qualitativeresearchjournal.com
Vol. 23 / No. 02 / 2023
ISSN: 1468-7941

Terracotta: A tool for conducting experimental Research on student learning

Benjamin Motz

Professor, (Department of Psychological and Brain Sciences), Indiana University: Bloomington, Indiana, US

Abstract : For researchers seeking to improve education, a common goal is to identify teaching practices that have causal benefits in classroom settings. To test whether an instructional practice exerts a causal influence on an outcome measure, the most straightforward and compelling method is to conduct an experiment. While experimentation is common in laboratory studies of learning, experimentation is increasingly rare in classroom settings, and to date, researchers have argued it is prohibitively expensive and difficult to conduct experiments on education in situ. To address this challenge, we present Terracotta (Tool for Education Research with RAndomized COnTrolled TriAls), an open-source web application that integrates with a learning management system to provide a comprehensive experimental research platform within an online class site. Terracotta automates randomization, informed consent, experimental manipulation of different versions of learning activities, and export of de-identified research data. Here we describe these features, and the results of a live classroom demonstration study using Terracotta, a preregistered replication of McDaniel et al. (Journal of Applied Research in Memory and Cognition, 1(1), 18–26, 2012). Using Terracotta, we experimentally manipulated online review assignments so that consenting students alternated, on a weekly basis, between taking multiple-choice quizzes (retrieval practice) and reading answers to these quizzes (restudy). Students' performance on subsequent exams was significantly improved

for items that had been in retrieval practice review assignments. This successful replication demonstrates that Terracotta can be used to experimentally manipulate consequential aspects of students' experiences in education settings.

Keywords: Terracotta, tool, conduct, experimental Research, student

Introduction

The productivity and resilience of society depends on our system of formal education, and this system currently suffers major challenges, including lagging achievement, low persistence in science, technology, engineering, and math (STEM) disciplines, and systematic inequities based on students' sociodemographic characteristics. These challenges are well documented across a range of national and international assessment instruments (e.g., NAEP, 2022; NCES, 2022; PISA, 2020). However, despite having strong instruments for measuring these issues, there has been a complete lack of tools for research on how these issues should be remedied.

Improving education involves identifying and promoting instructional practices that have causal benefits for student outcomes (National Science Foundation & Institute for Education Sciences, 2013; US Department of Education, 2016, 2017). To test whether an instructional practice exerts a causal influence on an outcome measure, the most straightforward and compelling research method is to conduct an experiment (National Research Council, 2002; Whitehurst, 2003), and in particular, to embed this experiment in an education setting, yielding causal inferences that are authentic to the contexts where they matter in practice (Koedinger et al., 2013; Motz et al., 2018). An experiment satisfies the strong requirements of causal inference by providing evidence that a change in behavior is attributable to a change in treatment, in a specific direction (ruling out reverse causality), and by minimizing (through random assignment) the possibility that it is explainable through other causal mechanisms. Combined,

these features constitute an ideal method for examining whether a treatment (e.g., a learning activity, an instructional strategy, a motivational intervention) causes improvements (Mosteller & Boruch, 2002; Shadish et al., 2001). Unfortunately, however, conventional education settings are not naturally conducive to experimentation.

Researchers have repeatedly argued that it is prohibitively difficult to conduct controlled experiments in education settings. Slavin (2002) notes that "randomized experiments of interventions applying to entire classrooms can be extremely difficult and expensive to do" (p. 17). Similarly, Sullivan (2011) comments on the "feasibility considerations" of conducting experiments in education, remarking that the cost "can be high and industry support for education trials rarely exists" (p. 285). Such concerns are also expressed by Levin (2005), who writes, "the requisite resources are generally far in excess of what most educational researchers could hope to amass in the absence of considerable extramural funding. Consequently, researchers elect to conduct more manageable, less ambitious, and typically, less carefully-controlled classroom-based investigations" (p. 19). Perhaps for these reasons, the relative frequency of experimental educational research studies has been in steady decline, from 47% of published education studies in 1983, to 40% in 1994, to 26% in 2004, to 20% in 2020, despite increasing prevalence of causal claims in the education research literature (Brady et al., 2022; Hsieh et al., 2005; Motz et al., 2023; Robinson et al., 2007).

Experimental psychologists have been filling this gap to some extent, conducting experimental studies on human learning, albeit primarily under controlled laboratory conditions, and then advocating for their applicability in education settings (Benassi et al., 2014; Pashler et al., 2007; Roediger & Pyc, 2012). Their advocacy, however, has been restrained; these same experimental psychologists, and others as well, affirm that research is needed to validate claims in practice (Daniel, 2012; Koedinger et al., 2013; Motz et

al., 2018). For example, in Dunlosky et al.'s (2013) extensive review of learning strategies from cognitive and educational psychology, the evidence from education settings is marked as "insufficient" for 8 out of the 10 strategies under investigation. This evidence is important for testing the generalizability of effects within the complexities of education settings (de Leeuw et al., 2022; Fyfe et al., 2021) and to build an evidence base that is more convincing and relevant to practitioners. However, despite advocating for the importance of this translational research, experimental and educational psychology have offered no fix for the methodological difficulties that have impeded experimentation in education for the past 40 years.

To be fair, experimental research in authentic, practical settings is difficult in all disciplines, and some researchers manage through it; the difficulties of experiments in education are not insurmountable, provided some struggle (Gueron, 2002). But a problematic feature of experimental research on student learning in practice, specifically, has been the complete lack of tools to support it (Schneider & Garg, 2020). This absence of experimental research infrastructure has been particularly vexing when considering that improving education is a critical research priority: all members of society are directly affected by our education system, and by its challenges.

Fortunately, things are starting to change. Enabled by the digital transformation of education, research tools for experimentally manipulating features of the student learning environment have begun to emerge (Baker et al., 2022; McCarthy et al., 2022). As student learning activities and student data are increasingly online, the technical infrastructure for supporting education can be leveraged for research purposes—digital learning platforms can support student learning, and also can be equipped with tools to support experimental research on education. In this article, we describe one such emerging tool, Terracotta, and we present a demonstration of Terracotta's capabilities in a class-embedded

preregistered replication experiment.

Overview of Terracotta

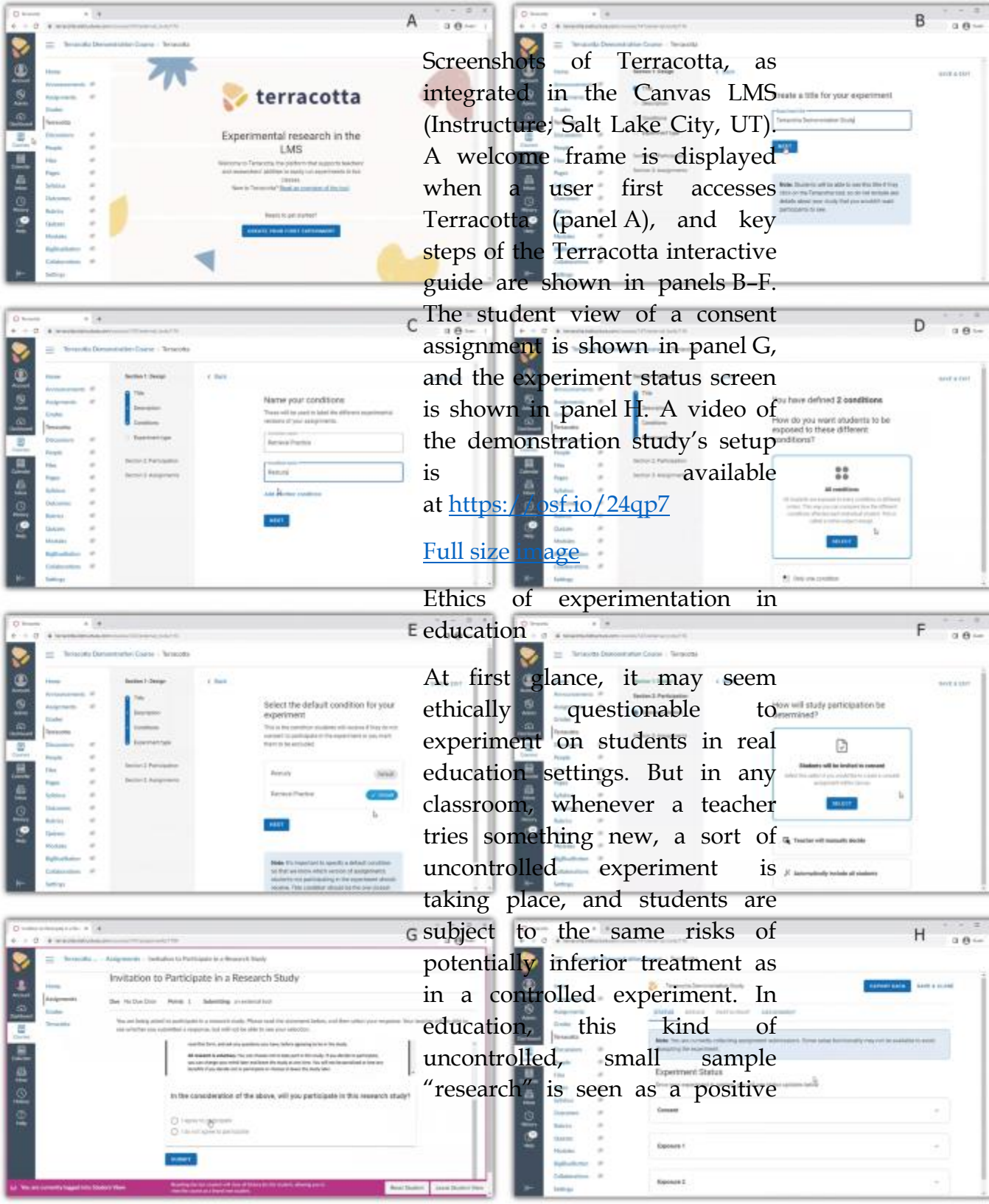
Terracotta (Tool for Education Research with RAndomized COnTrolled TriAls; https://terracotta.education) is an online experiment builder that allows researchers and teachers to rapidly design and deploy experiments directly within the learning management system (LMS). The LMS is now a prominent and integral component of formal education settings, central to routine practice akin to how the chalkboard was once the center of the classroom. It is a large-scale online platform that hosts secure websites for each class within a district or an institution and makes a suite of applications available for student and teacher use. These include organizational frameworks like customizable pages and modules, as well as file and multimedia storage, assignments, quizzes, announcements, calendars, private messaging, discussions, gradebooks, and other functions. Adoption of an LMS has for years been driven by education's ongoing digital transformation (Lonn & Teasley, 2009; Pomerantz & Brooks, 2017; Staker, 2011), but has accelerated rapidly due to COVID-19, particularly in K–12 environments where LMS adoption more than doubled between 2019 and 2020 (Hill, 2020). By centering in the LMS, Terracotta enables experimental research across education levels, student populations, and learning materials.

