB2eH	0.119	0.200	0.596	0.552		-0.113	0.196	-0.574	0.566	
derivationB	-0.079	0.113	-0.698	0.485		-0.066	0.109	-0.607	0.544	
derivationF	0.262	0.144	1.820	0.069	.	0.208	0.139	1.498	0.134	
derivationHPI	0.844	0.441	1.914	0.056	.	1.004	0.580	1.730	0.084	.
derivationH	-0.036	0.093	-0.392	0.695		-0.024	0.091	-0.270	0.787	
derivationU	0.056	0.436	0.128	0.898		-0.053	0.419	-0.126	0.900	
derivationW	0.013	0.087	0.151	0.880		0.012	0.086	0.135	0.893	
motheredlevel1	0.095	0.132	0.721	0.471		-0.036	0.024	-1.476	0.140	
motheredlevel2	0.109	0.122	0.890	0.373		-0.001	0.131	-0.008	0.993	
motheredlevel3	0.239	0.125	1.919	0.055	.	0.010	0.121	0.087	0.931	
motheredlevel4	0.296	0.123	2.395	0.017	*	0.169	0.123	1.380	0.168	
motheredlevel5	0.300	0.127	2.369	0.018	*	0.198	0.123	1.612	0.107	
motheredlevel6	0.354	0.135	2.623	0.009	**	0.141	0.125	1.125	0.261	
motheredlevelU	0.364	0.169	2.155	0.031	*	0.262	0.134	1.954	0.051	.
fatheredlevel1	0.043	0.138	0.310	0.757		0.153	0.166	0.922	0.356	
fatheredlevel2	-0.011	0.123	-0.092	0.927		0.008	0.137	0.056	0.956	
fatheredlevel3	-0.048	0.124	-0.389	0.697		-0.016	0.121	-0.130	0.897	
fatheredlevel4	-0.058	0.123	-0.473	0.636		-0.049	0.122	-0.401	0.689	
fatheredlevel5	-0.041	0.125	-0.333	0.739		0.000	0.121	0.001	0.999	
fatheredlevel6	-0.036	0.141	-0.259	0.795		0.048	0.123	0.390	0.697	
fatheredlevelU	-0.078	0.157	-0.497	0.619		0.004	0.139	0.030	0.976	
age	-0.036	0.026	-1.421	0.155		-0.028	0.153	-0.182	0.855	
class_zscore.x	0.559	0.020	27.363	<.001	***	0.585	0.020	28.655	<.001	***
CLASSIFICATION2	0.061	0.038	1.622	0.105		0.074	0.037	2.000	0.046	*
CLASSIFICATION3	0.059	0.063	0.936	0.349		0.036	0.062	0.580	0.562	
CLASSIFICATION4	-0.022	0.100	-0.219	0.826		0.117	0.098	1.189	0.235	
hspct2	0.205	0.120	1.709	0.088	.	0.193	0.118	1.643	0.101	
sexW	0.001	0.033	0.021	0.984		-0.045	0.032	-1.410	0.159	
transferredhours	0.000	0.002	0.022	0.983		-0.001	0.002	-0.515	0.607	
majorschool3	-0.101	0.106	-0.951	0.342		-0.056	0.104	-0.541	0.589	
majorschool4	-0.023	0.065	-0.349	0.727		-0.049	0.064	-0.775	0.438	
majorschool5	-0.076	0.098	-0.771	0.441		-0.070	0.095	-0.740	0.460	
majorschool9	-0.260	0.309	-0.843	0.399		-0.236	0.408	-0.578	0.563	
majorschoolC	0.031	0.068	0.447	0.655		-0.042	0.067	-0.634	0.526	
majorschoolE	0.028	0.058	0.492	0.623		-0.049	0.057	-0.871	0.384	
majorschoolJ	0.055	0.148	0.370	0.711		0.069	0.147	0.471	0.638	
majorschoolL	0.142	0.060	2.372	0.018	*	0.097	0.058	1.667	0.096	.
majorschoolN	-0.137	0.172	-0.799	0.425		-0.343	0.162	-2.122	0.034	*
majorschoolS	0.080	0.152	0.528	0.598		0.048	0.140	0.345	0.730	
majorschoolU	0.055	0.059	0.932	0.352		0.050	0.057	0.873	0.383	
C_HRS_UNDERTAKEN.y	0.014	0.007	1.975	0.048	*	0.012	0.007	1.745	0.081	.

