Registered Report

Advances in Methods and Practices in Psychological Science
July-September 2021, Vol. 4, No. 3, pp. 1–24

DOI: 10.1177/25152459211027575
www.psychologicalscience.org/AMPPS

ManyClasses 1: Assessing the Generalizable Effect of Immediate Feedback Versus Delayed Feedback Across Many College Classes

Emily R. Fyfe[1]*, Joshua R. de Leeuw[2]*, Paulo F. Carvalho[3]*, Robert L. Goldstone[1]*, Janelle Sherman[1]*, David Admiraal[4], Laura K. Alford[5], Alison Bonner[6], Chad E. Brassil[7], Christopher A. Brooks[8], Tracey Carbonetto[9], Sau Hou Chang[10], Laura Cruz[11], Melina Czymoniewicz-Klippel[12], Frances Daniel[13], Michelle Driessen[14], Noel Habashy[15], Carrie L. Hanson-Bradley[16], Edward R. Hirt[1], Virginia Hojas Carbonell[17], Daniel K. Jackson[18], Shay Jones[19], Jennifer L. Keagy[20], Brandi Keith[21], Sarah J. Malmquist[22], Barry McQuarrie[23], Kelsey J. Metzger[24], Maung K. Min[25], Sameer Patil[26], Ryan S. Patrick[27,28], Etienne Pelaprat[29], Maureen L. Petrunich-Rutherford[13], Meghan R. Porter[30], Kristina Prescott[22], Cathrine Reck[30], Terri Renner[31], Eric Robbins[32], Adam R. Smith[33], Phil Stuczynski[32], Jaye Thompson[34], Nikolaos Tsotakos[35], Judith K. Turk[36], Kyle Unruh[29], Jennifer D. Webb[37], Stephanie N. Whitehead[38], Elaine C. Wisniewski[39], Ke Anne Zhang[1], and Benjamin A. Motz[1]*

*Lead authors

[1]Department of Psychological and Brain Sciences, Indiana University Bloomington, Bloomington, Indiana, USA; [2]Department of Cognitive Science, Vassar College, Poughkeepsie, New York, USA; [3]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; [4]Department of Civil and Environmental Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, USA; [5]Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, Michigan, USA; [6]Department of Mathematics, Penn State University Lehigh Valley, Center Valley, Pennsylvania, USA; [7]School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, Nebraska, USA; [8]School of Information, University of Michigan, Ann Arbor, Michigan, USA; [9]Department of Engineering, Penn State University Lehigh Valley, Center Valley, Pennsylvania, USA; [10]School of Education, Indiana University Southeast, New Albany, Indiana, USA; [11]Schreyer Institute for Teaching Excellence, Penn State University, University Park, Pennsylvania, USA; [12]Department of Biobehavioral Health, Penn State University, University Park, Pennsylvania, USA; [13]Department of Psychology, Indiana University Northwest, Gary, Indiana, USA; [14]Department of Chemistry, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; [15]College of Agricultural Sciences, Penn State University, University Park, Pennsylvania, USA; [16]Department of Child, Youth and Family Studies, University of Nebraska-Lincoln, Lincoln, Nebraska, USA; [17]Department of Spanish and Portuguese, Indiana University Bloomington, Bloomington, Indiana, USA; [18]Department of Physics, Penn State University Lehigh Valley, Center Valley, Pennsylvania, USA; [19]Department of Humanities/Communications, Penn State University Harrisburg, Middletown, Pennsylvania, USA; [20]Center for Teaching Excellence, Penn State University Harrisburg, Middletown, Pennsylvania, USA; [21]Department of Sociology, Indiana University Kokomo, Kokomo, Indiana, USA; [22]Department of Biology Teaching & Learning, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; [23]Science and Math Division, University of Minnesota Morris, Morris, Minnesota, USA; [24]Center for Learning Innovation, University of Minnesota Rochester, Rochester, Minnesota, USA; [25]Department of Business, Penn State University Lehigh Valley, Center Valley, Pennsylvania, USA; [26]Luddy School of Informatics,

Corresponding Author:
Emily R. Fyfe, Department of Psychological and Brain Sciences, Indiana University
E-mail: efyfe@indiana.edu

Computing, and Engineering, Indiana University Bloomington, Bloomington, Indiana, USA; [27]Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, USA; [28]Department of Teaching, Learning and Teacher Education, University of Nebraska-Lincoln, Lincoln, Nebraska, USA; [29]Unizin, Austin, Texas, USA; [30]Department of Chemistry, Indiana University Bloomington, Bloomington, Indiana, USA; [31]O'Neill School of Public and Environmental Affairs, Indiana University Bloomington, Bloomington, Indiana, USA; [32]Black School of Business, Penn State University Behrend, Erie, Pennsylvania, USA; [33]School of Business, Indiana University Kokomo, Kokomo, Indiana, USA; [34]Department of Design, Housing, and Apparel, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA; [35]School of Science, Engineering, and Technology, Penn State University Harrisburg, Middletown, Pennsylvania, USA; [36]Conservation and Survey Division, School of Natural Resources, University of Nebraska-Lincoln, Lincoln, Nebraska, USA; [37]Department of Art & Design, University of Minnesota Duluth, Duluth, Minnesota, USA; [38]Department of Criminal Justice, Indiana University East, Richmond, Indiana, USA; and [39]College of Engineering, University of Michigan, Ann Arbor, Michigan, USA

Abstract

Psychology researchers have long attempted to identify educational practices that improve student learning. However, experimental research on these practices is often conducted in laboratory contexts or in a single course, which threatens the external validity of the results. In this article, we establish an experimental paradigm for evaluating the benefits of recommended practices across a variety of authentic educational contexts—a model we call *ManyClasses*. The core feature is that researchers examine the same research question and measure the same experimental effect across many classes spanning a range of topics, institutions, teacher implementations, and student populations. We report the first ManyClasses study, in which we examined how the timing of feedback on class assignments, either immediate or delayed by a few days, affected subsequent performance on class assessments. Across 38 classes, the overall estimate for the effect of feedback timing was 0.002 (95% highest density interval = [–0.05, 0.05]), which indicates that there was no effect of immediate feedback compared with delayed feedback on student learning that generalizes across classes. Furthermore, there were no credibly nonzero effects for 40 preregistered moderators related to class-level and student-level characteristics. Yet our results provide hints that in certain kinds of classes, which were undersampled in the current study, there may be modest advantages for delayed feedback. More broadly, these findings provide insights regarding the feasibility of conducting within-class randomized experiments across a range of naturally occurring learning environments.

Keywords

reproducibility, experiment, education, evidence-based practices, feedback

Received 5/7/19; Revision accepted 6/6/21

A teacher designing a lesson will encounter dozens of decision points: How should the lesson be sequenced? What kinds of visual aids should be incorporated? When should students receive feedback? A central goal of psychological research on human learning and memory is to provide answers to these kinds of questions, thereby improving teaching practices and student outcomes. This pursuit within psychological science aims to translate findings from basic research into educational contexts and provide an evidentiary base to support teaching practices in accordance with the understanding of how people learn (Benassi et al., 2014; National Academies of Sciences, Engineering, and Medicine, 2018). However, instructional strategies that appear to be effective in laboratory settings do not necessarily translate smoothly into classroom practice. Indeed, teachers sometimes discount the validity and applicability of empirical findings to their own courses (Andrews & Lemons, 2015), which may be one reason that teachers engage in more lecture-based and less active-learning methods than evidence indicates is merited (Freeman et al., 2014; Knight & Wood, 2005).

To provide a more ecologically valid evidence base for effective teaching and learning strategies, some psychologists have used experiments in classrooms to examine the benefits of, for example, sequencing study materials (e.g., Carvalho et al., 2016), practicing memory retrieval (e.g., Anderson et al., 2013; Gurung & Burns, 2019), and self-regulating study (e.g., Wakeling & Robertson, 2017), but the references listed above found evidence that diverges somewhat from the canon of established laboratory results. What should one make of these conflicting results? Do they suggest that laboratory findings have limited relevance to classroom practice?

Or instead, do they simply reveal that instructional practices will not work all the time in all situations?

Unfortunately, these applied educational experiments are often conducted in very narrow contexts (e.g., in single classrooms on a single topic with researchers who vigilantly curate and monitor the study to ensure compliance); thus, it can be easy to dismiss conflicting results as idiosyncratic to the specific context. Yet these idiosyncratic educational contexts are precisely those that psychological research aims to improve (Koedinger et al., 2013; Motz et al., 2018). To provide legitimate estimates of the benefits of recommended practices in authentic educational environments, rigorous, experimental research that extends beyond the bounds of any single class is needed. To that end, the goal of the current study is to establish a paradigm for evaluating the generalizability of recommended educational practices across a variety of educational contexts—a model we call *ManyClasses*.

The ManyClasses Model

To investigate the generalizable effects of an educational practice, one needs to collect independent samples from many different contexts—in this case, across many classes. Rather than conducting an education experiment that is embedded in just one course (e.g., introductory psychology), a ManyClasses study examines the same research question across multiple courses spanning a range of topics, institutions, teacher implementations, and student populations. To be clear, the goal is not merely to increase the sheer number of classrooms in which the experiment is conducted. There are existing examples of such studies; for instance, Rohrer et al. (2020) investigated the sequencing of study materials in 57 middle school classrooms, and Booth et al. (2015) tested the role of worked examples with students in 28 different algebra classrooms. These studies, although rigorous and on a larger scale than is typical, are still limited to examining relatively homogeneous contexts (e.g., high school algebra courses) with topic-specific materials that are created by or in conjunction with the research team.

In contrast, a ManyClasses study will investigate educational practices across a variety of class contexts with the goal of maintaining the rigor of a randomized experiment while also allowing teachers the flexibility to prepare materials that are authentic to their institutional and disciplinary norms. By drawing the same experimental contrast across many different educational implementations and then analyzing pooled results, we can assess the degree to which an experimental effect might yield benefits generalizing across educational settings, student populations, subject areas, and course types. This approach is intended to target three related design issues that contribute to understanding generalizability: replication (i.e., test in many independent samples), variation (i.e., test across contexts that vary on numerous dimensions), and ecological validity (i.e., test with authentic teacher-created materials).

In practice, a ManyClasses experiment is embedded in courses in which researchers manipulate a theoretically motivated variable (e.g., immediate vs. delayed feedback, worked examples vs. problems to solve). Participating teachers create assignments that are normative for their discipline and present them to their students as part of their normal classroom routine. Using random assignment at the student or class level, students receive different versions of assignments. Finally, teachers report relevant learning outcomes (e.g., exam scores) corresponding to the different assignments, and researchers analyze anonymized pooled results.