Presently, Terracotta's primary feature is to experimentally manipulate LMS assignments. Assignments instruct students to perform learning activities, and as such, assignments are the vehicles that connect students with practice, feedback, resources, interventions, and formative assessments. In many ways, assignments occupy a central element of education systems, because when students have autonomy to choose their own learning strategies, long-standing evidence shows that students' choices are often suboptimal (Kornell & Bjork, 2007; Pressley et al., 1989). Furthermore, assignments are

increasingly used to administer social and motivational interventions in education settings (Harackiewicz & Priniski, 2018; Yeager et al., 2019). Thus, what a teacher should assign, and how they should assign it, represent foundational priorities for education research (Benassi et al., 2014; Borman, 2002; Cohen et al., 2003; Dunlosky et al., 2013). At a high level, Terracotta makes it possible to create different versions of an LMS assignment (which might vary in instructions, contents, resources, etc.) and to randomly assign students to experience these different versions.

An experiment in Terracotta is created using an interactive guide (see screenshots in Fig. 1). Like the US Department of Education's Evidence-to-Insight (e2i) Coach (formerly RCE Coach; Office of Educational Technology, 2017), Terracotta walks the researcher through a sequence of design decisions, such as the number of treatment conditions, treatment design, informed consent, uploading materials, and so on. But unlike e2i Coach, which then leaves researchers to go out and conduct studies on their own, Terracotta embeds the experiment, as designed, automatically within the LMS. From the student's perspective, they complete assignments in the LMS as usual, but behind the scenes, Terracotta manages the details of presenting the appropriate experimental variations to different students and collecting data along the way.

Fig. 1



Screenshots of Terracotta, as integrated in the Canvas LMS (Instructure; Salt Lake City, UT). A welcome frame is displayed when a user first accesses Terracotta (panel A), and key steps of the Terracotta interactive guide are shown in panels B–F. The student view of a consent assignment is shown in panel G, and the experiment status screen is shown in panel H. A video of the demonstration study's setup is available at https://osf.io/24qp7

Full size image

Ethics of experimentation in education

At first glance, it may seem ethically questionable to experiment on students in real education settings. But in any classroom, whenever a teacher tries something new, a sort of uncontrolled experiment is taking place, and students are subject to the same risks of potentially inferior treatment as in a controlled experiment. In education, this kind of uncontrolled, small sample "research" is seen as a positive

feature of teachers' professional development and growth (Guskey & Huberman, 1995). However, this "research" (a teacher trying new things on a full student cohort) also has the drawbacks of being subject to the single unit confound (any improvement could be due to other properties of the cohort) and being subject to bias (whether new tactics "work" is only judged by subjective reflection of the teacher, who is also the creator of these tactics). Researchers can avoid these issues by randomly assigning students to experimental conditions. When doing so, it becomes possible to conduct more robust assessments of these instructional treatments, and to advance improvements that are known to benefit students.

Nevertheless, this kind of research is sensitive. When technologies unilaterally enroll their users in experiments, it has been met with strong public objection (Goel, 2014; Herold, 2018). The objection is not simply that experimental research is being performed; instead, such instances have revealed that people expect to be informed when they might be participating in research studies, and they want to be able to decide whether or not they will participate, even if such expectations are not codified in law or policy (Flick, 2015). What is supported by law, however, is the expectation that identifiable student records and research data must remain private. Terracotta's features are intended to achieve these normative expectations directly and systematically.

We believe that the benefits of transparency and agency generally outweigh any challenges that they may present to an embedded research study. In this regard, we advocate that students (and parents, in the case of research on minors) should be informed about an experimental research study that is occurring in Terracotta, and that Terracotta's informed consent feature should be used to provide them with a secure and private means for registering their choice of whether to participate. In situations when permission from a parent or guardian is required, it is possible to distribute permission forms, and then to

manually mark in Terracotta which students have returned these forms (or conversely, to mark which students' parents have opted out). The concern that students' knowledge of being in an experiment may change their behaviors, sometimes called the Hawthorne effect, has a shaky evidence base in education research (Adair, 1984). It is, however, possible that individuals from historically oppressed populations may be less likely to agree to participate (e.g., Li et al., 2022), which may yield a biased sample. These issues might best be remedied with more transparency during recruitment (Yancey et al., 2006), rather than forgoing consent to reduce sampling bias.

Another ethical concern is that an experiment may cause some students to receive inferior treatment. Simply by differentiating students' learning experiences, it stands to reason that one experimental variant might be better than or worse than another. To proactively mitigate this risk, and its potential effects, we recommend a variety of strategies (Motz et al., 2018). First, an experimental condition that is known to be inferior, such as a deprived or no-treatment control, should not be administered to students; instead, we recommend making comparisons with a business-as-usual control (Willingham & Daniel, 2021). Second, whenever possible, we recommend adopting a within-subjects (or delayed treatment) design. In doing so, all participants will experience all conditions, staggered in time, thus equating treatment between students overall. Third, we recommend that the scope of an embedded experiment should be modest—in research (as in routine instruction), large changes to a student's curricular experience may be inappropriate if the benefit of the change is unknown. When keeping an experiment's scope modest, if experimental differences are observed, they may nevertheless be negligible for any individual student's overall learning and achievement. Fourth, instructional assistance should never be withheld from students. If a student asks a question that is relevant to an experiment, it is more important for the student to

receive help than for the study's strictures of experimental control to be maintained. To help ensure that students are treated equitably, Terracotta does not reveal to the teacher the condition to which a student is assignedFootnote1.

Finally, we affirm that individual students' identities, their decisions to participate in a study (or not), their behaviors within Terracotta, and their learning performance are confidential and private. When data are exported from Terracotta, the data are de-identified (student identifiers are replaced by otherwise meaningless Terracotta-internal IDs), and data from nonparticipants are excluded from the export. Ideally, researchers will have no need for seeing identifiable research data. Furthermore, we recommend that researchers should clarify, both during recruitment and in an informed consent statement, how participant data will be used following the study.

Terracotta enables collaboration between teachers and researchers

When we tell people about Terracotta, they often ask: Who is supposed to use it? Teachers or researchers? The answer is "Both, together."

As digital tools for education research emerge, a known risk is that they will marginalize the teacher and neglect important details of the educational implementation, instead focusing myopically on the inner workings of the tool itself. Proponents of research within digital learning platforms advocate for greater involvement of teachers in the research process (Baker et al., 2022; McCarthy et al., 2022), as others have advocated more broadly (Joyce & Cartwright, 2019). This is also our goal with Terracotta.

In an LMS course site, as in a classroom, teachers are in control—they have express privileges to create assignments, course policies, resources, announcements, and so on. Historically, researchers who might want to manipulate these features of a class would need to partner, negotiate, and coordinate with teachers (unless the teachers are the researchers; Handelsman et al., 2004; Boyer, 1990). Similarly, Terracotta

relies on teacher involvement and enables researchers and teachers to collaborate on how an experimental study will be implemented in a class. A researcher who wants to conduct a study in Terracotta will need to partner with a teacher who is willing to embed the study in their class and collaborate on how an experimental contrast might be applied to the teacher's instructional materials. In practice, the teacher might invite the researcher into the class's LMS site to set up an experiment in Terracotta, or a teacher might create the experiment themselves, in collaboration with the researcher, using Terracotta's interactive guide. Terracotta automatically manages the details of informed consent, random assignment, experimental differentiation, data de-identification, and so on; accordingly, there is minimal effort required of the teacher once a study is initially set up in Terracotta. There are reciprocal benefits to such collaborations: this improves the external validity of a research study, while also involving teachers in building a more authentic and relevant evidence base of what works in routine contexts.