Note. *MSE = .53282;* $s_p^2 = 0.00434$; $s_s^2 = 0.01544$; *p < .05; **p < .01; ***p < .001

Table D21 Full regression output for cluster-robust standard errors models in the Government course sequence (median split) before and after adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*d.f.*	p-value		Estimate	*SE*	*d.f.*	p-value	
Intercept	-1.474	0.625	81.3	0.021	*	-1.439	0.771	10.71	0.090	.
High RP	-0.055	0.037	14.4	0.152		-0.072	0.043	11.06	0.121	
SAT equivalent	0.001	0.000	87.5	<0.001	***	0.001	0.000	7.97	<0.001	***
derivationAI	0.228	0.205	7.4	0.301		-0.393	0.721	1.30	0.664	
derivationA	0.022	0.080	64.1	0.789		0.001	0.096	12.03	0.990	
derivationB2eH	0.119	0.226	20.7	0.605		-0.085	0.202	4.96	0.691	
derivationB	-0.049	0.116	73.2	0.671		-0.106	0.164	9.92	0.532	
derivationF	0.282	0.147	50.4	0.061	.	0.217	0.208	11.39	0.318	
derivationHPI	0.783	0.174	2.3	0.036	*	0.886	0.311	2.55	0.079	.
derivationH	-0.017	0.086	61.1	0.844		-0.049	0.126	11.96	0.704	
derivationU	0.119	0.700	2.2	0.880		-0.114	0.650	1.66	0.879	
derivationW	0.038	0.070	55.5	0.587		0.011	0.089	11.69	0.902	
motheredlevel1	-0.035	0.026	78.9	0.185		-0.027	0.034	9.19	0.447	
motheredlevel2	0.091	0.166	50.6	0.586		0.006	0.180	5.56	0.976	
motheredlevel3	0.126	0.141	40.8	0.378		0.007	0.169	4.80	0.969	
motheredlevel4	0.263	0.139	41.0	0.066	.	0.160	0.152	5.21	0.339	
motheredlevel5	0.311	0.145	40.0	0.038	*	0.147	0.167	5.17	0.417	
motheredlevel6	0.321	0.141	42.4	0.028	*	0.124	0.167	5.37	0.491	
motheredlevelU	0.368	0.149	42.2	0.018	*	0.274	0.143	5.44	0.109	
fatheredlevel1	0.366	0.249	44.2	0.149		0.142	0.239	3.82	0.587	
fatheredlevel2	0.077	0.151	46.4	0.614		-0.039	0.140	6.12	0.787	
fatheredlevel3	-0.002	0.123	34.4	0.985		0.001	0.115	4.40	0.993	
fatheredlevel4	-0.040	0.131	37.3	0.764		-0.056	0.139	5.89	0.699	
fatheredlevel5	-0.057	0.124	34.5	0.652		0.020	0.135	5.20	0.885	
fatheredlevel6	-0.038	0.124	35.6	0.763		0.049	0.134	5.26	0.731	
fatheredlevelU	-0.026	0.146	36.6	0.859		0.075	0.172	6.02	0.679	
age	-0.074	0.195	39.2	0.708		-0.019	0.181	3.83	0.922	
class_zscore.x	0.558	0.027	82.6	<0.001	***	0.573	0.027	12.25	<0.001	***
CLASSIFICATION2	0.050	0.037	74.5	0.187		0.014	0.042	11.45	0.741	
CLASSIFICATION3	0.054	0.058	78.7	0.355		-0.031	0.078	11.23	0.702	
CLASSIFICATION4	-0.056	0.096	54.5	0.557		0.054	0.107	7.92	0.627	
hspct2	0.203	0.139	94.4	0.147		0.231	0.200	8.69	0.280	
sexW	-0.009	0.033	88.5	0.784		-0.059	0.046	10.51	0.225	
transferredhours	0.001	0.002	97.4	0.675		-0.001	0.002	14.58	0.800	
majorschool3	-0.085	0.114	53.4	0.460		-0.105	0.106	4.55	0.372	
majorschool4	-0.008	0.063	68.5	0.897		-0.027	0.058	9.93	0.652	
majorschool5	-0.042	0.095	78.1	0.658		-0.034	0.100	27.87	0.736	
majorschool9	-0.270	0.205	3.9	0.259		-0.231	0.211	4.30	0.331	
majorschoolC	0.044	0.060	65.7	0.470		0.006	0.082	11.78	0.943	
majorschoolE	0.045	0.054	59.0	0.409		-0.016	0.064	10.58	0.808	
majorschoolJ	0.035	0.151	26.5	0.818		0.036	0.117	9.22	0.763	
majorschoolL	0.153	0.058	63.3	0.011	*	0.156	0.063	12.35	0.029	*
majorschoolN	-0.084	0.290	17.9	0.776		-0.426	0.483	3.44	0.435	
majorschoolS	0.120	0.107	26.2	0.275		0.103	0.135	1.85	0.529	
majorschoolU	0.070	0.051	59.0	0.174		0.078	0.078	11.93	0.341	
C_HRS_UNDERTAKEN.y	0.014	0.008	85.2	0.080	.	0.013	0.010	9.07	0.202	