The ManyClasses model responds to the current call for prioritizing replicability in psychology in a unique way. As LeBel and colleagues (2017) noted, replications vary in how similar or dissimilar they are to the original observation. On one end of the continuum are near exact replications, which are the same on all features under the experimenter's control. However, ManyClasses lies on the other end of the continuum because the conditions in ManyClasses will be far from identical across different classes and will reflect natural variation in instructor preferences and disciplinary norms. Thus, although there will be a critical feature related to the research question that is manipulated in all classes (e.g., feedback presented immediately vs. a delay), there will be many varying contextual factors (e.g., class size, discipline, number of test items, test performance). Furthermore, some of these varying contextual factors will be directly related to how teachers choose to implement the target manipulation (e.g., frequency and value of treatment assignments), which is the type of variability typically removed via experimental control. A ManyClasses study embraces this variability because if the aim is to provide practical recommendations to support student learning in real classes, researchers ought to examine the generalizability of effects across a diversity of possible implementations.

In that respect, ManyClasses resembles the "metastudies" approach (Baribault et al., 2017), in which researchers test an experimental effect across many minor variations of the experimental setup. Rather than holding outside factors constant, a metastudies approach strategically varies them across many "microexperiments" to examine the generalizability of an effect over and above these variations. In ManyClasses, each class represents a sort of microexperiment (with factors that are not systematically varied but that naturally vary across classes), which allows us to estimate the generalizable effect size of a manipulation beyond each individual

classroom implementation. This emphasis on multiple independent samples is shared with other "Many" efforts in psychology, including Many Labs (Klein et al., 2014, 2018), Many Babies (Frank et al., 2017), and Many Primates (Bohn et al., 2019). For example, the initial Many Labs study examined the replicability of 13 experiments across 36 independent laboratories with more than 6,000 participants (Klein et al., 2014). The key difference is ManyClasses's explicit focus on heterogeneous samples and highly dissimilar replications.

ManyClasses 1 on the Effects of Feedback

In this article, we report the first ManyClasses study, which focused on a specific recommended educational practice: the provision of feedback on class assignments. Feedback is a common practice incorporated in nearly every class that often has positive effects on learning (e.g., Fyfe & Brown, 2018; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). However, there is considerable controversy over the optimal timing of feedback and the conditions under which immediate feedback is beneficial. It has long been assumed that feedback should be provided as soon as possible after a student response to best modify performance (e.g., Skinner, 1954). Furthermore, in a meta-analysis, it was concluded that immediate feedback was more effective than delayed feedback in classroom settings (Kulik & Kulik, 1988). Not surprisingly, many recommendations to educators specify that feedback should be provided immediately to have the greatest impact (e.g., Benassi et al., 2014; Booth et al., 2017). For example, one of the practice guides published by the What Works Clearinghouse for college instructors recommends "providing immediate feedback" with automated student response systems (Dabbagh et al., 2019).

However, recent arguments and data suggest that the benefits of immediate feedback may be limited to specific outcomes (e.g., speed of acquisition) and that delayed feedback may be optimal for knowledge retention (e.g., Mullet et al., 2014). Researchers in a recent report claimed to outline three key findings from the feedback literature that are "robust, well-replicated, and critical to understanding how people learn," and one was that "delaying feedback produces better learning and retention" than immediate feedback (Butler & Woodward, 2018, p. 23). Among the argued benefits of delayed feedback is that it provides spaced study (students study the content when they complete the assignment and when they receive feedback after a delay). Given the opposing nature of these recommendations, research is urgently needed to investigate the timing of feedback and the generalizability of the effects of immediate feedback on student learning outcomes across a variety of authentic classroom contexts.

The Current Study

In the current study, we compared the effects of immediate feedback (i.e., feedback provided immediately after an assignment is submitted) with delayed feedback (i.e., feedback provided several days after an assignment is submitted) on online homework assignments. However, when implemented in real classes, there is often a confound between immediate and delayed feedback: Viewing immediate feedback is automatic (because it appears immediately after submitting an online assignment), but viewing delayed feedback is optional (because it requires reaccessing an assignment). To account for this difference, we also compared the effects of incentivizing students to view the feedback with not incentivizing students to view the feedback. More broadly, our goal was to develop a model for conducting randomized experiments in a wide range of naturally occurring classes spanning a range of course types, institutions, teacher implementations, and student populations.

Disclosures

Preregistration

A time-stamped, independent, read-only registration of this article and protocol is available at https://osf.io/q84t7/, under the heading "Registrations" (dated July 22, 2019).

Data, materials, and online resources

The study materials, deidentified data, and analysis scripts are available at https://osf.io/q84t7/.

Reporting

Below we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical approval

The multisite experimental procedures, materials, and recruitment protocol were approved by the Indiana University Institutional Review Board. Furthermore, each participating institution provided a letter from a signatory official granting a Family Educational Rights and Privacy Act (FERPA) exception so that a researcher could access instructors' Canvas course sites and student enrollment data. All participating students provided informed consent electronically, and the study was

carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

Method

Participants

Researchers posted an online call for applications to interested instructors in Spring 2019. In addition, to facilitate institution-level approval, the researchers proactively sought applications and approval from institutional members of the Unizin Consortium. This included outreach on social media and university listservs, in-person visits to interested institutions, a presentation at the Unizin Summit in April 2019, and a series of informational webinars. These efforts occurred from February 2019 to July 2019. The call for applications described the goals and procedures of ManyClasses1 included a list of the requirements to apply, and contained a link to an online application. To be comparable with the scale of the inceptive Many Labs project (Klein et al., 2014), we aimed to recruit about 36 classes. College instructors were eligible to apply if they were teaching a for-credit undergraduate course in Fall 2019 that (a) used Canvas (Instructure, Salt Lake City, UT) as the online learning management system, (b) included at least two Canvas quiz assignments that were automatically scored, (c) included a measure of student learning that was administered after the quizzes and that assessed the content from each quiz using different items, and (d) had a projected enrollment of at least 20 students. We selected classes that were already using Canvas to lower the demands on instructors (e.g., they did not have to learn a new system) and to manipulate the timing of feedback using standard features that were available to them.

A total of 46 instructors applied to have their classes participate. Two instructors were removed for not meeting the stated criteria, five instructors did not reply to our follow-up e-mails, and one withdrew shortly after applying. We selected all the remaining 38 classes from 15 different institutions (15 campuses within five university systems), which resulted in a total enrollment of 2,917 students (76.8 per class on average). None of the selected classes withdrew from the study, leaving a final sample of 38 classes. See Table 1 for a brief overview of each class. Of the 2,917 enrolled students, 2,331 (79.9%) consented to release their course data to the research team by electronically signing a consent form as well as a FERPA waiver. Consenting students were included in the analysis if they completed at least one assignment with immediate feedback and at least one assignment with delayed feedback (thus receiving exposure to both treatments) and completed at least one posttest assessment in each condition. Of the consenting students, 250 (10.7%) did not complete at least one assignment in each condition and/or did not complete at least one posttest assessment in each condition. Of the included participants, 1,496 were in classes assigned to incentivized feedback, and 585 were in classes assigned to nonincentivized feedback. As shown in Table 1, participants were enrolled in classes across five university systems and across 15 different disciplines with varying class sizes.

Design

The study was a posttest-only randomized experiment with a 2 × 2 design that included a Feedback Timing (Immediate vs. Delayed) × Incentivized Feedback (Incentivized vs. Nonincentivized) interaction. Within each class, all enrolled students were randomly sorted into two groups on Canvas. Then, groups of students were randomly assigned to different treatment orders (assignments with immediate feedback first or assignments with delayed feedback first). Thus, feedback timing was manipulated as a within-subjects factor. This within-subjects design was selected to enhance power to detect effects and to maintain ethics when conducting research in classes (e.g., ensuring students were exposed to same treatments). In addition, classes were randomly assigned to incentivize or not incentivize students to look at the feedback (e.g., earn points on a follow-up assignment if they looked at the feedback and reported on it). Thus, incentivized feedback was manipulated as a between-subjects factor at the class level. This design ensured that within a class, all students were exposed to the same assignment variations but staggered in time (e.g., some students received immediate feedback on the first assignment and delayed feedback on the second assignment; some students received the reverse). The dependent variable was students' scores on an assessment (e.g., items from a course exam) that assessed the content knowledge from each assignment.

Procedure

Throughout the Fall 2019 semester, all students enrolled in participating classes completed their courses as they normally would. Courses varied in content, format, style, and so on according to instructor preference and disciplinary norms; however, each course included quiz assignments with feedback administered via Canvas. The course had to include a minimum of two treatment quiz assignments that were approximately matched in length and difficulty to ensure that each student was assigned at least one quiz with immediate feedback and one quiz with delayed feedback. The actual number of treatment quizzes in a class ranged from two to 18 (i.e., one to nine per feedback condition). At the beginning of the

Table 1. Brief Overview of Each Participating Class

Class ID	University	Discipline	Format	Total enrollment	Sample size
1	UMN	History	In person	81	57
2	PSU	Biology	Online	39	26
3	UMN	Biology	In person	227	187
4	UMN	Biology	In person	28	24
5	PSU	Chemistry	In person	27	19
6	IU	Chemistry	In person	448	373
7	PSU	Communication	In person	19	5
8	UMich	Chemistry	In person	24	20
9	UMN	Chemistry	Online	115	87
10	UN	Engineering	In person	87	66
11	UN	Computer science	In person	80	57
12	UN	Family studies	In person	130	83
13	PSU	Engineering	In person	27	24
14	UMich	Computer science	In person	631	451
15	PSU	Business	In person	21	12
16	PSU	Business	Hybrid	24	13
17	IU	Computer science	In person	38	19
18	PSU	Business	In person	24	21
19	PSU	Business	In person	45	24
20	UN	Biology	In person	108	85
21	UMN	Mathematics	In person	23	19
22	PSU	Mathematics	In person	51	8
23	IU	Chemistry	In person	111	89
24	IU	Criminal justice	Online	34	27
25	IU	Psychology	Hybrid	26	14
26	IU	Psychology	In person	55	36
27	IU	Psychology	In person	28	20
28	IU	Psychology	Online	28	17
29	IU	Psychology	In person	28	14
30	PSU	Physics	In person	16	15
31	PSU	Psychology	In person	32	13
32	UMN	Business	In person	44	32
33	IU	Sociology	Online	29	8
34	IU	Language	In person	22	17
35	UMich	Computer science	In person	43	25
36	UN	Ecology	Hybrid	33	9
37	IU	Business	In person	54	38
38	IU	Business	Online	37	27

Note: Enrollment is the total number of students in the participating class's Canvas site at the time when data collection commenced. Sample size is the number of students who provided consent and completed treatment assignment and posttests in each condition. IU = Indiana University; PSU = Penn State University; UMich = University of Michigan; UMN = University of Minnesota; UN = University of Nebraska.

semester, instructors announced the opportunity to participate in the study and made it clear to their students that participation in the research study would not change their experiences in the course but whether their course data would be provided to the researchers. All students were then assigned a survey in Canvas that presented the informed consent statement and the FERPA waiver. Students' responses indicated consent or not. These responses did not affect whether they got credit for this survey, and responses to these statements were encrypted so that instructors would not know which students opted to participate, which protected their privacy regarding their decision to include their data in this study from their instructors. The survey also had unlimited attempts so that students could change their consent status at any point during the semester.

