Title: Establishing a New Standard of Care for Calculus Students with Evidence from a Randomized Trial

Authors: Laird Kramer[1,4*], Edgar Fuller[1,2,3*], Charity Watson[1,2,3], Adam Castillo[1,2,3], Pablo Duran Oliva[1,2,5], Geoff Potvin[1,4]

Affiliations:

[1] STEM Transformation Institute, Florida International University

[2] Center for Transforming Teaching in Mathematics, Florida International University

[3] Department of Mathematics and Statistics, Florida International University

[4] Department of Physics, Florida International University

[5] Department of Mathematics and Applied Mathematics, Virginia Commonwealth University

* Correspondence to: Laird.Kramer@fiu.edu, Edgar.Fuller@fiu.edu

Abstract: Calculus, the study of change in processes and systems, serves as the foundation of many STEM disciplines. Traditional, lecture-based calculus instruction presents a persistent barrier for students seeking STEM degrees, limits access to STEM professions, and blocks their potential to address society's challenges. A large-scale pragmatic randomized trial was conducted to compare two calculus instruction styles: active student engagement (treatment condition) versus traditional, lecture-based instruction (control condition). A sample of 811 U.S. university students were studied across 32 sections taught by 19 instructors over three semesters at a large U.S. Hispanic-Serving Institution. Large effect sizes were consistently measured for student learning outcomes in the treatment condition, which demonstrated a new standard for calculus instruction and increased opportunities for completion of STEM degrees.

One-Sentence Summary: A new Standard of Care for Calculus Instruction is proposed, focused on student engagement and supported by experimentally confirmed evidence of substantially stronger learning outcomes and student success.

Main Text:

Calculus instruction needs significant transformation as it serves as a frequent barrier to STEM degree attainment, especially for traditionally underrepresented groups (*1–3*), depriving both individuals and society of the potential benefits of their inclusion. National calls for calculus transformation are numerous (*4*, *5*), as failing calculus can contribute to a student's departure from STEM degree programs. Only about 40% of students entering universities with STEM degree intentions actually graduate with a STEM degree (*6*). More concerning is that the odds of female students switching out of STEM after a calculus course is about 1.5 times higher than that of comparable male students (*3*). Furthermore, Hispanic and Black/African American students had more than 50% higher failure rates than White students in calculus (*7*, *8*).

Evidence-based instruction, implemented in many STEM disciplines, has reliably led to profound improvement in student success (*9–11*). However, common approaches to calculus instruction continue to rely on traditional, lecture-based practices, where students are passive learners in the classroom, expected to construct their knowledge mostly outside of the classroom, on homework, or in recitation sessions (*12*). Mathematics, as a discipline, thus needs to embrace its role in enabling STEM careers that will lead to prosperity for both individuals and society at large. "Calculus … must become a pump and not a filter" for the STEM pipeline, as noted by Robert White, President of the National Academy of Engineering in 1988 (*13*). Handlesman, et al*,* (*14*) recently argue that, "We must fix the classrooms where many students from historically excluded communities are discouraged from pursuing STEM" and that "…the continued exclusive use of lectures is malpractice at best, or an act of discrimination at worst." Thus, it is imperative that dramatic transformation in calculus instruction takes place to promote more equitable learning environments for all students.

We present a large-scale randomized trial carried out to rigorously compare an evidence-based, active student engagement calculus course to traditional, lecture-based calculus instruction. The work extends prior Calculus research investigations (*15–17*) by including random assignment of students to treatment and control sections as well as anonymized analysis of the identical end-of-semester learning outcomes. The study utilizes a pragmatic randomized trial (*18*) design to inform on the effectiveness of similar interventions at higher education institutions, reflecting real-world classroom constraints. In these contexts, blinding of the treatment and control conditions to both students and faculty is not possible, as blinding is only feasible when the treatment and control conditions remain unknown to the participants during the period of study (such as in a clinical trial drug study). As with some public health or sociological interventions, enrollment of participants in this study reveals some aspects of a cohort structure but it is still possible to maintain the essential aspects of random assignment, following a modified protocol as in Zwarenstein, et al (*18*). The treatment condition integrated a suite of coherent strategies that have been independently found (*19*, *20*) to improve student learning; thus, the treatment was a significant departure from traditional instruction, and it was not logically possible for the treatment condition to remain hidden from students or faculty after the treatment began. Random assignment of faculty to control or treatment conditions would not be possible because an individual faculty member's knowledge, philosophy, and experience with a variety of classroom strategies and instructional practices may intersect with the features of the treatment or control conditions. The experimental protocol thus included a group of instructors willing to adopt the instructional methods in the treatment condition. This comprehensive experimental approach was intended to secure the strongest possible evidence for critical stakeholders to sustain the treatment beyond the trial.

The treatment condition used the Modeling Practices in Calculus (MPC) curriculum and pedagogy, and the control condition represented the pre-existing, traditional instructional practices at the study institution. MPC integrates the practices of mathematicians as a central design tenet throughout the course. Instructors facilitate students application of mathematical "habits of mind" (*21*) that foster deeper understanding of calculus concepts including the identifying of patterns, hypothesis development and testing, making connections, and communicating ideas precisely to learn calculus throughout the course. Class time is devoted to students working collectively in small groups on pre-designed notes and learning activities developing their calculus understanding with minimal lecturing. Treatment included Learning Assistants (LAs) (*22*) who are undergraduate peers integrated within the instructional team to facilitate student learning and promote culturally responsive instruction. The curriculum promotes mathematical practices (sense-making, problem solving, argumentation, etc.) and established strategies to optimize student engagement: Cooperative Learning, Argumentation and Metacognition, Mathematical Fluency, and a Culturally Responsive Environment (*23*) (described in the Supplementary Materials (SM Section 2)). The MPC design builds on the SCALE-UP Calculus (*24*) model-and intentionally embodies well-established recommendations for calculus instruction including ambitious teaching practices and strategies promoted by national mathematics societies and national reports (*12*, *20*, *25–28*).

The study was carried out at Florida International University (FIU) in Miami, Florida, the fourth largest public research university in the United States, with 58,787 students, of which 41,795 are undergraduates (Fall 2019 (*29*)). FIU is a Hispanic-Serving Institution as 64% of students identify as Hispanic/Latino/a/. Moreover, 79% of the students identify as members of historically underrepresented racial/ethnic minority groups, and 57% are women. The institution's size provided a unique opportunity to carry out this study, as there are 18-34 40-student sections of Calculus 1 being taught each semester and primarily serving STEM majors. Furthermore, institutional conditions created urgency to transform calculus, as historic pass rates in introductory calculus averaged 55% (range of 13%–88%) over the six semesters prior to the project's pilot.

Research Design: A pragmatic randomized trial (*30–32*) of the MPC approach was carried out during the Fall 2018, Spring 2019, and Fall 2019 semesters to rigorously test student outcomes. Students were randomly assigned individually to treatment and control conditions at the beginning of the semester, after enrolling in sections based on their scheduling preferences using the institution's enrollment system. To accommodate the randomized assignments, each of the experimental sections doubled in size from the usual 40-seats to 80-seats prior to enrollment opening. Instructor names and section sizes were invisible to students throughout the enrollment phase. Just before each term, the 80-seat sections were split into two 40-seat sections by assigning each student at random to either a treatment or control section.

Once assigned, the treatment sections implemented the MPC approach while the control sections were unchanged. After assignment, students were free to change/drop/add course sections up until the regular institutional drop/add deadline (seven days after classes begin). To account for such changes, enrollments were monitored and only students who were randomly assigned to either a treatment or control section and remained in that section through the regular, non-penalty drop/add deadline were included in the data for the experimental study reported below. In total, 1,019 students were randomly assigned to either the treatment or control groups. Of these, 516

students were assigned to the treatment group and 417 remained in the section at the drop/add deadline. At the same time 503 students were assigned to the control group and 394 students remained in the section at the drop/add deadline. The study follows the Consolidated Standards for Reporting Trials (CONSORT) (*18*, *33*, *34*). The specifics of recruiting, enrollment, assignment, and completion for the trial are in SM Sections 1.3 and 3.1. The randomization process produced comparable groups by mathematical background and demographics; class sizes were typical for the course (SM Section 3).

Faculty participating in the study included seven individuals teaching 16 treatment sections along with 12 individuals teaching 16 control sections. Faculty recruited to teach the treatment sections indicated a willingness to adopt and implement the MPC approach, replicating the authentic condition of faculty reforming their classroom practice under the study design. To prepare for the new instructional approach, faculty participated in a two-day, pre-semester professional development workshop and were provided with the MPC curricular materials. Consistency of the MPC treatment was monitored through weekly preparation meetings where the course objectives and pacing were discussed. In-class monitoring by the project team was deemed overly intrusive and disruptive to classroom engagement. Control-section faculty were not guided to use any particular practices and chose their normal instructional practices, best described as traditional lecture format with at most limited student engagement. Potential effects of instructor differences on learning outcomes were investigated, presented in SM Section 3, and summarized below.

The student outcome measures reported include identical end-of-semester learning measures as well as course success data (i.e. course grades). The end-of-semester learning measures focused on evaluating learning using a set of identical assessment items (problems) developed by instructors spanning all calculus sections and spanning the major learning objectives of a Calculus 1 course. The aim was to determine how well students understood essential elements of, and exhibited fluency and technical competency in, calculus at course end. Assessment items aligned to both local and national standards(*35*), were embedded in a cumulative final exam, and were administered to all students in each treatment and control section. To ensure fidelity and fairness to both treatment and control sections, control and treatment faculty collaboratively developed a set of items to be administered to both conditions in identical format and wording. This set of identical items formed roughly two-thirds of the total final exam content, with the remaining items added by individual faculty in a separate section of the exam, allowing them to address their specific instructional goals. Furthermore, the exams and problems were formatted identically and without course section identifiers to allow completely anonymized evaluation during the subsequent comparative analysis. The identical items covered core calculus topics including evaluating limits, identifying extrema, curve sketching, related rates, and evaluating indefinite integrals. For the second and third semesters, additional items focusing on implicit differentiation and optimization were added to the identical set of items. Details are included in SM Section 3.3. Course success data (grades) reflect the overall assessment of students as assigned by each section's instructor. Course grade policies were established by individual instructors following departmental syllabus guidelines and were broadly consistent across sections and semesters.

Analysis of the end-of-course learning measures utilized a rubric for each problem, with five researchers testing the initial rubric on a subset of exams to establish inter-rater reliability. The final rubric represented consensus on all elements and accounted for initial ambiguity or disagreement. The analysis was carried out by a team of 10 trained evaluators, each of whom

evaluated a completely anonymous set of student solutions. An average of two evaluators reviewed each solution for correctness on a scale from 0-100%. The evaluators were very consistent with high inter-rater reliability (Cohen's kappa 0.827 in Fall 2018 and 0.797 in Spring 2019) (*36*, *37*). The same rubric was applied to the Fall 2019 data given its high degree of agreement. Once all problems were evaluated, the research team de-anonymized and sorted the results by treatment and control sections for the comparative analysis.

Results: The results indicate significant improvements in student learning for the MPC group across all three semesters. Students in the treatment group showed substantially higher scores on the identical end-of-semester learning outcomes: (Fall 2018: d=0.505, p<0.01; Spring 2019: d=0.748, p<0.001; Fall 2019: d=0.925, p<0.001) when compared to the control group. Combining results from all three semesters of trials (i.e., 32 sections and 811 total students), the overall difference between treatment and control is d=0.774 (95% confidence interval 0.618 to 0.930), a medium/large effect size (*36*, *37*). Overall, treatment group students show more consistency in applying the tools of calculus to optimization problems, using derivatives to sketch graphs of functions, evaluating limits, and evaluating integrals.

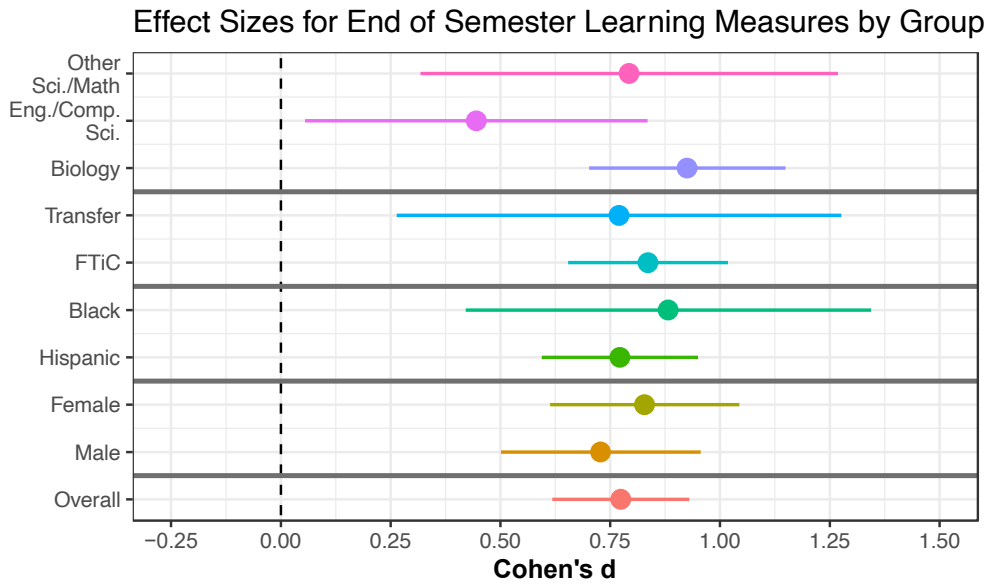


Fig. 1: Overall end-of-semester learning measures effect sizes broken out by major, race/ethnicity, and gender. Error bars indicate the 95% confidence interval for effect size for each group.

The success of the MPC intervention occurs across racial and ethnic groups, majors and academic pathways, and genders (Fig. 1). Similar medium/large overall effect sizes were observed for students in the treatment condition who identified as Black/African-American (d=0.882, p<0.001) or Hispanic/Latino/a (d=0.772, p<0.001) when directly comparing the identical learning measures to their counterparts in the control condition. While all STEM majors showed significantly improved learning, there were larger effect sizes for Biology majors in the treatment group (d=0.925, p<0.001). Students matriculating onto campus as both First Time in College (FTiC) and Transfer students showed medium/large effect sizes, and the majority were FTiC.

Furthermore, students of the MPC treatment condition had improved course grades. Average grades were significantly higher by ~0.4 points (4.0 grade point scale) in MPC sections across all semesters of the study (p<0.001, d=0.295). This translated to success rates (A/B/C grades) averaging 11% higher in MPC sections compared to traditional sections (p<0.001, d=0.251, Fig. 2). Outcomes were consistent across the three semesters of the experiment, Fig. 3. Moreover, the MPC sections also had lower course late drop rates (departure after the regular drop/add period ends) across all three semesters (p<0.05, d=0.141), suggesting students more clearly perceived they were likely to succeed in the course.

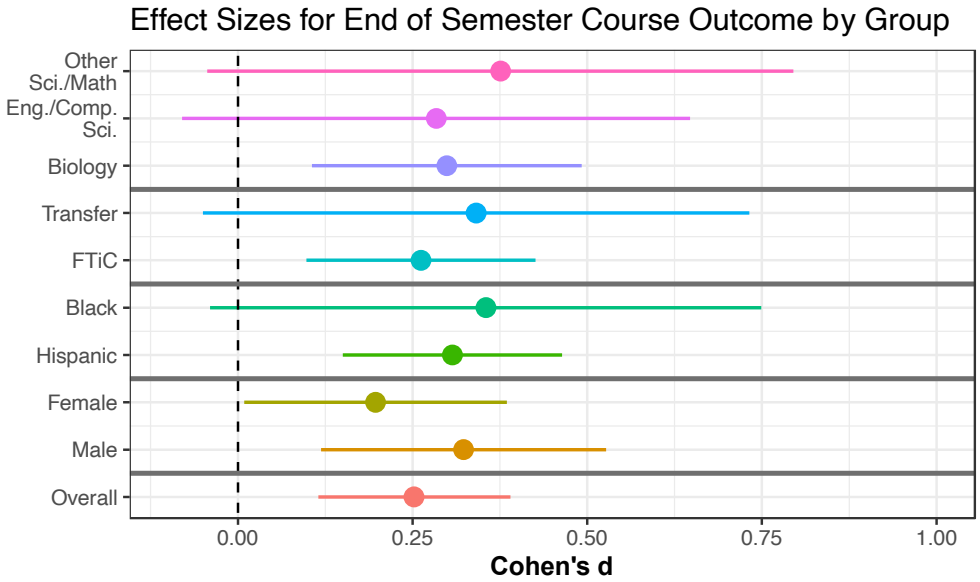


Fig. 2: Overall course success (i.e., earned grades of A, B, or C) effect sizes broken out by major, race/ethnicity, and gender. Error bars indicate the 95% confidence interval for effect size for each group.

The trend of improved outcomes in course success is also observed for demographic subgroups, seen in Fig. 2. A logistic regression model of success using gender identification, FTiC status, and Hispanic identification as independent variables showed the odds of a female-identified student in the treatment group passing the course to be 58% higher than the odds of a female-identified student in the control (b_1=0.46, p<0.05). Hispanic students' odds of passing the course were almost double that of their counterparts in the control (b_1=0.70, p<0.001). The likelihood of FTiC students in the treatment passing the course saw an increase by about 85% when compared to these students in the control (b_1=0.61, p<0.01). Details are included in SM Section 3.4.