Note. *p < .05; **p < .01; ***p < .001

Table D22 Full regression output for fixed-effects models in the Government course sequence (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.011	0.673	-1.503	0.133		-0.861	0.693	-1.243	0.214	
High RP	-0.044	0.065	-0.681	0.496		-0.026	0.061	-0.420	0.674	
SAT equivalent	0.001	0.000	7.654	<.001	***	0.001	0.000	7.222	<.001	***
derivationAI	0.113	0.303	0.374	0.708		-0.036	0.309	-0.118	0.906	
derivationA	0.022	0.098	0.225	0.822		0.089	0.096	0.922	0.357	
derivationB2eH	0.121	0.202	0.603	0.547		0.263	0.206	1.277	0.202	
derivationB	-0.085	0.114	-0.742	0.458		-0.075	0.113	-0.663	0.507	
derivationF	0.270	0.145	1.860	0.063	.	0.276	0.140	1.970	0.049	*
derivationHPI	0.954	0.444	2.150	0.032	*	0.954	0.570	1.675	0.094	.
derivationH	-0.037	0.094	-0.397	0.692		0.006	0.093	0.066	0.947	
derivationU	-0.022	0.444	-0.051	0.960		0.415	0.447	0.928	0.354	
derivationW	0.010	0.089	0.115	0.908		0.071	0.087	0.819	0.413	
motheredlevel1	0.059	0.134	0.442	0.659		0.066	0.133	0.492	0.623	
motheredlevel2	0.065	0.123	0.525	0.600		0.062	0.126	0.495	0.621	
motheredlevel3	0.186	0.126	1.471	0.141		0.192	0.128	1.497	0.134	
motheredlevel4	0.242	0.125	1.937	0.053	.	0.224	0.127	1.771	0.077	.
motheredlevel5	0.255	0.128	1.994	0.046	*	0.248	0.130	1.907	0.057	.
motheredlevel6	0.310	0.136	2.274	0.023	*	0.318	0.138	2.308	0.021	*
motheredlevelU	0.324	0.170	1.901	0.057	.	0.438	0.171	2.565	0.010	*
fatheredlevel1	-0.013	0.140	-0.096	0.923		0.001	0.143	0.004	0.997	
fatheredlevel2	0.002	0.124	0.017	0.986		0.059	0.128	0.461	0.645	
fatheredlevel3	-0.029	0.126	-0.226	0.821		-0.034	0.130	-0.259	0.796	
fatheredlevel4	-0.040	0.124	-0.325	0.745		-0.023	0.128	-0.178	0.859	
fatheredlevel5	-0.017	0.126	-0.137	0.891		-0.003	0.130	-0.020	0.984	
fatheredlevel6	-0.035	0.142	-0.245	0.807		0.017	0.146	0.116	0.907	
fatheredlevelU	-0.062	0.159	-0.391	0.696		-0.124	0.161	-0.772	0.440	
age	-0.043	0.026	-1.659	0.097	.	-0.050	0.026	-1.902	0.057	.
class_zscore.x	0.562	0.021	26.969	<.001	***	0.563	0.021	26.988	<.001	***
CLASSIFICATION2	0.078	0.039	2.010	0.045	*	0.088	0.039	2.255	0.024	*
CLASSIFICATION3	0.083	0.065	1.268	0.205		0.110	0.065	1.682	0.093	.
CLASSIFICATION4	0.059	0.103	0.571	0.568		0.121	0.104	1.167	0.243	
hspct2	0.141	0.122	1.154	0.249		0.065	0.122	0.534	0.594	
sexW	0.004	0.033	0.120	0.904		0.018	0.033	0.527	0.598	
transferredhours	0.000	0.002	0.047	0.963		0.000	0.002	-0.062	0.951	
majorschool3	-0.100	0.108	-0.932	0.351		-0.116	0.106	-1.093	0.274	
majorschool4	-0.041	0.066	-0.622	0.534		-0.026	0.066	-0.392	0.695	
majorschool5	-0.113	0.100	-1.126	0.260		-0.128	0.100	-1.277	0.202	
majorschool9	-0.252	0.314	-0.804	0.422		-0.228	0.305	-0.746	0.456	
majorschoolC	0.009	0.069	0.132	0.895		0.020	0.070	0.288	0.773	
majorschoolE	0.008	0.059	0.134	0.894		0.018	0.059	0.307	0.759	
majorschoolJ	0.082	0.149	0.551	0.581		0.076	0.148	0.515	0.607	
majorschoolL	0.127	0.061	2.088	0.037	*	0.127	0.060	2.101	0.036	*
majorschoolN	-0.231	0.174	-1.329	0.184		-0.175	0.175	-0.999	0.318	
majorschoolS	0.075	0.154	0.484	0.629		0.096	0.156	0.616	0.538	
majorschoolU	0.033	0.060	0.559	0.576		0.028	0.059	0.476	0.634	
C_HRS_UNDERTAKEN.y	0.016	0.007	2.245	0.025	*	0.018	0.007	2.452	0.014	*

Note. $n = 2363$; *RMSE* = 0.7294; $df = 2217$, $R^2 = .4490$; *$p < .05$; **$p < .01$; ***$p < .001$