Key features

Informed consent that conceals responses from the instructor

Informed consent is a cornerstone in the ethical conduct of research with humans. While consent may not be legally required of research with Terracotta (current code provides an exemption for research on normal educational practices; Exempt Research, § 46.104(d)(1), 2018), consent (or assent, in the case of research with minors) nevertheless improves the transparency of research and provides agency to participants. To realize these benefits, Terracotta implements an LMS assignment that presents an informed consent statement followed by a simple consent prompt, enabling students to mark whether they agree to participate in a research study. Like any assignment in Terracotta, this consent assignment can have a deadline, and it can be configured to give students credit for responding. However, there is no correct answer to the consent prompt, and students will receive

submission credit regardless of their response. When a student marks their consent, they become a participant, eligible to be randomly assigned to experimental conditions and to have their de-identified data included in the experiment's data export. When a student does not provide consent (either by providing a negative response or by not responding at all), they will not become a participant, they will not be randomly assigned to experimental conditions (they will only receive whichever condition the researcher has marked as "default"), and their data will not be included in the experiment's data export.

An important aspect of Terracotta's consent process is that students' consent responses are never revealed to the teacher or the researcher. This is because many review boards are sensitive that a teacher is in a position of power, and teachers might be perceived as coercing students to participate if they were able to see individual students' consent responses. Terracotta does allow teachers to see whether a student has responded to the consent assignment, but it conceals any information about how students have responded.

Assignment of students to experimental conditions

An experiment's fundamental feature is the introduction of different experimental conditions to different subjects. By default in Terracotta, participants are assigned to a condition (in a between-subjects design) or to a schedule of conditions (in a within-subjects design) at the time when they first access an experimentally manipulated activity in Terracotta. We prefer this trickle assignment approach (Riecken & Boruch, 1974) rather than batch assignment, because class enrollment is not necessarily static—students may be added to the LMS course site after the start of the academic term, possibly after an experiment is created, and these students should not be automatically ineligible to participate, nor should they have instructional resources withheld. However, it is also possible in Terracotta to manually assign participants to conditions in a single batch during experiment

setup, overriding Terracotta's default behavior.

While pure random assignment may be considered the de facto method for assigning participants to conditions, Terracotta implements a hybrid random/sequential assignment algorithm that aims to balance the quantity of participants assigned in each condition. When a participant first accesses an experimentally manipulated activity in Terracotta, they are randomly assigned to a condition. When a subsequent participant accesses an activity in Terracotta, they are assigned to whichever condition has the least number of participants; if two or more conditions have the least number of participants, they are randomly assigned among them. If all conditions have the same number of participants, again, inbound participants are randomly assigned among them. The benefit of this approach is that it optimizes balance in the number of participants assigned to different experimental conditions, while still preserving the lack of bias that is characteristic of random assignment. Balance is a priority in class-embedded experiments, where sample size is limited by the class's enrollment, and it is important to actively avoid the rare possibility that pure random assignment will yield disproportionate group sizes. However, in a between-subjects design, if imbalance is desirable, Terracotta allows the researcher to specify a custom distribution scheme for different experimental conditions (e.g., 75% of participants could be assigned to condition A, and 25% to condition B), and the algorithm described previously will aim to achieve these custom proportions. And again, a researcher could override these processes entirely, and could manually assign participants to conditions when setting up the experiment.

Repeated treatments and within-subject crossovers

An occasional criticism of embedded experiments in education is that they are short-term manipulations, "limited to testing the impact of pulling a single lever at a time" (Schanzenbach, 2012). Such criticisms are misleading for at

least two reasons. First, some specific levers, even in their transience, can be consequential in education (Walton & Cohen, 2011; Yeager et al., 2019), but second and more broadly, it is simply not true that embedded experiments need to be short-term. Terracotta allows an experimental contrast to be applied across many treatments. A researcher could manipulate a single assignment or could manipulate months-long sequences of assignments (as we did in the Demonstration Study, below). Doing so may increase the "dosage" of a treatment regimen, while also providing a more authentic measure of the regimen's effect if it were routinized in normal instructional practice. Moreover, by manipulating multiple treatments, the effect of an experimental manipulation may be claimed to generalize beyond the nuances of any single assignment.

In allowing repeated treatments, Terracotta also supports the ability for participants to change conditions across treatments, in a within-subjects design. A within-subjects design has important statistical advantages and ethical benefits, but it also introduces the risk of carry-over effects, and the decision to use a within-subject design should be made with an awareness of these tradeoffs (Greenwald, 1976). Once assignments have been created in Terracotta, the timing of these assignments (the open and due date) is configured in the LMS, so the schedule of when participants are exposed to each condition is highly customizable. For example, a participant could experience condition A for four assignments, then cross over to experience condition B for four more assignments (AAAABBBB; a single crossover). Alternatively, a participant could alternate between A and B repetitively (ABABABAB; as we did in the Demonstration Study, below), with multiple crossovers.

Mapping outcomes to experimental treatments

In routine educational practice, a teacher measures student learning outcomes (e.g., exam scores) and other relevant behavioral outcomes (e.g., attendance, classroom conduct, participation). In an experiment,

these might be relevant as pretest or posttest measures. To assess whether an experimental manipulation in Terracotta affects these distal measures, there needs to be a mapping from these distal measures onto students' research profiles. Without Terracotta, mapping outcomes to experimental treatments can be a sensitive task, as it involves joining identifiable data from different sources (research data and student data). To reduce the risk of loss of confidentiality in the research process, Terracotta includes a feature where outcome data can be identified directly in the LMS gradebook following each treatment exposure (for within-subject designs) or at the end of an experiment (for between-subject designs). Alternatively, if the research is targeting outcome measures that are not in the LMS gradebook (or a sub-score of a gradebook item), Terracotta allows outcomes to be manually entered into a simple class roster for the full class, preventing any exposure of participant's consent decisions or condition assignments. De-identified outcome data, whether selected from the LMS gradebook or manually added, are then included in the Terracotta data export, with nonparticipants removed.

Export of de-identified data

At the end of the experiment, Terracotta produces an export of all study data, with student identifiers replaced with random codes, and with non-consenting students removed. This export includes condition assignments, responses and scores on manipulated learning activities, granular clickstream data for interactions with Terracotta assignments, and outcomes data as specified by the teacher. By joining these data, de-identifying it, and excluding non-consenting participants, Terracotta prepares a data export that is shareable with research collaborators (includes no personally identifiable information) and that meets ethical requirements (excludes data from nonparticipants).

Technical description

Terracotta is an open-source web application; the full source code is available at https://github.com/terracotta

-education/terracotta under a permissive Apache 2.0 license (Sinclair, 2010). The Terracotta backend architecture uses a model-view-controller (MVC) pattern written in Java using the Spring framework (https://spring.io). The Terracotta frontend is written in JavaScript using the Vue framework (https://vuejs.org).

Terracotta integrates with an LMS using current learning tool interoperability (LTI; version 1.3) standards (1EdTech, 2022a; Unicon Inc., 2019). User authentication takes place within the LMS, and the LMS provides Terracotta with an encrypted LTI token when a user launches Terracotta. This token identifies the user, their role (learner, instructor), and their context (the LMS course number), so that when a user requests a resource within Terracotta (e.g., a student attempts to complete an assignment), Terracotta can respond with the appropriate experimental treatment (or display the teacher interface, if appropriate). However, the LTI standards do not currently provide all the requisite endpoints to support Terracotta's features, so Terracotta also uses the LMS's native application programming interface (API) to make functional requests within the LMS course site that are outside the scope of LTI 1.3. These API calls are made on behalf of the instructor, using an OAuth 2.0 token explicitly granted by the instructor when first accessing Terracotta. While contemporary LMSes have the same general API endpoints, there are modest differences in them, and thus Terracotta is not universally interoperable by default. At this time, Terracotta is designed for integration with the Canvas LMS (Instructure; Salt Lake City, UT), but we anticipate extending support to other LMSes over time.

Timestamps collected by Terracotta (the time when a student provided consent, started an assignment, clicked submit, etc.) are logged on the server side using the thread-safe Java .now() method.

Terracotta's user interface uses the open-source Material Design system (Google, 2022) so that the user experience is familiar to users who are accustomed to

common web interfaces. To benefit the broadest community possible, every frame in Terracotta is screened using the axe DevTools extension to test for common accessibility issues. Terracotta aspires to comply with Web Content Accessibility Guidelines (WCAG) 2.1 AA standards both within the tool, on its website, and in accompanying documentation.

As an open-source web application, one could self-host their own instance of Terracotta on their own infrastructure. However, Indiana University, Terracotta's home institution, hosts a multi-tenant service in the Amazon Web Services (AWS) Cloud, with elastic scaling and industrial-grade security, currently provided free of charge to US-based districts and institutions who are interested in supporting experimental education research.

Terracotta data and vocabulary

At the conclusion of an experiment in Terracotta, a de-identified data export can be downloaded from the Terracotta web interface as a zipped archive. This archive contains a set of CSV (comma-separated value) files that describe all aspects of the experiment, as well as a JSON (JavaScript object notation) file that contains participants' timestamped interactions with Terracotta as an event stream. The JSON file is formatted according to the Caliper standard (1EdTech, 2022b). However, to our knowledge, there is currently no common data format for representing complex experimental research studies, and thus, the CSVs contained in Terracotta's data export adopt a novel data structure. Some of these elements can be mapped to the Common Education Data Standards (US Department of Education, 2022), and to facilitate this mapping, an alignment tool is provided with Terracotta's data dictionary. Key Terracotta vocabulary is described in Table 1, along with specific examples from our Demonstration Study, below. Many of the concepts in Terracotta's data vocabulary will be familiar to those with experience with LMS assignments and experimental research. However, we introduce one novel concept to organize

and articulate the experiment's structure: an exposure set.