Potential biases arising in the random student assignment and faculty selections were investigated for hidden level effects or confounders to establish limitations of the study (see SM Section 3). The randomization process showed equivariance in the demographics of student allocation. Analyses showed that allowing students to drop/add sections during the open registration period after the initial assignment did not impact the measured outcomes. Faculty characteristics were compared and found to be similar in both background and prior course student grade distributions. A mixed effects model with student fixed effects and random cluster effects due to section and instructor levels was fit (SM Section 3.2.4.1) with tests of fixed effects computed using Satterthwaite approximations to control for Type I errors. The explanatory

power of the model was found to be high (conditional R^2=0.39) and the portion related to the fixed effects was 0.303. The effect of Treatment was statistically significant, and explanatory of 0.119 (semi-partial R^2) of the outcome variance. This implies an estimated effect size of Cohen's f=0.371 with covariates and cluster level effects present. Random effects explain 0.0852 of outcome variance with an intraclass correlation of 0.14. A sensitivity analysis showed (SM Section 3.2.4.2) that unmeasured confounders would need to be four times more powerful than any measured covariate including student mathematics background to be responsible for the observed effect.

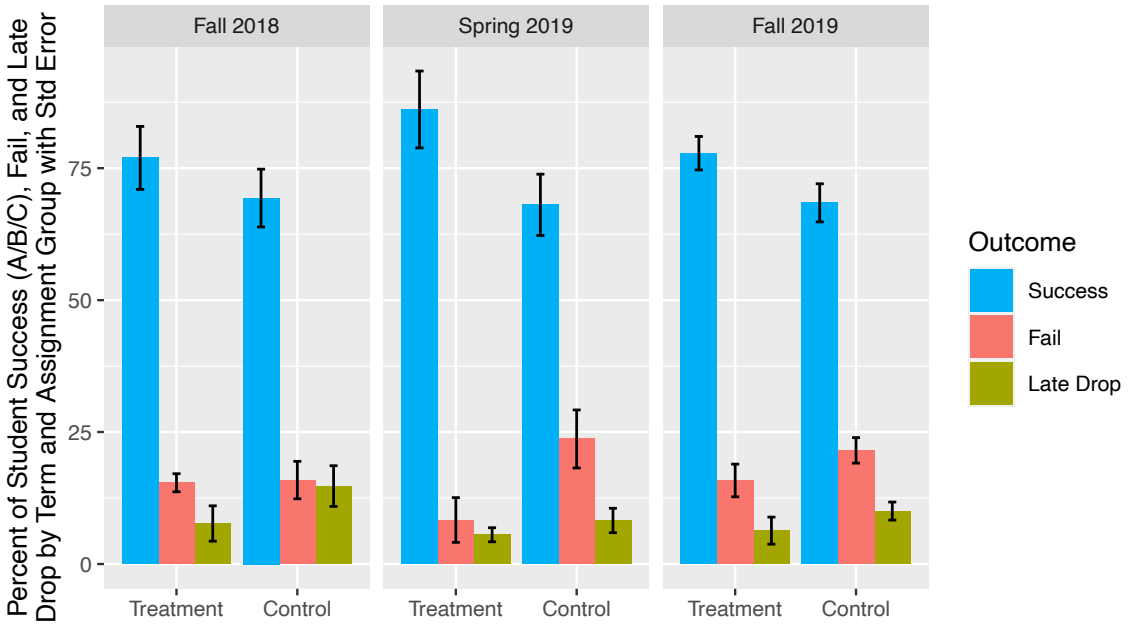


Fig. 3: Final course grade outcomes broken out by term and curriculum, including Success (earned grades of A, B, or C), Fail (earned grades of D or F), and Late Drops (withdrawals or drops after the institution's drop/add deadline). Vertical scale is percent by outcome, error bars indicate the 95% confidence interval for the mean percentage of students in each outcome group over all sections in a term.

Discussion and Conclusion: This pragmatic randomized trial demonstrates that student learning outcomes were significantly improved in the treatment condition. Contrary to previous research (*38*), this study shows that when students are expected to engage with calculus concepts collaboratively, using intentional, evidence-based teaching strategies, they develop a better understanding of calculus concepts and techniques. Importantly, the benefits of the MPC curriculum and pedagogy are realized regardless of racial/ethnic group, gender, or major/academic pathway. These trends suggest that the treatment includes culturally responsive and equitable strategies. Specifically, the MPC learning environment is designed to promote learning communities that provide ongoing support for learning mathematics through collaborative engagement and ongoing formative feedback. This aims to promote inclusion and increases access for students with different mathematical backgrounds, different cultural identities, and different life experiences by allowing them to utilize their mathematics skills in a supportive, non-threatening environment.

The improved learning and course success for Modeling Practices in Calculus reported in this study have profound implications for calculus instruction. This study demonstrates the substantial benefit to students of the MPC approach designed around established, evidence-based principles and should motivate educators in mathematics and other STEM disciplines to adopt the same or similar approaches and conduct similar studies to replicate these findings. Improved student success also leads to more efficient student progress to graduation and boosts institutional effectiveness. Applying this study's 11% average improvement in pass rate to all 2,000 first-time calculus students at FIU, would translate to 220 additional students succeeding in calculus annually and reducing the instructional load by five sections annually. Extending this strategy to the roughly 300,000 students across the nation taking Calculus 1 each year, these results translate to a potential of an additional 33,000 students passing calculus each year, saving students an estimated $23.9M in tuition (based on a 3-credit course at the average public college/university tuition rate of $242/credit (*39*, *40*). Pragmatic randomized trials provide guidance on what can be achieved by engaging faculty willing to change their instruction. These results potentially represent a lower bound on the long-term effects, as faculty likely develop additional expertise through continued instruction and realize improved outcomes. The measured effect size provides rationale to stop the control due to treatment benefit, if one follows medical research protocols (*41*, *42*).

The experimental methodology establishes a new Standard of Care for calculus instruction and a high standard of evidence to bear on understanding the impacts on student learning. Improved learning of calculus aims to foster higher success in future STEM courses and develop the STEM "habits of mind" students take with them into their future careers. Further, MPC shows potential to address the disparities that differentially impact historically underrepresented groups, thus offering a mechanism to address Handelsman, et al (*14*)'s call to promote the success of historically excluded communities. We envision a mathematics experience for all students built on this approach and advocate that active student engagement must be deployed across all STEM disciplines to improve our development of future STEM professionals from all backgrounds.

References and Notes:

1. U. Treisman, Studying Students Studying Calculus: A Look at the Lives of Minority Mathematics Students in College. *The College Mathematics Journal*. **23**, 362-372, (1992).

2. B. B. Alexander, A. C. Burda, S. B. Millar, A community approach to learning calculus: Fostering success for underrepresented ethnic minorities in an emerging scholars program. *Journal of Women and Minorities in Science and Engineering*. **3** (1997).

3. J. Ellis, B. K. Fosdick, C. Rasmussen, Women 1.5 Times More Likely to Leave STEM Pipeline after Calculus Compared to Men: Lack of Mathematical Confidence a Potential Culprit. *PLOS ONE*. **11, e0157447** (2016).

4. S. L. Ganter, D. J. Lewis, D. Hughes-Hallett, Calculus reform. *Science*. **279**, 2019–2025 (1998).

5. S. Olson, D. G. Riordan, *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President* (Executive Office of the President, 2012).

6. E. Seymour, N. M. Hewitt, *Talking about leaving* (Westview Press, Boulder, CO, 1997).

7. E. Seymour, A.-B. Hunter, Eds., *Talking about Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education* (Springer International Publishing, Cham, 2019; http://link.springer.com/10.1007/978-3-030-25304-2).

8. A. K. Koch, B. Drake, Digging into the disciplines: The impact of gateway courses in accounting, calculus, and chemistry on student success. *Chemistry*. **34**, 29–4 (2018).

9. M. J. Graham, J. Frederick, A. Byars-Winston, A. B. Hunter, J. Handelsman, Increasing persistence of college students in STEM. *Science*. **341**, 1455–1456 (2013).

10. M. Stains, J. Harshman, M. K. Barker, S. V. Chasteen, R. Cole, S. E. DeChenne-Peters, A. M. Young, Anatomy of STEM teaching in North American universities. *Science*. **359**, 1468–1470 (2018).

11. C. Henderson, M. Connolly, E. L. Dolan, N. Finkelstein, S. Franklin, S. Malcom, K. S. John, Towards the STEM DBER alliance: Why we need a discipline-based STEM education research community. *International Journal of Research in Undergraduate Mathematics Education*. **3**, 247–254 (2017).

12. C. Rasmussen, N. Apkarian, J. E. Hagman, E. Johnson, S. Larsen, D. Bressoud, Brief Report: Characteristics of Precalculus Through Calculus 2 Programs: Insights From a National Census Survey. *Journal for Research in Mathematics Education*. **50**, 98–111 (2019).

13. L. A. Steen, *Calculus for a New Century: A Pump, Not a Filter. Papers Presented at a Colloquium (Washington, DC, October 28-29, 1987). MAA Notes Number 8.* (ERIC, 1988).

14. J. Handelsman, S. Elgin, M. Estrada, S. Hays, T. Johnson, S. Miller, V. Mingo, C. Shaffer, J. Williams, Achieving STEM diversity: Fix the classrooms. *Science*. **376**, 1057–1059 (2022).

15. J. Bookman, C. P. Friedman, "The Evaluation of Project Calc at Duke University, 1989-1994" in *Assessment Practices in Undergraduate Mathematics*, B. Gold, S. Keith, W. A. Marion, Eds. (MATHEMATICAL ASSOCIATION OF AMERICA, 1999), *MAA Notes*, p. 253.

16. M. Brown, "Planning and change: The Michigan calculus project" in *Calculus: The Dynamics of Change*, W. Roberts, Ed. (MATHEMATICAL ASSOCIATION OF AMERICA, 1996), vol. 39 of *MAA Notes*, pp. 52–58.

17. E. Tidmore, A comparison of calculus materials used at Baylor University. *Focus on Calculus*. **7**, 5–6 (1994).

18. M. Zwarenstein, S. Treweek, J. J. Gagnier, D. G. Altman, S. Tunis, B. Haynes, A. D. Oxman, D. Moher, for the CONSORT and Pragmatic Trials in Healthcare (Practihc) groups, Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ*. **337**, a2390–a2390 (2008).

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

20. M. Abell, L. Braddy, D. Ensley, L. Ludwig, H. Soto-Johnson, *MAA Instructional Practices Guide* (Mathematical Association of America, Washington, DC, 2018).

21. K. Saxe, L. Braddy, J. Bailer, R. Farinelli, T. Holm, V. Mesa, U. Treisman, P. Turner, *A common vision for undergraduate mathematical sciences programs in 2025* (Mathematical Association of America, Washington, DC, 2015).

22. V. Otero, S. Pollock, N. Finkelstein, A physics department's role in preparing physics teachers: The Colorado learning assistant model. *American Journal of Physics*. **78**, 1218–1224 (2010).

23. G. Ladson-Billings, But that's just good teaching! The case for culturally relevant pedagogy. *Theory into practice*. **34**, 159–165 (1995).

24. L. Benson, S. Biggers, W. Moss, M. W. Ohland, M. K. Orr, S. D. Schiff, *Adapting and implementing the scale-up approach in statics, dynamics, and multivariable calculus* (2009).

25. A. Kezar, A. C. Chambers, J. C. Burkhardt, Eds., *Higher education for the public good: Emerging voices from a national movement* (John Wiley & Sons, 2015).

26. D. Bressoud, Insights from the MAA National Study of College Calculus. *The Mathematics Teacher*. **109**, 179–185 (2015).

27. G. Sonnert, P. M. Sadler, S. M. Sadler, D. M. Bressoud, The impact of instructor pedagogy on college calculus students' attitude toward mathematics. *International Journal of Mathematical Education in Science and Technology*. **46**, 370–387 (2015).

28. P. Zorn, E. Bressoud, David, W. Pearson, Michael, Response to the PCAST Report to the President, Engage to Excel (2012), (available at https://www.maa.org/sites/default/files/pdf/sciencepolicy/MAA-ResponsePCAST-E2E.pdf).

29. D. Burrows, *Common Data Set 2019-2020* (Florida International University CDS Archive, 2020; https://opir.fiu.edu/CDS/CDS2019.pdf).

30. C. J. Torgerson, Randomised controlled trials in education research: a case study of an individually randomised pragmatic trial. *Education 3–13*. **37**, 313–321 (2009).

31. M. Roland, D. J. Torgerson, Understanding controlled trials: What are pragmatic trials? *Bmj*. **316**, 285 (1998).

32. B. Styles, C. Torgerson, Randomised controlled trials (RCTs) in education research – methodological debates, questions, challenges. *Educational Research*. **60**, 255–264 (2018).

33. K. F. Schulz, D. G. Altman, D. Moher, the CONSORT Group, CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Trials*. **11**, 32 (2010).

34. D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, D. G. Altman, CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *International journal of surgery*. **10**, 28–55 (2012).

35. M. A. Tallman, M. P. Carlson, D. M. Bressoud, M. Pearson, A characterization of calculus I final exams in US colleges and universities. *International Journal of Research in Undergraduate Mathematics Education*. **2**, 105–133 (2016).

36. J. Cohen, *Statistical power analysis for the behavioral sciences* (2nd edn. Á/L., 1988).

37. J. C. Valentine, H. Cooper, *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes* (What Works Clearinghouse, Washington, DC, 2003).

38. M. T. Hora, Limitations in experimental design mean that the jury is still out on lecturing. *Proceedings of the National Academy of Sciences*. **111**, 3024–3024 (2014).

39. D. M. Bressoud, M. P. Carlson, V. Mesa, C. Rasmussen, The calculus student: insights from the Mathematical Association of America national study. *International Journal of Mathematical Education in Science and Technology*. **44**, 685–698 (2013).

40. T. Snyder, C. de Brey, S. Dillow, Digest of Education Statistics 2018 (NCES 2020-009) (2019), (available at https://nces.ed.gov/programs/digest/d19/tables/dt19_330.10.asp).

41. V. M. Montori, P. J. Devereaux, N. K. Adhikari, K. E. Burns, C. H. Eggert, M. Briel, H. C. Bucher, Randomized trials stopped early for benefit: a systematic review. *Jama*. **294**, 2203–2209 (2005).

42. S. J. Pocock, When (not) to stop a clinical trial for benefit. *Jama*. **294**, 2228–2230 (2005).

43. L. C. Moore, D. A. Smith, R. G. Douglas, Toward a Lean and Lively Calculus. *The College Mathematics Journal*. **18**, 439 (1987).

44. L. A. Steen, *Heeding the Call for Change* (MAA, Washington, DC, 1992).

45. C. C. Narasimhan, CALCULUS REFORM FOR THE NON-SCIENCE CLIENT DISCIPLINES*. *PRIMUS*. **3**, 254–262 (1993).

46. S. Ganter, "An evaluation of calculus reform: A preliminary report of a national study" in *Assessment Practices in Undergraduate Mathematics*, B. Gold, S. Keith, W. A. Marion, Eds. (Mathematical Association of America, 1999), vol. 49 of *MAA Notes*, pp. 233–236.

47. J. F. Hurley, U. Koehn, S. L. Ganter, Effects of Calculus Reform: Local and National. *The American Mathematical Monthly*. **106**, 800–811 (1999).

48. D. H. Hallett, "What Have We Learned from Calculus Reform? The Road to Conceptual Understanding" in *A Fresh Start for Collegiate Mathematics: Rethinking the Courses below Calculus*, N. B. Hastings, Ed. (Mathematical Association of America, 2006; https://www.cambridge.org/core/books/fresh-start-for-collegiate-mathematics/what-have-we-

learned-from-calculus-reform-the-road-to-conceptual-understanding/9A22C7E53A5FC5FFD4FB3329C2A06A8B), pp. 43–45.

49. E. Miller, J. Fowler, C. Johns, J. Johnson, B. Ramsey, B. Snapp, Increasing Active Learning in Large, Tightly Coordinated Calculus Courses. *null*. **31**, 371–392 (2021).

50. W. M. Smith, M. Voigt, A. Strom, D. Webb, G. Martin, Eds., *Transformational change efforts: student engagement in mathematics through an institutional network for active learning* (American Mathematical Society, Providence, Rhode Island, 2021).

51. M. Borrego, C. Henderson, Increasing the use of evidence-based teaching in STEM higher education: A comparison of eight change strategies. *Journal of Engineering Education*. **103**, 220–252 (2014).

52. D. W. Johnson, R. T. Johnson, K. A. Smith, Cooperative learning returns to college what evidence is there that it works? *Change: the magazine of higher learning*. **30**, 26–35 (1998).

53. S. L. Laursen, C. Rasmussen, I on the Prize: Inquiry Approaches in Undergraduate Mathematics. *Int. J. Res. Undergrad. Math. Ed.* **5**, 129–146 (2019).