Table D23 Full regression output for random-effects models in the Government course sequence (mean split) before and after propensity-score adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*t*	p-value		Estimate	*SE*	*t*	p-value	
Intercept	-1.421	0.594	-2.391	0.017	*	-1.235	0.599	-2.063	0.039	*
High RP	-0.057	0.047	-1.222	0.243		-0.051	0.047	-1.099	0.296	
SAT equivalent	0.001	0.000	7.704	<.001	***	0.001	0.000	7.268	<.001	***
derivationAI	0.157	0.294	0.533	0.594		-0.136	0.290	-0.470	0.639	
derivationA	0.013	0.096	0.131	0.896		0.082	0.095	0.870	0.384	
derivationB2eH	0.119	0.200	0.597	0.551		0.276	0.204	1.350	0.177	
derivationB	-0.076	0.113	-0.670	0.503		-0.057	0.111	-0.514	0.607	
derivationF	0.265	0.144	1.839	0.066	.	0.283	0.139	2.035	0.042	*
derivationHPI	0.825	0.441	1.868	0.062	.	0.858	0.570	1.506	0.132	
derivationH	-0.036	0.093	-0.386	0.699		0.009	0.091	0.104	0.917	
derivationU	0.056	0.436	0.128	0.898		0.448	0.440	1.018	0.309	
derivationW	0.014	0.087	0.165	0.869		0.079	0.086	0.922	0.357	
motheredlevel1	0.096	0.132	0.728	0.467		-0.045	0.026	-1.753	0.080	.
motheredlevel2	0.111	0.122	0.914	0.361		0.107	0.132	0.809	0.419	
motheredlevel3	0.243	0.125	1.950	0.051	.	0.112	0.124	0.899	0.369	
motheredlevel4	0.299	0.123	2.421	0.016	*	0.251	0.126	1.986	0.047	*
motheredlevel5	0.304	0.127	2.398	0.017	*	0.287	0.125	2.293	0.022	*
motheredlevel6	0.360	0.135	2.669	0.008	**	0.306	0.128	2.386	0.017	*
motheredlevelU	0.367	0.169	2.176	0.030	*	0.375	0.137	2.749	0.006	**
fatheredlevel1	0.044	0.138	0.320	0.749		0.481	0.169	2.840	0.005	**
fatheredlevel2	-0.012	0.123	-0.098	0.922		0.063	0.140	0.448	0.654	
fatheredlevel3	-0.048	0.124	-0.386	0.699		0.039	0.126	0.311	0.755	
fatheredlevel4	-0.059	0.123	-0.484	0.628		-0.053	0.128	-0.412	0.681	
fatheredlevel5	-0.041	0.125	-0.331	0.741		-0.048	0.126	-0.379	0.705	
fatheredlevel6	-0.039	0.141	-0.278	0.781		-0.032	0.128	-0.254	0.800	
fatheredlevelU	-0.081	0.157	-0.518	0.605		0.008	0.144	0.057	0.955	
age	-0.036	0.026	-1.416	0.157		-0.142	0.159	-0.896	0.370	
class_zscore.x	0.559	0.020	27.368	<.001	***	0.561	0.020	27.465	<.001	***
CLASSIFICATION2	0.062	0.037	1.648	0.100	.	0.075	0.038	1.989	0.047	*
CLASSIFICATION3	0.058	0.063	0.923	0.356		0.087	0.063	1.374	0.170	
CLASSIFICATION4	-0.025	0.099	-0.250	0.803		0.046	0.100	0.458	0.647	
hspct2	0.205	0.120	1.709	0.088	.	0.135	0.120	1.128	0.260	
sexW	0.000	0.033	0.004	0.997		0.014	0.033	0.427	0.669	
transferredhours	0.000	0.002	0.018	0.986		0.000	0.002	0.031	0.976	
majorschool3	-0.103	0.106	-0.972	0.331		-0.115	0.104	-1.100	0.272	
majorschool4	-0.023	0.065	-0.355	0.722		-0.009	0.064	-0.143	0.886	
majorschool5	-0.077	0.098	-0.787	0.431		-0.086	0.098	-0.876	0.381	
majorschool9	-0.262	0.309	-0.849	0.396		-0.244	0.299	-0.817	0.414	
majorschoolC	0.029	0.068	0.430	0.668		0.041	0.069	0.604	0.546	
majorschoolE	0.029	0.058	0.493	0.622		0.040	0.058	0.703	0.482	
majorschoolJ	0.050	0.148	0.339	0.734		0.059	0.147	0.402	0.688	
majorschoolL	0.140	0.060	2.353	0.019	*	0.147	0.059	2.481	0.013	*
majorschoolN	-0.137	0.172	-0.797	0.425		-0.097	0.174	-0.557	0.578	
majorschoolS	0.074	0.152	0.483	0.629		0.095	0.154	0.621	0.535	
majorschoolU	0.055	0.059	0.935	0.350		0.052	0.058	0.893	0.372	
C_HRS_UNDERTAKEN.y	0.014	0.007	1.989	0.047	*	0.016	0.007	2.215	0.027	*

Note. *MSE = .53304;* $s_p^2 = 0.003631$; $s_s^2 = 0.01480$; *p < .05; **p < .01; ***p < .001

Table D24 Full regression output for cluster-robust standard errors models in the Government course sequence (mean split) before and after adjustment