Table 1 Terracotta vocabulary

Full size table

An exposure set is a set of assignments in which a participant experiences one condition. In a between-subjects design, there will be only one exposure set, because each participant will only experience one condition throughout the entirety of the study. However, in a within-subjects design, participants will experience the same number of exposure sets as the number of conditions. For example, imagine a within-subjects design that has two conditions, A and B, and that has four separate assignments with one crossover: half the participants will experience AABB, and the other half will experience BBAA. In the language of repeated measures designs, this experiment has four periods, because there are four different treatment opportunities. However, this experiment only has two different exposure sets, because there are two different sets of assignments corresponding to the two conditions: one exposure set for the first two assignments, and another exposure set for the latter two assignments. Within any exposure set, a researcher can add multiple class assignments (and while it is often desirable to balance the number of treatments in each exposure set, it is also possible for them to be imbalanced in Terracotta), and Terracotta will ensure that the right students see the right versions of each assignment (according to the student's treatment condition in that exposure set).

The advantage of structuring an experiment around an exposure set is that the researcher can specify experimental outcomes at the level of the exposure set, rather than for each period. For example, let us imagine that in the AABB/BBAA experiment above, the first two assignments both focus on mitosis (exposure set 1), and then the subsequent two assignments both focus on meiosis (exposure set 2). To contrast the effect of A and B on students' understanding of these concepts on a later class exam, it should only be necessary to measure students' knowledge

separately in each of the two exposure sets. In other words, the researcher should only need to measure two scores (questions about mitosis and questions about meiosis), and it would not be necessary to have one outcome measure for each of the four assignments separately. However, if it is desirable, Terracotta allows many outcome measures to be added to an exposure set.

Limitations of Terracotta

By manipulating LMS assignments, Terracotta enables experimentation on a wide range of student learning activities and interventions, and these represent important targets for education research. At this time, a Terracotta assignment can contain multiple-choice questions, short answer questions, and file upload response formats, which enables research ranging from well-structured tasks (how students learn science facts) to ill-structured tasks (how students learn literary argumentation; McCarthy et al., 2022). The assignments themselves can contain rich text, links, images, and embedded media. Further, Terracotta enables experimental manipulation of submission policies, grading policies, and feedback policies on these assignments. Nevertheless, Terracotta's scope is limited to the LMS, and this clearly restricts the range of educationally relevant variables that can be manipulated by Terracotta. Additionally, Terracotta's requisite LTI integration with the LMS can also present an obstacle to adoption and recruitment.

LTI tools, like Terracotta, typically require administrative support and endorsement before they can be integrated with a district or institution's LMS—teachers and researchers may request that a tool should be integrated, but the integration is typically approved and managed by administrators. For this reason, schools, districts, and institutions are gatekeepers, and may need to be convinced of the benefits of Terracotta, of embedded experimentation, and of Terracotta's commitment to security and privacy. These are important and beneficial conversations to have, but these may limit the speed or scale with

which a researcher might deploy a study. So, while Terracotta automates many of the mechanics of an experimental research study, it still relies on researchers, teachers, and administrators to form partnerships. Once such partnerships are made, however, Terracotta minimizes the effort involved in carrying out a rigorous and responsible experimental research study within a formal education setting.

Unlike laboratory research where participants are typically isolated from one another, student participants in a Terracotta experiment are classmates who are not isolated from one another. Students do communicate modestly about schoolwork with their classmates outside of class, commonly by sharing answers, artifacts, and summaries (Asterhan & Bouton, 2017; Bouton et al., 2021). If participants communicate about experimentally manipulated assignments, and this communication exposes them to treatments that were outside their assigned condition, contamination has occurred. Cross-treatment contamination is nothing new in education research (Cook, 2007), and while Terracotta differentiates the treatments that students can access in the LMS, it cannot prevent students from talking with one another. This possibility reinforces the importance of being transparent with student participants: letting them know that, should they agree to participate in a research study within Terracotta, they may have slightly different learning experiences than their classmates, and that they should avoid talking with each other about these experiences. Nevertheless, should cross-treatment contamination occur, this will blur the intended contrast between conditions, and at worst, the consequence would be an underestimate of the effect of an experimental manipulation. In general, researchers should be aware that experimental control is more challenging in the real world, and that there is a risk of observing smaller effect sizes than in the laboratory (Hulleman & Cordray, 2009; Vanhove & Harms, 2015), although sometimes such differences are not observed (Mitchell, 2012).

Demonstration Study

Terracotta makes it possible to experimentally manipulate consequential aspects of students' educational experiences, to embed complex crossover designs, and to collect streamlined data on these manipulations and their effects. To demonstrate these features, we used Terracotta to conduct a preregistered replication of McDaniel et al. (2012), a well-cited experimental demonstration of the benefits of retrieval practice in a college class using authentic class materials.

In learning contexts, retrieval is the process of accessing knowledge—getting the information out of memory—and it is often associated with quizzes or exams. Although these activities are frequently used to measure how much a student has already learned, the retrieval of information on quizzes or exams may also produce learning, not just measure it (e.g., Roediger & Karpicke, 2006). The act of "getting the information out" requires mental effort and the reconstruction of knowledge, which can lead to robust learning (Roediger & Butler, 2011). Many studies have demonstrated that retrieval practice improves long-term retention, relative to re-reading the same material (Agarwal et al., 2021; Dunlosky et al., 2013; Moreira et al., 2019; Yang et al., 2021). But the McDaniel et al. (2012) study, in particular, had key features that make it ideal for demonstrating Terracotta's capabilities: it manipulated the format of online quiz assignments, it used a within-subjects design, it had repeating treatment periods with multiple crossovers, and it measured the effect of these treatments on students' subsequent exam performance—features that are all supported by Terracotta.

Method

This demonstration study was approved by the Indiana University Institutional Review Board (IRB) and was publicly preregistered prior to data collection at https://osf.io/juq7n/. All materials, data, and analyses are publicly available at https://osf.io/yrbhe/.

Education context

We embedded this study in one section of PSY-P335 Cognitive Psychology during the Fall 2022 academic term at Indiana University Bloomington. P335 is a required course for undergraduate students majoring in psychology. This was an in-person full semester (16-week) section, which had weekly online "reviews" throughout the term, intended to help students learn the material prior to taking in-class exams. The current study was implemented during the second half of the semester (weeks 8 through 15), which is when key features of the current method became available in Terracotta (e.g., cumulative grading for multiple submissions, see Procedure below). Total enrollment in the course was 106 students. This course used the Canvas LMS (Instructure, Inc.; Salt Lake City, UT), into which Terracotta was integrated. The instructor of this course is an author of the current study.

Participants

Using IRB-approved in-person and email announcements, the instructor invited students to volunteer to participate during the seventh week of the semester. Students were invited to provide consent in an online assignment entitled "Invitation to Participate in a Research Study" (see Fig. 1G), and they were encouraged to complete the assignment within a week; specifically, before the first manipulated review assignment. All students received a small amount of course credit for responding to this assignment prior to the deadline, regardless of whether they chose to provide consent or not.

Ultimately, 77 students submitted responses to this assignment, and among those who provided a response, 39 provided affirmative consent, and these students are considered as participants. Students who did not agree or who did not respond to this assignment were considered nonparticipants, and are excluded from further analysis. The number of students who provided consent was lower than we anticipated in our preregistration. This may have been because consent was administered mid-semester, and students may have been less inclined toward any modification

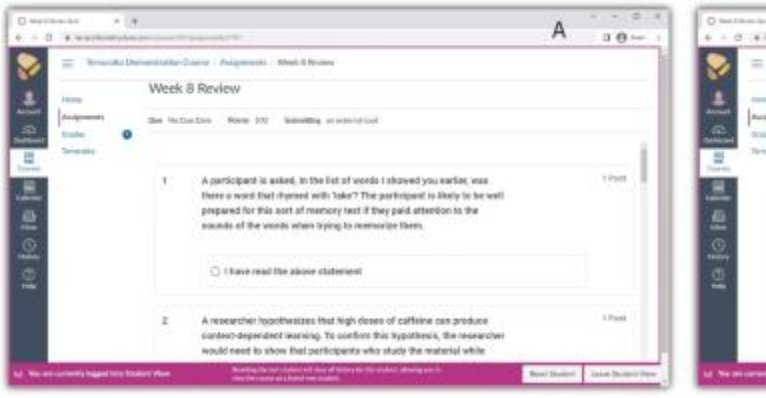
of an established routine. Terracotta does not provide opportunities to directly incentivize students to provide consent (e.g., giving extra credit points only to students who agree to participate), as this would disclose private consent responses to the teacher and possibly create inequities. More research is needed to examine students' consent decisions and how to increase participation in experimental education research.

Materials

The instructor of the course created weekly online reviews for students to revise the course material from a given week. Eight reviews (Reviews 8–15) were included in the current study, and these reviews targeted the following areas of cognitive psychology: Memory (Reviews 8 and 9), Concepts (Reviews 10 and 11), Language (Reviews 11 and 12), Mental Imagery (Review 12), Judgment and Reasoning (Review 13), and Intelligence (Reviews 14 and 15). All reviews contained ten items, with the exception of the last two reviews, which contained six and seven items, respectively.

There were two versions of each review, corresponding to the two conditions: retrieval practice and restudy. If a participant was assigned to the retrieval practice condition on a review, they answered multiple-choice questions; if they were assigned to the restudy condition on the same review, they read correct statements that were the answers to the questions in the retrieval practice version (see Fig. 2).