54. E. J. Theobald, M. J. Hill, E. Tran, S. Agrawal, E. N. Arroyo, S. Behling, N. Chambwe, D. L. Cintrón, J. D. Cooper, G. Dunster, J. A. Grummer, K. Hennessey, J. Hsiao, N. Iranon, L. Jones, H. Jordt, M. Keller, M. E. Lacey, C. E. Littlefield, A. Lowe, S. Newman, V. Okolo, S. Olroyd, B. R. Peecook, S. B. Pickett, D. L. Slager, I. W. Caviedes-Solis, K. E. Stanchak, V. Sundaravardan, C. Valdebenito, C. R. Williams, K. Zinsli, S. Freeman, Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proc Natl Acad Sci USA*. **117**, 6476 (2020).

55. F. Carreon, S. DeBacker, P. Kessenich, A. Kubena, P. G. LaRose, What is Old is New Again: A Systemic Approach to the Challenges of Calculus Instruction. *PRIMUS*. **28**, 476–507 (2018).

56. W. G. McCallum, D. Hughes-Hallett, A. M. Gleason, *Multivariable calculus* (Wiley, 1997).

57. T. L. Lovelace, C. K. McKnight, The Effects of Reading Instruction on Calculus Students' Problem Solving. *Journal of Reading*. **23**, 305–308 (1980).

58. U. D. of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. *Beginning reading interventions report* (2007).

59. D. McNeish, K. Kelley, Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*. **24**, 20 (2019).

60. D. M. McNeish, L. M. Stapleton, The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*. **28**, 295–314 (2016).

61. D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group, others, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International journal of surgery (London, England)*. **8**, 336–341 (2010).

62. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria; https://www.R-project.org).

63. S. Garfunkel, M. Montgomery, *GAIMME: guidelines for assessment & instruction in mathematical modeling education* (2016).

64. National Research Council, *Adding It Up: Helping Children Learn Mathematics* (The National Academies Press, Washington, DC, 2001; https://www.nap.edu/catalog/9822/adding-it-up-helping-children-learn-mathematics).

65. K. H. Lim, A. Selden, "Mathematical habits of mind" in *Proceedings of the thirty-first annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (2009), pp. 1576–1583.

66. D. W. Johnson, R. T. Johnson, K. A. Smith, Cooperative learning: Improving university instruction by basing practice on validated theory. *Journal on Excellence in University Teaching*. **25**, 1–26 (2014).

67. D. W. Johnson, R. T. Johnson, Cooperative learning: The foundation for active learning. *Active Learning—Beyond the Future* (2018).

68. L. Springer, M. E. Stanne, S. S. Donovan, Effects of Small-Group Learning on Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis. *Review of Educational Research*. **69**, 21–51 (1999).

69. S. B. Wilson, P. Varma-Nelson, Small Groups, Significant Impact: A Review of Peer-Led Team Learning Research with Implications for STEM Education Researchers and Faculty. *J. Chem. Educ.* **93**, 1686 (2016).

70. M. M. Chiu, Flowing toward Correct Contributions during Group Problem Solving: A Statistical Discourse Analysis. *The Journal of the Learning Sciences*. **17**, 415–463 (2008).

71. C. Rumsey, C. W. Langrall, Promoting Mathematical Argumentation. *Teaching Children Mathematics*. **22**, 412–419 (2016).

72. P. J. Riccomini, G. W. Smith, E. M. Hughes, K. M. Fries, The language of mathematics: The importance of teaching and learning mathematical vocabulary. *Reading & Writing Quarterly*. **31**, 235–252 (2015).

73. E. Yackel, P. Cobb, Sociomathematical Norms, Argumentation, and Autonomy in Mathematics. *Journal for Research in Mathematics Education*. **27**, 458–477 (1996).

74. M. B. Ginsberg, R. J. Wlodkowski, Professional learning to promote motivation and academic performance among diverse adults. *Learning Never Ends*. **23** (2009).

75. G. Lawrie, E. Marquis, E. Fuller, T. Newman, M. Qiu, M. Nomikoudis, F. Roelofs, L. van Dam, Moving towards inclusive learning and teaching: A synthesis of recent literature. *T&LI*. **5** (2017), doi:10.20343/teachlearninqu.5.1.3.

76. A. S. Lillard, J. Taggart, D. Yonas, M. N. Seale, An alternative to “no excuses”: Considering Montessori as culturally responsive pedagogy. *J. Negro Educ* (2021).

77. G. Ladson-Billings, "Three Decades of Culturally Relevant, Responsive, & Sustaining Pedagogy: What Lies Ahead?" in *The Educational Forum* (Taylor & Francis, 2021), vol. 85, pp. 351–354.

78. X. Wu, K. Rambo-Hernandez, E. Fuller, J. Deshler, Predicting STEM persistence from mathematics affect, pedagogical perceptions and Calculus I setting. *International Journal of Mathematical Education in Science and Technology*, 1–22 (2022).

79. E. Brewe, Modeling theory applied: Modeling Instruction in introductory physics. *American Journal of physics*. **76**, 1155–1160 (2008).

80. R. M. Goertzen, E. Brewe, L. Kramer, Expanded Markers of Success in Introductory University Physics. *International Journal of Science Education*. **35**, 262–288 (2013).

81. E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O’Brien, I. Rodriguez, P. Pamelá, Toward equity through participation in Modeling Instruction in introductory university physics. *Physical Review Special Topics-Physics Education Research*. **6**, 010106 (2010).

82. E. Brewe, A. Traxler, J. De La Garza, L. H. Kramer, Extending positive CLASS results across multiple instructors and multiple classes of Modeling Instruction. *Physical Review Special Topics-Physics Education Research*. **9**, 020116 (2013).

83. E. Brewe, L. Kramer, G. O’Brien, Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS. *Physical Review Special Topics-Physics Education Research*. **5**, 013102 (2009).

84. I. Rodriguez, G. Potvin, L. H. Kramer, How gender and reformed introductory physics impacts student success in advanced physics courses and continuation in the physics major. *Physical Review Physics Education Research*. **12**, 020118 (2016).

85. M. Lampert, H. Beasley, H. Ghousseini, E. Kazemi, M. Franke, "Using designed instructional activities to enable novices to manage ambitious mathematics teaching" in *Instructional explanations in the disciplines* (Springer, 2010), pp. 129–141.

86. D. Bressoud, C. Rasmussen, Seven characteristics of successful calculus programs. *Notices of the AMS*. **62**, 144–146 (2015).

87. S. Larsen, E. Glover, K. Melhuish, Beyond good teaching. *insights*, 93 (2015).

88. J. Gleason, S. Bagley, M. Thomas, L. Rice, D. White, The calculus concept inventory: a psychometric analysis and implications for use. *International Journal of Mathematical Education in Science and Technology*. **50**, 825–838 (2019).

89. R. M. Talbot, L. M. Hartley, K. Marzetta, B. S. Wee, Transforming undergraduate science education with learning assistants: Student satisfaction in large-enrollment courses. *Journal of College Science Teaching*. **44**, 24–30 (2015).

90. J. L. Alzen, L. S. Langdon, V. K. Otero, A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. *International Journal of STEM Education*. **5**, 56 (2018).

91. M. Carlson, M. Oehrtman, N. Engelke, The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*. **28**, 113–145 (2010).

92. A. Justel, D. Peña, R. Zamar, A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & probability letters*. **35**, 251–259 (1997).

93. R. F. Woolson, Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, 1–3 (2007).

94. S. W. Raudenbush, X. Liu, Statistical power and optimal design for multisite randomized trials. *Psychological methods*. **5**, 199 (2000).

95. S. W. Raudenbush, Analyzing effect sizes: Random-effects models. *The handbook of research synthesis and meta-analysis*. **2**, 295–316 (2009).

96. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. J Stat Softw (2015).

97. A. Kuznetsova, P. B. Brockhoff, R. H. Christensen, lmerTest package: tests in linear mixed effects models. *Journal of statistical software*. **82**, 1–26 (2017).

98. S. G. Luke, Evaluating significance in linear mixed-effects models in R. *Behavior research methods*. **49**, 1494–1502 (2017).

99. A. S. Selya, J. S. Rose, L. C. Dierker, D. Hedeker, R. J. Mermelstein, A practical guide to calculating Cohen'sf 2, a measure of local effect size, from PROC MIXED. *Frontiers in psychology*. **3**, 111 (2012).

100. B. C. Jaeger, L. J. Edwards, K. Das, P. K. Sen, An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*. **44**, 1086–1105 (2017).

101. M. A. Stoffel, S. Nakagawa, H. Schielzeth, partR2: partitioning R2 in generalized linear mixed models. *PeerJ*. **9**, e11414 (2021).

102. V. Dorie, M. Harada, N. B. Carnegie, J. Hill, A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*. **35**, 3453–3470 (2016).

103. J. Hass, C. Heil, M. D. Weir, *Thomas' Calculus* (Pearson, 2018; https://books.google.com/books?id=RRlAvgAACAAJ).

104. C. Schumacher, M. J. Siegel, *2015 CUPM Curriculum Guide to Majors in the Mathematical Sciences* (2015; https://works.bepress.com/carol_schumacher/1/).

105. C.-Y. J. Peng, K. L. Lee, G. M. Ingersoll, An introduction to logistic regression analysis and reporting. *The journal of educational research*. **96**, 3–14 (2002).

Acknowledgments: We thank the FIU administration and the State of Florida for initial support that made this project possible. The work would not be possible without the faculty, undergraduate Learning Assistants and students participating in the study, they have earned our deep gratitude. We thank Dr. Karen Rambo-Hernandez for her insightful statistical analysis discussions. We are indebted to the reviewers and editors for their insight and guidance. The research conducted for this study under an IRB protocol approved by the Florida International University Institutional Review Board (Approval # IRB-18-0211-AM02). Students participating in the study consented to participation at the beginning of each semester. Only students aged eighteen or over were included. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Funding:

National Science Foundation under Grant No. DUE 1832450 (LK, EF, GP, AC, CW)

Author contributions:

Conceptualization: LK, EF, GP, AC, CW

Methodology: LK, EF, GP, AC, CW, PDO

Investigation: LK, EF, GP, AC, CW, PDO

Funding acquisition: LK, EF, GP

Project administration: LK, EF

Writing – original draft: LK, EF, GP, AC, CW, PDO

Writing – review & editing: LK, EF, GP, AC, CW, PDO

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All experimental data are archived at the Open Science Framework and available at https://osf.io/8x472/.

Science
AAAS

Establishing a New Standard of Care for Calculus Students with Evidence from a Randomized Trial

Laird Kramer, Edgar Fuller, Charity Watson, Adam Castillo, Pablo Duran Oliva, Geoff Potvin

Correspondence to: Laird.Kramer@fiu.edu, Edgar.Fuller@fiu.edu

This PDF file includes:

Materials and Methods

1: Experimental Design

1.1: Historic Calculus Development and Approach

Calculus courses and their impact on student progress in post-secondary settings have been a significant focus of pedagogical study and policy debate for several decades. This study builds on prior calculus work and advances the understanding of the impact of instructional change on student learning by randomizing a large sample of students at the individual level into comparable control and treatment groups over three semesters, utilizing a comprehensive set of curricular and pedagogic instructional materials in the treatment condition, and then carrying out blinded evaluation of the student outcomes. The study design and pedagogic strategies are motivated by and built on the experiences emanating from the Tulane conference (*43*, *44*) that inspired efforts to reform calculus in the 1990s (*4*, *15–17*, *45–48*) and continuing into the work of the CSPCC (*26*, *39*), PtC (*12*), and SEMINAL projects more recently(*49*, *50*). These concerns have persisted (*5*, *51*) and a great deal of research has identified group work (*52*), inquiry-based learning (IBL) (*53*), and active learning approaches in mathematics (ALM) (*19*, *50*, *54*) as primary levers of change that could lead to the significant improvements in student learning generating outcomes that national calls have sought. Much of the prior work refocused the calculus curriculum on conceptual foundations. These include Project Calc at Duke (*15*) the calculus projects at DePaul University (*45*), the University of Michigan (*16*, *55*), and Baylor University (*17*) used the text developed by the Harvard Consortium (*56*). Some incorporated strong technological supports (*15*) and others incorporated laboratory or recitation components (*16*).

Randomization at the student level was not a common strategy in prior studies. One study that included randomization (*57*) had 37 total participants taught in two sections by the same instructor and did not control for potential internal bias of the instructor towards either condition. There are no registered post-secondary calculus studies in the What Works Clearinghouse (*58*) used by the U.S. Department of Education to evaluate instructional effectiveness and experimental techniques. A central motivation for this study was to carry out a rigorous randomized trial and address the concerns that have limited the propagation of the broader calculus instructional improvement efforts and that would meet the WWC “without reservations” standard (*58*). In developing the experimental protocol, the randomization process aimed to ensure equivariantly distributed student characteristics as well as address concerns such as time of day / day of week levels which had been critiques of prior studies.

1.2: Pragmatic Randomized Trial Approach

This study intentionally utilized a pragmatic randomized trial approach (*30*, *31*) designed to replicate real-world conditions in order to inform institutions of higher education considering similar interventions and adapting to the inherent nature of classroom education research interventions. The study protocol integrates random assignment of students to treatment and control sections, as well as blind analysis of the end-of-semester learning measures. Random assignment of students to treatment and control sections removes effects that could arise from students intentionally selecting treatment or control conditions. The random assignment occurred after students selected a day/time meeting pattern to remove any potential bias due to time of day. As the complete blinding to study participants is not possible in an education study, the outcome measures were objective and not open to subjective interpretation, following (*18*, *30*,

31). Analysis of the end-of-semester learning measures, the outcome measures, was blinded to the researchers to minimize bias.

The aim of the intervention was to establish a novel calculus instructional paradigm, one that promoted significantly improved student learning and course achievement; thus, the design integrated multiple research-driven strategies into a coherent classroom approach. These strategies included the curriculum (class notes, in-class learning activities, homework assignments, and exams) as well as the pedagogic practices (group work, white-boarding sessions, instructor and Learning Assistants (*22*) in-class facilitation. and other classroom norms), with coherence arising through the consistent themes across all curricular elements and classroom language that mutually reinforce each other. This intervention contrasts with a piecemeal approach to intervention where individual elements are changed over time leading to limited impact, cognitive dissonance among students, ambiguous results and/or overlooked potential synergies across course elements. The approach mimics a real-world application where an institution brings together multiple promising practices with the goal of dramatically improving student outcomes, as a large effect is a mechanism that enables sustained change.

Randomly assigning faculty to control or treatment conditions was not feasible or appropriate, as it could implicitly or explicitly introduce biases in favor/against the treatment or control conditions if faculty were forced to teach using strategies that conflicted with their instructional preferences. An individual faculty member's knowledge, philosophy, and experience with a variety of classroom strategies and instructional practices may intersect with the features of the treatment or control conditions. Further, a design that incorporates the same instructor teaching both treatment and control sections in the same semester could introduce similar implicit or explicit biases. Recognizing that biases arising from either of these two strategies could not be reliably measured, thus neither of the approaches were utilized in the study. The potential limitations due to faculty awareness of the intervention and not being assigned both conditions simultaneously are consistent with related investigations of public health or sociological interventions. The experimental protocols aimed to reduce the impact of instructor biases and investigated as detailed in Section 3 below. The protocol intentionally compared instructors willing to change instructional methods, with a range of prior active learning experience, as it more genuinely replicates the state of faculty in mathematics departments across the nation.

Students were randomly assigned to treatment or control conditions in the same meeting pattern just prior to each semester, but treatment students likely realized their course was reformed in their initial class meeting. In educational research studies, both the students and the instructor are keenly aware of the historic instructional norms in classrooms. College students are most likely to experience traditional univocal/direct instruction throughout their pre-college studies as well as in the majority of their college courses, thus students would easily detect the obvious differences in a treatment section and could never be 'blind' to the treatment. Students dropping the course or switching to a different section of the course (by institution's official drop deadline) were removed from the analysis, preventing bias from their choices while following the pragmatic educational constraints. It would have been unethical and impractical to require students to remain in the assigned sections for the duration of the semester, as their schedules need to accommodate changes in other courses, work schedules, and family responsibilities. Overall, 80% of the randomly assigned students remained in their sections beyond the drop date and are included in the study. Roughly 12% of the randomly assigned students switching sections prior to the drop date, the majority of which selected a different class meeting pattern.

Less than 3% of the randomly assigned students swapped treatment for control, or vice versa, with roughly equal swaps from treatment to control as control to treatment and were excluded from the study. The remaining 8% of the randomly assigned students dropped the course for the semester before the drop date. Complete details of the student enrollment patterns are in Table S1. Potential biases arising in the student participant allocations were investigated for hidden level effects(*59*, *60*) or confounders to establish limitations of the study and documented in Section 3 below.