Variable	Unadjusted					Adjusted				
	Estimate	*SE*	*d.f.*	p-value		Estimate	*SE*	*d.f.*	p-value	
Intercept	-1.471	0.624	82.0	0.021	*	-0.929	0.811	18.27	0.267	
High RP	-0.023	0.044	26.8	0.608		-0.034	0.047	22.14	0.470	
SAT equivalent	0.001	0.000	88.1	<0.001	***	0.001	0.000	11.27	0.002	**
derivationAI	0.229	0.209	7.4	0.307		-0.264	0.605	1.50	0.718	
derivationA	0.023	0.080	64.2	0.775		0.020	0.084	20.37	0.818	
derivationB2eH	0.118	0.227	20.7	0.608		-0.104	0.220	4.07	0.660	
derivationB	-0.050	0.116	73.1	0.668		-0.041	0.134	13.66	0.765	
derivationF	0.284	0.147	50.4	0.059	.	0.260	0.149	12.41	0.105	
derivationHPI	0.789	0.174	2.3	0.035	*	0.876	0.240	2.46	0.049	*
derivationH	-0.017	0.086	61.1	0.845		-0.014	0.102	20.43	0.892	
derivationU	0.120	0.705	2.2	0.880		0.027	0.688	2.28	0.972	
derivationW	0.039	0.070	55.5	0.581		0.022	0.075	18.86	0.776	
motheredlevel1	-0.036	0.027	79.5	0.186		-0.044	0.033	15.94	0.198	
motheredlevel2	0.091	0.166	50.6	0.587		-0.011	0.161	7.91	0.946	
motheredlevel3	0.127	0.141	40.8	0.374		0.039	0.148	5.73	0.801	
motheredlevel4	0.263	0.139	41.0	0.065	.	0.195	0.129	6.42	0.178	
motheredlevel5	0.311	0.145	40.0	0.038	*	0.214	0.142	6.05	0.181	
motheredlevel6	0.321	0.141	42.3	0.028	*	0.165	0.145	6.46	0.295	
motheredlevelU	0.368	0.149	42.2	0.018	*	0.280	0.127	7.00	0.063	.
fatheredlevel1	0.369	0.249	44.1	0.145		0.174	0.219	5.59	0.460	
fatheredlevel2	0.080	0.152	46.4	0.603		0.062	0.153	8.81	0.695	
fatheredlevel3	0.001	0.123	34.5	0.992		0.001	0.111	5.72	0.994	
fatheredlevel4	-0.038	0.131	37.3	0.776		-0.034	0.128	7.29	0.800	
fatheredlevel5	-0.052	0.125	34.5	0.678		0.028	0.134	6.58	0.839	
fatheredlevel6	-0.035	0.124	35.6	0.781		0.068	0.133	6.65	0.626	
fatheredlevelU	-0.024	0.146	36.6	0.868		0.024	0.162	7.87	0.885	
age	-0.073	0.195	39.2	0.710		-0.030	0.167	4.83	0.865	
class_zscore.x	0.557	0.027	82.5	<0.001	***	0.581	0.026	19.58	<0.001	***
CLASSIFICATION2	0.053	0.038	78.0	0.163		0.074	0.048	25.32	0.136	
CLASSIFICATION3	0.056	0.058	79.4	0.341		0.025	0.077	17.85	0.745	
CLASSIFICATION4	-0.056	0.095	55.1	0.557		0.096	0.103	9.62	0.378	
hspct2	0.200	0.140	94.3	0.155		0.198	0.180	12.24	0.293	
sexW	-0.009	0.033	88.5	0.798		-0.056	0.041	22.25	0.186	
transferredhours	0.001	0.002	97.4	0.687		-0.001	0.002	27.66	0.820	
majorschool3	-0.081	0.114	53.3	0.482		-0.015	0.136	12.25	0.914	
majorschool4	-0.007	0.063	68.6	0.915		-0.028	0.061	17.53	0.647	
majorschool5	-0.047	0.095	78.3	0.626		-0.046	0.102	32.36	0.651	
majorschool9	-0.265	0.206	3.9	0.270		-0.237	0.212	4.18	0.325	
majorschoolC	0.043	0.060	65.5	0.476		-0.040	0.076	17.68	0.603	
majorschoolE	0.046	0.054	59.0	0.401		-0.031	0.061	19.18	0.614	
majorschoolJ	0.033	0.151	26.5	0.828		0.048	0.130	16.53	0.719	
majorschoolL	0.153	0.059	63.4	0.011	*	0.110	0.065	24.08	0.103	
majorschoolN	-0.085	0.291	17.9	0.774		-0.296	0.364	3.60	0.465	
majorschoolS	0.122	0.107	26.3	0.267		0.056	0.152	1.71	0.752	
majorschoolU	0.070	0.051	59.0	0.174		0.064	0.065	21.04	0.338	
C_HRS_UNDERTAKEN.y	0.014	0.008	85.1	0.075	.	0.013	0.010	19.58	0.192	

Note. *p < .05; **p < .01; ***p < .001

References

Abelson, Hal. "The creation of OpenCourseWare at MIT." Journal of Science Education and Technology 17.2 (2008): 164-174.

Act of June 19, 2009, 81st Leg., R.S., ch. 681 (H.B. 2504), § 3, 2009 Tex. Gen. Laws. 1732-1733

Alpay, E., & Verschoor, R. (2014). The teaching researcher: Faculty attitudes towards the teaching and research roles. *European Journal of Engineering Education*, *39*(4), 365-376.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770-1780.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, *28*(25), 3083-3107.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, *34*(28), 3661-3679.