The treatment quizzes that students completed in Canvas could be automatically graded so that the feedback could be provided immediately in the Canvas course site. On each quiz assignment, students either received

feedback immediately after submitting the assignment or after a several-day delay (the exact number of days was selected by the instructor, range = 1–5 days). If feedback was delayed, a Canvas message notified students when the feedback was available to view. The default feedback in Canvas presented the quiz items, the student's responses, and correct/incorrect indications. Instructors could choose, via the options available in Canvas, to include additional information in the feedback message (e.g., correct answers, instructional explanations) or to present more limited information in the feedback message (e.g., grade only). Thus, all enrolled students accessed and completed the quizzes as they normally would and received the same type of feedback. However, the timing of feedback varied from one quiz to the next.

Some classes were assigned to the incentivized feedback condition. Students in these classes were assigned "follow-up" assignments after each quiz they completed. These follow-up assignments required students to view the feedback to their original quiz responses and report specific information from the feedback (e.g., number of items they missed, the correct answers to the missed items). The specific content and how many points were associated with completing these follow-up assignments were determined by the instructor, as with any course requirement.

Finally, all enrolled students completed some form of posttest that assessed their learning from the different quiz assignments. Frequently, this posttest was a set of items on an exam that occurred after the quiz, but the specific type and number of questions on the posttest varied by class and depended on instructor preference and disciplinary norms. Note that each student had two posttest scores; one comprised items that corresponded to their learning from quiz assignments with immediate feedback, and one comprised items that corresponded to their learning from quiz assignments with delayed feedback.

Data collection

Data were collected from the Canvas course site of each participating class. The primary dependent variable was student performance on posttests that assessed learning from the treatment assignments. Instructors either indicated which of the existing gradebook items measured relevant learning performance following each treatment period or uploaded custom outcome measures that were, most frequently, scores on subsets of exam items (e.g., report scores on the first 10 exam items as measuring learning from one treatment quiz, report scores on the next 10 exam items as measuring learning from a different treatment quiz, and not report scores on the last 10 items because they were not relevant to the material on any treatment quiz).

We collected additional information on a variety of potential moderators related to student activity (e.g., whether they accessed assignments, scores on exams), course content (e.g., number of assignments, days between assignment due date and delayed feedback), and general course information (e.g., discipline, class size). We were primarily interested in characterizing aspects of the assignments on which feedback was provided as well as aspects of the exams that assessed learning from those assignments. Most of the student-level moderator values were measured from activity logs recorded within the Canvas learning management system and made available through a framework developed by the Unizin Consortium. One participating university did not permit access to these activity logs at the time of data collection, so most of the student-level moderator values are missing for 11 classes (180 participating students).

We also obtained publicly available information about the institutions from College Scorecard and National Center for Education Statistics College Navigator. See Table 2 for a list of the moderators. A detailed description of data sources and preprocessing for all measures is available at https://osf.io/p2csf/. In our primary analyses, the institution-level variables were considered class-level variables; there were not enough classes per institution to estimate separate institution-level effects.

Results

Descriptive statistics

Treatment characteristics. Averaging the class averages across all 38 participating classes, we found that the mean number of items on quiz assignments was 10.15 (*SD* = 7.00) and that the mean score was 80.76% (*SD* = 9.45%). The average percentage of items that were retrieval-based (e.g., fill-in-the-blank, numerical response) rather than recognition-based (e.g., multiple-choice) was 6.76% (*SD* = 23.60%, range = 0%–100%), and 42.1% of assignments included explanation feedback beyond the correct answer. Across all classes, the average number of days between immediate feedback and delayed feedback was 2.58 (*SD* = 1.13, range = 1–5).

For those classes assigned to incentivized feedback (i.e., administered a follow-up assignment about quiz feedback to incentivize students to view the feedback), an average of 85.1% of students completed follow-up assignments when feedback was immediate (*SD* = 8.9%, range = 68%–96%), and 73.6% of students completed follow-up assignments when feedback was delayed (*SD* = 16.1%, range = 35%–100%). According to students' Canvas activity in the 27 classes whose institutions provided access to the activity logs, most of the feedback on treatment quizzes was viewed, and this was somewhat higher in the incentivized classes. The mean percentage of

Table 2. List of Moderator Variables

Level	Moderator	Scale	Data source
Student	Accessed all treatment assignments (yes/no)	Categorical	Canvas
Student	Accessed all delayed feedback (yes/no)	Categorical	Canvas
Student	Accessed all immediate feedback (yes/no)	Categorical	Canvas
Student	Cumulative time spent on treatment assignments	Continuous	Canvas
Student	Cumulative time spent viewing feedback	Continuous	Canvas
Student	Average time spent on treatment assignments*	Continuous	Canvas
Student	Average time spent viewing feedback*	Continuous	Canvas
Student	Number of treatment assignments with feedback view*	Continuous	Canvas
Student	Cumulative Canvas grade (percentage correct)	Continuous	Canvas
Student	Average number of days before due date that treatment assignments were submitted	Continuous	Canvas
Student	Average number of days after treatment assignments were submitted that delayed feedback was received	Continuous	Canvas
Class	Discipline (STEM vs. non-STEM)	Categorical	Canvas
Class	Format (in-class, online, hybrid)	Categorical	Canvas
Class	Class size	Continuous	Canvas
Class	Proportion of class that is lecture-based	Continuous	Canvas
Class	Class level (introductory, immediate, advanced)	Categorical	Canvas
Class	Number of exams in class	Continuous	Canvas
Class	Number of treatment assignments	Continuous	Canvas
Class	Cumulative number of questions in assignments	Continuous	Canvas
Class	Assignment question presentation (one at a time, all at once)	Categorical	Canvas
Class	Proportion of retrieval-based items (e.g., numerical response, fill-in-the-blank) in assignments	Continuous	Canvas
Class	Assignment difficulty (percentage correct; averaged across assignments)	Continuous	Canvas
Class	Number of days between assignment due date and provision of delayed feedback	Continuous	Canvas
Class	Type of feedback content on assignments (verification only, correct answer, explanation)	Categorical	Canvas
Class	Assignment value (percentage of class points)	Continuous	Canvas
Class	Follow-up assignment value (percentage of class points)	Continuous	Canvas
Class	Time constraint on assignments (yes/no)	Categorical	Canvas
Class	Number of days between due date of assignments and exam (average by treatment)	Continuous	Canvas
Class	Number of exam questions that correspond to assignments (average by treatment)	Continuous	Canvas
Class	Proportion of retrieval-based item (e.g., numerical response, fill-in-the-blank) in exams	Continuous	Canvas
Class	Exam difficulty (percentage correct; averaged across exams)	Continuous	Canvas
Class	Exam type (in class vs. take-home)	Categorical	Canvas
Class	Exam value (percentage of class points)	Continuous	Canvas
Class	Exam question mapping to assignment (exact same as assignment questions, not exact same)	Categorical	Canvas
Class	Consent rate*	Continuous	Canvas
Class	Quizzes combined in outcome scores*	Categorical	Canvas
Institution	Admission rate	Continuous	College Navigator
Institution	Percent part-time faculty	Continuous	College Navigator
Institution	Annual cost of attendance	Continuous	College Scorecard
Institution	Graduation rate	Continuous	College Scorecard
Institution	Percentage White	Continuous	College Scorecard
Institution	Percentage of students receiving federal loans	Continuous	College Scorecard
Institution	Percentage of students returning after first year	Continuous	College Scorecard
Institution	Percentage of full-time students at institution	Continuous	College Scorecard
Institution	Percentage of students receiving income-based Pell grants	Continuous	College Scorecard

Note: For more details about the measurement of each moderator value, see https://osf.io/p2csf/. Moderator variables marked with an asterisk (*) were not preregistered and are exploratory. For a justification of our inclusion of these variables, see https://osf.io/q97wa/. STEM = science, technology, engineering, and math.

Box 1. In Detail: Base Model

The base model assumes that change in z (Δz) values are normally distributed within each class, c, and estimates the mean, μ_c, and standard deviation, σ_c, for each class's distribution. (The subscript c_s refers to the classroom that student s is in.)

$$\Delta z_s \sim N(\mu_{c_s}, \sigma_{c_s})$$

The means of the class-level distributions, μ_c, are also assumed to be normally distributed within each incentive condition, i, and the model estimates the mean, γ_i, and standard deviation, τ_i, of these condition-level distributions for both incentivized and nonincentivized classes.

$$\mu_c \sim N(\gamma_{i_c}, \tau_{i_c})$$

The standard deviations of the class-level distributions, σ_c, are γ-distributed across all classes, and the model estimates the mode and standard deviation of this γ distribution.

$$\sigma_c \sim G(mode = \alpha, SD = \varphi)$$

Priors for the model are weakly informative, according to the expected scale of the data.

$$\gamma_i \sim N(0,1)$$

$$\tau_i \sim G(mode = 1, SD = 2)$$

$$\alpha \sim G(mode = 0.5, SD = 1)$$

$$\varphi \sim G(mode = 1, SD = 2)$$

For accessible and thorough explanations of the analysis methods used in this research, see Kruschke (2014) and Kruschke and Liddell (2018a, 2018b).

feedback viewed for the incentivized classes was 82.7% (*SD* = 8.6%, range = 67%–96%; 13 classes) and was 76.5% (*SD* = 12.5, range = 55%–96%; 14 classes) for the nonincentivized classes. Further exploratory analysis related to feedback viewing is available at https://osf.io/t73rp/.

Assessment characteristics. The average class had 25.39 (*SD* = 23.35) assessment questions that were relevant to the quiz feedback and were therefore included in our measure of posttest performance. The mean score on posttest assessments was 79.29% (*SD* = 9.28%).

Condition differences on performance

Quantifying the effect of immediate feedback and delayed feedback. Because instructors controlled all aspects of the posttest assessments (e.g., number of items, how they were scored, item difficulty), we standardized students' performance within each posttest assessment using z scores. We used only assessment items that were relevant to the feedback the student had previously received. If instructors reported multiple outcome scores within each treatment period, we calculated the average z score for each student separately for outcomes following immediate and delayed feedback. We then calculated the difference between a student's average z score on posttest assessments for which the student had received immediate feedback and posttest assessments for which the student had received delayed feedback. We refer to this measurement as *change in* z (Δz). A positive Δz indicates that the student tended to perform better, relative to peers on the same posttest assessment, after receiving immediate feedback on prior quiz assignments. A negative Δz indicates that the student tended to perform better, relative to peers on the same posttest assessment, after receiving delayed feedback on prior quiz assignments.

Box 2. In Detail: Markov Chain Monte Carlo Sampling

We used JAGS (Version 4.3.0; Plummer, 2003) and the R package *runjags* (Denwood, 2016) for Markov chain Monte Carlo (MCMC) sampling. The JAGS specifications of the base model and the four moderator models are available at https://osf.io/q84t7/.