Fig. 2



Screenshots of example assignments in Terracotta. The left panel (A) shows a Restudy version of the Week 8 Review, and the right panel (B) shows a Retrieval Practice version of the Week 9 Review

Full size image

Three items from each review were selected to appear on the course exam at the end of a unit. Selected items from Reviews 8 and 9 appeared on Exam 2, selected items from Reviews 10–13 appeared on Exam 3, and selected items from Reviews 14 and 15 appeared on Exam 4 (the final exam). Thus, there were a total of 24 multiple-choice questions that were repeated verbatim from retrieval practice versions of the reviews across three course exams. All review items (from retrieval practice and restudy versions both) and a list of those items that appeared on course exams can be found at https://osf.io/yrbhe/.

Procedure

During the seventh week of the term, students completed an online assignment that asked for their consent to participate in the current study, and during the eighth week, participants received their first manipulated review (Review 8) as retrieval practice (quiz) or restudy. Participants had a new manipulated review every week until the end of the term, and the format of these alternated on a weekly basis (e.g., retrieval practice one week, restudy the next week, retrieval practice the following week). Students who did not provide consent completed retrieval practice reviews (quizzes) for the remainder of the semester. This was consistent with how reviews were implemented in the first half of the semester (prior to the start of the current study), where all students answered multiple-choice questions on these reviews.

There were eight reviews included in the current study (Reviews 8–15), and each contained ten items, except for Reviews 14 and 15 (which had six and seven items, respectively). Items corresponded to multiple-choice questions in retrieval practice reviews and to correct statements in restudy reviews. Therefore, if participants were assigned a retrieval practice review, they answered multiple-choice questions. If, however, they were assigned a restudy review, they read correct statements that were the answers to the corresponding questions in the retrieval practice review. To ensure their engagement in the

task, participants assigned a restudy review checked a box after each statement, verifying that they had read it.

We created two exposure sets on Terracotta with four reviews each: the first exposure set included Reviews 8, 10, 12, and 14, and the second exposure set included Reviews 9, 11, 13, and 15. Participants were assigned to the retrieval practice condition for reviews in one exposure set, and to the restudy condition for reviews in the other exposure set. That is, for roughly half of the participants, the first exposure set made up the retrieval practice condition and the second exposure set made up the restudy condition, and this was reversed for the remaining half of the participants. Thus, all participants were given four retrieval practice reviews and four restudy reviews, and review format alternated on a weekly basis.

Aside from the difference in review format for participants (retrieval practice or restudy), reviews had the same structure for all students enrolled in the course for the duration of the current study. The instructor posted a given week's review on Canvas each Friday, which students had to complete by class time the following TuesdayFootnote2. All review items were presented on the same page and in the same order. Students could complete a review up to four times, and they had unlimited time on each attempt. Once they submitted an attempt, students were provided the correct answer and their accuracy on each item (though these were informative only on the retrieval practice reviews). This correct-answer feedback feature, however, was not available on Terracotta until the third week of the current study; instead, students were only provided their accuracy on each item for Reviews 8 and 9. Between attempts, students were required to wait at least two hours to prevent completion of these attempts back-to-back. Students earned 2.5 points each time they completed an attempt, and they could earn a maximum of 10 points if they completed all four attemptsFootnote3. That is, grading of reviews was cumulative based on number of

completions, rather than accuracy of responses. The cumulative grading feature became available on Terracotta starting the second week of the current study; until then, students' grades were adjusted manually to reflect the number of completed attempts. Reviews (including the ones from the first half of the term) made up about 20% of students' final grades in the course, and all were based on completion (not accuracy).

Over the course of the academic term, students took four exams (three unit exams, and one cumulative final exam), which included some questions repeated from the reviews. The time between a review and a course exam varied, such that the retention interval could have ranged from less than a day (for the last review due prior to an exam) up to 34 days (for the first review following a course exam). Three items from each of Reviews 8–15 appeared on the last three exams. Specifically, items selected from Reviews 8 and 9 appeared on Exam 2, items from Reviews 10–13 appeared on Exam 3, and items from Reviews 14 and 15 appeared on Exam 4 (the final exam). Thus, there were a total of 24 multiple-choice review items repeated on course exams, and scores on these questions were used to compare the mnemonic benefits of retrieval practice and restudy.

Of note, the instructor made all multiple-choice review items (and the answer key to these review items) available to students prior to a course exam. This was done to facilitate exam preparation and to ensure that all students had equitable access regardless of experimental treatment; and the instructor encouraged students to incorporate these practice questions to their study. Put differently, even if participants had experienced some restudy reviews, they still had access to the retrieval practice version before each exam. Considering that this would result in cross-treatment contamination, we expect that this study would yield a more muted estimate of the potential benefits of retrieval practice on memory retention and performance.

Students' review responses were automatically graded in

Terracotta and students' in-class exam responses were graded using Akindi (a web-based system that automates grading of multiple-choice assessments). After data collection ended, we aggregated scores on the 24 exam questions (those that were repeated from reviews) by student and for each review. Because there were three multiple-choice items repeated from each review, values ranged from 0 to 3, and students who missed a course exam did not have a calculated score corresponding to some of the reviews. We created eight outcomes on Terracotta (four for each exposure set) to manually enter the scores described above. We chose to create an outcome for each review, rather than an outcome for each of the two conditions (retrieval practice and restudy), to allow the analysis of participant behavior across the eight reviews.

Statistical analysis

We opted to use Bayesian estimation methods for statistical analyses in the current study. Bayesian estimation provides a framework for making inferences about experimental effects, given observed data and our prior assumptions about these effects. The general advantages of Bayesian inference have been discussed elsewhere (Kruschke, 2011; Vandekerckhove et al., 2018), but the specific benefits for this study include the ability to define a custom analytical model appropriate to the structure of the observed data (e.g., a hierarchical within-subject logistic model) and the ability to deal with unbalanced data (not all students complete all assignments). Also, rather than merely yielding a p-value, Bayesian estimation methods produce an informative posterior distribution. The posterior distribution is a direct estimate of the tendency and uncertainty of a parameter in an analytical model, given the observed data and the priors. In this study, despite having knowledge of the expected effect (McDaniel et al., 2012), we elected to use uninformed and vague priors (wide normal distributions centered on zero), so that our replication would provide convincing evidence for skeptical

audiences. We characterize the posterior distribution by its modal estimate and by the 95% highest density interval (HDI), which is the range of the most likely parameter values. If the 95% HDI does not include zero or values close to zero, we may infer a credibly nonzero experimental effect.

We sampled the posterior distribution using JAGS (Plummer, 2003) and the runjags package (Denwood, 2016) for R. This was performed using Markov chain Monte Carlo (MCMC) sampling with four independent chains each sampled for at least 30,000 iterations and thinned to every fifth step, following 500 adaptation steps and 1000 burn-in steps. For all parameters of interest, the Gelman–Rubin $\hat{R}$ statistic (Gelman & Rubin, 1992) was less than 1.01, and the effective sample size (ESS) was greater than 20,000. Detailed model specifications are available at https://osf.io/b7cwa.

Results and discussion

Treatment characteristics

Each participant was assigned to complete eight different reviews; four were retrieval practice (quizzes) and four were restudy. On average, participants completed 7.74 reviews (SD = 0.44); 3.90 were retrieval practice (SD = 0.31) and 3.85 were restudy (SD = 0.37). Further, participants were incentivized to complete these reviews multiple times with accumulating completion credit up to four submissions. On average, participants made 3.52 submissions to each review (SD = 0.43); 3.53 for retrieval practice (SD = 0.60), and 3.5 for restudy (SD = 0.50). Thus, there was rough equivalence in the number of times participants were exposed to reviews in the two conditions.

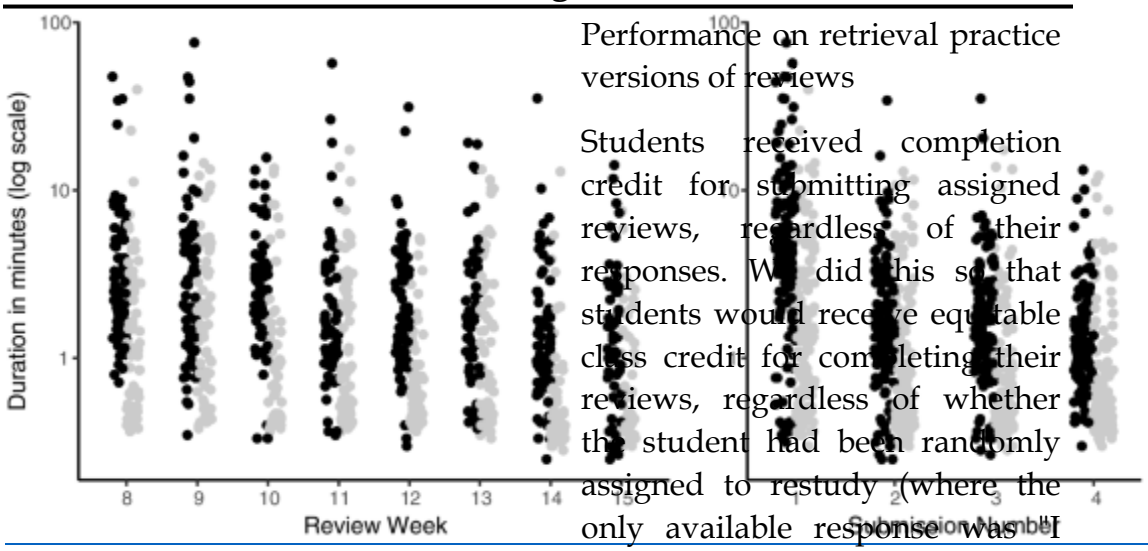
However, participants spent more time on reviews when they were in the retrieval practice condition than in the restudy condition. Participants spent an average of 3.71 minutes (SD = 2.73) per attempt on retrieval practice reviews, and 1.87 minutes (SD = 1.50) per attempt on restudy reviews (bear in mind, however, that attempt duration data have a large positive skew).