The pragmatic randomized trial strategy extended to the outcome measures as well. The assessment of student learning outcomes strove to generate knowledge on students' understanding of the overall calculus course, following the coherent instruction strategies. The end-of-semester learning measures assessed the learning objectives common to many courses across the nation, documented in (*26*, *35*) and described in the End-of-Semester Learning Measures Overview section below. The end-of-semester learning measures are therefore limited to providing insight on the whole of the course instruction and are the most valuable for institutions seeking to implement similar transformations. Ascribing effect to any of the individual instructional strategies with these measures is ineffective given the grain size of the measure as well as likely interference effects coming from their combination. Further, accounting for the complexities of human experiences and interactions prior to, and during the 15-week intervention (both within the class and external to the class), would significantly limit the conclusions drawn from a more precise investigation.

1.3: CONSORT Protocol Discussion

In enrolling and assigning students to the study, the Consolidated Standards for Reporting Trials (CONSORT) (*18*, *33*, *34*, *61*) framework was followed for pragmatic randomized trials and collected measures for that protocol. Following the checklist for that framework, the title and abstract (Item 1) references to the randomized nature of the trial and a brief description of the intervention.

Introduction. The background for the study (Item 2) is described in the main body of this paper as well as in the methods section. It focuses on determining whether or not the use of the comprehensive Modeling Practices of Calculus active learning approaches in a calculus classroom results in increases in student learning over traditional instructional methods that can be measured.

Methods. In the methods section the setting for the study is described as an urban, Carnegie classification Research-Very High, minority serving institution with a large population of students in general as well as in the mainstream Calculus 1 course offered on campus. Data were collected from both institutional research (for demographics and enrollment data) as well as from administered assessment tools that were offered to students in their classroom settings. Students who chose to enroll in Calculus 1 in a given term and who had no registration holds or other administrative issues that prevented participation were enrolled in the trial. These students were then randomly assigned to either the treatment group or the control group (Item 3).

Students in the treatment group were assigned to sections of the course at the same day and time as their original choice mirroring those in the control group. The treatment sections were then conducted using the active-learning curriculum and pedagogy implemented in the Modeling Practices in Calculus approach as described in detail in the Methods section. Instructors were

selected for these treatment sections and provided professional development to support the implementation of this curriculum. Control students received instruction using the historical methods employed by the instructors who were assigned to their sections using normal departmental processes (Item 4). The objective of this process was to provide students access to different instructional practices that would improve their student learning outcomes and to measure these outcomes using a blinded assessment protocol (Item 5). The hypothesis is that the treatment curriculum will improve student learning.

Outcomes were measured by refining a set of identical end-of-semester learning measures included as embedded final exam questions based first on the departmental practice of administering an identical final exam to all students in Calculus 1. The identical measures were based on items developed collaboratively by department faculty that were aligned to existing learning outcomes established for calculus and used in university accreditation processes. Historical exam questions were reviewed by a group of instructors drawn from both treatment and control sections during the project and sets of assessment items were agreed upon to be used in a commonly administered final exam. Exam items were aligned to national standards. Exams were administered to all students in calculus sections and so any student in the trial who completed the semester in their assigned sections were administered an exam that included the blocks of items and these questions were identical for both groups. The exams administered to treatment and control students were then blinded and assessed using a rubric by multiple independent evaluators (Item 6). Evaluators cross-calibrated their scoring and the end-of-semester learning measures scores from all evaluators were recorded and averaged. Additional outcomes were measured using pre- and post-surveys administered in the classrooms.

In total, 1,058 students were assessed for eligibility after having enrolled in a calculus 1 section designated for randomization. Of these, 1,019 were able to be included in randomization and were assigned to either the treatment or control groups. Subsequently 516 students were assigned to the treatment group and 417 remained in the treatment section at the drop/add deadline while 99 left either for a different calculus section or left calculus altogether before the date when students could change schedules with no impact on grade assignment. At the same time 503 students were assigned to the control group with 394 remained in the treatment section at the drop/add deadline in the control condition. Finally, 44 students in treatment group did not complete a final exam and are designated “lost to follow-up,” leaving 373 students with analyzable outcomes. In the control group, 84 students did not complete a final exam and are designated “lost to follow-up,” leaving 310 students with analyzable outcomes in the control group. The sample size was dependent on student enrollment for the initial population of eligible participants and was limited by administrative factors such as academic holds. After assignment, treatment and control populations were limited by student departure from the calculus course altogether or student movement to other sections of calculus. These data are reported in the Methods section and in the flowchart in Fig. S1 (item 7). Randomization was performed by a team member not involved in course administration, content or measure development, or student enrollment and was performed using *R* (*62*) after obtaining lists of students who had enrolled in offered calculus sections in a given term. No restrictions were applied. The randomization resulted in enrollment lists provided to the registrar for assignment to treatment sections. These lists were hidden from research team members and from instructors until after classes began. Once enrolled in their courses, instructors and students were no longer blind to the intervention as it is impossible to mask the actual curriculum during implementation (Items 8, 9, 10 and 11).

Outcomes from the study were analyzed to determine effect size (Cohen's *d*) for any difference in the identical end-of-semester learning measure item scores as well as using a fixed-effects model and comparison of paired section means to identify any level effects related to course or instructor levels within the study (Item 12).

Results. Participant flow, recruitment, baseline data (Items 13, 14, and 15) are described in the methods section. The number of participants in each group included (Item 16) in the analysis are also in the Methods section as well in the Supplementary Materials information below with the outcomes and estimation (Item 17) along with subgroup analysis (Item 18). Any adverse events, mostly related to the assignment of students who had academic holds or other departure events, are described in the Methods section (Item 19).

Discussion. A discussion of the main results, generalizability and overall evidence (Items 20, 21, and 22) are provided in the main paper and in more detail in Sections 3.2.1 to 3.2.4 of the Supplementary Materials, below.

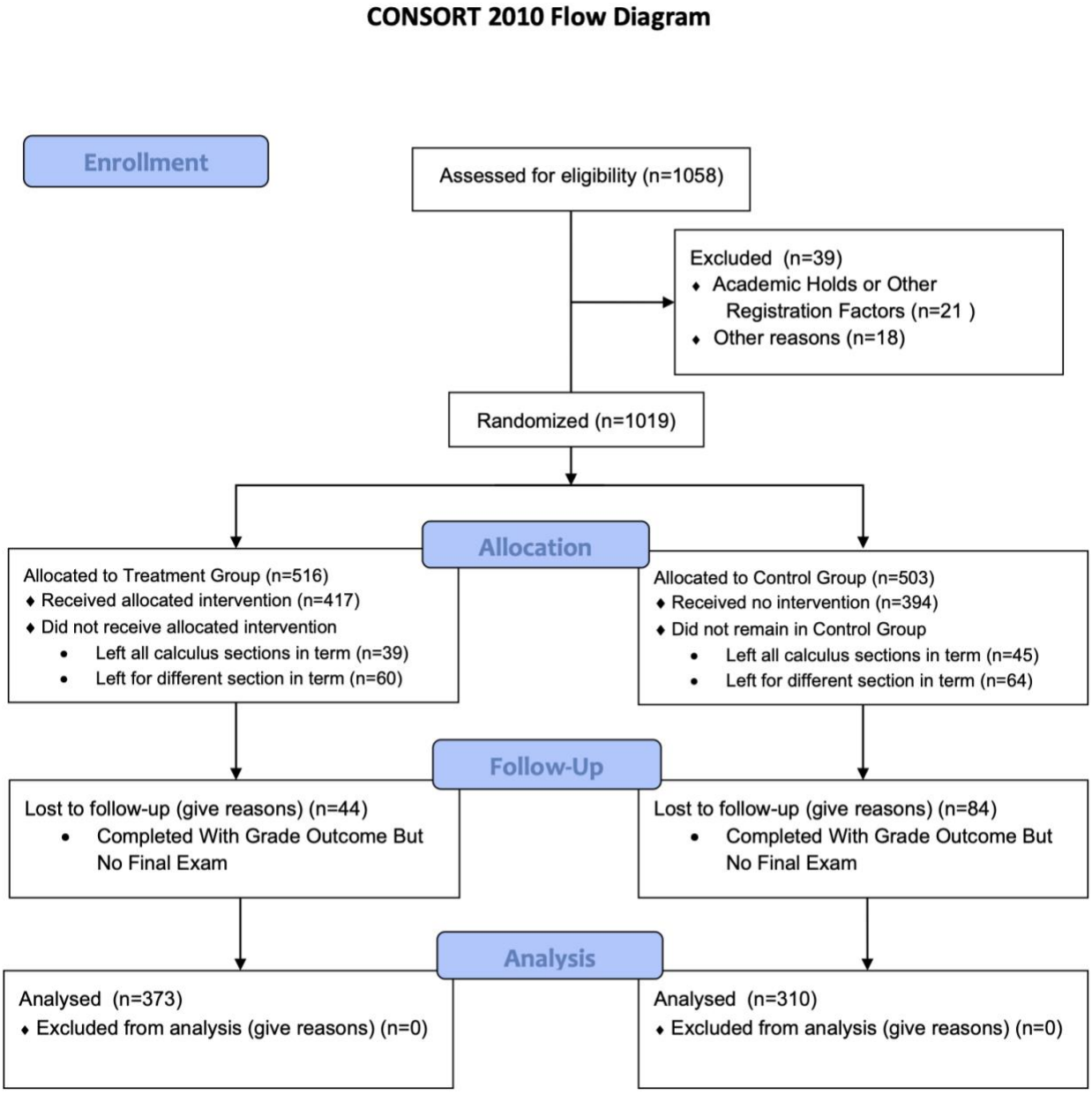


Fig. S1: CONSORT 2010 Flow Diagram with Trial Enrollments, Assignments, and Analysis Outcomes for Participants

2: Materials

2.1: Modeling Practices in Calculus Pedagogy and Curriculum Overview

The study utilized a newly developed collection of pedagogic strategies and classroom materials that implemented active learning centered classroom practices in the treatment intervention. The Modeling Practices in Calculus (MPC) approach is designed to bring the authentic practices of mathematicians into the classroom, by facilitating active student engagement in the practices of mathematicians to learn calculus in a student-centered environment. The pedagogy and curriculum followed recommendations established by the major mathematical organizations (*21*, *63*) and draws on best practices from across the mathematical education research spectrum, discussed below. The MPC approach is a conceptual framework for learning introductory calculus. Students begin with a fundamental model of mathematical behavior, such as limits, and continually develop and expand their model based on additional considerations, such as continuity. The MPC approach incorporates five essential features:

1. **Practices of Mathematicians.** The core of the MPC approach is the process of students developing their understanding of calculus by engaging in the practices of mathematicians, including: sense making and constructing an understanding of mathematical concepts; solving mathematical problems; adaptive reasoning; modeling with mathematics; using appropriate tools strategically; building mathematical communication skills; and connecting mathematics with other disciplines. These practices of mathematicians are centered around the published recommendations for curricula and pedagogy from professional mathematical associations (*21*, *64*, *65*).
2. **Cooperative Learning.** Students work in small groups cooperatively to accomplish shared learning goals while providing each other with formative feedback (*52*, *66–70*). Students work together to complete structured learning activities that involve sharing ideas, improving skills, developing interpersonal skills, and evaluating group performance.
3. **Argumentation/Metacognition.** Students engage in mathematical argumentation on course problems and topics, a process of dynamic and meaningful social discourse for discovering new ideas, providing justifications, convincing others, and evaluating claims in both group and whole-class discussions (*71–73*). The inclusion of instruction promoting mathematical argumentation can provide a deeper understanding of mathematics as students become generators of knowledge out of their reasoning and sense-making (*73*).
4. **Mathematical Fluency.** Mathematical fluency includes being able to solve problems accurately, efficiently, and with flexibility (*64*). Students build fluency, by noticing mathematical relationships and using strategies through the study and small group / whole class discussions of various concepts in course learning activities, as well as through tasks that promote reasoning and problem-solving.
5. **Culturally Responsive Environment.** The MPC model is centered around Ginsberg and Wlodkowski's (*74*) four motivational conditions for culturally-responsive teaching: establishing inclusion(*75*), developing attitude, enhancing meaning, and engendering competence in all proposed activities. Students are provided with an immersive, transformation learning experience that allows them to construct their understanding by working with each other (*76*), Learning Assistants (LAs), and faculty. LAs, undergraduate near peers prepared to foster learning (*22*, *74*), are integrated into the classroom to facilitate learning with the groups. The MPC approach offers a classroom

environment conducive to learning by allowing students to try out their ideas in a low-stakes environment with peers, receive ongoing formative assessment, and participate in a learning community(*23*, *77*). The LAs are natural agents of this learning community, as their demographics are that of the students (i.e., all LAs are undergraduates), who provide insights and connections from the point of view of a recent participant in the course. As former students successful in calculus, LAs use their own backgrounds and experiences to promote students' success(*78*).

There are two immediate antecedents of the Modeling Practices in Calculus approach: Modeling Instruction in Physics (*79*) and SCALE-UP Calculus (*24*). Modeling Instruction in Physics in university physics instruction has been taught by numerous faculty at Florida International University since 2003. Modeling Instruction in Physics is organized as an integrated, studio-based, lecture-free environment where students develop their understanding of physics by modeling the practices of physicists. These practices include carrying out experiments to discover the underlying physics, comparing results with others to form consensus on the rules and laws of physics, negotiating shared meaning (including terminology, physical concepts and quantities, and relations) and then refining understanding through additional practice and experimentation. Modeling Instruction in Physics has been institutionally sustained for almost two decades due to evidence of its profound impact on students including those from historically underrepresented groups. Evidence includes: 1) significantly improved conceptual understanding and course outcomes overall and across gender, race, and ethnicity groups, when compared to traditional instruction (*80*, *81*); 2) the first improved favorable attitudes towards physics and physics learning measured in an introductory physics course (*82*, *83*); 3) and increased access to physics degree by underrepresented groups (*84*). It was also part of programmatic efforts that lead to dramatic increases in the number of physics majors and graduates at FIU, thus serving as inspiration for the current project. The other root of the MPC approach is SCALE-UP Calculus (*24*) which was developed and taught for many years at Clemson University (including two of the current authors). SCALE-UP Calculus also uses a studio-based approach complementary to Modeling, while relying on mini, or targeted, lectures. The development of the MPC curriculum began with much of the SCALE-UP topical coverage and integrated Modeling Instruction-based pedagogical approaches into the curriculum. Refinement of MPC based on student experience and formative feedback from instructors and LAs has continued every semester since the project began.

MPC implements ambitious teaching practices (*20*, *26*, *85*) for developing conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive dispositions and as outlined specifically for Calculus in the Mathematics Association of America's 2015 report *MAA National Study of College Calculus* (*26*). Elements of both good and ambitious teaching practices are outlined and observed in the MAA *Characteristics of Successful Programs of College Calculus* study (*26*), Bressoud and Rasmussen's *Seven Characteristics of Successful Calculus Programs* (*86*), and additional studies (*27*, *87*, *88*). Many of these practices form the core of the MAA Instructional Practices Guide (*20*). The development and implementation of the MPC reflect an integration of multiple aspects from these studies as well as multiple semesters of development and revision.

2.2: Modeling Practices in Calculus: Classroom Learning Strategies

The MPC model was developed to engage students in the practices of mathematicians and potentially experience the joy of mathematics. The curriculum is offered in studio classroom

environments with minimal lecturing, with most of the class time devoted to students working in groups at small tables to collectively develop their understanding through guided notes, complete pre-designed learning activities, write-up solutions on whiteboards, and conduct board meetings. The curriculum is based on Scale Up Calculus (*24*), which includes guided inquiry notes and structured learning activities but modified to promote practices of mathematicians as well as a culturally responsive environment.

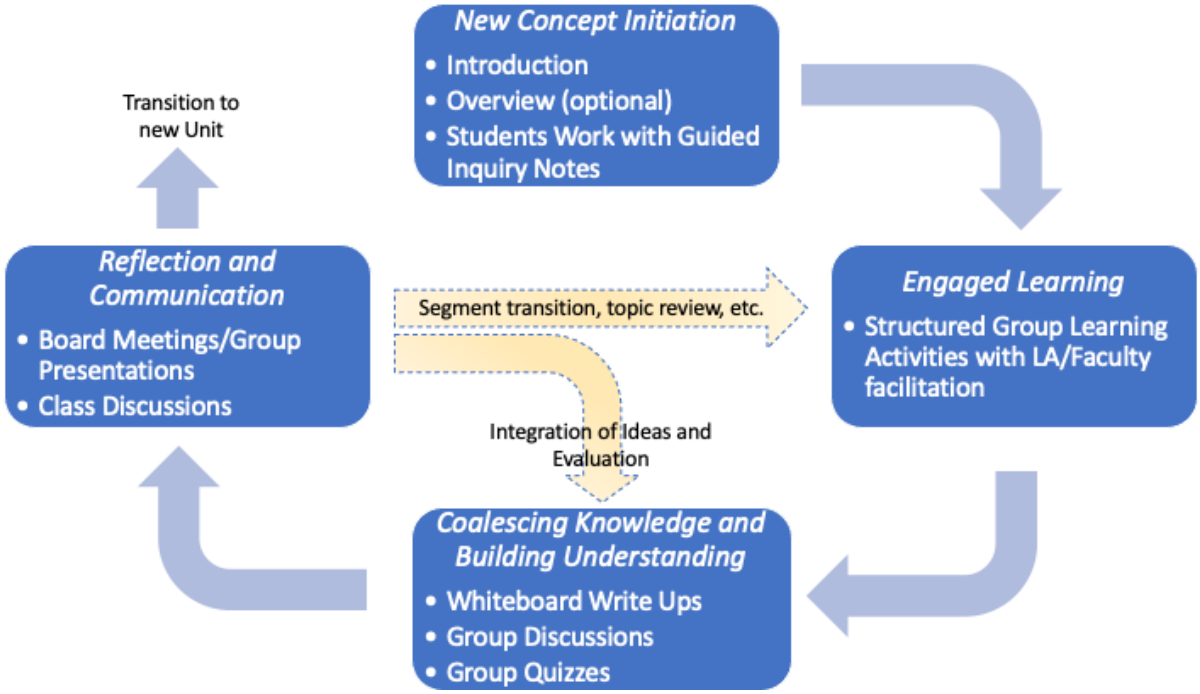


Fig. S2: The Modeling Practices of Calculus Learning Cycle

The MPC curriculum is organized as a set of units divided into learning cycles which span multiple class sessions. The learning cycle is illustrated in Fig. S2 and described in the following. MPC units begin with a *New Concept Initiation* phase (top of Fig. S2, that can include a brief introduction of a new concept or a review of a previous class topic and may include a set of warm-up question for the students. Then, students, working in groups, actively work through a set of guided inquiry notes that develop their understanding of the core calculus concepts, such as limits, rates of change, related rates, optimization, and integration. The notes are presented as a series of mathematical investigations, where students proceed through concept development with designed questions and problems that lead them to important insights and challenge (and build) their mathematical toolset.

In the *Engaged Learning* phase (right of Fig. S2), students work through Learning Activities in groups in order to build their knowledge while working through a set of problems designed to develop mathematical practices and skills. During these learning activities, students are asked to reflect on the mathematical concepts of the day, describe these concepts to their peers in their own words, choose appropriate problem-solving strategies, and validate peers' ideas within their groups. Thus, the MPC pedagogy promotes social metacognition through meaningful and

structured learning activities that encourage discussions about misconceptions and construction of shared knowledge.

In the *Coalescing Knowledge and Building Understanding* phase (bottom of Fig. S2), students summarize the knowledge they have developed on portable whiteboards, participate in group discussions and/or test their knowledge on group quizzes. This coalescing of knowledge allows students to work collaboratively to write up and present solutions they developed on whiteboards, a practice perceived by students as analogous to preparing a publication. The social metacognition continues as group members must monitor each other's thinking and make suggestions to prepare their group whiteboard.

In the *Reflection and Communication* phase (left of Fig. S2), students present their whiteboards to the whole class and/or participate in faculty-led class discussions. Board meetings are where groups present their findings to the whole class, akin to publishing or presenting results to the community. Lessons learned are codified through the board meetings when multiple groups come to consensus on their findings. Spurious or unexpected results lead to dynamic conversations that can identify common misconceptions or prime for future learning. For instance, when students transition from evaluating limits with a graphical representation to computing limits without a visual representation, board meetings allow groups to check and validate other group members' way of thinking and writing up solutions, since misconceptions regarding limit notation, algebra and simplification often arise when computing limits.

Immersed throughout all phases of the learning cycle is persistent guidance and formative feedback, thus promoting a culturally responsive learning environment. Students become accustomed to trying out their ideas in a low-stakes environment, receive ongoing formative feedback from their peers and the instructional team, and participate in a community of learners. Grades are assigned based on an absolute scale, with no curve, thus it is in the best interest of all students to develop their knowledge through cooperative learning. The instructor promotes the safe learning environment by regularly messaging the value of making mistakes and asking questions as a central element of learning mathematics. The low-stakes environment is also enhanced as Learning Assistants (LAs), or trained undergraduate classroom facilitators, are integrated into the classroom to support learning with groups and provide valuable information to instructors about student interactions (*22*, *89*). LAs help to center mathematical discussions in and between groups, as they constantly interact with groups and have multiple opportunities to redirect students' questions and comments to the groups. LAs are natural agents of this culturally appropriate model, as their demographics are that of the students, who provide insights and connections from the point of view of a former student in the course. LAs serve to mitigate 'blind spots' that experienced mathematicians bring into dialogues, which helps to increase the flow of ideas from the students to instructors, so that discussions are more strongly centered on students' points of view. They also help students develop skills, such as creating and defending ideas, making connections between concepts, and solving conceptual problems (*90*). The classroom strategies draw out student interaction in this way intentionally to enhance the connection of mathematical thought and concept development to the student experience. Through the ongoing dialogues, faculty and LAs have a portal into student ideas and are constantly adapting to their needs.

It is useful to note that not all class sessions follow the same pattern, as there are times the faculty recognizes or identifies a need to delve deeper into a topic and adds additional learning

activities before proceeding into a new unit. Faculty may also want to refine knowledge and have students expand upon their whiteboard summaries and re-discuss one or more topics. These adaptations are illustrated as the dashed line bounded segments in the middle of Fig. S2. In these transitions a facilitator may choose to regroup around an idea, have students revise their understanding and try again, or bridge from one portion of a concept to another. There are also accommodations related to the scheduled class periods, including adjusting activities to fit within the allocated time slot. A Learning Activity may be split into two parts so that the first is completed one day and the second is completed in the following class meeting.

Faculty in the treatment group participated in professional development activities to prepare for their MPC implementation. They participated in a 2-day summer professional development workshop that highlighted the MPC approach, provided the curricular structure and summarized the pedagogy. They were given access to the full set of MPC learning materials including basic day-to-day pacing guides, a sample course syllabus with learning outcomes and a course schedule, guided instructor notes for student learning facilitation, and learning activities to build skills and understanding within topic areas. Weekly preparation sessions were held to provide ongoing guidance on instructional strategies, following the usual practice for FIU faculty using active learning. Faculty met to discuss ongoing progress, prepare for upcoming topics, collaborate on assessments, and adapt to changing course demands during the semester. MPC materials were situated in an online repository with authenticated access provided to faculty individually each term. Materials were provided to students in hard copy form during class time on a daily basis depending on the scheduling of topics. The project placed no restrictions on preparation or classroom materials for faculty in the control group. They were free to prepare as individuals or in groups and utilize any learning materials of their choosing. They were not provided with access to the MPC materials.

3: Methods

This study carried out a large-scale pragmatic randomized trial to establish a new standard of care for calculus instruction. As described in the main body of the paper, the treatment group used the Modeling Practices of Calculus (MPC) pedagogy and curriculum with the control group employing pre-existing instructional practices (primarily traditional lecture). Identical end-of-semester learning measures were used to rigorously assess student learning outcomes in the two conditions in the randomized trial. Sets of open answer, learning measure questions were collectively developed (see Section 3.3.1) and embedded as part of the identical end-of-semester learning measures given each semester to the treatment and control sections in the MPC Calculus 1 trials in order to assess end-of-course student learning in each of the groups.

The Methods section includes investigations of potential bias in the student participants and instructors, end-of-semester learning outcome measures with samples and analyses, and odds of success ratios calculations for students by group in the course, The participant investigations include: 1) student participant randomization and enrollment patterns, 2) equivalence of student populations 3) comparison of instructor participant characteristics, 4) student and instructor differences within clusters, 5) a full mixed effect model to estimate variance in learning outcome measures, and 6) a sensitivity analysis to investigate possible unmeasured confounders.

3.1: Student Participant Randomization and Enrollment

Student participants in the study were students that enrolled in a number of pre-designated sections of the Calculus 1 course (3, 5 and 8 sections over the 3-semester experiment) using the Institution's class registration system. To accommodate the randomized assignments, each of the experimental sections doubled in size from the usual 40-seats to 80-seats prior to enrollment opening. Instructor names and section sizes were invisible to students throughout the enrollment phase. Just prior to the start of the semester, students that were enrolled in the 80-student sections were randomly assigned to the treatment and control conditions by randomly selecting half of the enrolled students in each designated section and assigning them to treatment sections (with new section numbers), with the original sections serving as control. Enrollment capacity in both treatment and control sections followed the institutional standards for the course. After randomized assignments were completed, all sections were open to additional enrollment and course changes as is the usual, customary institutional practice. Students changing sections or leaving the course prior to the institutional drop deadline (7 days into the semester) were excluded from the study.

Randomization of assignment was performed for all students in each 80-seat section who did not have an academic or other hold preventing a registration change. Students with holds or other constraints were excluded as noted in the CONSORT flow chart for participants (Fig. S1) and summarized in Table S1. If a student was excluded, additional students were randomly chosen to replace them in their assignment in order to balance the populations according to the number currently enrolled in the open 80 seat section. In the Fall of 2019, a group of students were randomly chosen for assignment to the treatment group to compensate for students who were excluded in this way while the total enrolled population decreased in the 80-seat sections. This resulted in a small differential between the treatment (N=270) and control (N=257) populations for that semester. These students were included in the assigned treatment group as they were randomly assigned to treatment from the total population.

In Fall 2018, 115 students were randomly assigned to the control (3 sections) group and 115 students were randomly assigned to the treatment (3 sections) group. Of those, 88 students in the control group and 91 in the treatment group remained enrolled at the drop/add deadline and had a grade outcome on the class roster of their respective section. In the Spring semester of 2019, 130 students were randomly assigned to the control (5 sections) group, and 131 to the treatment group (5 sections). Of the students in this term, 97 students in the control group and 108 in the treatment group remained enrolled at the drop/add deadline and had a grade outcome on the class roster of their respective section. In the Fall semester of 2019, 257 students were randomly assigned to the control (8 sections) group, and 270 to the treatment group (8 sections). In this last term of the study, 209 students in the control group and 218 in the treatment group enrolled at the drop/add deadline and had a grade outcome on the class roster of their respective section.

Across all three semesters of study, the within-groups outcomes (e.g. within the treatment or control groups) were highly consistent and were also consistent with the complete section data (e.g. counting all students who enrolled in each section, regardless of trial eligibility, looks largely the same as the trial-only analysis reported here), indicating that allowing students to drop/add sections during the regular, open registration period after the initial assignment did not impact the measured outcomes. Specific details of enrollment, exclusion, assignment, and completion for both groups are included in the CONSORT protocol discussion in Section 1.3 above.

Instructors were assigned to the treatment and control course sections before randomization of student enrollment. Treatment section instructors were faculty open to adopting and implementing the MPC curriculum and pedagogy. They were assigned to treatment sections, participated in professional development, and had access to the full MPC curricular materials. Identification of control instructors was only dependent on instructor preference of course time, with control faculty using their traditional instructional practices.

Table S1 summarizes the randomized trial allocations, enrollment patterns and success summary observed for students in the three terms of the study. This clarification is provided as registration procedures may vary across institutions. The institutional procedures allow for registration changes, i.e., drops and adds, freely and without penalty for 7 days after the first day of the semester. After that date, students may request enrollment changes and approved changes are designated as late drops or withdrawals on the course roster and student's transcript. There were 811 Study Participants, i.e., students who were randomly assigned to a treatment or control group and then remained in their assigned section through the institution's regular drop/add deadline. They either received a course grade or were assigned a late drop or withdrawal indicator for leaving the course after the institutional drop/add period deadline. The remainder of the students dropped their assigned section during the institution's drop/add period. Students who switched to another Calculus 1 section in the same term during the drop/add period are identified as switchers. Students who departed calculus for the semester after being assigned to treatment or control conditions are identified as departers.

Table S1 illustrates that no unusual enrollment patterns existed for treatment or control sections. Roughly 80% of all students assigned to treatment or control sections remained in those sections past the drop/add deadline. Twelve percent of the students in treatment and control sections switched to a section in the same semester, 75% of which selected a different meeting pattern.

The remaining 8% of students in treatment or control sections left their section for the semester. These enrollment patterns are typical for the institution's introductory courses and are often the result of switching into other classes, job schedule changes, and/or adapting to family responsibilities. Treatment and control sections showed similar enrollment patterns, thus it is not likely the enrollment patterns biased the results.

Randomized Allocations, Enrollment Patterns and Success Summary			
	Treatment	Control	Total
Enrolled at Randomized Allocation	-	-	1058
Excluded from Assignment	-	-	39
Randomly Allocated to Treatment or Control	516	503	1019
Study Participants: Remained in Assigned Section at Add/Drop Deadline	417 (81%)	394 (78%)	811 (80%)
Succeeded with A/B/C in Course	332 (80%)	270 (69%)	602 (74%)
Left Assigned Section for Another in Same Term (Switcher)	60 (12%)	64 (13%)	124 (12%)
Left Assigned Section for Different Day/Time	47	49	96
Left Assigned Section for Same Day/Time	13	15	28
Left All Calculus 1 Sections in Term (Departer)	39 (8%)	45 (9%)	84 (8%)

Table S1: Table of Randomized Trial Allocations, Enrollment Patterns, and Success Summary

3.2: Equivalence of Student Populations and Investigation of Potential Demographic, Section or Instructor Level Effects on Outcomes

Recognizing that student enrollment may change after randomization, faculty preferences regarding teaching must be taken into consideration for assignments, as well as the inability in a pragmatic randomized trial to blind the control or treatment conditions to students or instructors could introduce unexpected biases / contamination into the study, investigations into possible sources of unexpected biases were carried out. Demographics and academic backgrounds of treatment and control participants were evaluated for significant differences between populations that might have impacted outcomes, and characteristics of instructors in control and treatment groups were compared in an effort to identify any significant differences. Student learning outcomes in the study were measured using a collection of embedded end-of-semester learning measure questions as discussed in SM Section 3.3. These questions were scored anonymously, and student outcomes converted to a scaled score from 0 to 100 percent. This measure, *LearningOutcome,* is then used as the dependent variable to analyze student learning outcomes in the study. Multiple investigation of level effects related to individual sections were carried out from both the student and instructor perspectives including paired section mean analyses and

mixed-effect analyses as described here. No significant differences were found between groups, as detailed below.

3.2.1: Comparison of Student Demographics and Academic Backgrounds

Randomizing the assignment of participants to treatment and control groups seeks to ensure that underlying population characteristics are equivalent in each group. In this section, distributions of participant gender, race, ethnicity and university level in the treatment and control groups are examined for statistically significant differences (presented in Tables S2 and S3). In addition, participants' incoming mathematics backgrounds as represented by high school grade point average (HSGPA), SAT mathematics score, ACT mathematics score, and university assigned Mathematics Placement Score (MPS) were examined and presented in Table S4. Equivalence was found in all measures across treatment and control groups.

Demographics of Randomly Assigned Students				
	Treatment		Control	
	Count	Percent	Count	Percent
Gender				
Female	283	55%	263	52%
Male	225	44%	235	47%
No Data Available	8	2%	5	1%
Race or Ethnic Group				
Black or African American	59	11%	62	12%
Asian Or Pacific Islander	34	7%	32	6%
Hispanic	389	75%	380	76%
Non-Resident Alien	24	5%	23	5%
White / Not of Hispanic Origin	55	11%	52	10%
Other	15	3%	9	2%
No Data Available	13	3%	12	2%
University Level				
College First Year	80	8%	106	10%
College Sophomore	200	20%	188	18%
College Junior	150	15%	123	12%
College Senior	82	8%	79	8%

Table S2: Table of Demographics (Gender, Race/Ethnicity, Class Standing) for Randomly Assigned Students in Treatment and Control Groups

To confirm the similarity of the various demographic and background indicators within the assigned populations, the demographics of those students assigned to treatment and control groups of the trial were compiled. The two groups were compared in total and across each semester with respect to demographic variables of gender, academic standing, and race/ethnicity (Tables S2 and S3). No significant differences were found in the assignments for control and treatment in any category, implying that the randomization introduced no unexpected bias.

	Fall 2018				Spring 2019				Fall 2019				Total			
	Treatment		Control		Treatment		Control		Treatment		Control		Treatment		Control	
Race or Ethnicity	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Asian	7	8	5	6	4	4	5	5	12	6	11	5	23	6	21	5
Black or African American	12	13	11	13	16	15	13	13	23	11	30	14	51	12	54	14
Hispanic	74	81	70	80	83	77	74	76	164	75	158	76	321	77	302	77
White / Not of Hispanic Origin	6	7	5	6	11	10	15	16	32	15	20	10	49	12	40	10
Gender																
Female	47	52	50	57	63	58	44	45	124	57	108	52	234	56	202	51
Male	44	48	38	43	45	42	53	55	94	43	101	48	183	44	192	49
University Level																
College First Year	22	24	23	26	6	6	10	10	40	18	52	25	68	16	85	22
College Sophomore	32	35	36	41	42	39	35	36	91	42	78	37	165	40	149	38
College Junior	20	22	19	22	40	37	27	28	62	28	53	25	122	29	99	25
College Senior	17	19	10	11	20	19	25	26	25	12	26	12	62	15	61	16
Total	91		88		108		97		218		209		417		394	

Table S3: Table of Demographics (Gender, Class Standing, Race/Ethnicity) by Semester for students in the study (i.e., enrolled at the end of the drop/add period on day seven of the semester).

The university computes an institutional Math Placement Score (MPS) to guide course placement. This score is developed as a regression on student outcomes in mathematics courses and other student demographics including Pell grant eligibility, transfer student status, prior mathematics course outcomes, SAT and ACT mathematics subscores, and other available mathematics placement scores such as the ALEKS placement test. The MPS score assigns a student to a course where the regression model predicts at least a 70% likelihood of success (A/B/C) in the course. This data is included in Table S4.

To measure student mathematics backgrounds using factors specific to externally comparable academic outcomes, a comprehensive Mathematics Background Score (MBS) was computed for students in the treatment and control groups. Here, the combined academic background data from SAT mathematics and ACT mathematics subscores with unweighted high school GPA and the institutional MPS by first scaling these scores to a range from 0 to 100 based on the

population mean and range of each measure and then taking the mean of all available scaled scores, ignoring data that was not available for a given student. The Mathematics Background Score was computed for 95% of student participants in the control and treatment groups and is included in Table S4.

Mathematics Background of Randomly Assigned Students		
	Treatment	Control
Mean HSGPA Math	3.42	3.45
% students with score	89	86
# students with score	455	434
Mean Math Placement Score	65.00	67.07
% students with score	62	62
# students with score	320	314
Mean Math Background Score	65.02	65.97
% students with score	95	95
# students with score	490	478
Total Possible Students	516	503

Table S4: Table of Mathematics Background Information for Students in the Trial

Finally, to develop a more comprehensive understanding of student mathematics backgrounds as they entered one of the study sections, in the Spring and Fall 2019 semesters, the Precalculus Concept Assessment (PCA) inventory (*91*) was collected at the beginning of the semester in both the treatment and control sections. The mean scores of that pre-assessment for both groups are also reported in Table S5. Within these data, students in the treatment and control groups have almost identical demographic characteristics as well as mathematical backgrounds, as expected with a randomized study. This consistency is evident in the total populations from all three semesters combined as well as within each semester.

	Fall 18			**Spring 19**			**Fall 19**		
	CN	**TR**	***p*-value**	**CN**	**TR**	***p*-value**	**CN**	**TR**	***p*-value**
Unweighted HSGPA	3.46	3.45	$p = 0.8976$	3.38	3.32	$p = 0.1647$	3.48	3.46	$p = 0.5915$
PCA Prescore				9.3	8.7	$p = 0.2783$	9.8	9.5	$p = 0.4379$

Table S5: Term-by-term comparisons (and the associated *p*-values from *t*-test comparisons) of control and treatment groups on students' academic background information indicating mathematical preparation. The information includes means of their reported unweighted high school GPAs, and prescores on the PCA inventory on a 0 to 25 point scale. (CN = control sections; TR = treatment sections)

Examining all measures, no statistically significant differences between control and treatment populations are observed in the overall or semester-by-semester demographic or mathematics background data sets. No statistically significant differences are seen in reported unweighted High School GPAs, institutional Math Placement Scores, Mathematics Background Scores or PCA prescores. Finding no significant differences in the allocations for control and treatment in any category leads to the conclusion that the randomization introduced no unexpected bias. In the section below discussing the paired means of student mathematical background data, investigations of the differences in the mathematics backgrounds of students are examined and show no statistically significant difference even at the paired section level.

3.2.2: Comparison of Instructor Characteristics

Characteristics for faculty teaching both control and treatment sections were compared as groups and to the total group of faculty teaching calculus prior to the beginning of the trial to identify any potential biases. Instructors in treatment and control sections included mathematics department faculty and all were allowed to request specific course sections, following the department's customary practice. Prior to beginning of each semester, both control and treatment instructors were assigned to sections based on their availability, other teaching obligations, and scheduling preferences, but not added to the course rosters. Faculty willing to implement the Modeling Practices of Calculus curriculum and pedagogy were recruited and prepared to teach the treatment sections, as noted above. The experimental sections were double the standard capacity and split into two sections at standard capacity (one for the control and one for the treatment section). Once the randomization of students was complete and these sections were split, both treatment and control faculty names were added to the section rosters. This process produced comparable faculty groups in both treatment and control conditions, when compared across a number of faculty characteristics.

Faculty in both groups included full-time tenured/tenure-track (TT) and non tenure-track (NTT). All of the control faculty were full-time TT and NTT faculty. The treatment group consisted of full-time TT and NTT faculty and one part-time adjunct faculty. Faculty backgrounds were collected from surveys and institutional sources to develop categories that expressed the teaching experience, gender, active learning experience, and rank and are summarized in Table S6. Also tracked was the number of times a faculty member taught in each condition. There was one instance of one person teaching in the treatment group after having taught in a control section, but no faculty member who taught in the treatment group subsequently taught in the control group again. To examine for outliers in faculty instructor practice, historic student grade outcomes were compared across the treatment and control sections. Historic student success data, represented as the percent of assigned A/B/C grades in a course, for either Precalculus, Calculus I, or Business Calculus were collected for all faculty (treatment, control, or any other group) for the four semesters prior to the experiment (fall and spring semesters Fall 2016 to Spring 2018). These three courses were used to provide a representative student response to faculty effectiveness at the time of the experiment using similar level mathematics courses. Student outcome data from the prior 4 semesters and three courses was used to establish a comparative index for all faculty.

	Treatment	Control
Teaching Experience		
Less Than Five Years	11	3
Five to Ten Years	0	2
More Than Ten Years	5	11
Gender		
Female	4	9
Male	12	7
Active Learning Experience		
None	5	3
Low	4	9
Medium	4	4
High	3	0
Rank		
Adjunct	1	0
Non tenure-track (NTT)	12	11
Tenured or Tenure-Track	3	5
Total	16	16
Student Success (A/B/C grades) percentage in Precalculus, Calculus I, and Business Calculus, Fall 2016 to Spring 2018	54% (N=643)	55% (N=2,514)

Table S6: Treatment and Control Faculty Characteristics for the 16 treatment and 16 control sections. Instructor characteristics represent accumulated individual instances of instruction for each condition over the three-semester experiment counting repeated instances of faculty. Student success in prior courses is computed as an A/B/C grade in an instance of Precalculus, Calculus or Business Calculus for an instructor during Fall 2016, Spring 2017, Fall 2017, and Spring 2018 prior to the beginning of the trial.

Overall, the faculty characteristics are generally similar. The most significant departure is the number of years of experience teaching. Faculty teaching treatment sections were typically less experienced, but both groups included a sample of experienced as well as comparable numbers of tenured research faculty. This may suggest that faculty new to the field are more open to trying new instructional techniques, though two treatment group faculty had substantial prior teaching experience. Student success rates for faculty in treatment and control were not statistically different, indicating neither group had unusually high or low passing rate histories. An investigation of instructor level effects on student outcomes was carried out and presented in Section 3.2.4 below.

3.2.3: Student and Instructor Differences Within Clusters

Analyses of section level differences for student backgrounds and student end-of-semester learning measures was carried out to check for equivalence at the section level. The first analysis in Section 3.2.3.1 explores the variability of the student mathematics backgrounds within time of day and day of week section clusters, and the second in Section 3.2.3.2 examines the learning measure data within those same course pairings that resulted from the random allocation of students to different conditions.

Section 3.2.3.3 compares end-of-semester learning measures data to the prior course student success data for instructors teaching in the study arms collected from institutional data during the two years prior to the study in Precalculus, Calculus I, and Business Calculus. These relationships are then investigated for covariation with outcomes before moving to the full mixed effects model in Section 3.2.4.

3.2.3.1: Paired Means of Student Demographic Variables

In a randomized trial, randomization is expected to create equal variation of the characteristics of the populations in each group, and as noted earlier in Section 3.2.1, the treatment and control student groups have equivalent distributions of all measured demographics an no unusual patterns of enrollment are observed. As an initial investigation, the High School Grade Point Averages (HSGPA) and Mathematics Background Scores (MBS) of students in paired sections created by the randomization process were analyzed to determine how the student populations were distributed within that level. The paired means of the unweighted high school GPA (reported in Table S4) were examined and no significant difference was found ($t(15)=-0.76$, $p=0.46$, $n=16$) in the incoming backgrounds of students in the treatment and control groups by section. These paired means are shown in Fig. S3 (left).

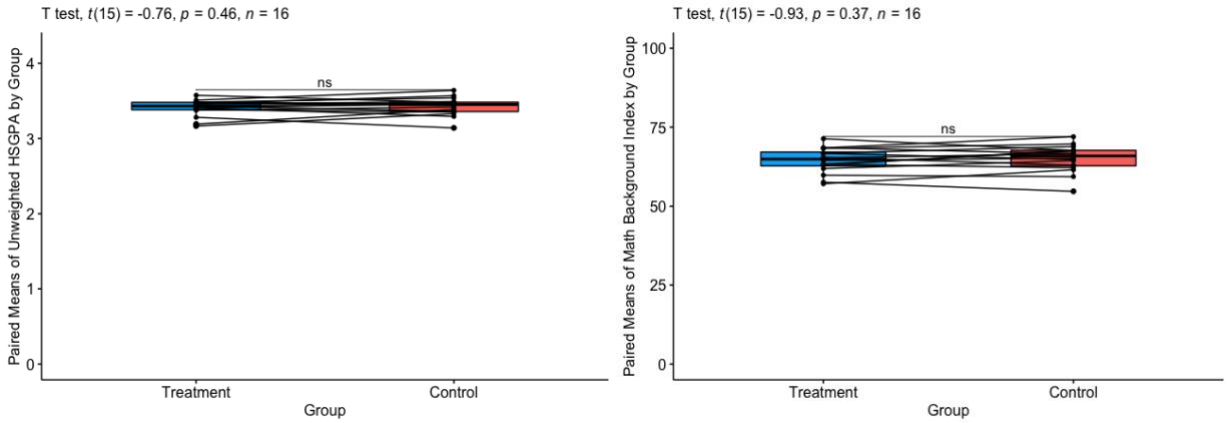


Fig. S3: Paired Section Means for Unweighted HSGPA (left) and Math Background Scores (right) by Treatment Group

Similarly, paired means of the scaled composite Math Background Scores (reported in Table S4) were examined and no significant difference was found ($t(15)=-0.93$, $p=0.37$, $n=16$) for students in the treatment and control groups by section. These paired means are shown in Fig. S3 (right).

Additionally, sample distributions for student mathematics background score (MBS) data computed (above) were compared using Kolmogorov-Smirnov tests and Wilcoxon rank sum tests (*92*, *93*) with continuity correction for both the total populations in treatment and control as

well as the subpopulations in paired sections. In all cases, both tests found the sample distributions to be the same for treatment and control. Indeed, the paired samples were found to be more similar distributionally due to their structural association.

Direct comparisons as well as Kolmogorov-Smirov and Wilcoxon tests show that randomization of treatment and control section assignments allocated students equivariantly both within those constraints, as well as across the entire study population, as intended. This confirms that student mathematics backgrounds were equivalent at the time of condition assignment, which would be expected for randomized assignment.

3.2.3.2: Paired Means of Section Level Student Outcomes

A visual inspection of the paired section means of the end-of-semester learning measures for the treatment and control groups, along with a *t*-test of the within pair section means, confirms the overall differences in outcomes for the two populations. The paired section means were found to be statistically significantly different ($t(15)=5.57$, $p<0.0001$, $n=16$) and the within-pair section means reflect the overall general trend found in the data set as shown in Fig. S4.

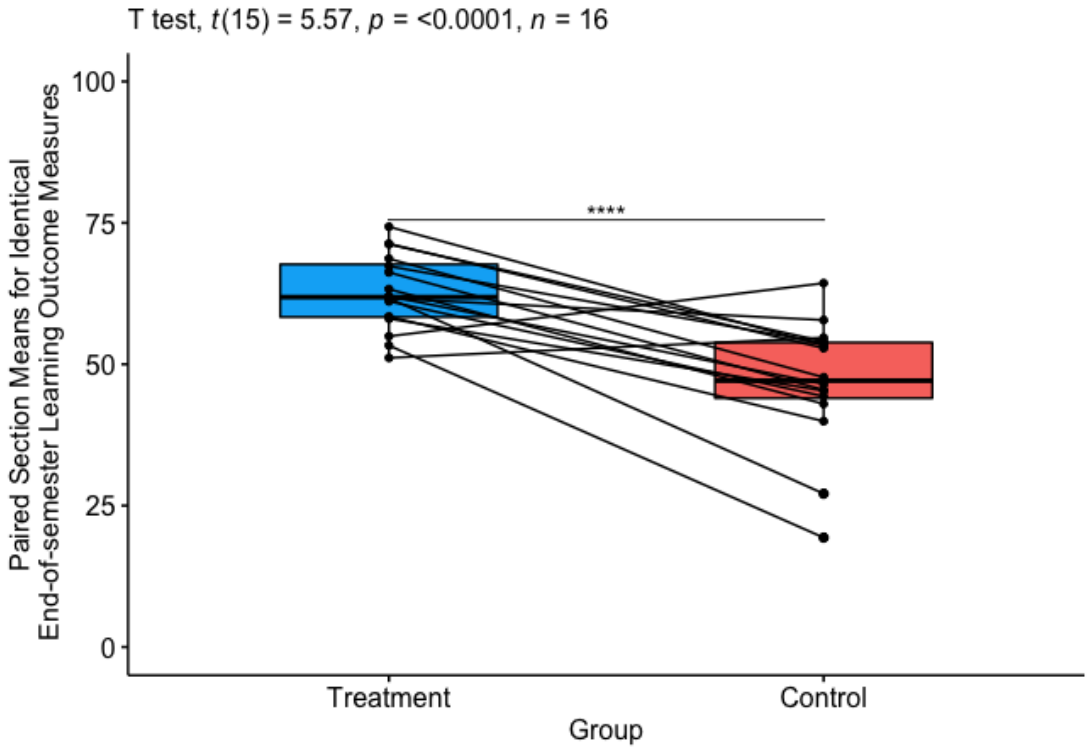


Fig. S4: Paired Section Means for Identical End-of-Semester Learning Outcome Measures, with significant difference indicated by the top line.

3.2.3.3: Instructor Level Effects on Outcomes

Instructor instructional practices and their potential impact on outcomes were also investigated to determine whether or not instructors in the treatment or control conditions exhibited historical instructional patterns that were consistent or not with each other and with historical departmental patterns. Historic student success rates (grade of A, B or C in the course) were used as a proxy

for faculty instructional impact on students at the time of the experiment. Institutional data were collected for all Calculus 1 instructors' Precalculus, Calculus I, or Business Calculus sections for the fall and spring semesters from Fall 2016 to Spring 2018. During this time, 51 distinct instructors taught 9,095 students in 90 sections over four semesters. For comparison, 9 of the 12 control group instructors taught 2,514 students in 37 sections while 5 of the 7 distinct treatment group instructors taught 643 students across 10 sections.

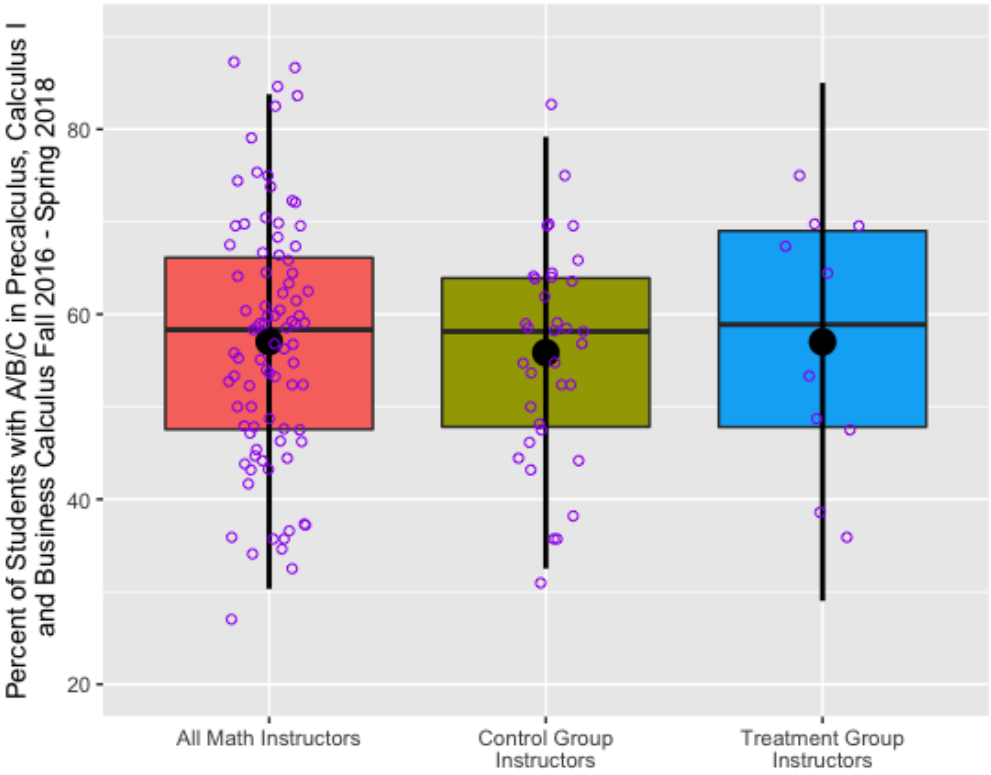


Fig S5: Historic percentage of students with A/B/C grades (passing rate) in sections of Precalculus, Calculus I and Business Calculus for all math faculty, treatment, and control instructors prior to the beginning of the trial. Lower and upper box boundaries correspond to 25th and 75th percentiles. Mean for groups indicated by solid dot with vertical error bars equal to two standard deviations.

The box with scatter plot shown in Fig. S5 illustrates the historic relationship of the student course success outcomes for the three groups. Note that both treatment and control section instructors were representative of the existing instructional outcomes in these mathematics courses. To quantify any potential relationships between the historic student success rates in Precalculus, Calculus 1, and Business Calculus sections for these instructors with the end-of-semester learning measures score outcomes in the assignment group of students in the trial, the difference, *SuccessVar*, between each instructor's success percentage in each section they taught during this time period and the overall departmental success rate in the three courses (μ=57.1%, *sd*=15.4) was computed. No significant differences ($t(524.72) = 1.2716$, $p = 0.2041$) between the success rates of the treatment (μ=54.8%, *sd*=13.6) and control group (μ=56.4%, *sd*=10.2) instructors is observed, and no statistically significant effects of this variable, *SuccessVar*, are observed on the outcome variable *LearningOutcome* ($F(226,1) = 0.5097$, $p = 0.4757$).

3.2.4: Full Models of Study Outcomes and Effect Analyses

The primary goals of the study included collecting measures of student learning that could be used to quantify the extent to which the treatment condition impacts those outcomes represented by the anonymously scored end-of-semester learning measures. These data also provide insight into the extent to which that impact might be expected to be repeatable in other circumstances (*94*, *95*).

The main document reports the effect size measured for the study taken as a whole. The following provides models and analyses of variance to investigate the ways in which covariates and other regressors impact the relationship of the learning outcome measure with the treatment condition. Models are presented that analyze student learning outcomes measured using the identical end-of-semester student learning measures by constructing mixed-effects models of that outcome dependent on existing student demographic data with random effects from section and instructor level factors. Sensitivity analyses are provided to characterize possible confounding within these results.

3.2.4.1 Mixed-Effects Model of Student Learning Outcome

Institutional data were collected and used to construct the Mixed-Effects Model (MEM) below with student demographics as fixed effects, the treatment condition as a fixed effect, and the section levels (*Section)*, time of day/day of week levels (*SecPair)*, and instructor levels (*TIDw*) as random effects. Student demographic variables included student mathematics background score (*MBS*), gender (*Gender*), and race and ethnicity (*RE*). A description of all the variables involved the model is shown in Table S7.

Interactions between *Treatment* and *MBS* along with *Treatment* and *Gender* were initially included to control for possible dependence on those factors. Both interactions were found to be statistically nonsignificant ($t(658.9)= –0.250$, $p=0.803$ and $t(646.0)= –0.099$, $p=0.921$, respectively), and so the model with no interactions was used. Linearity, normality, and homogeneity of variance assumptions were all assessed through visual inspection of the residuals, finding no significant violations. Homogeneity of variance was further confirmed using a Levene's test ($F(31) = 0.991$, $p = 0.483$). Considering the cluster level effects of *Section, SecPair,* and *TIDw* as random effects, the mixed effects model below was constructed to explore cluster level random effects in the study data as it models section clustering, time of day/day of week effects and instructor effects in the data:

$$LearningMeasure_{i,jkl} \sim \left(\beta_0 + \gamma_{Section_{ij}} + \gamma_{SecPair_{ik}} + \gamma_{TIDw_{il}}\right) + \beta_1 Treatment_i + \beta_2 MBS_i + \beta_3 Gender_i + \beta_{4m_i} RE_{m_i} + \epsilon_i$$

$LearningOutcome_i$ is the $n_i x 1$ vector of students' scores. Table S8 shows the model structure for the full model with all demographic covariates on the left and the model with only the *Treatment* fixed effect on the right for comparison.

Field	Description	Type
$LearningMeasure$	Anonymously scored measures of calculus understanding from the end-of-semester learning measure scaled to [0,100]	Outcome; continuous variable
$Treatment$	Assigned group as 0=control or 1=treatment	Randomized Assignment; categorical with two levels
MBS	Math Background Score scaled to [0,100]	Demographic Continuous
$Gender$	Self-Identified Gender as 0=male or 1=female if reported	Demographic categorical with two levels
RE_u	Race or Ethnicity encoded as 0 or 1 for RE_1=Black or African American, RE_2=Hispanic or Latino/Latina, RE_3=White, RE_4=Others Combined Due to Low Numbers	Demographic categorical with four levels
$Section_j$	Section assignment induced by Course Split Random Assignment with $Section_j$ = 0 or 1 depending on student presence in cluster level j=1 to 32	Cluster; Categorical with 32 levels
$SecPair_k$	Section pairing induced by Time of Day/Day of week chosen for registration by student prior to course split random assignment with $SecPair_k$ = 0 or 1 depending on student presence in cluster level k=1 to 16	Cluster; Categorical with 16 levels
$TIDw_l$	Instructor of record for course with $TIDw_l$ = 0 or 1 depending on student presence in cluster level l=1 to 19	Cluster; Categorical with 19 levels

Table S7: Demographic and Other Study Data utilized in MEM Analyses

	Full Model			Reduced Model		
Predictors	*Estimates*	*std. Error*	*p*	*Estimates*	*std. Error*	*p*
(Intercept)	0.69 (-9.39 – 10.78)	5.14	0.892	47.75 (43.11 – 52.40)	2.37	**<0.001**
Treatment [TR]	15.79 (10.36 – 21.22)	2.77	**<0.001**	15.37 (9.44 – 21.29)	3.02	**<0.001**
MBS	0.78 (0.66 – 0.90)	0.06	**<0.001**			
Gender [F]	-3.80 (-6.44 – -1.15)	1.35	**0.005**			
RE [2]	-9.30 (-16.14 – -2.47)	3.48	**0.008**			
RE [3]	-2.26 (-7.27 – 2.74)	2.55	0.374			
RE [4]	-0.29 (-6.71 – 6.14)	3.27	0.930			
Random Effects						
σ^2	288.60			365.39		
τ_{00}	17.36 Section 13.38 TIDw 12.44 SecPair			27.76 Section 11.85 TIDw 21.73 SecPair		
ICC	0.13			0.14		
N	32 Section 19 TIDw 16 SecPair			32 Section 19 TIDw 16 SecPair		
Observations	671			671		
Marginal R^2 / Conditional R^2	0.303 / 0.394			0.121 / 0.247		
AIC	5746.537			5911.859		
log-Likelihood	-2862.269			-2949.929		

Table S8: Mixed Effects Model with Random Intercepts for *Section* clusters – Full Model with all Demographic Covariates (left) and Reduced Model with Treatment Fixed Effect only (right)

Restricted maximum likelihood fitting was used for these models with the `lme4` (*96*) package in R. Tests of fixed effects were conducted using *t*-tests with Satterthwaite degrees of freedom approximations computed with the `lmerTest` package (*97*). Satterthwaite degrees of freedom were used to control for Type I error rates in the multilevel models employed (*98*). The total explanatory power of the model is substantial (conditional $R^2 = 0.39$) and the part related to the fixed effects alone (marginal R^2) was found to be 0.30. Within the model, the effect of *Treatment* [TR] is statistically significant and positive ($\beta = 15.79$, 95% CI [10.36, 21.22], $t(660) = 5.71$, $p < 0.001$), in the presence of the effect of *MBS* (statistically significant and positive, $\beta = 0.78$, 95% CI [0.66,0.90], $t(660) = 12.68$, $p < 0.001$), the effect of *Gender* [F] (statistically significant

and negative, β = -3.80, 95% CI [-6.44, -1.15], $t(660) = -2.82$, $p = 0.005$), and the effect of *RE* [2] (statistically significant and negative, β= -9.30, 95% CI [-16.14, -2.47], $t(660) = -2.67$, $p = 0.008$). Confidence Intervals (CIs) and *p*-values were computed using a Wald *t*-distribution approximation.

In the reduced model, the intercept in the *Treatment* only fixed effects is 47.75 and the coefficient of *Treatment* is found to be 15.37. The random effect variance describing how section levels vary is 365.39. The value of the adjusted R^2 for the fixed effects portion of the full model was computed as 0.303 which implies (*99*) an effect size of Cohen's $f = 0.659$, close to that of the total study effect observed. The adjusted R^2 for the random effects in the full model is 0.085 which implies an effect size of 0.305 for the random portion. Computing the same terms for the reduced model to approximate the portion of the variance related to the *Treatment* effect, the model indicates an adjusted $R^2 = 0.121$ for the fixed effects portion which implies an estimated effect size of 0.371 (small-medium) in the presence of random section level intercepts. Using the methods from (*99*, *100*) and (*101*) to compute the adjusted R^2 values for the fixed effects in the model, the variance explained by the fixed effects are shown in Table S9.

Term	Partial R^2 estimate	CI lower	CI upper
Full	0.303	0.23	0.396
Treatment	0.119	0.04	0.213
MBS	0.150	0.07	0.244
Gender	0.010	0	0.116
RE	0.012	0	0.118

Table S9: Estimates of partial R^2 for Fixed Effects Terms in Full Model

Using these values, the effect of treatment within the full mixed effects model is found to be Cohen's $f = 0.368$ with 95% CI [0.197,0.520], consistent with the reduced model. The study is powered to detect a difference of means (*t*-test) to an effect size of Cohen's $f = 0.109$ with the unbalanced sample sizes observed and so there is a 95% likelihood that the effect of the treatment will be observed in this interval.

In models with the type of cluster levels observed here, it can be a concern that the variance in the independent variable may differ greatly in different cluster levels. To determine if this type of heteroscedasticity was a concern, the variances of the residuals for the full model between cluster levels were compared and tested using Levene's test. Variances of model residuals for the total model were not significant within the Section levels ($F(31) = 0.991$, $p = 0.483$). Overall, variance of the model residuals between the two *Treatment* levels were also not significantly different ($F(1) = 1.406$, p=0.236) and so the model estimates can be considered reliable computed with the random effects for the cluster levels included. Quantile-Quantile plots of the random effects of the section levels and the residuals confirm the fit of the model.

The model fit confirms the significance of *Treatment*. The variance due to random and fixed effects, and the portion of this variance that is due to the *Treatment* effect in the presence of the random instructor and section effects suggest that the effect is significant and that a substantial portion of that effect is due to other factors beyond instructor and time of day/day of week effects.

3.2.4.3 Sensitivity Analysis and Unmeasured Confounders

To check for the possible existence of an unmeasured confounder, a sensitivity analysis (SA) was conducted. Let *U* be a confounder that could be responsible for the effect of *Treatment* measured on *LearningMeasure* in the study. Using the approach to identify the impact of unmeasured confounders developed in (*102*), the SA was implemented in the R package `treatSens` with a linear analysis of a possible confounder *U* whose output is provided in Fig. S6.

In the Fig. S6 plot, the level curves show that, for an unmeasured confounder with a coefficient of $\zeta^{Z} = 1$ in the model developed in (*102*) as

$$LearningMeasure|X,\ U,\ Treatment \sim N(\beta^{LM}X + \zeta^{LM}U + \tau Treatment, \sigma_y^2)$$

$$Treatment\ |\ X,U \sim \text{Bernoulli}(\Phi(X\beta^{Z} + \zeta^{Z}U))$$

where *U* is the unmeasured confounder, *X* is the matrix of the covariates of *Treatment*, and τ is the treatment effect, the outcome variable would need to have a coefficient $\zeta^{LM} >> 35$ in the model to reduce the effect of *Treatment* on that variable *LM* = *LearningMeasure* to insignificant levels. If ζ^{Z} is closer to the observed values in this plot less than 0.25, the outcome coefficient would need to be close to the maximum value of *LM*, that is near or greater than 100. Both of these configurations are inconsistent with the observed data. All other covariates including the appear near the vertical axis in this plot with horizontal components close to zero and with no vertical components larger than 7. This output indicates that the hypothetical unmeasured confounder would need to more than four times larger than all other measured effects, including student mathematical background even if collinear with *Treatment*.

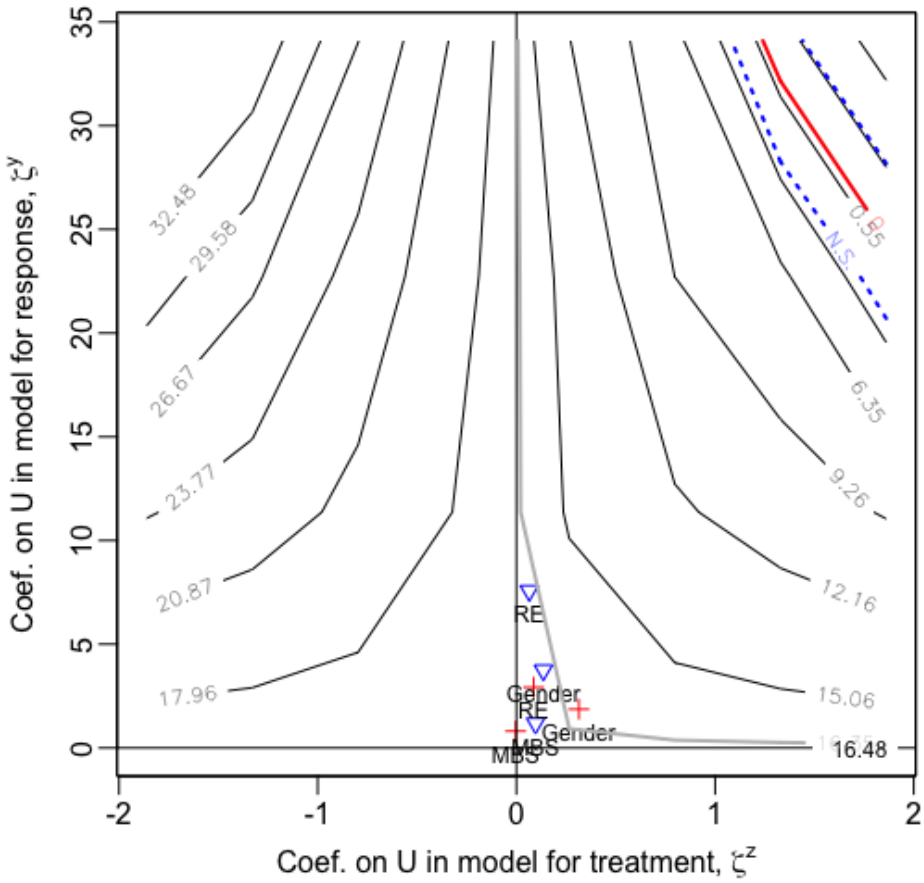


Fig. S6: Sensitivity Analysis for Fixed Effects in MEM

3.3: End-of-Semester Learning Measures Overview

The end-of-semester learning measures were designed to span the major learning objectives of a Calculus 1 course to determine how well students understood essential elements of, and could exhibit fluency and technical competency in, calculus at the end of the course. The assessment strategy followed the national consensus on best assessment practices. Gleason et al., (*48*) note that introductory calculus is rich in conceptual topics derived from the interactions between algebraic, tabular, and graphical representations of function and data such as limits, continuity, integration, and the concept of the derivative itself. Substantial prior work in the teaching and learning of calculus has reinforced the importance of the development of conceptual understanding as a primary goal of calculus (*49*). To further understand what students should know and be able to do upon completion of introductory calculus, (*50*) developed a framework of essential end goals based on the shared views of 24 calculus experts' views on what it means to understand first-year calculus. Collectively, these experts agreed that first-year calculus students must demonstrate the mastery of fundamental calculus concepts and skills, build connections and relationships between these concepts and skills, and the ability to apply calculus ideas to solve problems.

The identical learning measure questions used in this study consisted of six questions for the initial Fall 2018 semester and eight questions in the following two semesters. The topics of the questions included the following: interpreting a graph, evaluating limits, implicit differentiation, related rates, absolute extrema, applied optimization, using derivatives to sketch a curve, and

evaluating an integral. Selection of these topics was driven by both the local conditions and national consensus. All sections of calculus follow a department established list of concepts and topics along with a similar set of assessment practices, established by the department, thus ensuring that the topics are included in both treatment and control sections. The introductory Calculus 1 course is based on George Thomas' *Calculus and Analytical Geometry* (*103*) outline which has been widely used since the 1950s and focuses on limits and continuity, derivatives, applications of derivatives, and integration (*35*, *39*). The items in the study's identical assessment block align with the findings of the experts interviewed by (*104*) who shared complete agreement that limits, derivatives, applications of the derivative, and integrals are core concepts for introductory calculus students. The analysis showed less agreement between experts that approximation and sequences and series serve as fundamental calculus concepts, thus assessment of approximation and sequences topics was not included in the end-of-semester learning measures.

3.3.1: Development of Identical End-of-Semester Learning Measure Questions

The identical end-of-semester learning measures were questions collectively developed and embedded in identical cumulative final examinations across both treatment and control sections. Development of the identical questions seamlessly integrated into established departmental practices. The introductory calculus courses were collectively organized by a committee of instructors responsible for choosing an official textbook and homework platform, preparing a course outline with topics, and developing the final exam.

The learning measure questions were developed initially by a final exam working group consisting of a subset of instructors currently teaching the course and representing both the treatment and control conditions. After agreeing to topics for each of the identical questions, individual members would develop questions and share with the working group for review and feedback. Revisions were then made and shared with the entire group of faculty teaching the course for feedback iteratively until all instructors approved of the questions. This process ensured that the language in the questions was appropriate for students in all sections of the course. Once all questions were finalized, multiple versions of each question were created by making minor modifications to the questions that did not change their essential difficulty in order to accommodate departmental exam administration protocols. Questions shared across all versions and sections represented roughly two-thirds of the final exams, allowing individual faculty to cater the remainder of the exams to their instructional goals. Further, the exams and questions were formatted consistently and without course section identifiers to allow completely anonymized evaluation of these portions during the subsequent comparative analysis.

3.3.2: Characteristics of End-of-Semester Learning Measure Questions

As noted above, end-of-semester measures of student learning outcomes were embedded in course final exams. The basic framework for the identical final exam questions used came from department course assessment practices aligned to content required for all Calculus 1 courses in the state. The embedded exam questions use arose from normal question types developed as a departmental norm over time and refined by the instructional team that included both treatment and control section instructors. In order to ensure that the assessments would maintain the prior emphasis on traditional calculus skills and concepts while also providing feedback on student learning that aligned with nationally accepted standards for calculus, faculty teaching calculus across all sections worked to refine the items and determine their correspondence with the framework established by Tallman, et al in (*35*). Table S10 provides the characteristics of the

end-of-semester learning measure questions based the Exam Characterization Framework (*35*) there which characterizes exam items by three attributes: item orientation, item representation, and item format. The item orientation dimension contains seven categories of intellectual behaviors needs to respond to an item: remember, recall and apply procedure, understand, apply understanding, analyze, evaluate, and create. Items were also coded by their representation type: applied/modeling, symbolic, tabular, graphical, definition/theorem, proof, example/counterexamples, and explanation. The third dimension characterizes exam items by their format based on three categories: multiple choice, short answer, or broad open-ended. Under the item format dimension, items were also subcategorized based on whether the item required students to provide justification, an explanation, or solve a word problem. Research team members reviewed the identical embedded exam items used in all versions and sections during the study and coded them using the ECF. Multiple team members evaluated the questions and aligned results to ensure consistent characterizations within the framework.

The results from coding the end-of-semester learning measure using the item orientation taxonomy showed that three of the exam items require students to apply or demonstrate their understanding. One of the eight items required students to create a graph using their understanding of derivatives. Four out of eight of the exam items require students to evaluate or recall and apply a procedure. This is evidence that the final exam requires students to apply a range of cognitive skills to solve the problem.

The results from coding the identical question items using the item representation taxonomy revealed that the majority of items (63%) were stated symbolically. No items asked students for information in the form of a table. These results are consistent with Tallman et al.'s (*35*) findings. It is important to note that the remaining three items were stated exclusively as "applied modeling" or "graphical" items. Increased emphasis was placed on these items since they covered more than one calculus concept. For example, the related rates item involves the application of the derivative, but a foundational understanding of implicit differentiation is also needed.

The results from coding the identical question items using the item format taxonomy revealed that three-fourths of the items were coded as short answer. While a majority of the items were coded as short answer, students were still required to justify their solution on four of six items in the "short answer" format category. No exam items were coded as multiple-choice format items. Two of the items, related rates and applied optimization, were coded as "broad, open-ended" items because these problems were stated in a real-world context. It is noteworthy that the set of question items meets the Tallman et al.'s (*35*) coding threshold of having more than 10% of the exam items classified as a word problem in the "broad, open-ended" category. The learning measure questions presented opportunities for students to demonstrate their ability to apply their understanding of derivatives to real-world applications, something that Tallman et al. (*35*) noted as lacking when examining over 150 randomly-selected Calculus I final exams studied from various post-secondary U.S. institutions. This data shows that the set of identical questions embedded in the final exam used in this study does not focus solely on students' ability to memorize and apply a procedure.

Learning Measure Item Topic	Item Orientation	Item Representation	Item Format (sub-code)
Interpreting a graph	Understand	Symbolic	Short answer
Evaluating limits	Evaluate	Symbolic	Short answer (justify)
Implicit differentiation	Recall and apply procedure	Symbolic	Short answer
Related rates	Apply understanding	Applied/modeling	Broad open-ended (word problem)
Absolute extrema	Recall and apply procedure	Symbolic	Short answer (justify)
Applied optimization	Apply understanding	Applied/modeling	Broad open-ended (word problem)
Sketching a curve	Create	Graphical	Short answer (justify)
Evaluating an integral	Evaluate	Symbolic	Short answer (justify)

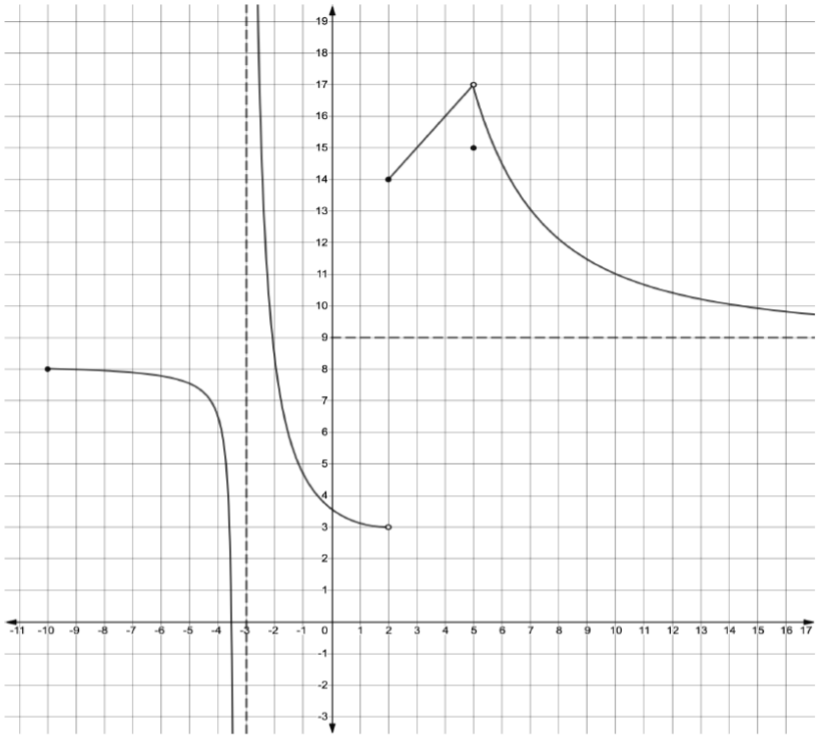
Table S10: Characteristics of the learning measure questions based on Tallman et al.'s Exam Characterization Framework (*35*).

3.3.3: Sample End-of-Semester Learning Measure Questions by Topic

Figs. S7 – S9 present a representative sample of the identical end-of semester learning measure questions by topic. These are typical of the types of problems seen by students in Calculus 1 courses. Question format, terminology, and notation were agreed upon by instructors of control and treatment sections to ensure fairness across all sections. Also, to discourage and detect any academic dishonesty, the questions were versioned by varying the order of the problems and making minor modifications (e.g., changing constants, values or functions) while keeping the difficulty of the problem consistent.

Interpreting a graph

Consider the function $y = f(x)$ graphed below.



- Find the following limits. (If a limit does not exist write DNE.)

$\lim_{x \to -3^+} f(x) =$ $\lim_{x \to -3^-} f(x) =$ $\lim_{x \to -3} f(x) =$ $\lim_{x \to 2^+} f(x) =$ $\lim_{x \to 2^-} f(x) =$

$\lim_{x \to 2} f(x) =$ $\lim_{x \to 5^+} f(x) =$ $\lim_{x \to 5^-} f(x) =$ $\lim_{x \to 5} f(x) =$ $\lim_{x \to \infty} f(x) =$

- State the domain of $f(x)$ as a union of intervals.
- State the range of $f(x)$.
- State the equations of the horizontal and vertical asymptotes (if any).
- List the x-values where the function is discontinuous.

Fig. S7: End-of-semester Learning Measures Item - Interpreting a Graph

Evaluating limits

- $\lim_{x \to 1} \frac{1 - x^2}{x^3 \ln x}$

- $\lim_{x \to \infty} \frac{3x^5 - 8}{6 - 2x}$

Implicit differentiation

Use implicit differentiation to find $\frac{dy}{dx}$ for $3x^4 + xy^2 = 2y^3 + 5$.

Related rates

Two trains leave the station at the same time. Train A travels east at 6 kilometers per hour, while Train B travels north at 8 kilometers per hour. At what rate is the distance between the two trains changing at the moment Train A has traveled 3 kilometers and Train B has traveled 4 kilometers?

Absolute extrema

Find the absolute (global) maximum and minimum, values and locations, of the function $f(x) = \left(x^2 - 2x\right)^{1/3}$ over the closed interval $[0, 4]$

Applied optimization

A contractor is tasked with enclosing a 32-m^2 rectangular patch by a fence divided into three identical and adjacent plots separated by two parallel fences. *See figure below.* What dimensions for the outer rectangle will require the smallest total length of fencing material needed?

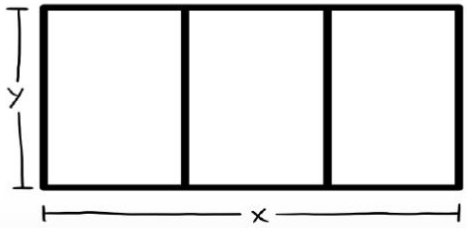


Fig. S8: End-of-semester Learning Measures Items - Evaluating Limits, Implicit Differentiation, Related Rates, Absolute Extrema and Applied Optimization

Sketching a curve

Given $f(x) = x^3 - 3x^2$:

- State the domain of $f(x)$.
- State the $x-$ and $y-$intercepts of $f(x)$.
- Determine the intervals on which the function is increasing/decreasing. Also, identify any relative maxima and minima.
- Determine the intervals on which the function is concave up/down. Also, list any inflection points.
- Find the limits at the ends of the domain (end behavior) and state any asymptotes of the function. If an asymptote does not exist, state "NA".
- Sketch $f(x)$. **Label all relative minima, relative maxima and inflection points.**

Evaluating an integral

Evaluate the following integral. If substitution is used, be sure to clearly indicate u and du.

$$\int 30x^2 e^{(5x^3+4)} dx$$

Fig. S9: End-of-semester Learning Measures Items - Curve Sketching, and Evaluating an Integral

3.3.4: End-of-Semester Learning Measure Questions - Individual Item Results

Tables S11 – S13 provide the performance of students assigned to the treatment and control sections on the individual embedded identical end-of-semester learning measures items for all three semesters of the study. Note that results reported below are in all cases consistent with the results for all students enrolled in these course sections. That is, when including students who were not randomly assigned but enrolled in the treatment and control sections after the randomized assignment, the findings are identical in nature.

Fall 2018	Maximum Score	Treatment n = 83	Control n = 63	*p*-value	Effect Size	95% Confidence Interval for eff. size
Absolute extrema	5	62%	38%	$p < 0.001$	$d = 0.739$	(0.398,1.08)
Evaluating an integral	6	50%	54%	n/s	$d = -0.095$	(-0.426,0.235)
Evaluating limits	5	65%	58%	n/s	$d = 0.233$	(-0.098,0.565)
Interpreting a graph	20	85%	76%	$p < 0.01$	$d = 0.469$	(0.135,0.804)
Related rates	6	11%	10%	n/s	$d = 0.086$	(-0.244,0.417)
Sketching a curve	23	66%	52%	$p<0.01$	$d = 0.520$	(0.184,0.855)
Overall	**65**	**65%**	**55%**	***p* <0.01**	***d* = 0.502**	**(0.166,0.837)**

Table S11: Individual item results for Fall 2018 treatment and control sections, with percent correct for treatment and control sections, as well as the associated *p*-value, effect size and confidence intervals.

Spring 2019	Maximum Score	Treatment n = 100	Control n = 79	*p*-value	Effect Size	95% Confidence Interval for effect size
Absolute extrema	6	48%	25%	$p < 0.001$	$d = 0.757$	(0.449,1.064)
Applied optimization	9	25%	9%	$p < 0.001$	$d = 0.686$	(0.38,0.991)
Evaluating an integral	5	47%	31%	$p < 0.01$	$d = 0.441$	(0.141,0.742)
Evaluating limits	8	53%	44%	n/s	$d = 0.263$	(-0.0354,0.561)
Implicit differentiation	5	68%	59%	n/s	$d = 0.291$	(-0.008,0.589)
Interpreting a graph	17	75%	72%	n/s	$d = 0.169$	(-0.129,0.467)
Related rates	8	60%	41%	$p < 0.001$	$d = 0.617$	(0.313,0.921)
Sketching a curve	18	66%	45%	$p < 0.001$	$d = 0.726$	(0.42,1.033)
Overall	**76**	**57%**	**42%**	$\mathbf{p < 0.001}$	$\mathbf{d = 0.748}$	**(0.44,1.05)**

Table S12: Individual item results for Spring 2019 treatment and control sections, with percent correct for treatment and control sections, as well as the associated *p*-value, effect size and confidence intervals.

	Maximum Score	Treatment n = 193	Control n = 168	*p*-value	Effect Size	95% CI for eff. size
Absolute Extrema	6	49%	25%	$p < 0.001$	$d = 0.839$	(0.622,1.055)
Applied Optimization	9	37%	15%	$p < 0.001$	$d = 0.806$	(0.591,1.022)
Evaluating an Integral	6	69%	38%	$p < 0.001$	$d = 0.850$	(0.633,1.066)
Evaluating Limits	8	62%	44%	$p < 0.001$	$d = 0.627$	(0.414,0.839)
Implicit Differentiation	5	76%	69%	$p = 0.05646$	$d = 0.203$	(-0.005,0.411)
Interpreting a Graph	17	82%	73%	$p < 0.001$	$d = 0.512$	(0.301,0.722)
Related Rates	8	55%	38%	$p < 0.001$	$d = 0.438$	(0.228,0.648)
Sketching a Curve	18	80%	60%	$p < 0.001$	$d = 0.750$	(0.536,0.965)
Overall	**77**	**66%**	**48%**	$\mathbf{p < 0.001}$	$\mathbf{d = 0.925}$	**(0.707,1.143)**

Table S13: Individual item results for Fall 2019 treatment and control sections, with percent correct for treatment and control sections, as well as the associated *p*-value, effect size and confidence intervals.

3.4: Course success odds ratios

Logistic regression models were carried out to predict odds of success in the course (1: Pass; 0: DFW). Treatment (1:MPC; 0:non-MPC) was the only independent variable included in the model. The assumption that the conditional mean of the course success variable was binomial was considered to be robust given the random nature of the sample (*105*). Odds ratios were calculated using this model separately for three different groups of students: female, Hispanic-identified, and First-Time-in-College (FTiC) students. When examining each group, the overall models were significant when compared to the null models (Female-identified: $\chi 2\ (1) = 4.55$, $p < 0.05$; Hispanic-identified: $\chi 2\ (1) = 14.64$, $p < 0.001$; FTiC: $\chi 2\ (1) = 9.28$, $p < 0.01$;). This

indicates that our models fit the data better than intercept-only models. In addition, given the significance of the Treatment coefficient in each model, it is concluded that this variable is reliable in predicting course success.

Estimates with error measures and statistical significance are presented for each group in Tables S14 – S16. The standard errors indicate the variability associated with the estimates and the *z*-values are calculated by dividing the coefficient estimate by the standard error. The estimates given in the tables represent the average change in the log odds of the response variable (course success) related to the Treatment variable. For example, for the female-identified model, being in the MPC group is associated with an average increase of 0.4587 in the logs odds of successfully completing the course. In other words, being in the MPC group is associated with having a higher likelihood of passing the course for female-identified students.

The odds ratios between the MPC and non-MPC groups were obtained using the coefficient estimates for each model. For female-identified students, the odds ratio is $e^{0.4587} = 1.582$. This translates to the odds for a female-identified student in the MPC group passing the course being about 58% higher than the odds for a female-identified student in the non-MPC group. Similarly, the odds ratios for the Hispanic-identified and FTiC students are 2.021 and 1.845, respectively. So, the odds of a Hispanic-identified student passing the course is about 100% higher if they were enrolled in an MPC section and 85% higher for FTiC students in the MPC group.

	Estimate	Std. Error	*z*-value	*p*-value
(Intercept)	0.7673	0.1501	5.112	3.19e-07***
Group	0.4587	0.2156	2.128	0.0334*

Signif. codes: ***0.001 **0.01 *0.05

Table S14: Female-identified students

	Estimate	Std. Error	*z*-value	*p*-value
(Intercept)	0.7328	0.1223	5.992	2.07e-09***
Group	0.7038	0.1860	3.784	0.000154***

Signif. codes: ***0.001 **0.01 *0.05

Table S15: Hispanic-identified students

	Estimate	Std. Error	*z*-value	*p*-value
(Intercept)	0.9487	0.1325	7.159	8.13e-13 ***
Group	0.6126	0.2032	3.015	0.00257**

Signif. codes: ***0.001 **0.01 *0.05

Table S16: First-time-in-college students