We assessed model convergence visually and through the potential scale reduction factor, commonly known as the $\widehat{R}$ statistic (Gelman & Rubin, 1992). $\widehat{R}$ was less than 1.005 for each parameter (values of 1.00 are ideal). The effective sample size for each parameter, which estimates the number of independent samples of the model posterior accounting for autocorrelation of the sampler, was at least 10,000. To meet these goals, we preregistered a plan for model fitting, which is available at https://osf.io/m38c2/. We fit the models using 48 chains, 5,000 steps of burn-in, and thinning the chain by four steps for every one step kept. For 41 of the 46 models, we reached our target effective sample size and $\widehat{R}$ goals after an initial sample of 3,000 steps per chain. Four models required 9,000 steps per chain, and one model required 81,000 steps per chain.

Effect of immediate feedback compared with delayed feedback. We used a hierarchical Bayesian model (see Boxes 1 and 2) to estimate the effect of immediate feedback compared with delayed feedback within each class and to estimate the effect of immediate feedback compared with delayed feedback across all classes within each incentive condition.

Figure 1 shows the model's estimates of the average Δz score for each individual class as well as the two condition-level estimates (i.e., an estimate of the mean of the classes within the incentivized condition and an estimate of the mean of the classes in the nonincentivized condition). The overall estimate for the average Δz across classes was 0.002 (95% highest density interval [HDI] = [−0.05, 0.05]), which indicates that there was no overall effect of feedback timing across classes. In classes with incentive to view feedback, the estimated average Δz was 0.00 (95% HDI = [−0.06, 0.06]). In classes with no incentive to view feedback, the estimated average Δz was 0.00 (95% HDI = [−0.08, 0.08]). The estimated difference in average Δz for incentivized classes relative to nonincentivized classes was 0.00 (95% HDI = [−0.10, 0.10]). In sum, there is no overall effect of feedback timing, and this does not depend on incentive condition.

Heterogeneity analysis. To describe the heterogeneity of the effect of immediate feedback compared with delayed feedback across classes, we relied on the visual display of the data as well as estimated measures of the distribution, which have advantages over conventional heterogeneity statistics (Borenstein et al., 2017; Rücker et al., 2008). The model estimates the heterogeneity of the effect of immediate feedback compared with delayed feedback across classes in each condition (condition-level variance, τ_i^2). These two parameters, one for each incentive condition, describe the variance in average Δz scores between classes. In the incentivized feedback condition, the standard deviation between classes' average Δz scores was 0.06 (95% HDI = [0.01, 0.14]). In the nonincentivized feedback condition, the standard deviation between classes' average Δz scores was 0.06 (95% HDI = [0.0004, 0.16]). Thus, according to the model's estimates, there was not large heterogeneity in the effect of feedback timing across classes. In Figure 2, we visualize the model's estimate of the distribution of classes in each condition, which can be used to infer the expected effect of immediate feedback compared with delayed feedback, and the uncertainty of the effect in new classes.

Moderator analyses. To explore the degree to which the effect of immediate feedback compared with delayed feedback depended on characteristics of the class or student, we estimated the relation between each moderator and Δz scores using a series of hierarchical Bayesian models (for a list of the moderators, see Table 2). For each moderator, one of four different models was selected depending on whether the moderator was measured on a metric or nominal scale and whether the moderator was at the class level or student level. All four models shared the same hierarchical structure as the base model described above but with additional parameters to account for potential effects of the moderator on either student-level means or class-level means (see Box 3).

Figure 3 shows the estimated coefficients for all class- and student-level moderators. Because we modeled only one moderator at a time (and thus ignore any possible interactions), this analysis is primarily intended to generate candidate moderators and not to definitively compare the relative strength of moderators. Furthermore, we emphasize that the moderators are observed and not manipulated, so the usual caveats about correlations apply. We found that the estimated 95% HDI contained zero for all moderators. In other words, there were no moderators that demonstrated a consistent effect on Δz scores.

There are several possible explanations of these results, which we cover in the Discussion section, but one class of moderators that is worth a closer look is

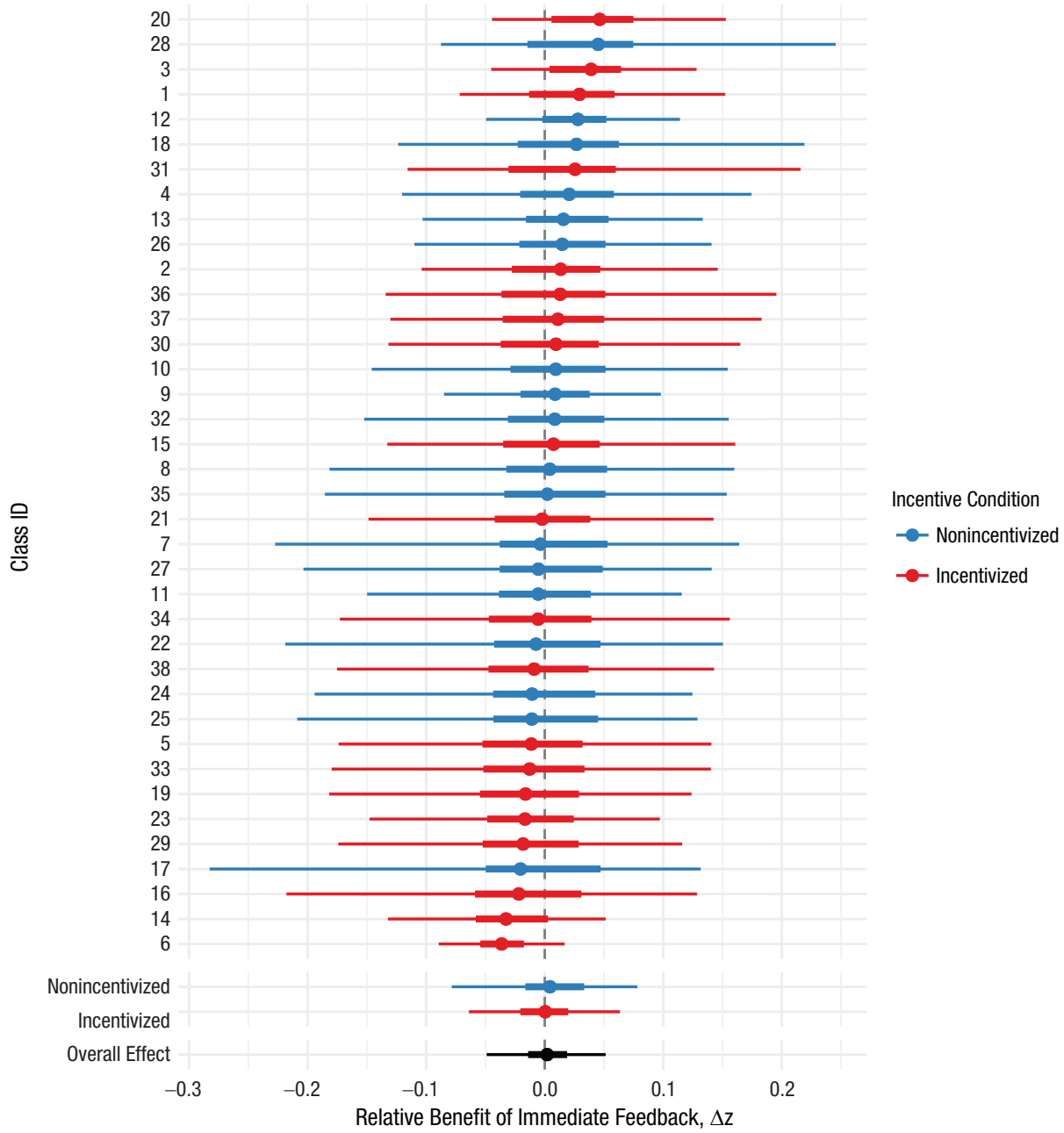

Fig. 1. Class- and condition-level estimates of the relative benefit of immediate feedback. The width of each bar represents the 95% highest density interval of the posterior estimate for a class. The thicker bar represents the 50% highest density interval, and the point represents the median.

moderators related to dosage of the treatment quizzes. A unique piece of the ManyClasses paradigm is that teachers can choose how to implement the target manipulation. For example, the teachers in this study decided on (a) the number of treatment quizzes, (b) the number of questions per quiz, and (c) the length of time between immediate and delayed feedback. These decisions have the effect of picking out a portion or region of the possible space of experimental designs. The benefit of this approach is that the experiments we ran in each class represented the teachers' authentic choices for how to use these quizzes in practice and thus arguably represent a more realistic estimate of the effects in practice, whereas the drawback is that there are portions of the design space that, had they been better covered, may have produced a more powerful experimental test.

In the following paragraphs, we highlight these three moderators that are directly related to the dosage of the manipulation. The goal is not to make definitive yes/no claims about whether the moderators matter or whether

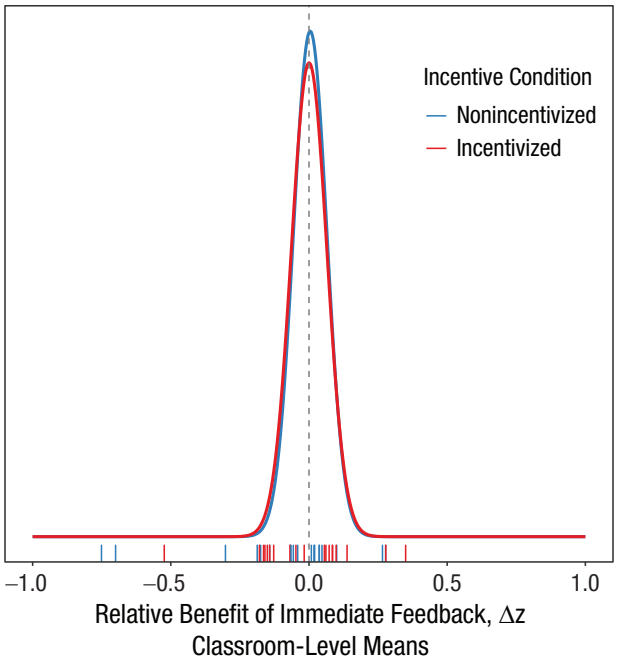

Fig. 2. Estimated distribution of class-level means in both incentive conditions. The observed mean of each class is indicated by the tick marks on the horizontal axis. Some of the observed means fall well outside the model's estimated distribution of means because the model estimates that the true mean of the class is much closer to zero.

there was or was not a statistically significant effect (Wasserstein et al., 2019). Statistically, there were no moderators that demonstrated a consistent effect on Δz scores. Rather, the goal is to showcase the findings of preregistered contrasts and how our observations were not evenly distributed across moderator values, which resulted in differential coverage of the design space because of the teacher's authentic choices.

First, the number of treatment quizzes varied from just two (one per feedback condition) to 18 (nine per feedback condition). Out of the 38 classes, 17 had either one or two quizzes per feedback condition. Figure 4 shows the pattern of class-level results across different levels of the moderator. The majority of classes are in the low-dosage region, and the estimates in these classes all hover tightly around zero. But there is an indication that uncertainty is high in the higher dosage region of the design space (e.g., when more quizzes were administered). Again, there are no credibly nonzero effects of this moderator, and it does not interact with incentive condition, but the trends in the data suggest that the higher dosage region is one in which feedback timing may have practical effects.