Austin, S. D. M. "A study in logical memory." *The American Journal of Psychology* (1921): 370-403.

Babcock, P., & Marks, M. (2011). The falling time cost of college: Evidence from half a century of time use data. *Review of The Economics course sequence and Statistics*, *93*(2), 468-478.

Bahrick, H. P. (2000). Long-term maintenance of knowledge. *The Oxford handbook of memory*, 347-362.

Bahrick, H. P. (1983). The Cognitive Map of a City: Fifty Years of Learning and Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 17, pp. 125–163).

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General*, *113*(1), 1–29.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of Foreign Language Vocabulary and the Spacing Effect. *Psychological Science*, *4*(5), 316–321.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, *104*(1), 54–75.

Bahrick, H. P., & Hall, L. K. (1991). Lifetime maintenance of high school mathematics content. *Journal of Experimental Psychology: General*, *120*(1), 20–33.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, *2*, 35-67.

Butcher, K. F., McEwan, P. J., & Weerapana, A. (2014). The effects of an anti-grade-inflation policy at Wellesley College. *Journal of Economic Perspectives*, *28*(3), 189-204.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133.

Butler, A. C., & Raley, N. D. (2015). The Future of Medical Education: Assessing the Impact of Interventions on Long-Term Retention and Clinical Care. *Journal of Graduate Medical Education*, *7*(3), 483–485.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking Transfer: A Simple Proposal with Multiple Implications. *Review of Research in Education*, *24*, 61–100.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, *8*.

Brown, J., & Kurzweil, M. (2017). *The Complex Universe of Alternative Postsecondary Credentials and Pathways*. Cambridge, Mass.: American Academy of Arts & Sciences.

Cadez, S., Dimovski, V., & Zaman Groff, M. (2017). Research, teaching and performance evaluation in academia: the salience of quality. *Studies in Higher Education*, *42*(8), 1455-1473.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317-372.

Caplan, B. (2018). *The Case Against Education: Why the Education System is a Waste of Time and Money*. Princeton University Press.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using Spacing to Enhance Diverse Forms of Learning: Review of Recent Research and Implications for Instruction. *Educational Psychology Review*, *24*(3), 369–378.

Carpenter, S. K. (2012). Testing Enhances the Transfer of Learning. *Current Directions in Psychological Science*, *21*(5), 279–283.

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review*, *19*(3), 443–448.

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.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642.

Carson, S. (2009). The unwalled garden: growth of the OpenCourseWare Consortium, 2001–2008. Open Learning, 24(1), 23-29.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380.

Cheah, B. C. (2009). Clustering standard errors or modeling multilevel data. *University of Columbia*, 2-4.

Chen, C. Y. (2015). A study showing research has been valued over teaching in higher education. *Journal of the Scholarship of Teaching and Learning*, *15*(3), 15-32.

Cheng, K. K., Thacker, B. A., Cardenas, R. L., & Crouch, C. (2004). Using an online homework system enhances students' learning of physics concepts in an introductory physics course. *American journal of physics*, *72*(11), 1447-1453.

Chickering, A. W., & Gamson, Z. F. (1987). Seven Principles for Good Practice in Undergraduate Education. AAHE Bulletin.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295-313.

Coley, R. J., Goodman, M. J., & Sands, A. M. (2015). *America's skills challenge: Millennials and the future*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/s/research/30079/asc-millennials-and-the-future.pdf

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of educational research*, *66*(3), 227-268.

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of educational research*, *76*(1), 1-62.

Coppola, B. P., & Krajcik, J. S. (2013). Discipline-centered post-secondary science education research: Understanding university level science learning. *Journal of Research in Science Teaching*, *50*(6), 627-638.

Corlu, M. S. (2013). Insights into STEM Education Praxis: An Assessment Scheme for Course Syllabi. *Educational Sciences: Theory and Practice*, *13*(4), 2477–2485.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. Psychological science, 15, 68-93.Cullen, R.., & Harris, M. (2009). Assessing learner-centredness through course syllabi. *Assessment & Evaluation in Higher Education*, *34*(1), 115–125.

Cullen, R., & Harris, M. (2009). Assessing learner-centredness through course syllabi. *Assessment & Evaluation in Higher Education*, *34*(1), 115–125.

Custers, E. J. (2010). Long-term retention of basic science knowledge: a review study. *Advances in Health Sciences Education*, *15*(1), 109-128.

Cribb, T. K., & Bunting, J. (2008). Our fading heritage: Americans fail a basic test on their history and institutions. *Intercollegiate Studies Institute*.

Crispe, I. (2017, June 28). Coding Bootcamps With Job Guarantees [Web log post]. Retrieved from https://www.coursereport.com/blog/guide-to-coding-bootcamps-with-job-guarantees

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627–634.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved Learning in a Large-Enrollment Physics Class. *Science*, *332*(6031), 862–864.

Dolan, B. M., Yialamas, M. A., & McMahon, G. T. (2015). A Randomized Educational Intervention Trial to Determine the Effect of Online Education on the Quality of Resident-Delivered Care. *Journal of Graduate Medical Education*, *7*(3), 376–381.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14(1), 4-58.