In an exploratory analysis, we estimated the parameters of a hierarchical linear model, with the log-transformed duration of each submission as the dependent variable. There were three independent variables: condition (retrieval practice and restudy), assignment (eight different assignments), and submission number (up to four submissions per assignment); coefficients for these variables were estimated for individual subjects and at the group level.

The group-level effect of condition was credibly greater than zero (estimate: 0.35; 95% HDI: 0.25 to 0.44), confirming that participants spent more time on retrieval practice reviews than restudy reviews. McDaniel et al. (2012) did not report time on task; however, laboratory research has similarly observed more time spent completing retrieval practice activities relative to analogous restudy activities when time constraints are not imposed (Üner & Roediger, 2018). We are cautious to interpret the additional time spent on retrieval practice reviews as a source of potential memory benefits, as past laboratory research has shown that additional time spent on a task does not always enhance learning, particularly on rereading tasks (Callender & McDaniel, 2009; Rawson & Kintsch, 2005).

The effect of assignment on submission duration was credibly lower than zero, indicating that students spent less time on reviews as the semester progressed (estimate: −0.30; 95% HDI: −0.37 to −0.22). And additionally, the effect of submission number was credibly lower than zero, indicating that students tended to spend less time on each subsequent submissions of the same quiz (estimate: −0.32; 95% HDI: −0.40 to −0.24). These effects are shown in Fig. 3.

Fig. 3



Time spent on reviews. Each dot is an individual student submission, with minor horizontal jitter added. Retrieval practice submissions are shown in black, and restudy submissions are shown in gray. Overall, students spent more time completing retrieval practice reviews than restudy reviews, as is evident in both panels (black dots are higher than grey dots). The left panel shows the duration spent on each individual quiz assignment, with decreasing time across over the semester. The right panel shows the duration spent on each submission, showing decreasing time on each subsequent submission

Full size image

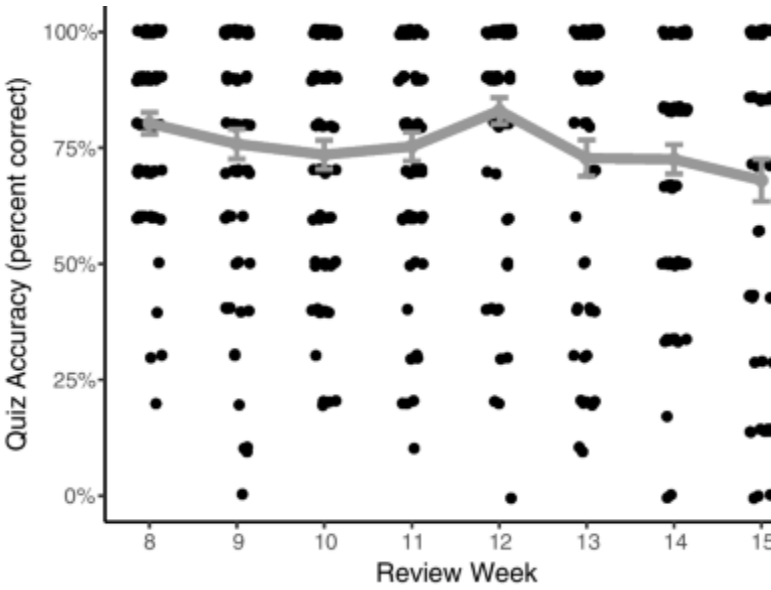
Performance on retrieval practice versions of reviews

Students received completion credit for submitting assigned reviews, regardless of their responses. We did this so that students would receive equitable class credit for completing their reviews, regardless of whether the student had been randomly assigned to restudy (where the only available response was "I have read the above statement") or retrieval practice (which had four different response options, one of which was correct). However, even while students received the same credit for any response, Terracotta still stored students' selections on multiple-choice questions, and it is possible to analyze the accuracy of these responses.

Averaging across all submissions in the retrieval practice condition, participants got 75.4% of items correct (SD = 27.0%). To examine how accuracy in the retrieval practice condition changed over the semester, and also how accuracy changed as students made multiple submissions to each assignment, we again conducted an exploratory

analysis using a hierarchical linear model, with a logistic response variable (number of correct responses out of number of questions). We found that there was no credible linear change in accuracy over the course of the semester (estimate: −0.07; 95% HDI: −0.25 to 0.11). However, students' accuracy tended to increase in repeated submissions of each review (estimate: 0.88; 95% HDI: 0.69 to 1.08), suggesting that students in the retrieval practice condition were using the reviews as an opportunity to learn and improve, despite receiving full credit for any submission. These data are shown in Fig. 4.

Fig. 4



Accuracy on reviews in the retrieval practice condition (quizzes). Each dot is an individual student submission, with minor horizontal jitter added. Gray markers indicate the mean accuracy, and error bars indicate ± one standard error. The left panel shows accuracy on each individual quiz assignment. The right panel shows the accuracy on each submission, with improving accuracy on each subsequent submission

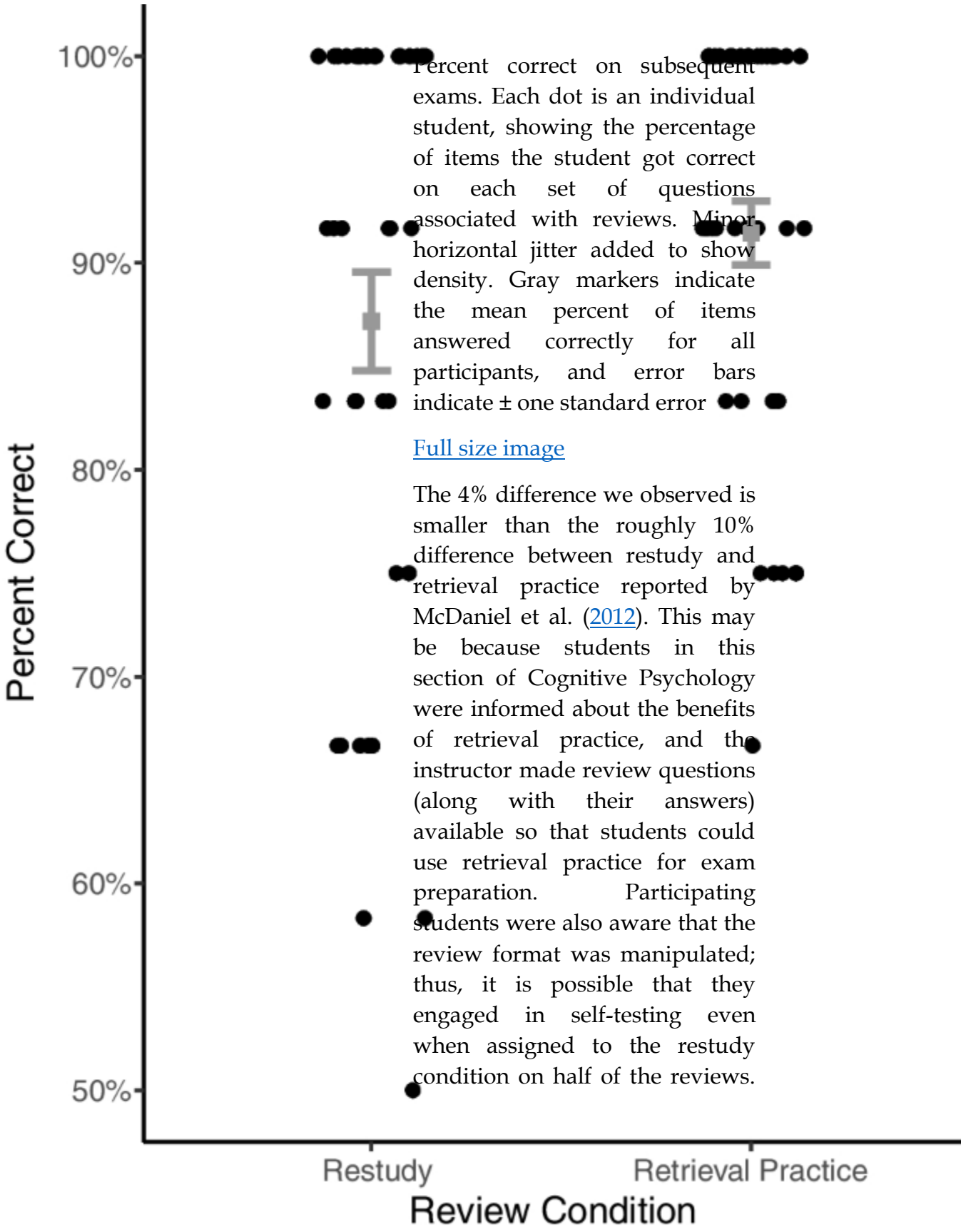
Full size image

Learning outcomes

Three questions from each individual review were included

in subsequent in-class exams. Student's scores on these questions, specific to their associated review condition (retrieval practice or restudy), were manually added to Terracotta by the instructor. Overall, participants performed well on these questions, getting an average of 2.68 correct out of 3 (89.3%; SD = 0.32). However, students were more likely to answer questions correctly if they had previously been included in retrieval practice reviews (2.74 correct; 91.5%; SD = 0.29), compared with questions previously included in restudy reviews as correct statements (2.62 correct; 87.2%; SD = 0.45), and our preregistered analysis estimated this difference between retrieval practice and restudy conditions to be credibly greater than zero (estimate: 0.51; 95% HDI: 0.015 to 1.05), as shown in Fig. 5.

Fig. 5



Percent correct on subsequent exams. Each dot is an individual student, showing the percentage of items the student got correct on each set of questions associated with reviews. Minor horizontal jitter added to show density. Gray markers indicate the mean percent of items answered correctly for all participants, and error bars indicate ± one standard error

Full size image

The 4% difference we observed is smaller than the roughly 10% difference between restudy and retrieval practice reported by McDaniel et al. (2012). This may be because students in this section of Cognitive Psychology were informed about the benefits of retrieval practice, and the instructor made review questions (along with their answers) available so that students could use retrieval practice for exam preparation. Participating students were also aware that the review format was manipulated; thus, it is possible that they engaged in self-testing even when assigned to the restudy condition on half of the reviews.

Furthermore, the reduced effect size in the current study may also be due to differences in the scope of the studies' implementations, as the current study manipulated eight quizzes and measured outcomes on 24 exam questions, whereas the original study manipulated 15 quizzes and measured 84 exam questions. Furthermore, given the low consent rate observed in the current study (~37%), sample bias and ceiling effects may be affecting these estimates (exam scores in the current study are higher than in McDaniel et al., 2012). Despite these considerations, however, a 4% difference in performance on exam questions is non-negligible, and demonstrates that manipulations within Terracotta can have meaningful consequences for student learning.

Summary of demonstration study

Successfully replicating McDaniel et al. (2012), we observed a credible improvement in exam performance in an authentic education setting, after participants had practiced retrieving the material (retrieval practice) compared with rereading the material (restudy). This experimental manipulation took place entirely within Terracotta, after students were informed about the research in full transparency and given agency in deciding whether to participate.

We do not claim that this demonstration study is emblematic of all research that could be conducted within Terracotta. Terracotta's feature set enables a wide range of possible experimental manipulations and designs, so that any single study would be unrepresentative on its own. Nevertheless, beyond its successful replication, two things are particularly noteworthy about this demonstration: (1) the ease with which we were able to embed this experiment into the class, and (2) Terracotta's ability to automatically collect granular data on student behaviors when interacting with manipulated assignments. As to the ease of embedding, we estimate that experiment setup took a total of 2 minutes, assignment construction took roughly 10 minutes per assignment, and manual entry of

outcomes took about an hour. A video of the setup and creation of two assignments is available at https://osf.io/24qp7.

Anecdotally, for an embedded experiment with informed consent, a within-subject design, eight treatments, alternating crossovers, and exam scores mapped to treatments, this represents a dramatic time savings.

As to Terracotta's granular data collection, the current study extends past research conducted in laboratory settings (Üner & Roediger, 2018), observing in particular that participants in authentic education settings spend more time performing retrieval practice activities compared with restudy activities. This is noteworthy because, in the current study, participants in the restudy condition had academic incentives to learn the material, which is not necessarily the case in the lab. Indeed, supplemental reading in authentic education settings is correlated with improved performance (Carvalho et al., 2018). Nevertheless, the current study demonstrates that assignments to restudy result in quantitively less time on task, and less educational benefit, than assignments to practice retrieving the material. Analyses of time spent on self-regulated learning activities are potentially fruitful avenues for future research (Carvalho et al., 2022; Knight et al., 2017; Son & Kornell, 2009).

We are careful to label this a demonstration study and not a validation study. As with any behavioral research tool, measurement validity in Terracotta is principally determined by how it is used and is not an invariant property of the tool itself. A hammer might be a valid means for driving a nail, but not for turning a screw—and similarly, not all education research studies are suitable for implementation in Terracotta, or in any single tool. Nevertheless, the current study demonstrates that Terracotta can be used to experimentally manipulate consequential aspects of students' experiences in education settings, to embed complex crossover designs, and to collect streamlined data on these manipulations and their effects.

Conclusion

Terracotta enables a wide range of experimental research manipulations and designs, which aim to match the wide range of instructional decisions and interventions that might be implemented with LMS assignments. We have demonstrated how Terracotta might be used to easily test a cognitive manipulation of LMS assignments, but there are other research approaches in education that can also benefit from having Terracotta in their methodological toolkits. For example, Terracotta might also be used to test social and motivational interventions, and to test the effectiveness of learning resources and instructional strategies that are presented to students in assignments. Moreover, by eliminating many of the difficulties of implementing an experiment in a single class, we further hope that Terracotta might be used to deploy experiments that are distributed across many classes, thus building our understanding not only of what works in education, but also where it works (Churches et al., 2020; de Leeuw et al., 2022; Fyfe et al., 2021).

Terracotta is a research tool and is not intended for routine practice—which means that once an effective learning strategy is identified within Terracotta, it will require additional work to disseminate this finding and to improve practice more broadly. How research findings can be used to affect routine practice is a steep challenge in education (and in all disciplines) that is unlikely to be addressed by any single method or tool. Nevertheless, Terracotta might make inroads. By lowering the barriers to experimental research in authentic class settings, and by involving teachers in the practice of experimental research, the distance segregating research findings from education practice may be reduced (Mace & Critchfield, 2013; National Research Council, 1999).

Terracotta's specific goal is to lower the practical barriers to easy, accessible, responsible, and rigorous experimental research across education levels, student populations, and learning materials. By allowing

researchers to embed studies in the LMS, where many learning activities already take place, Terracotta can help advance our understanding of what works in student learning on a broad scale and help build stronger evidence of how to improve education.

References

1EdTech (2022a). 1EdTech LTI 1.3 and LTI Advantage. From https://www.imsglobal.org/activity/learning-tools-interoperability. Accessed 16 June 2023.

1EdTech (2022b). Caliper Analytics. From https://www.imsglobal.org/activity/caliper

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. Journal of Applied Psychology, 334-345. https://doi.org/10.1037/0021-9010.69.2.334

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. Educational Psychology Review, 33, 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Asterhan, C. S., & Bouton, E. (2017). Teenage peer-to-peer knowl- edge sharing through social network sites in secondary schools. Computers & Education, 110, 16–34. https://doi.org/10.1016/j.compedu.2017.03.007

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning Engineering: A View on Where the Field Is at, Where It's Going, and the Research Needed. Technology, Mind, and Behavior, 3(1). https://doi.org/10.1037/tmb0000058

Benassi, V. A., Overson, C., & Hakala, C. M. (2014). Applying science of learning in education: Infusing psychological science into the curriculum. Society for the Teaching of Psychology. From http://teachpsych.org/ebooks/asle2014/index.php. Accessed 16 June 2023.

Borman, G. D. (2002). Experiments for educational evaluation and improvement. Peabody Journal of Education, 77(4), 7–27 From

https://www.jstor.org/stable/1493216

Bouton, E., Tal, S. B., & Asterhan, C. S. (2021). Students, social network technology and learning in higher education: Visions of collaborative knowledge construction vs. the reality of knowledge sharing. The Internet and Higher Education, 49. https://doi.org/10.1016/j.iheduc.2020.100787

Boyer, E. L. (1990). Scholarship reconsidered: Priorities of the profes- sorate. The Carnegie Foundation for the Advancement of Teaching. Brady, A. C., Griffin, M. M., Lewis, A. R., Fong, C. J., & Robinson,

D. H. (2022). The increasing trend of inferring causality from correlation in educational psychology journals. OSF Preprints. https://doi.org/10.31219/osf.io/24dfm

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. Contemporary Educational Psychol- ogy, 34(1), 30–41. https://doi.org/10.1016/j.cedpsych.2008.07.001

Carvalho, P. F., Gao, M., Motz, B. A., & Koedinger, K. R. (2018). Analyzing the relative learning benefits of completing required activities and optional readings in online courses. Proceedings of the 11th International Conference on Educational Data Min- ing (pp. 418-423). Buffalo, NY: International Educational Data Mining Society.

Carvalho, P. F., McLaughlin, E. A., & Koedinger, K. R. (2022). Varied practice testing is associated with better learning outcomes in self- regulated online learning. Journal of Educational Psychology, 114(8), 1723–1742. https://doi.org/10.1037/edu0000754

Churches, R., Dommett, E. J., Devonshire, I. M., Hall, R., Higgins, S., & Korin, A. (2020). Translating laboratory evidence into class- room practice with teacher-led randomized controlled trials: A perspective and meta-analysis. Mind, Brain, and Education, 14(3), 292–302. https://doi.org/10.1111/mbe.12243

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research.

Educational evaluation and policy analysis, 25(2), 119–142. https://doi.org/10.3102/01623737025002119

Cook, T. D. (2007). Randomized experiments in education: Assessing the objections to doing them. Economics of Innovation and New Technology, 16(5), 331–355. https://doi.org/10.1080/10438590600982335

Daniel, D. B. (2012). Promising principles: Translating the science of learning to educational practice. Journal of Applied Research in Memory and Cognition, 1(4), 251–253. https://doi.org/10.1016/j.jarmac.2012.10.004

de Leeuw, J. R., Motz, B. A., Fyfe, E. R., Carvalho, P. F., & Goldstone,

R. L. (2022). Generalizability, transferability, and the practice-to- practice gap. Behavioral and Brain Sciences, 45, e11. https://doi.org/10.1017/S0140525X21000406

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

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willing- ham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and edu- cational psychology. Psychological Science in the Public Interest, 14(1), 4–58. https://doi.org/10.1177/1529100612453266

Exempt Research, § 46.104(d)(1). (2018). US Department of Health and Human Services.

Flick, C. (2015). Informed consent and the Facebook emotional manip- ulation study. Research Ethics, 12(1), 14–28. https://doi.org/10.1177/1747016115599568

Fyfe, E. R., de Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sher- man, J., Admiraal, D., et al. (2021). ManyClasses 1: Assessing the generalizable effect of immediate feedback versus

delayed feedback across many college classes. Advances in Methods and Practices. Psychological Science, 4(3). https://doi.org/10.1177/25152459211027

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

Goel, V. (2014). Facebook tinkers with users' emotions in news feed experiment, stirring outcry. New York Times. From https://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users- emotions-in-news-feed-experiment-stirring-outcry.html. Accessed 16 June 2023.

Google. (2022). Material Design. From https://m3.material.io/ Greenwald, A. G. (1976). Within-subjects designs: To use or not to

use? Psychological Bulletin, 83(2), 314–320. https://doi.org/10.1037/0033-2909.83.2.314

Gueron, J. M. (2002). The politics of random assignment implement- ing studies and impacting policy. In F. Mosteller, & R. F. Boruch, Evidence matters: Randomized trials in education research (pp. 15-49).

Guskey, T. R., & Huberman, M. (1995). Professional Development in Education: New Paradigms and Practices. Teachers College Press.

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., Wood, W. B. (2004). Scientific teaching. Science, 304(5670), 521-522. https://doi.org/10.1126/science.1096022

Harackiewicz, J. M., & Priniski, S. J. (2018). Improving student outcomes in higher education: The science of targeted interventions. Annual Review of Psychology, 69, 409–435. https://doi.org/10.1146/annurev-psych-122216-011725

Herold, B. (2018). Pearson tested "social-psychological" messages in learning software, with mixed results. New York: EducationWeek. From https://www.edweek.org/technology/pearson-tested-social-psychological-messages-in-learning-software-with-mixed-

results/ 2018/04. Accessed 16 June 2023.

Hill, P. (2020). LMS Market Acceleration: An initial view in North America. PhilOnEdTech https://philonedtech.com/lms-market- acceleration-an-initial-view-in-north-america/. Accessed 16 June 2023.

Hsieh, P., Acee, T., Chung, W. -H., Hseih, Y. -P., Kim, H., Thomas, G. D., Robinson, D. H. (2005). Is educational intervention research on the decline? Journal of Educational Psychology, 97(4).

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

Joyce, K. E., & Cartwright, N. (2019). Bridging the gap between research and practice: Predicting what will work locally. Ameri- can Education Research Journal, 57(3), 1045–1082. https://doi.org/10.3102/0002831219866687

Knight, S., Wise, A. F., & Chen, B. (2017). Time for change: Why Learning Analytics needs temporal analysis. Journal of Learn- ing Analytics, 4(3), 7–17. https://doi.org/10.18608/jla.2017.43.2

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

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. Psychonomic Bulletin & Review, 14, 219–224. https://doi.org/10.3758/BF03194055

Kruschke, J. K. (2011). Doing Bayesian data analysis: A tutorial with R and BUGS. Academic Press.

Levin, J. R. (2005). Randomized classroom trials on trial. In G. D. Phye,

D. H. Robinson, & J. R. Levin (Eds.), Empirical Methods for Evalu- ating Educational Interventions (pp. 3–27). Academic Press.

Li, W., Sun, K., Schaub, F., & Brooks, C. (2022). Disparities in stu- dents' propensity to consent

to learning analytics. International Journal of Artificial Intelligence in Education, 32, 564–608. https://doi.org/10.1007/s40593-021-00254-2

Lonn, S., & Teasley, S. (2009). Saving time or innovating practice: Investigating perceptions and uses of Learning Management Sys- tems. Computers & Education, 53(3), 686–694.

Mace, F. C., & Critchfield, T. S. (2013). Translational research in behavior analysis: Historical traditions and imperative for the future. Journal of the Experimental Analysis of Behavior, 93(3), 293–312. https://doi.org/10.1901/jeab.2010.93-293

McCarthy, K. S., Crossley, S. A., Meyers, K., Boser, U., Allen, L. K., Chaudhri, V. K., et al. (2022). Toward more effective and equita- ble learning: Identifying barriers and solutions for the future of online education. Technology, Mind, & Behavior, 3(1). https://doi.org/10.1037/tmb0000063

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web- based class: An experimental study. Journal of Applied Research in Memory and Cognition, 1(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychology laboratory. Perspectives on Psy- chological

Science, 7(2), 109–117. https://doi.org/10.1177/17456

91611432343

Moreira, B. F., Pinto, T. S., Starling, D. S., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. Frontiers in Education, 4. https://doi.org/10.3389/feduc.2019.00005

Mosteller, F., & Boruch, R. (2002). Randomized Trials in Education Research. The Brookings Institution.

Motz, B. A., Bergner, Y., Brooks, C. A., Gladden, A., Gray, G., Lang, C., Quick, J. D. (2023). A LAK of direction: Misalignment between the goals of learning analytics and its research scholarship. Journal of Learning Analytics. https://doi.org/10.18608/jla.2023.7913

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

NAEP. (2022). National Assessment of Educational Progress. US Department of Education, National Center for Education Statis- tics From https://nces.ed.gov/nationsreportcard/. Accessed 16 June 2023.

National Research Council. (1999). Improving Student Learning: A Strategic Plan for Education Research and Its Utilization. The National Academies Press. https://doi.org/10.17226/6488

National Research Council. (2002). Scientific Research in Education. The National Academies Press. https://doi.org/10.17226/10236

National Science Foundation, & Institute for Education Sciences (2013). Common guidelines for education research and devel- opment. Washington, DC. From https://www.nsf.gov/pubs/2013/ nsf13126/nsf13126.pdf. Accessed 16 June 2023.

NCES (Ed.). (2022). Condition of Education. US Department of Edu- cation, National Center for Education Statistics From

https://nces.
ed.gov/pubsearch/pubsinfo.asp?pubid=2022144

Office of Educational Technology. (2017). Reimagining the Role of Technology in Education: 2017 National Educational Technology Plan Update. US Department of Education.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing instruction and study to improve student learning. Washington, DC: National Center for Education Research, Institute of Education Sciences, US Department of Education. From http://ncer.ed.gov. Accessed 16 June 2023.

PISA. (2020). Highlights of US PISA 2018 Results Web Report (NCES 2020-166 and NCES 2020-072). US Department of Educa-tion From https://nces.ed.gov/surveys/pisa/pisa2018/index.asp. Accessed 16 June 2023.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graph- ical models using Gibbs sampling. Proceedings of the 3rd Inter- national Workshop on Distributed Statistical Computing. Vienna. Pomerantz, J., & Brooks, D. C. (2017). ECAR Study of Faculty and Information Technology. Louisville, CO: EDUCAUSE Center

for Analysis and Research. From https://library.educause.edu/-/media/files/library/2017/10/facultyitstudy2017.pdf. Accessed 16

June 2023.

Pressley, M., Goodchild, F., Fleet, J., Zajchowski, R., & Evans, E.

D. (1989). The challenges of classroom strategy instruction. The Elementary School Journal, 89(3), 301–342. https://doi.org/10. 1086/461578

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. Journal of Educational Psychology, 97(1), 70–80. https://doi.org/10.1037/0022-0663.97.1.70

Riecken, H. W., & Boruch, R. F. (1974). Social Experimentation: A Method for Planning and Evaluating Social Intervention. Aca- demic Press.

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn,

S. (2007). The Incidence of "Causal" Statements in Teaching-and-Learning Research Journals. American Educational Research Journal, 44(2), 400–413.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. Trends in Cognitive Sciences, 15(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17, 249–255. https://doi.org/10.1111/j.1467-9280. 2006.01693.x

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. Journal of Applied Research in Memory and Cognition, 1(4), 242–248. https://doi.org/10.1016/j.jarmac.

2012.09.002

Schanzenbach, D. W. (2012). Limitations of experiments in education research. Education Finance & Policy, 7(2), 219–232. https://doi.org/10.1162/EDFP_a_00063

Schneider, M., & Garg, K. (2020). Medical researchers find cures by conducting many studies and failing fast. We need to do the same for education. The 74. From https://www.the74million.org/article le/schneider-garg-medical-researchers-find-cures-by-conducting- many-studies-and-failing-fast-we-need-to-do-the-same-for-educa tion/. Accessed 16 June 2023.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Sinclair, A. (2010). License profile: Apache License, Version 2.0. International Free and Open Source Software Law Review, 2(2), 107–114. https://doi.org/10.5033/ifosslr.v2i2.42

Slavin, R. E. (2002). Evidence-based education policies:

Transform- ing Educational Practice and Research. Educational Researcher, 31(7), 15–21.

Son, L. K., & Kornell, N. (2009). Simultaneous decisions at study: time allocation, ordering, and spacing. Metacognition and Learning, 4, 237–248. https://doi.org/10.1007/s11409-009-9049-1

Staker, H. (2011). The rise of K-12 blended learning: Profiles of emerg- ing models. San Mateo, CA: Innosight Institute. From https://eric. ed.gov/?id=ed535181

Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. Journal of Graduate Medical Education, 3(3), 285–289.