Second, along similar lines, the cumulative number of quiz questions across the full semester ranged from

Box 3. In Detail: Moderator Models

Each of the four moderator models includes all the parameters of the base model (see Box 1). Here we detail the additional parameters added to the base model to capture the relationship between the moderating variable and student-level or class-level means.

For student-level moderators, the model included an additional parameter, β_c, in the estimate of change in z_s. The level of the moderator for student s is x_s. When the moderator is continuous, a single value of β_c is estimated for each class, and the model is a linear regression. When the moderator is categorical, separate values of β_c are estimated for each distinct level of x in each class. The β_c values are assumed to be normally distributed in each incentive condition i, with mean θ_i and standard deviation ω_i. These values provide an estimate of the overall moderator effect and its consistency across classes.

$$\Delta z_s \sim N(\mu_{c_s} + \beta_{c_s} x_s, \sigma_{c_s})$$

$$\beta_c \sim N(\theta_{i_c}, \omega_{i_c})$$

Weakly informative priors were placed on the condition-level mean and standard deviation.

$$\theta_i \sim N(0,1)$$

$$\omega_i \sim G(\text{mode} = 0.5, SD = 1)$$

For class-level moderators, the β_i parameter is introduced in the estimate of the class-level means, μ_c, with i representing the two incentive conditions. The same principles of the student-level moderator apply: x_c is the level of the moderator for the class, and one or more β values are estimated depending on whether the moderator is continuous or categorical. In this model, there is no hierarchical structure to the estimate of β_i, and so a prior is placed directly on β_i.

$$\mu_c \sim N(\gamma_{i_c} + \beta_{i_c} x_c, \tau_{i_c})$$

$$\beta_i \sim N(0,1)$$

Fig. 3. Estimated coefficients for moderators on Δz scores. Positive coefficients mean that an increase in the value of the moderator (continuous moderators) or a shift to this level of the moderator relative to other levels of the moderator (discrete moderators) is correlated with an increase in the relative benefit of immediate feedback. Lines span the 95% highest density interval, and the dot represents the median. Moderators with [E] designation were not preregistered.

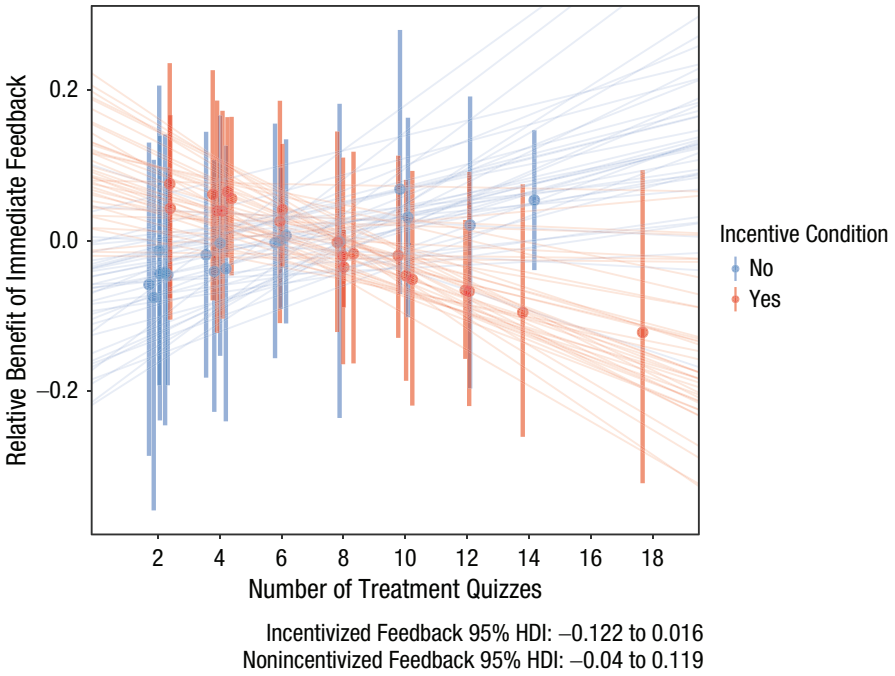

Fig. 4. Relationship between number of treatment quizzes and the effect of feedback timing. The model's median estimate of the mean Δz score for each class is shown as a circle. The vertical lines show the 95% highest density interval estimate for each class. Sample regression lines from the model's posterior distribution are shown in the background. These lines represent plausible fits. We show a sample of these lines to visualize the model's uncertainty. A small amount of horizontal jitter has been added to the points to improve the clarity of the visualization.

just eight (four questions per condition) to 198 (99 questions per condition). This represents substantial variation in how much potential feedback students were exposed to throughout the class. Figure 5 shows the pattern at the class level as the cumulative number of quiz questions varies. Naturally, this moderator is correlated with the number of treatment quizzes (Spearman's $\rho = .79$), and the pattern of results is similar.

Third, teachers also varied the length of delay between immediate and delayed feedback, and it ranged from 1 to 5 days. The modal choice was 3 days, but most teachers (32 of 38) opted for no more than 3 days of delay. Figure 6 shows the pattern of class-level results across different delay periods. Visually, the preponderance of negative slopes suggests that as the delay increases, the relative benefit of delayed feedback increases. However, as with the other moderators, there is sufficient uncertainty in the estimates that we cannot make strong claims here, and there are indications that we have undersampled the region of the design space in which the manipulation might have had a stronger influence on student performance. In this case, we also ended up, by chance, with no classes in the incentivized feedback condition with a delay longer than 3 days. All six classes with 4- or 5-day delays were in the nonincentivized feedback condition.

In addition to decisions about dosage, teachers also controlled features of the posttests that were used to measure student learning from the feedback-timing manipulation. One salient aspect of the posttest exams was the kinds of assessment questions that teachers chose to use. We categorized the questions as retrieval-based or not and calculated the proportion of retrieval-based questions on the exams. Unlike the moderators directly related to dosage of the treatment quizzes, here the natural choices of teachers were nearly optimal for a contrast between low and high use of retrieval-based questions. Most teachers used either all retrieval-based questions or all non-retrieval-based questions on the posttest exams. Perhaps because of this, the estimates for the effect of this moderator were the closest to reaching our decision threshold for statistical credibility (and, in fact, do *barely* cross this threshold if we compute a posterior estimate for an overall moderator coefficient averaging across the coefficient estimates for both incentive conditions; 95% HDI = [−0.113, −0.002]). Figure 7 shows the pattern of class-level results across the different levels of the moderator, and the negative slopes suggest potential benefits of delayed feedback when learning is assessed with more retrieval-based items.

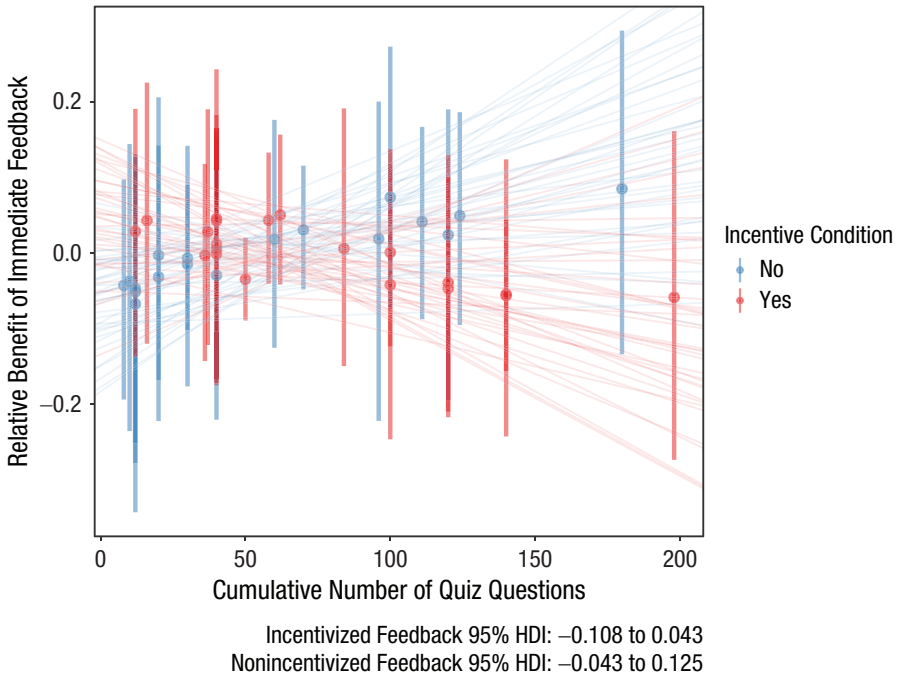

Fig. 5. Relationship between number of cumulative number of quiz questions and the effect of feedback timing. The model's median estimate of the mean Δz score for each class is shown as a circle. The vertical lines show the 95% highest density interval estimate for each class. Sample regression lines from the model's posterior distribution are shown in the background. These lines represent plausible fits. We show a sample of these lines to visualize the model's uncertainty.

Discussion

The effect of feedback timing on student learning

This first iteration of ManyClasses compared the effects of immediate feedback with delayed feedback on Canvas quizzes throughout the course of the Fall 2019 semester. It included data from 38 classes with a total of 2,081 participating students. The results indicate the global effect of feedback timing on learning activities is close to zero. We infer that under routine implementations such as those measured in the current study, there is no broadly generalizable difference in learning performance on educationally relevant outcomes when students receive immediate feedback on their learning activities compared with delayed feedback.

Our observation of no main effect of feedback timing on student performance in more than three dozen classes provides a prominent benchmark in research on feedback in educational settings. Many recommendations for the benefits of immediate feedback stem from the meta-analysis by Kulik and Kulik (1988), which reported small to moderate advantages for immediate feedback in 10 out of 11 studies conducted in classroom settings that often lacked experimental designs. Recent views point to the potential benefits of delayed feedback (e.g., Butler & Woodward, 2018; Mullaney et al., 2014), and two classroom experiments provide empirical support for delaying feedback on classroom quizzes (Mullet et al., 2014). Our results suggest that these past findings from a small number of classrooms may have limited external validity because we see no indication of a single global effect of feedback timing that generalizes across classrooms. Note that it is not the case that we observe high uncertainty in this estimate; rather, our model estimated a main effect of feedback timing that was tight around zero.

The next question is whether the effect of feedback timing changed systematically with different kinds of classes, students, or implementations used by the teachers in this study. Preregistered analyses of 40 different candidate moderators found no strong evidence of systematic differences in the effects of feedback timing between students or classes. The few classes in which the effect of feedback timing appeared to deviate from zero (shown in Figure 2) had small numbers of students and thus did not exert strong influence on these estimates. We also examined whether the effect of feedback timing was influenced by incentives for students to view the feedback and found no overall interaction between feedback timing and these incentives on learning performance.

We do, however, observe suggestive evidence of small moderator influences in the current study, but uncertainty in our estimates of moderator effects prevents us

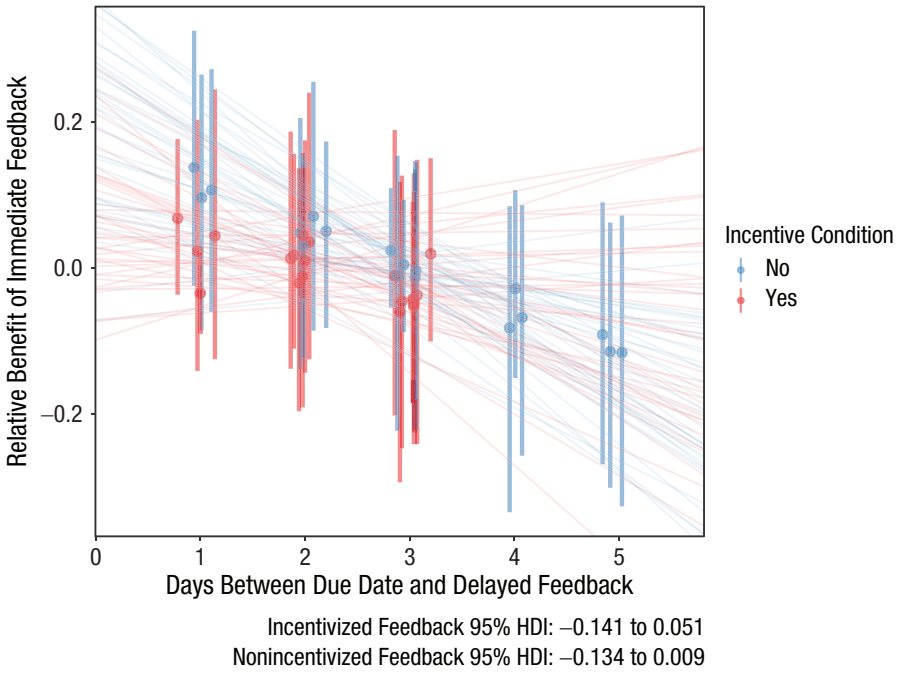

Fig. 6. Relationship between length of the delay for delayed feedback and the effect of feedback timing. The model's median estimate of the mean Δz score for each class is shown as a circle. The vertical lines show the 95% highest density interval estimate for each class. Sample regression lines from the model's posterior distribution are shown in the background. These lines represent plausible fits. We show a sample of these lines to visualize the model's uncertainty. A small amount of horizontal jitter has been added to the points to improve the clarity of the visualization.

from drawing clear inferences about them. Specifically, there was a trend for students to perform better following delayed feedback in classes in which posttest exam items were retrieval-based (e.g., fill-in-the-blank rather than multiple-choice). In addition, primarily in classes in which viewing feedback was incentivized, measures related to the "dosage" of treatment at the class level (e.g., number of quizzes, cumulative number of questions, length of feedback delay) all suggest, from their consistent directional trends, that increasing the amount of feedback and the length of feedback delay may improve performance following delayed feedback relative to immediate feedback. Such trends are consistent with Mullet et al.'s (2014) observation of advantages for delayed feedback under a classroom protocol that involved a retrieval step during outcome testing, incentives for looking at feedback, a 7-day delay for the release of feedback, and a large amount of feedback (18 practice quizzes with more than 200 questions total).

These statements suggesting that increasing delayed feedback dosage may be associated with possible benefits of delayed feedback for retrieval tasks are highly speculative and must be clearly caveated. They are based on trends that are consistent with a particular theoretical interpretation but that did not achieve our threshold for making credible inferences. We mention these trends primarily because the amount of feedback that teachers administered for the current study was modest. Only one class in the current study had levels of exposure to delayed feedback that were comparable with Mullet et al.'s (2014) study, and none of the classes in which viewing feedback was incentivized had delays greater than 3 days. Given the consistent trends across several moderators and the current study's sparse coverage of classes with high exposure to delayed feedback, we feel that it merits speculation that benefits for delayed feedback may yet exist in these undersampled circumstances. However, our evidence is convincing that inferences drawn from such circumstances do not generalize to improvements under more routine settings in which the current ManyClasses study was conducted.

The benefits and challenges of the ManyClasses methodology

Authenticity to routine educational practice is both the current study's primary advantage and a disadvantage for our ability to estimate moderating effects. Our experiment was distributed across 38 college classes and required minimal qualifying criteria—the class needed to include at least two automatically graded online quizzes in Canvas. This ease of recruitment provided beneficial

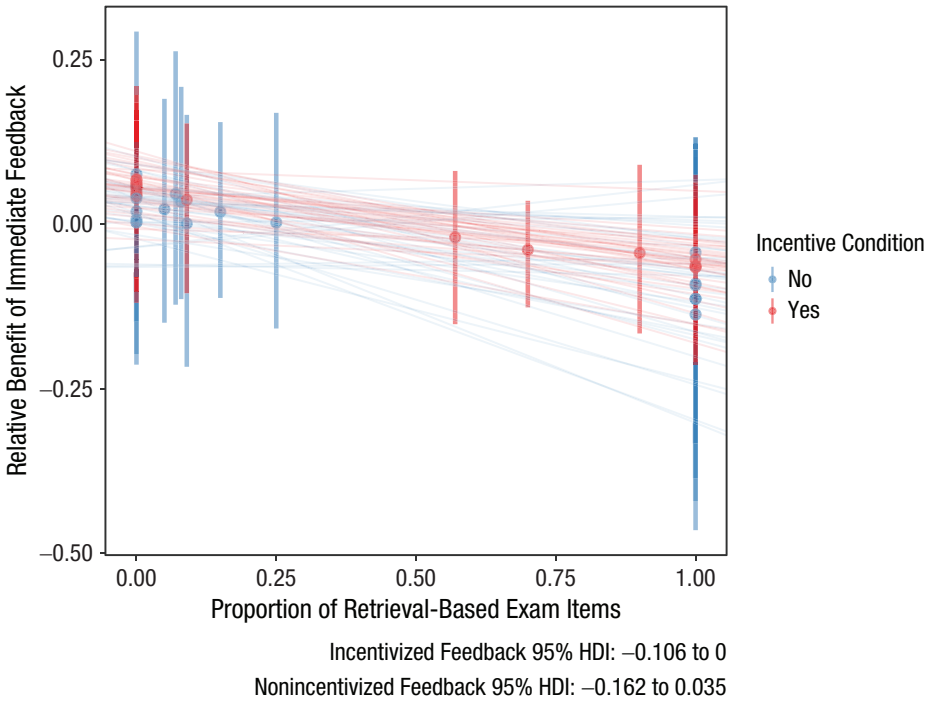

Fig. 7. Relationship between the proportion of retrieval-based exam items and the effect of feedback timing. The model's median estimate of the mean Δz score for each class is shown as a circle. The vertical lines show the 95% highest density interval estimate for each class. Sample regression lines from the model's posterior distribution are shown in the background. These lines represent plausible fits. We show a sample of these lines to visualize the model's uncertainty.

features, such as the ability to assess the effect of feedback timing's replicability (i.e., we were able to recruit many independent samples), robustness to variation (i.e., we were able to recruit diverse classes that varied on numerous dimensions), and ecological validity (i.e., we were able to recruit classes that were representative of typical practice). But this design space (the 38 classes in which we embedded the experiment) was not uniformly sampled from all possible class designs (as would be the case in a metastudy; Baribault et al., 2017). Instead, it was biased toward typical practice in contemporary college learning settings. Thus, the current study is ill-suited to determine whether immediate or delayed feedback timing *could* be beneficial under special circumstances but is particularly well suited to determine whether feedback timing affects student learning under authentic circumstances, which it does not (see Tipton & Hedges, 2017). We believe these contingencies, whether and how cognitive principles of learning translate to improvements in common educational situations, are where the psychological science of learning has room to improve.

Forty-five years ago, Cronbach (1975) argued that any effect in social science, and particularly in student instruction, should not be assumed to be stable but, rather, will vary across situations; and he repeatedly lamented the difficulty of collecting the "enormous volumes of data . . . required to pin down higher interactions as significant." Cronbach went on to write, "It is rarely practical to obtain information in a large number of situations. And the statistical estimates typically describe the gross aggregation of conditions instead of pinning down just what joint action of situational variables produces a particular effect" (p. 124). Our current findings are consistent with Cronbach's insights, and we also endorse his articulation of the practical difficulties of this kind of research. Despite the current unprecedented collaboration in education research (Makel et al., 2019), with 38 classes, we have scarcely begun to approach the sample size required to clearly "pin down" these effects. Part of this challenge is attributable to our ManyClasses model, in which the degree of exposure to experimental manipulations and the precision of the outcome measures are permitted to vary across sites, which contrasts with other multisite studies that measure invariant interventions and objective outcomes (for which our current sample likely would have been sufficiently powered for detecting moderator effects; Bloom & Spybrook, 2017). Future ManyClasses studies will need to cast a wider net across the design space of classes if we are to convincingly detail what works for whom in what context.

Our study also revealed the practical challenges of conducting this type of experimental education research. The teachers who participated in this collaborative effort should be lauded for their time, their expertise, and their willingness to embed an ambitious project in their course. But the coordination between educational practice and

experimental research multiplied the complexities of both. For example, during the semester-long period of data collection, due dates were changed without updating the feedback release dates, students contacted the research team with questions about course content, and more. We were often able to promptly remedy these issues, but at other times, we were forced to exclude treatment assignments (and then rebalance the design for that class) because they had been compromised. Future ManyClasses studies might avoid such issues if a tool were available for systematically facilitating experimental research in online learning settings, which we are currently pursuing (https://terracotta.education). Differences also were apparent in instructors' beliefs about feedback and the ways they communicated the research to their students, which were unmeasured for the current study and likely added variance (Hulleman & Cordray, 2009). Finally, we also found that reporting outcome scores separately for each treatment was particularly challenging in some cases. Some instructors' assessments were more cumulative across content and not easily separable between the content learned from quizzes with immediate feedback and content learned from quizzes with delayed feedback, which resulted in less precise outcome measures than would be desirable.

Nevertheless, these varied, suboptimal situations are precisely the classroom settings that require evidence-based insights for how to improve student learning (Koedinger et al., 2013; Motz et al., 2018). Psychological science is often looked to for these types of abstract principles of learning (Benassi et al., 2014), but our study reveals one example in which a psychologically relevant variable is inconsequential for routine practice at a global level. These findings imply that repeated A/B testing of a global effect in isolated settings will likely yield conflicting findings with limited external validity and perpetuate opacity about effective practices for improving student learning (see also, Yarkoni, 2020). If we, as researchers, hope to effectively translate theory into practice, we should be conducting a totally different kind of science, like the current study, that takes into account natural variation between settings. Our findings reveal that this kind of science needs to be massively scaled, perhaps an order of magnitude larger than ManyClasses1, if we are going to convincingly answer questions of what works for whom in what settings.

Limitations and Future Directions

There are trade-offs inherent in any research design. For example, a researcher might conduct a study in a sterile lab setting to control or optimize critical features, such as the strength of the manipulation and the precision of the outcome measure. In contrast, a different researcher might conduct a study in a class setting in which the findings gain external validity at the cost of these tight controls. The ManyClasses model is no exception. In fact, by conducting an experiment across many different classes, our study shines a spotlight on sources of variability that are relevant to experimental design but rarely given much consideration when studies draw from only one sample. When an experiment is conducted in a single class, the study may carry the guise of controlling the classroom context, but in reality, these contextual features vary widely in normative practice. We have argued for the necessity of incorporating these contextual classroom features into the research design, and even though these features are no more consequential in our study than in any single-class study, they are more visible here. We highlight two categories of issues to inform future research using the ManyClasses paradigm.

First, aspects of the current research design may have improved external validity but also may have diluted the experimental manipulation and made it harder to detect effects. For example, it is common in lab studies to control the frequency and duration of participants' exposure to the study materials. We did not control students' behaviors related to the feedback message but let them vary in natural ways. This means that students may have accessed their feedback multiple times (e.g., immediately and 3 days later) or shared their feedback with other students. Both of these behaviors would reduce the differences between the immediate and delayed feedback conditions. It is also common in lab studies to control how closely the content to be learned matches the final outcome measure. We did not control teachers' decisions about the contents of treatment quizzes or their assessments. Because of this, the concepts that students learned on one quiz may have overlapped with the concepts on another quiz, again, potentially diluting the experimental contrast. Likewise, the match between the contents of the treatment quizzes and the teachers' reported outcomes may have varied between classes in unmeasured ways. In total, the current study clearly lacked the control that would be characteristic of laboratory research.

Second, our method of recruiting classes may limit the generalizability of these results to other samples. We used an open recruitment model in which we advertised the study widely and invited interested teachers to apply to participate. Although we leveraged campus teaching centers for our initial callout, our recruitment model largely contrasts with a top-down approach in which a researcher purposively selects one or more classes to be included in a research project. Our open model had a variety of advantages, which included producing a motivated, diverse group of teachers from outside our circles who were willing to collaborate with our team. Yet this approach revealed another trade-off in which we achieved breadth at the cost of control. For example, our

Box 4. Lessons Learned

Plan on it taking twice as long as you think it will take.

- This project was hard and took longer than expected. The primary team gathered in May 2018, and data collection launched in Fall 2019—more than a full year of planning. It helped to streamline all communication (e.g., single contact person for teachers, prerecorded training videos) and to have a clear checklist of ordered tasks (e.g., institutional approval, then recruitment).

Collaboration is key for success.

- The author and acknowledgments lists are a tribute to the team-based nature of this study. Our core team had diverse areas of expertise (e.g., education research, big data, cognitive theory) that complemented each other. We also leveraged our connections with Unizin to facilitate recruitment and identify a set of enthusiastic teachers and administrators at each location to champion outreach efforts.

Teachers' active contributions to the research made this project possible.

- We formed true researcher-teacher partnerships with the 38 participating teachers, and teachers took on key responsibilities in the research design. They contributed by working with us to discuss their instructional materials, orient us to their Canvas course sites, and implement the manipulation in a way that ensured we obtained the best data possible within the constraints of their class context.

Flexibility is necessary when working across institutions.

- Our project required approval of a multisite protocol, and the agreement process looked slightly different at each participating institution. We had to be flexible in terms of the specific personnel who needed to be included, the precise order of steps for approval, and the timeline. Multiinstitutional education research would benefit from the standardization of data-sharing agreements.

Transparency and open science practices made our science better.

- We prioritized transparency—administrators knew the precise data we were going to collect, teachers knew how their assignments would be shaped by the study, students consented to share their data, and all materials and analyses were preregistered. These practices increased the buy-in from stakeholders, facilitated the data-sharing agreements with institutions, and enhanced the credibility of our results.

There are trade-offs to control compared with authenticity.

- Our strategy was to maximize teacher choice and authenticity to their class norms within the context of an experiment. This strategy resulted in ecologically valid settings, but at the cost of some control over specific features (e.g., treatment dosage, precision of outcomes). It is key to plan each decision (e.g., minimum dosage, open recruitment) in a way that fits with the goals of the research project.

We need technology that enables experimentation in diverse classrooms.

- This project was largely completed manually; a researcher manually created groups of students in Canvas, manually released feedback at the appropriate delay, and manually recorded outcome scores mapped to each treatment quiz. Experimental education research would benefit from streamlined technology that automates these processes seamlessly within the learning management system.

ManyClasses projects may need to include many more classes.

- We worked in 38 classes, but realistically, more classes are needed to test the effects of class-level moderators. This is especially true when teacher choice is maximized and there are unpredictable distributions of moderator values. The power of the statistical test is different if almost all teachers choose the same value than if teachers choose well-distributed values across the design space.

sample did not have strong representation from classes with large amounts of online quizzes, which might be normative in some disciplines. In addition, we recruited classes of college students, and it remains unknown whether the findings would generalize to classes with less advanced student populations. Likewise, we worked exclusively with teachers who volunteered to be included in a research study, and it is certainly possible that they differ in consequential ways (e.g., prior experience, classroom management style) from teachers who would not volunteer. Ideally, perhaps, we might have used a targeted recruitment strategy to obtain a representative

sample with proportionate representation of all kinds of classes, pedagogies, teachers, and students, but the concept of a representative sample of learning environments is currently undefined.

For researchers considering the use of the ManyClasses paradigm, or any field experiment in education, these two issues highlight the need to think critically about the trade-offs inherent in classroom-based research (also see Box 4). Specifically, what is gained in authenticity is lost in control. We have advanced a model that estimates differences in student learning during routine educational practice, when teachers manipulate a single instructional variable, effectively simulating what happens when teachers adopt an instructional recommendation. We believe this model has value but comes at the cost of our ability to maximize, via experimental control, the potential size of the measured effect. For this reason, we remind readers that feedback timing may still affect student learning in some contexts. However, instructional recommendations drawn from such limited contexts do not generalize broadly.

Conclusion

Given these limitations and challenges, what should one conclude from this study? First, we have observed evidence that there is no single, invariable benefit to receiving feedback immediately after a learning activity or when this feedback is delayed by a few days. Across typical college educational settings, the estimate of such a main effect is confidently close to zero. Second, our efforts to clearly identify moderating effects, situations in which the effect of feedback timing might deviate from zero, may have been hindered by a limited sample across the relevant design space and perhaps by low precision in our outcome measures. The current results suggest that future ManyClasses efforts will require yet grander scales with wider samples than the current study. Even so, our current results provide hints that in certain kinds of classes, which were undersampled in the current study, there may be modest advantages for delayed feedback. Third and finally, despite its obvious difficulties, one should conclude that this kind of experimental research is feasible in educational settings. The current ManyClasses study stands as a proof of concept that it is possible to test diverse implementations of an instructional recommendation and to assess the efficacy of these implementations for improving authentic measures of student learning.

Appendix: Moderator R^2 Analysis

Overview

Our initial, preregistered plan for summarizing the effect of each moderator on the student-level and class-level means included generating an additional metric beyond simply reporting the credible intervals of the coefficients. This metric is based on the increase in the proportion of variance explained with the addition of the moderator over the base model, and we describe it below. We planned to use this to identify which moderators were worth exploring in more detail and to provide a high-level summary of the moderators. However, in practice, we found the metric unhelpful, especially at the class level. There was simply too much uncertainty in the estimates to extract useful information. We ultimately decided to just directly present the posteriors of the moderator coefficients. In this Appendix, we present the planned analysis and the results (Fig. A1).

Summary of metric

One way to estimate the strength of the relationship between moderators and student-level or class-level means is to estimate the proportion of variance of the means explained by each model. The models that contain a moderator that is predictive of Δz scores will tend to explain a larger portion of the student-level or class-level variance. To do this, we measured the ratio of explained variance over explained variance plus residual variance. This measure is closely related to R^2 but is adapted for a Bayesian framework (Gelman et al., 2019). The values range between zero and one, and larger values indicate a greater proportion of variance explained. For class-level moderators, the predicted average for a class was the sum of the condition-level estimate and moderator effect ($\gamma_{i_c} + \beta_{i_c} x_c$), and the residual variance was the difference between this prediction and μ_c. For student-level moderators, the predicted Δz for a student was the sum of the class mean and moderator effect ($\mu_{c_s} + \beta_{c_s} x_s$), and the residual variance was the difference between this prediction and the observed Δz.

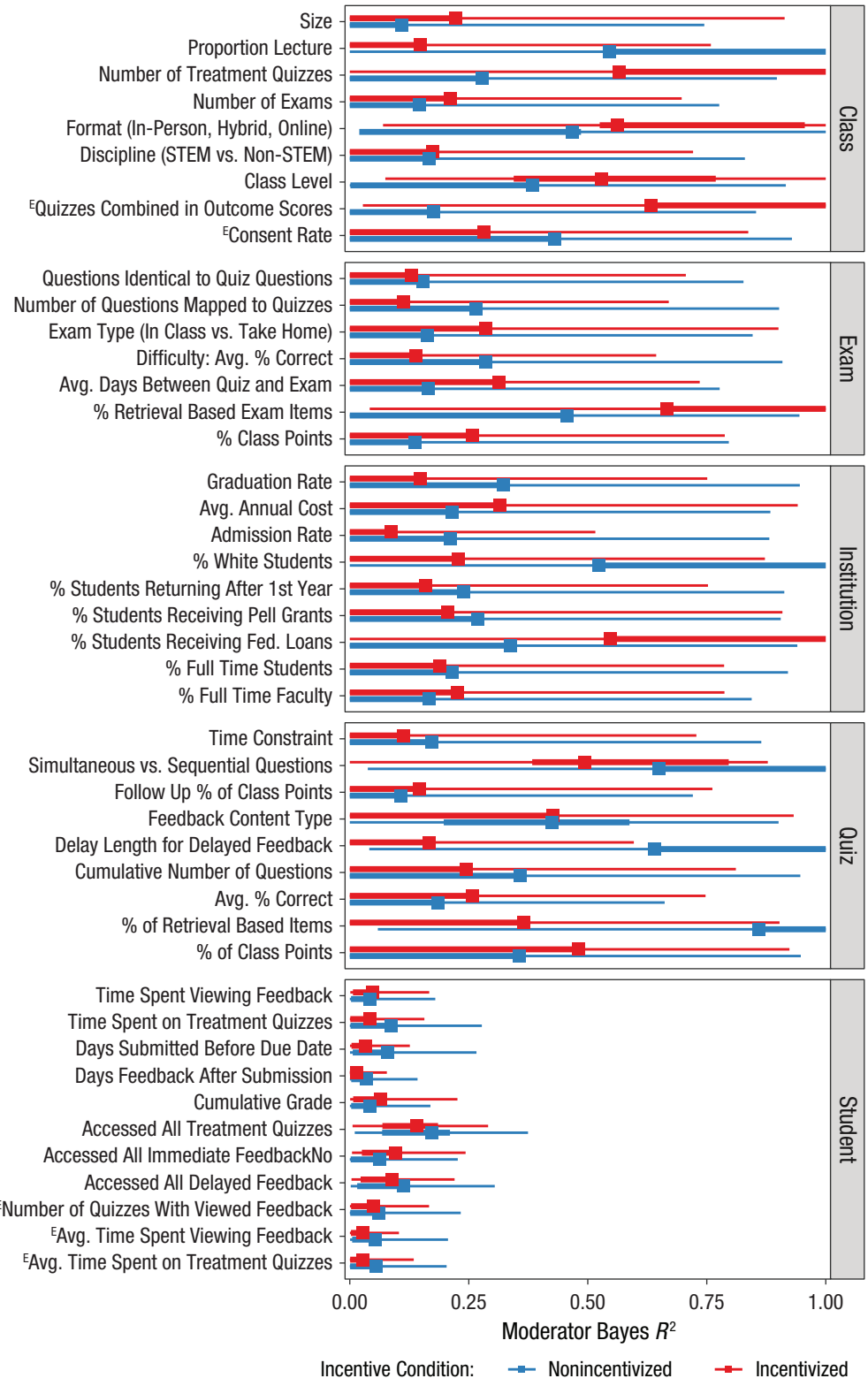

Fig. A1. Posterior estimates for R^2 for each moderator. The thin lines span the 95% highest density interval (HDI). The thicker line spans the 50% HDI, and the box is centered on the median. Moderators with [E] designation were not preregistered.

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons

Author Contributions

E. R. Fyfe, J. R. de Leeuw, P. F. Carvalho, R. L. Goldstone, J. Sherman, and B. A. Motz are lead authors. E. R. Fyfe, J. R. de Leeuw, P. F. Carvalho, R. L. Goldstone, and B. A. Motz jointly generated the idea for the study. E. R. Fyfe and J. R. de Leeuw wrote the first draft of the predata manuscript, and all six lead authors participated in edits and revisions. All six lead authors approved the final submitted version of the manuscript. E. R. Fyfe and B. A. Motz completed and submitted the application for approval from the Institutional Review Board at Indiana University. B. A. Motz led recruitment efforts and obtained approval from participating institutions in collaboration with L. Cruz at Penn State, and all six lead authors contributed to recruiting instructors. P. F. Carvalho wrote the first draft of the guidebook for participating instructors, and all six lead authors critically edited it. J. R. de Leeuw wrote the analysis code. J. Sherman met with participating instructors, facilitated the preparation of materials, and implemented the experiment (including random assignment) in the participating classes. B. A. Motz, E. R. Fyfe, and J. Sherman collected data. E. Pelaprat and K. Unruh provided data from Unizin. B. A. Motz processed these data, and J. R. de Leeuw analyzed the processed data. B. A. Motz and J. R. de Leeuw produced a first draft of the postdata manuscript. All six lead authors participated in edits and revisions. All of the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This study was supported with supplemental funding from the Department of Psychological and Brain Sciences and University Information Technology Services's division of Learning Technologies at Indiana University Bloomington.

Open Practices

Open Data: https://osf.io/q7avj
Open Materials: https://osf.io/q7avj
Preregistration: https://osf.io/sdqwm

All data and materials have been made publicly available via OSF and can be accessed at https://osf.io/q7avj. The protocol and analysis plans were preregistered via OSF and can be accessed at https://osf.io/sdqwm. Changes to the preregistered analyses are described in the text. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Emily R. Fyfe https://orcid.org/0000-0002-5420-360X
Joshua R. de Leeuw https://orcid.org/0000-0003-4815-2364
Janelle Sherman https://orcid.org/0000-0001-7940-3205
Benjamin A. Motz https://orcid.org/0000-0002-0379-2184

Acknowledgments

We thank John K. Kruschke for feedback on the analysis plan and Andrew C. Butler and two anonymous reviewers for their comments on an earlier draft of this article. We also acknowledge and thank the many people who assisted with this study, including Aaron Neal (Unizin), Jill Buban (Unizin), Stephan Nicklow (Unizin), Kara Armstrong (Unizin), David Goodrum (Oregon State), Sol Bermann (University of Michigan [UMich]), Sean DeMonner (UMich), Paul Robinson (UMich), Kelly Cruz (UMich), James Hilton (UMich), Matthew Kaplan (UMich), Lisa Emery (UMich), Angela Linse (Penn State University), Stacy Morrone (Indiana University [IU]), Erik Scull (IU), Greg Siering (IU), John Gosney (IU), Andrew Korty (IU), Andrew Nill (IU), Juliet Aders (IU), Katie Morris (IU), Jeffrey Goetz (IU), LeAnna Faubion (IU), Ryan Ballard (IU), Bethany Johnson (IU), Emily Oakes (IU), Sara Chambers (IU), Julie Lorah (IU), Dubravka Svetina (IU), Amy Goodburn (University of Nebraska Lincoln [UNL]), Heath Tuttle (UNL), Matt Morton (UNL), Sydney Brown (UNL), Tammie Herrington (UNL), Donalee Attardo (University of Minnesota [UMN]), Robert Alberti (UMN), Karen Hanson (UMN), Emily Ronning (UMN), Lauren Marsh (UMN), Paul Savereide (UMN), and Brian Dahlin (UMN).

References

Anderson, L. S., Healy, A. F., Kole, J. A., & Bourne, L. E., Jr. (2013). The clicker technique: Cultivating efficient teaching and successful learning. *Applied Cognitive Psychology*, *27*, 222–234. https://doi.org/10.1002/acp.2899

Andrews, T. C., & Lemons, P. P. (2015). It's personal: Biology instructors prioritize personal evidence over empirical evidence in teaching decisions. *CBE - Life Sciences Education*, *14*, 1–18. https://doi.org/10.1187/cbe.14-05-0084

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2017). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences, USA*, *11*, 2607–2612. https://doi.org/10.1083/pnas.1708285114

Benassi, V. A., Overson, C. E., & Hakala, C. M. (2014). *Applying science of learning in education: Infusing psychological science into the curriculum*. American Psychological Association.

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *4*, 877–902. https://doi.org/10.1080/19345747.2016.1271069

Bohn, M., Schmitt, V., Sanchez-Amaro, A., Keupp, S., Hopper, L., Völter, C., Altschul, D., Fischer, J., Fichtel, C., Beran, M. J., Kano, F., Call, J., Watzek, J., Joly, M., & Hernandez-Aguilar, R. A. (2019). *Establishing an infrastructure for collaboration in primate cognition research*. OSF. https://doi.org/10.31234/osf.io/3xu7q

Booth, J. L., McGinn, K. M., Barbieri, C., Begolli, K. N., Chang, B., Miller-Coto, D., Young, L. K., & Davenport, J. (2017). Evidence for cognitive science principles that impact learning in mathematics. In D. C. Geary, D. B. Berch, R.

Ochsendorf, & K. M. Koepke (Eds.), *Acquisition of complex arithmetic skills and higher-order mathematics concepts* (Vol. 3, pp. 297–325). Academic Press.

Booth, J. L., Oyer, M., Paré-Blagoev, J., Elliot, A. J., Barbieri, C., Augustine, A., & Koedinger, K. (2015). Learning algebra by example in real-world classrooms. *Journal of Research on Educational Effectiveness*, *8*, 530–551. https://doi.org/10.1080/19345747.2015.1055636

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Butler, A. C., & Woodward, N. R. (2018). Towards a consilience in the use of task-level feedback to promote learning. In K. D. Federmeier (Ed.), *Psychology of learning and motivation* (pp. 1–38). Academic Press. https://doi.org/10.1016/bs.plm.2018.09.001

Carvalho, P. F., Braithwaite, D. W., de Leeuw, J. R., Motz, B. A., & Goldstone, R. L. (2016). An in vivo study of self-regulated study sequencing in introductory psychology courses. *PLOS ONE*, *11*, Article e0152115. https://doi.org/10.1371/journal.pone.0152115

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116–127.

Dabbagh, N., Bass, R., Bishop, M., Costelloe, S., Cummings, K., Freeman, B., Frye, M., Picciano, A. G., Porowski, A., Sparrow, J., & Wilson, S. J. (2019). *Using technology to support postsecondary student learning: A practice guide for college and university administrators, advisors, and faculty*. What Works Clearinghouse, National Center for Education Evaluation and Regional Assistance.

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25. https://doi.org/10.18637/jss.v071.i09

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*, 421–435. https://doi.org/10.1111/infa.12182

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, USA*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Fyfe, E. R., & Brown, S. A. (2018). Feedback influences children's reasoning about math equivalence: A meta-analytic review. *Thinking and Reasoning*, *24*, 157–178. https://doi.org/10.1080/13546783.2017.1359208

Gelman, A., Goodrich, B., Garby, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gurung, R. A., & Burns, K. (2019). Putting evidence-based claims to the test: A multi-site classroom study of retrieval practice and spaced practice. *Applied Cognitive Psychology*, *33*(5), 732–743. https://doi.org/10.1002/acp.3507

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. https://doi.org/10.3102/003465430298487

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88–110. https://doi.org/10.1080/19345740802539325

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. https://doi.org/10.1177/2515245918810225

Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Knight, J. K., & Wood, W. B. (2005). Teaching more by lecturing less. *Cell Biology Education*, *4*, 298–310.

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, *342*(6161), 935–937.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*, 79–97. https://doi.org/10.3102/00346543058001079

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*, 254–261. https://doi.org/10.1037/pspi0000106

Makel, M. C., Smith, K. N., McBee, M. T., Peters, S. J., & Miller, E. M. (2019). A path to greater credibility: Large-scale collaborative education research. *AERA Open*, *5*(4), 1–15. https://doi.org/10.1177/2332858419891963

Motz, B. A., Carvalho, P. F., de Leeuw, J. R., & Goldstone, R. L. (2018). Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics*, *5*, 47–59. https://doi.org/10.18608/jla.2018.52.4

Mullaney, K. M., Carpenter, S. K., Grotenhuis, C., & Burianek, S. (2014). Waiting for feedback helps if you want to know the answer: The role of curiosity in the delay-of-feedback benefit. *Memory & Cognition*, *42*(8), 1273–1284. https://doi.org/10.3758/s13421-014-0441-y

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, *3*, 222–229. https://doi.org/10.1016/j.jarmac.2014.05.001

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. The National Academies Press. https://doi.org/10.17226/24783

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop of Distributed Statistical Computing (DSC 2003)* (Vol. 124, pp. 1–10). Vienna, Austria: Technische Universität Wien.

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2020). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*(1), 40–52. https://doi.org/10.1037/edu0000367

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*(1), Article 79. https://doi.org/10.1186/1471-2288-8-79

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review*, *24*, 86–97. https://doi.org/1955-02985-001

Tipton, E., & Hedges, L. V. (2017). The role of the sample in estimating and explaining treatment effect heterogeneity. *Journal of Research on Educational Effectiveness*, *10*(4), 903–906. https://doi.org/10.1080/19345747.2017.1364563

Wakeling, V., & Robertson, P. R. (2017). A comparison of student behavior and performance between an instructor-regulated versus student-regulated online undergraduate finance course. *American Journal of Educational Research*, *5*, 863–870. https://doi.org/10.12691/education-5-8-5

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05." *The American Statistician*, *73*(Suppl. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*. Advance online publication. https://doi.org/10.1017/S0140525X20001685