Ebbinghaus, H. (1913). *Memory; a contribution to experimental psychology*. New York city: Teachers college, Columbia university.

Educational Testing Service. (2016). *America's skills challenge: Millennials and the future*. Retrieved from https://www.ets.org/s/research/29836/

Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on ratemyprofessors.com. *Assessment & Evaluation in Higher Education*, *33*(1), 45-61.

Foerster, N. (1937). *The American state university, its relation to democracy*. The University of North Carolina Press.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410-8415.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American journal of epidemiology*, *173*(7), 761-767.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models* (Vol. 1). New York, NY, USA: Cambridge University Press.

Gerhard, D. (1955). The emergence of the credit system in American education considered as a problem of social and intellectual history. *Bulletin of the American Association of University Professors (1915-1955)*, *41*(4), 647-668.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355.

Gower J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857-872.

Graves, R., Hyland, T., & Samuels, B. M. (2010). Undergraduate Writing Assignments: An Analysis of Syllabi at One Canadian College. *Written Communication*, *27*(3), 293–317.

Greifer, N. (2017) cobalt: Covariate Balance Tables and Plots. R package version 3.1.0. https://CRAN.R-project.org/package=cobalt

Halpern, D. F., & Hakel, M. D. (2003). Applying the Science of Learning to the University and Beyond: Teaching for Long-Term Retention and Transfer. *Change: The Magazine of Higher Learning*, *35*(4), 36–41.

Hart Research Associates. (2010). *Raising the bar: Employers' views on college learning in the wake of the economic downturn*. Washington, DC: Association of American Colleges and Universities.

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

Helwig, N. E. (2017). Adding bias to reduce variance in psychological results: A tutorial on penalized regression. *The Quantitative Methods for Psychology*, *13*(1), 1-19.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199-236.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1-28.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Hu, M. & Liu, B. (2004) Mining and Summarizing Customer Reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining*, 21-34.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171(2), 481-502.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706-710.

Jacobs, J. A. (2004, March). Presidential address: The faculty time divide. In *Sociological Forum* (Vol. 19, No. 1, pp. 3-27). Kluwer Academic Publishers-Plenum Publishers.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun use reflects standings in social hierarchies. Journal of Language and Social Psychology, 33, 125143.

Kang, S. H. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 12-19.

Kaufman, L., & Rousseeuw, P. (1987). Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical Data Analysis Based on L1 Norm and Related Methods* (405–416). North-Holland.

Kerfoot, B. P., Fu, Y., Baker, H., Connelly, D., Ritchey, M. L., & Genega, E. M. (2010). Online Spaced Education Generates Transfer and Improves Long-Term Retention of Diagnostic Skills: A Randomized Controlled Trial. *Journal of the American College of Surgeons*, *211*(3), 331–337.

King, A. (1993). From Sage on the Stage to Guide on the Side. *College Teaching*, *41*(1), 30–35.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *The Economics course sequence of Education Review*, *47*, 180-195.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97.

Kutner, M., Greenberg, E., & Baer, J. (2006). A First Look at the Literacy of America's Adults in the 21st Century. NCES 2006-470. *National Center for Education Statistics*.

Lacy, S. (2011). Peter Thiel: we're in a bubble and it's not the internet. It's higher education. *Tech Crunch*, *10*.

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, *6*(3), e18174.

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in medicine*, *32*(19), 3373-3387.

Lough, J. R. (1997). The Carnegie Professors of the Year: Models for teaching success. *Inspiring Teaching: Carnegie Professors of the Year Speak*, 212–225.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, *23*(19), 2937-2960.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9 (Nov), 2579-2605.

Mann, C. J. (2003). Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency medicine journal*, *20*(1), 54-60.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects

of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414.

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied*, *21*(4), 370–382.

McKeachie, W., & Svinicki, M. (2013). *McKeachie's Teaching Tips*. Cengage Learning.

Mohammad, S., & Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29 (3), 436-465.

National Academies of Sciences, Engineering, and Medicine. (2018). *Indicators for monitoring undergraduate STEM education*. National Academies Press.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. Personality and Social Psychology Bulletin, 29, 665675.

Nielson, A.F. (2013). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: 718 CEUR Workshop Proceedings: 93-98.

Nilson, L. B. (2010). *Teaching at Its Best: A Research-Based Resource for College Instructors*. John Wiley & Sons.

Oakes, J. M. (2004). The (mis) estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Social science & medicine*, *58*(10), 1929-1952.

Pagès, J. (2014). *Multiple factor analysis by example using R*. Chapman and Hall/CRC.

Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of academically adrift?. *Change: The Magazine of Higher Learning*, *43*(3), 20-24.

Pennebaker, J.W. (2011). The Secret Life of Pronouns: What Our Words Say About Us (NY: Bloomsbury).

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily Online Testing in Large Classes: Boosting College Performance while Reducing Achievement Gaps. *PLoS ONE*, *8*(11), e79774.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. PloS One, 9, 110.

Peter D. Hart Research Associated, Inc. How Should Colleges Assess and Improve Student Learning? Employers Views on the Accountability Challenge. Washington, DC: Peter D. Hart Research Associated, Inc., 2008.

Pitt, R. (2015). Mainstreaming open textbooks: Educator perspectives on the impact of openstax college open textbooks. *The International Review of Research in Open and Distributed Learning*, *16*(4).

Podani, J. (1999). Extending Gower's general coefficient of similarity to ordinal characters. Taxon, 331-340.

Popham, W. J. (1978). As Always, Provocative. *Journal of Educational Measurement*, *15*(4), 297–300.

Primo, D. M., Jacobsmeier, M. L., & Milyo, J. (2007). Estimating the impact of state policies and institutions with mixed-level data. *State Politics & Policy Quarterly*, *7*(4), 446-459.

Pustejovsky, J. (2017). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. *R package version 0.2*, *3*.

Pustejovsky, J. E., & Tipton, E. (2014). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics*, 1-12.

Pyle, W. H. (1913). Economical learning. *Journal of Educational Psychology*, *4*(3), 148–158.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rawson, K. A., & Kintsch, W. (2005). Rereading Effects Depend on Time of Test. *Journal of Educational Psychology*, *97*(1), 70–80.

Remler, D. K., & Pema, E. (2009). *Why do institutions of higher education reward research while selling education?* (No. w14974). National Bureau of Economic Research.

Revelle, W. R. (2017). psych: Procedures for personality and psychological research.

Richards-Babb, M., Drelick, J., Henry, Z., & Robertson-Honecker, J. (2011). Online homework, help or hindrance? What students think and how they perform. *Journal of College Science Teaching*, *40*(4), 81.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239.

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review*, *27*(4), 635-643.

Romano, A. (2011). How Ignorant Are Americans? *Newsweek, March*, *20*.

Rosenbaum, P. R. (1984). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 656-666.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of The Economics course sequence*, *125*(1), 175-214.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, D. B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169–188.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322-331.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, *2*(6), 110-114.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, *13*(4), 279.

Shulman, L. S. (2005). Signature pedagogies in the professions. *Daedalus*, *134*(3), 52–59.

Shulman, L. S. (2013). Those Who Understand: Knowledge Growth in Teaching. *Journal of Education*, *193*(3), 1–11.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The Validity of Student Evaluation of Teaching in Higher Education: Love me, love my lectures? *Assessment & Evaluation in Higher Education*, *25*(4), 397–405.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, *24*(4), 35-43.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. *CBE-Life Sciences Education*, *12*(4), 618-627.

Snyder, T. D., de Brey, C., & Dillow, S. A. (2018). Digest of Education Statistics 2016, NCES 2017-094. *National Center for Education Statistics*.

Snyder, T. D. (Ed.). (1993). *120 years of American education: A statistical portrait*. DIANE Publishing.

Spellings, M. (2006). *A test of leadership: Charting the future of US higher education*. US Department of Education.

Stanny, C., Gonzalez, M., & McGowan, B. (2015). Assessing the culture of teaching and learning through a syllabus review. *Assessment & Evaluation in Higher Education*, *40*(7), 898–913.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., ... & Levis-Fitzgerald, M. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468-1470.

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

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science, 25*(1), 1–21.

Talbot, R. M., Hartley, L. M., Marzetta, K., & Wee, B. S. (2015). Transforming undergraduate science education with learning assistants: Student satisfaction in large-enrollment courses. *Journal of College Science Teaching*, *44*(5), 24-30.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Association analysis: basic concepts and algorithms. *Introduction to Data mining*, 327-414.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* (*Methodological)*, 267-288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology)*, *73*(3), 273-282.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research*, *46*(1), 90-118.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, *55*(4), 525-534.

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in medicine*, *27*(11), 1934-1943.

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

Wellman, J. (2005). The student credit hour counting what counts. *Change: The Magazine of Higher Learning*, *37*(4), 18–23.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wickham, H. (2016). tidyr: Easily Tidy Data with spread() and gather() Functions. *Version 0.6.0*.

Wickham, H., Francois, R., Henry, L., & Müller, K. (2015). dplyr: A grammar of data manipulation. *R package version 0.4*, *3*.

Wieman, C., & Gilbert, S. (2014). The teaching practices inventory: A new tool for characterizing college and university teaching in mathematics and science. *CBE-Life Sciences Education,* 13(3), 552-569.

Wieman, C. (2015). A Better Way to Evaluate Undergraduate Teaching. *Change: The Magazine of Higher Learning*, *47*(1), 6–15.

Williams, C. T., Walter, E. M., Henderson, C., & Beach, A. L. (2015). Describing undergraduate STEM teaching practices: a comparison of instructor self-report instruments. *International Journal of STEM Education*, *2*(1), 18.

Windemuth, A. (2014, October 6). After faculty vote, grade deflation policy officially dead. Daily Princetonian. Retrieved from http://dailyprincetonian.com/news/2014/10/breaking-after-faculty-vote-grade-deflation-policy-officially-dead/

Workforce-Skills Preparedness Report. (n.d.). Retrieved February 5, 2018, from https://www.payscale.com/data-packages/job-skills

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.

Young, P. (2006). Out of balance: Lecturers' perceptions of differential status and rewards in relation to teaching and research. Teaching in higher education, 11(2), 191-202.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators.