EENE Two-Stage Exams I2D Proposal

Project Overview

The purpose of this study is to investigate the extent to which two-stage exams result in improved learning, as measured by student retention of material over time, relative to standard course designs that employ single-stage exams. Although there is a large literature on the ability of two-stage exams to generate improved student retention of course content, most prior studies derive these results from relatively small samples in which they compare average scores of subgroups of students. Furthermore, among those that do employ crossover designs, they are unable to leverage the power of this experimental technique because the design is often implemented within a single section of a course in which the observations from individual students cannot be treated as independent.

Thus, despite a large literature investigating this pedagogical technique, we still do not know whether two-stage exams "work." If a two-stage exam "works," then it should result in additional student learning, above and beyond the learning they would have achieved individually based on their ability and preparation (Zipp 2007, p. 65). Knowing whether two-stage exams "work" in the sense defined by Zipp (2007) requires a study design that can examine the effect of two-stage exams on individual student performance while controlling for individual student ability and effort as well as the effects of clustering in the data.

This proposal aims to fill this gap in knowledge. Filling this gap will enable instructors to determine whether the potential gains in student learning from the intervention are worth the additional costs associated with its implementation. In a two-stage exam, students first complete an exam individually. After the individual exams are collected, students are assigned to groups and complete all or a portion of the same exam as a group. A student's score on the exam is a weighted average of the individual and second-stage scores, with a student's individual exam score typically constituting 80-90% of their final grade on the exam. In addition, weighting schemes often include a correction so that students' grades are not lowered by taking the group exam. Two-stage exams are hypothesized to improve learning through a number of mechanisms. First, the use of the two-stage exam format leverages the testing effect (Roediger and Karpicke 2006), a well-established result from cognitive psychology that demonstrates that individuals learn and retain information better if they are tested on it early and repeatedly after initially encoding the material. Additionally, this format may help student learning by reducing test anxiety and providing an opportunity for self-assessment with immediate feedback during the group exam. We summarize the literature on two-stage exams in more detail below.

This project will evaluate whether a two-stage exam improves retention of class content (as measured by performance on selected final exam questions) relative to a single-stage individual exam format using a multi-site within-course crossover design. This design improves on prior literature in a number of ways. A multi-site design combined with a fixed effects econometric model addresses the problems of small sample size and clustering by leveraging the large numbers of potential students and sections who can participate in a study. The crossover design combined with a fixed effects model addresses the problem of lack of random assignment by randomly assigning each student to treatment and control conditions over the course of the semester. Thus, each student serves as his or her own control.

Background

We divide the literature into two groups of studies - those that assign a two-stage exam treatment at the section level, and those that assign treatment at the student level (within section) by employing a crossover design. (We use “section” to refer to an instructor-course unit for which a student could enroll.) Studies in the first category include Zipp (2007); Knerim, Turner and Davis (2015); Ford (2019); Shaffer (2020); two of three studies reported in Kinnear (2021); Meaders and Vega (2023); and Callaghan et al. (2025). These studies either assign the treatment of a two-stage exam to a single section of a course (Zipp, 2007; Shaffer, 2020; Meaders and Vega, 2023; Callaghan et al., 2025), or assign alternative versions of the treatment or treatment and control to different sections (Ford, 2019; Kinnear, 2021; Knerim, Turner, and Davis, 2015). Both Ford (2019) and Knerim, Turner, and Davis (2015) find that average scores on the follow-on exam are higher for sections who receive the two-stage exam treatment. Kinnear (2021) finds mixed results. Shaffer (2020) finds that, controlling for student grade point average, those students whose grades benefited the most from the group exam did the worst on the final exam. Other outcomes include student perceptions of learning from the two-stage exam (Shaffer, 2020; Callaghan et al., 2025), differences in student performance on the individual exam versus the group exam (Shaffer, 2020; Kinnear, 2021; Meaders and Vega, 2023), or a measure of retention such as the difference in average scores on a baseline individual exam and similar questions on an individual final exam (Knerim, Turner, and Davis, 2015; Kinnear, 2021) or scores on a standardized test (Ford, 2019).

Sample sizes in these studies range from a low of less than 20 students assigned to a treated group to a high of over 200. In all of these studies, observations for individual students were treated as independent despite students being enrolled in the same section of a course.

The second group of studies randomly assigns treatment and control at the individual student level using a crossover design. These designs generally fall into two categories, which we term Design A and Design B. Studies that use Design A typically have at least two midterm exams during the semester, as well as either a third midterm exam or a final exam. Figure 1 in the Appendix illustrates a typical Design A study. In this design, students are randomly assigned to take the two-stage exam for either the first or the second midterm exam of the semester. The retention effects of the two-stage exam relative to the individual exam are measured by a student’s performance on matched questions on a follow-on (either the next midterm or the final) exam.

Figure 2 in the Appendix illustrates a typical Design B study. For the second stage, groups of students are randomly assigned to different subsets of the first-stage test questions. Although these subsets may be on different topics, ideally, they are equally challenging. In some versions of Design B, students also take an individual second stage that contains a different randomly assigned subset of test questions. Researchers argue that including this individual second stage helps control for the testing effect.

Studies that employ Design A crossover designs include Cortright et al. (2003), Leight et al. (2012), and Cao and Porter (2017). Both Cortright et al. (2003) and Leight et al. (2012) employ the crossover design within a single section of a course containing 38 and 250 students, respectively. Cortright et al. (2003) find that average follow-on scores were higher for students who completed those questions as a group. Leight et al. (2012) find no improvement in performance on the follow-on exam for students who completed a group exam. Although Cao and Porter (2017) employ the crossover design in two sections (n = 228), they treat the data as

though it comes from a single section for the purpose of analysis. In their variation of Design A, students are randomly assigned to take either an individual or group retest for the first midterm exam. They are assigned to the other group for the second midterm. The authors use the individual retest as the control data, and the group retest as the treatment. They find that for Exam 1, the group retest showed more improvement on the follow-on exam than the control. For Exam 2, both groups did worse on the follow-on exam, but the performance of those in the treatment group declined by less than that of those in the control. They find no difference in treatment or control performance on linked questions on the final exam.

Studies that employ Design B include Gilley and Clarkston (2014); Vogler and Robinson (2016); Cooke, Weir, and Clarkston (2019); Newton et al. (2019); and one study reported in Kinnear (2021). Gilley and Clarkston (2014); Cooke, Weir, and Clarkston (2019); and Kinnear (2021) employ the crossover design in a single section of a course ranging in size from 98 to 301 students. Newton et al. (2019) and Vogler and Robinson (2016) repeat the crossover design in two different sections containing 39 to 102 students. All studies treat individual student observations as independent, despite the fact that they are clustered into sections. The results of these studies are mixed. The only study that definitively finds that the group exam results in increased learning is Gilley and Clarkston (2014). Cooke, Weir, and Clarkston (2019) find better performance for those who worked in groups on one set of questions, but no difference between the individual and group treatments for another set. Newton et al. (2019) find that although the group exam seems to prevent short term decay in knowledge, they find no significant effects of collaborative exams six weeks after treatment. Similarly, Kinnear (2021) finds that retention from group exams at later stages of the semester is highly variable. Vogler and Robinson (2016) find the opposite. They find no effects of participating in group exams on intermediate exams, but small positive effects of being in the group treatment on final exam performance. The conflicting results of these studies point to the importance of controlling for both individual ability and effort as well as the time between treatment and assessment in the analysis.

All in all, the data on whether or not two-stage exams “work” is decidedly mixed, despite the fact that they are highly recommended by prominent researchers in STEM education (Wieman, Rieger, and Heiner 2014).

Methodology

a. What is the population being studied?

The study population will consist of undergraduate economics students enrolled in in-person courses in the United States; We omit non-U.S. institutions due to variations in rules about the types of interventions that can be studied in a course, particularly when it comes to testing (e.g. the United Kingdom). Because Design B is more flexible and has the potential to allow more study sites to participate, the study will employ Design B, where students will be randomly assigned in the second stage to a group exam on topic X versus topic Y, and both topics will be reassessed for all students in the next (midterm or final) exam. We will omit re-testing the other topic individually in the second stage as it adds unnecessary complexity and potential confusion for students. Thus, the study will employ a version of Design B that does not include an individual second-stage exam. The purpose of this study is to examine whether two-stage exams “work,” not why they work. Thus, we see no need to add an additional stage of individual testing to the design.

In order for an institution to participate in the study, their course(s) will need to satisfy the following criteria:

1. (REQUIRED) The course is an in-person, US-based course with at least two exams. These exams can be a midterm and a final or two midterms.
2. (REQUIRED) Site participants must be willing to test topics covered on the treatment exam on the individual portion of a subsequent exam.
3. (REQUIRED) All exams are closed book and taken in-person.
4. (REQUIRED) Students are required to take both the individual and group exams. Both exam scores count towards a student's final grade.
5. (REQUIRED) Students are required to take the follow-on exam which counts towards the student's final grade.
6. (REQUIRED) Site leads are willing to follow study guidelines, which are based on best practices from the literature (LoGiuduce, Pachai, and Kim, 2015). These guidelines are discussed in more detail in section c of the Methodology section (pp. 5-6) below.

Figure 3 reports the screening questions that the study leads will distribute to those who indicated on the June/July 2026 EENE questionnaire that they will use two-stage exams in one or more of their courses during the 2026-2027 academic year.

b. How will you sample from this population? How large will your sample need to be to precisely identify hypothesized effects?

The study team will recruit sites to participate through the EENE network. The call will include a description of the design criteria and the study guidelines.

Following Wozny (2017) and Schochet (2005, 2008), we solve for the minimum detectable effect size (MDE) assuming that (1) observations are clustered at the section level of the; (2) average section size is 30 students; and (3) the variance of y (the dependent variable) is normalized to one. Under these assumptions, the minimum detectable effect size is:

$$MDE = \sigma_y \sqrt{\frac{1-R^2_{y,x}}{p(1-p)}} \left(\frac{\rho}{\sqrt{c}} + \frac{1-\rho}{\sqrt{N}}\right) \left[\Phi^{-1}\left(\frac{1-\alpha}{2}\right) - \Phi^{-1}(1-\beta)\right] \quad (1)$$

Where:

- σ_y is the standard deviation of the dependent variable, which in the calculations is normalized to one.
- $R^2_{y,x}$ is the r-squared from the estimating equation
- *p* is the probability of a student being assigned to a treatment condition, which in the calculations is set to 0.5.
- $\boldsymbol{\rho}$ is the inter-cluster correlation
- *c* is the number of clusters (lecture sections) participating in the study
- *N* is the number of individuals
- $\boldsymbol{\alpha}$ is the level of significance used to reject the null hypothesis of no treatment effect, which in the calculations is set to 0.05
- $\boldsymbol{\beta}$ is the probability of a Type II error, which is set to 0.2 in the calculations. As a result, the probability of correctly rejecting the null hypothesis, or the power of the test $(1-\boldsymbol{\beta})$ is 0.8.
- Φ^{-1} is the inverse of the cumulative normal density function

Table 1. Number of 30-Student Lecture Sections Required by MDE, Inter-cluster Correlation (ICC), and R^2

	MDE (in SDs)			
	0.1	0.2	0.25	0.33
ICC = 0.2				
R^2 = 0.2	150	37	25	14
R^2 = 0.6	75	19	12	7
ICC = 0.5				
R^2 = 0.2	425	110	70	40
R^2 = 0.6	213	55	35	20
ICC = 0.8				
R^2 = 0.2	850	220	140	80
R^2 = 0.6	425	110	70	40

In this study, the units of observation are individual student-test section combinations. For each student in the sample, there will be two of these observations - one treatment and one control. Therefore, the appropriate value of *N* is twice the number of individual students in the sample. Assuming a level of significance (α) of 0.05 and a power (1 - β) of 0.8, the value of the term in square brackets for sufficiently large samples is 2.81.

Following Schochet (2005), Table 1 reports the number of sections of average size 30 required to reach MDEs ranging from 0.1 to 0.33 for small (0.2), moderate (0.5), and large (0.8) values of the inter-cluster correlation and R^2 values of 0.2 and 0.6. It is clear from this table that it is probably infeasible to obtain an MDE of 0.1, given the large number of sections (and thus participating sites) that would be required. Assuming the most optimistic values for R^2 and ρ, an MDE of 0.2 is obtainable with as few as 19 sections, something that might be obtainable with as few as three sites if each site has multiple sections and data are collected over multiple semesters. In order to avoid being underpowered, we recommend between 25 and 40 sections.

c. Provide details of the intervention

LoGiudice, Pachai, and Kim (2015) suggest that researchers studying two-stage exams employ the following best practices: use a crossover design, randomly assign students to groups, use individual testing as the baseline, track performance on individual questions over time, and calculate the percent of students answering each question correctly, as well as the discrimination and distractor coefficients for questions to ensure that the questions being

analyzed are capable of eliciting the intended retention effects to begin with. Thus, to ensure treatment fidelity, we will standardize practices across sites in the following ways:

1. (REQUIRED) Only question types that can be scored as correct or incorrect (e.g. multiple choice or short answer questions with unambiguously correct answers) will be part of the two-stage portion of each exam. We refer to this portion as the treatment questions below.
2. (REQUIRED) In order to create the potential for student improvement on the second stage and follow-on exam, instructors should use questions for the two-stage exam and follow-on exam that are commonly missed by students.
3. (REQUIRED) Questions in the follow-on exam should cover the same topic and use wording that is as similar as possible to the treatment questions, but use (for example) different numbers so that the questions are not identical.
4. (REQUIRED) The group exam should comprise at least 10% of a student's exam grade and at most 20%, unless a student's individual exam grade is higher than his or her combined individual & group exam grades.
5. Restrictions on how groups will be created. In particular:
    a. (FLEXIBLE) Instructors will strive to form groups of four students for the two-stage exam.
    b. (REQUIRED) If a course uses fixed groups for group work, instructors will use the pre-established groups for the two-stage exam; OR
    c. (REQUIRED) If groups are formed only for the purpose of the two-stage exam, instructors will randomly assign students to groups.
6. Restrictions on the amount of time between the individual and group portions of the two-stage exam. In particular:
    a. (FLEXIBLE) Instructors will hold the two-stage exam immediately after the first stage without a time gap.
7. (REQUIRED) Data requirements of the study, including:
    a. Data about groups and students' group membership: test-only or semester-long; student-chosen or instructor-created; assignment method (e.g. random or other); number of students per group; students' group membership.
    b. Test question and performance data: Individual and group performance on each test question included in the study; each student's individual exam score (to enable calculation of a discrimination index for each question); for multiple choice questions in the study, individual and group responses to each question (to enable calculation of distractor coefficients).
    c. Student demographic survey responses.
    d. The amount of time that elapsed between the individual and group portions of the exam.

After using the survey in Figure 3 of the Appendix to identify instructors who meet the minimum requirements for the study, the study team will hold a Zoom meeting (or meetings) with interested and eligible instructors to go over the requirements listed above and answer instructors' questions.

After the Zoom meeting(s), the study team will send a follow-up survey to determine which instructors think that they can structure their course to satisfy the study requirements and to determine how much variation in related study parameters exists. These related study parameters include:

1. How much weight they typically give the group exam in a student's exam grade.
2. The fraction of individual exam questions instructors include on the group exam or the minimum number of questions from the individual exam instructors include on the group

exam. This information will be used to generate guidelines about the fraction of/number of questions from the individual exam to use in the group exams and the follow-on exam.
3. The ratio of group exam time to individual exam time. Example: If students have 50 minutes to complete the individual exam and 25 minutes to complete the group exam, then this ratio is 50%.
4. What resources (e.g. calculator, formula sheet, notes sheet) instructors allow students to use on exams.
5. Whether instructors provide feedback to students about exam results or correct answers beyond simply returning graded exams electronically or in-person.
6. Whether instructors typically find that students are time-constrained on the group exam (LoGiuduce et al., 2015).
7. Whether instructors allow students to retain the individual exam to use as a reference while they complete the group exam.
8. Whether students use groups only on the group exams or throughout the semester. If the latter, are the groups instructor-created or student-selected?

An example draft is provided in Figure 4.

d. What are the important outcomes and how will you measure outcomes? Propose an empirical strategy; i.e., How will you analyze the data you collect?

In the basic fixed effects design, the outcome of interest is the student's score on the individual portions of the follow-on exam that correspond to the control (individual) and treatment (group) sections of the previous exam. The regression takes the following basic form:

$$\Delta Y_{isq} = \beta_0 + \beta_1 T_{sq} + \alpha_i + u_{isq} \qquad (2)$$

Where ΔY_{isq} is the change in standardized score of student i in section s on test questions q, T_{sq} is an indicator variable for when group of questions q in section s are treatment questions for the student, α_i is an individual student fixed effect, and u_{isq} is a random error term. The coefficient of interest is β_1, which measures the effect of having taken a two-stage exam on a subset of questions on subsequent exam performance on those questions. If the study is sufficiently large, we may also be able to include interaction terms in which we examine whether certain subgroups of students (e.g. underrepresented minorities, first-generation college students) differentially benefit from the treatment of a two-stage exam.

e. What are the biggest technical challenges and how will you mitigate them?

There are a number of other design questions to consider. They include:
1. What other covariates does it make sense to collect?
2. What instructor-level information do we need to collect? Examples of potentially useful information include instructors' use of group work outside of exams and, if used, how they form these groups.
3. Should we require instructors to use consistent procedures for providing feedback about exam performance? Some studies (e.g. Zipp 2007) have found that providing feedback about correct and incorrect answers can benefit learning in the same way that participating in a two-stage exam does. Thus, it will be important to control this aspect of the study design.

Data Collection/ Submission: In order to ensure that sites are fully aware of the data they should track during the semester, we will create an excel template to facilitate the reporting of data and provide it to sites at the beginning of the study. In addition, we will have a Research Assistant who will administer and keep track of the data collection on a timeline. We will employ a Google survey (draft in Figures 2 and 3 in the Appendix) which we will use to collect information about the two-stage exam administration at the site level. Figure 5 is a draft demographic survey for the students.

Timeline

The proposed timeline is as follows:

- Fall 2026: Finalize study design parameters and begin IRB approval process at Cornell.
- Fall 2026: Begin study site recruitment and finalize Cornell IRB approval. Complete IRB submissions for study sites as needed.
- Fall 2026 (tentative): Run pilot version of study at 1-2 sites.
- Spring 2027: Begin running study at other sites.
- Summer/ Fall 2027: Begin data analysis and continue data collection at sites as needed.
- Spring/ Summer 2028: Finalize data analysis and write-up of results. Begin submission of study to journals.

Research team

Katherine Silz-Carson (*Project Lead*) earned her Ph.D. in Economics from the University of Colorado, Boulder, in 1997. She currently serves as a Professor of Economics and Deputy Department Head of the Department of Economics and Geosciences at the United States Air Force Academy. Most of her published research papers are laboratory experimental studies examining questions relevant to environmental and public economics. She has also served as the lead analyst on a multi-site, randomized controlled trial of the Team-Based Learning pedagogy and as a Fellow for the U.S. Office of Evaluation Sciences, where she contributed to the development of multiple large-scale evaluations in support of President Biden's climate and equity agendas. In addition to contributing to the study design, she will support this study by helping to coordinate the IRB submission process at Cornell and other sites (in conjunction with EENE leadership), preregistering the study on the American Economic Association's Randomized-Controlled Trial Registry, serving as a site coordinator at the Air Force Academy (if it chooses to participate as a site), and by conducting the econometric analysis of the results. Her relevant publications are:

- Carson, K.S., Gonzalez-Ramirez, J., Heinicke, C., Maier, M., Ruder, P., Simkins, S.P., Adams, H. and Malakar, C.L., 2025. Results of a randomized evaluation of team-based learning exercises. *International Review of Economics Education*, *49*, p.100316.
- Carson, K.S., Adams, H., Gonzalez-Ramirez, J., Heinicke, C., Latham, J.M., Maier, M., Malakar, C.L., Ruder, P. and Simkins, S.P., 2021. Challenges and lessons: Design and implementation of a multi-site evaluation of team-based learning. *The Journal of Economic Education*, *52*(3), pp. 241-248.
- Carson, K.S., Chilton, S.M., Hutchinson, W.G. and Scarpa, R., 2020. Public resource allocation, strategic behavior, and status quo bias in choice experiments. *Public Choice*, *185*(1), pp.1-19.
- Carson, K.S., 2013. Incentive compatible mechanisms for providing environmental public goods, in *Handbook on Experimental Economics and the Environment,* eds. J.A. List and M. Price (Edward Elgar Publishing).

- Burton, A.C., Carson, K.S., Chilton, S.M. and Hutchinson, WG., 2009. Why do people non-demand reveal in hypothetical double referenda for public goods?. *Applied Economics*, *41*(27), pp.3561-3569.
- Carson, K.S., Chilton, S.M. and Hutchinson, W.G., 2009. Necessary conditions for demand revelation in double referenda. *Journal of Environmental Economics and Management*, *57*(2), pp.219-225.
- Carson, K.S., Chilton, S.M., and Hutchinson, W.G., 2008. Consequentiality and demand revelation in double referenda, in *Environmental Economics, Experimental Methods*, eds. T.L. Cherry, S. Kroll, and J.F. Shogren (Oxford, UK: Routledge Press), 407-423
- Burton, A.C., Carson, K.S., Chilton, S.M. and Hutchinson, W.G., 2007. Resolving questions about bias in real and hypothetical referenda. *Environmental and Resource Economics*, *38*(4), pp.513-525.
- Burton, A.C., Carson, K.S., Chilton, S.M. and Hutchinson, W.G., 2003. An experimental investigation of explanations for inconsistencies in responses to second offers in double referenda. *Journal of Environmental Economics and Management*, *46*(3), pp.472-489.

Sylvia Kuo (*Analyst*) earned her PhD in Economics at the University of Wisconsin, Madison, in 2001 and is currently a Teaching Professor of Economics at Brown University. Prior to taking the teaching position in Economics in 2012, she conducted government contract and grant-funded health services research, involving grantwriting, submitting IRB review, large dataset analysis and manuscript writing. Most recently, she published “The Impact of Distributed Homework on Student Outcomes,” in the International Review of Economics Education in 2022 [https://doi.org/10.1016/j.iree.2022.100243] which found that repeating the most problematic homework problem in the next homework (with different numbers) led to higher final exam scores within her large advanced economics class. Prior work in the health services research used propensity score models that resulted in the most rigorous evidence to date on negative health outcomes resulting from the use of feeding tubes among elderly nursing home residents with advanced dementia.[doi:10.1001/archinternmed.2012.1200].

Sylvia will be responsible for hiring undergraduate RAs (the site coordinator and analyst), provide regular oversight over the analyst, and oversee data collection and data cleaning and baseline summary statistics and modeling, using the TIER protocol to manage the data files for reproducibility [https://www.projecttier.org/tier-protocol/protocol-4-0/].

Derek Lougee (*Coordinator*) holds an MA in Economics from Cornell University, where he was also formerly a Ph.D. candidate in Economics. Derek is an economics instructor at Clackamas Community College in Oregon City, Oregon, and currently serves as Department Chair of the Social Sciences Department. In addition to being a study participant, Derek will serve as the main point of contact for study participants and will coordinate communication and data collection for the project.

References

Callaghan, Kristina, Tim Milbourne, Anna Klales, Greg Kestin, Carlso Arguelles, Logan McCarty, and Louis Deslauriers. "Two-stage final exams: An assessment strategy for enhanced collaborative learning and reduced student stress." *arXiv preprint arXiv:2504.04281* (2025).

Cao, Yingjun, and Leo Porter. "Evaluating student learning from collaborative group tests in introductory computing." In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on computer science education*, pp. 99-104. 2017.

Cooke, James E., Laura Weir, and Bridgette Clarkston. "Retention following two-stage collaborative exams depends on timing and student performance." *CBE—Life Sciences Education* 18, no. 2 (2019): ar12.

Cortright, Ronald N., Heidi L. Collins, David W. Rodenbaugh, and Stephen E. DiCarlo. "Student retention of course content is improved by collaborative-group testing." *Advances in Physiology Education* 27, no. 3 (2003): 102-108.

Ford, Dayton. "A three-year retrospective analysis comparing student retention of human physiology concepts in flipped, lecture, and two-stage cooperative testing classrooms." *Journal of College Science Teaching* 49, no. 1 (2019): 59-63.

Hollis Gilley, Brett, and Bridgette Clarkston. "Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students." *Journal of College Science Teaching* 43, no. 3 (2014): 83-91.

Kinnear, George. "Two-stage collaborative exams have little impact on subsequent exam performance in undergraduate mathematics." *International Journal of Research in Undergraduate Mathematics Education* 7, no. 1 (2021): 33-60.

Knierim, Katherine, Henry Turner, and Ralph K. Davis. "Two-stage exams improve student learning in an introductory geology course: Logistics, attendance, and grades." *Journal of Geoscience Education* 63, no. 2 (2015): 157-164.

Leight, Hayley, Cheston Saunders, Robin Calkins, and Michelle Withers. "Collaborative testing improves performance but not content retention in a large-enrollment introductory biology class." *CBE—Life Sciences Education* 11, no. 4 (2012): 392-401.

LoGiudice, Andrew B., Amy A. Pachai, and Joseph A. Kim. "Testing together: When do students learn more through collaborative tests?." *Scholarship of Teaching and Learning in Psychology* 1, no. 4 (2015): 377.

Meaders, Clara L., and Yalila Vega. "Collaborative two-stage exams benefit students in a biology laboratory course." *Journal of Microbiology & Biology Education* 24, no. 1 (2023): e00138-22.

Newton, Genevieve, Rebecca Rajakaruna, Verena Kulak, William Albabish, Brett H. Gilley, and Kerry Ritchie. "Two-stage (collaborative) testing in science teaching: does it improve grades on short-answer questions and retention of material?." *Journal of College Science Teaching* 48, no. 4 (2019): 64-73.

Roediger III, Henry L., and Jeffrey D. Karpicke. "The power of testing memory: Basic research and implications for educational practice." *Perspectives on psychological science* 1, no. 3 (2006): 181-210.

Schochet, Peter Z. “Statistical power for random assignment evaluations of education programs.” Mathematica Policy Research Report submitted to U.S Department of Education Institute of Education Sciences, Contract no. ED-01-CO-0038-0009. (2005): 1-59.

Schochet, Peter Z. “Statistical power for random assignment evaluations of education programs.” *Journal of Educational and Behavioral Statistics* 33, no. 1 (2008): 62-87.

Shaffer, Justin Franklin. "Student performance on and perceptions of collaborative two-stage exams in a material and energy balances course." *Chemical engineering education* 54, no. 2 (2020): 52-58.

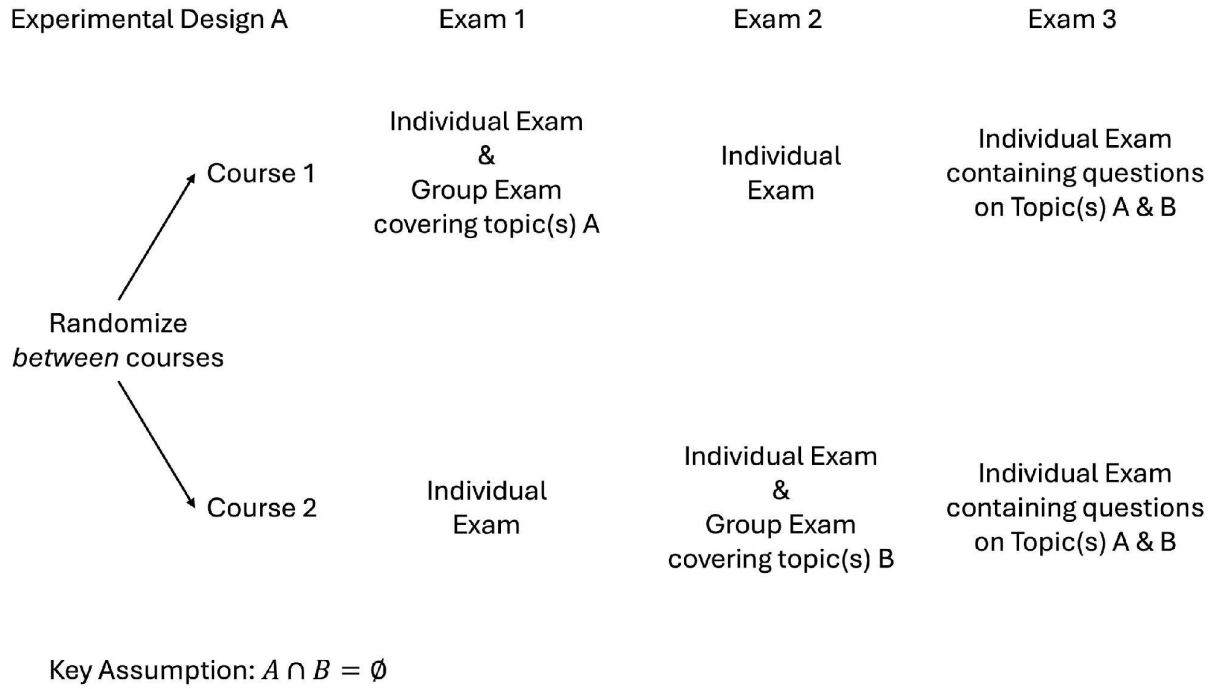
Wozny, Nathan N. “Power calculation notes.” Personal communication (2017): 1-4.

Vogler, Jane S., and Daniel H. Robinson. "Team-based testing improves individual learning." *The Journal of Experimental Education* 84, no. 4 (2016): 787-803.

Zipp, John F. "Learning by exams: The impact of two-stage cooperative tests." *Teaching Sociology* 35, no. 1 (2007): 62-76.

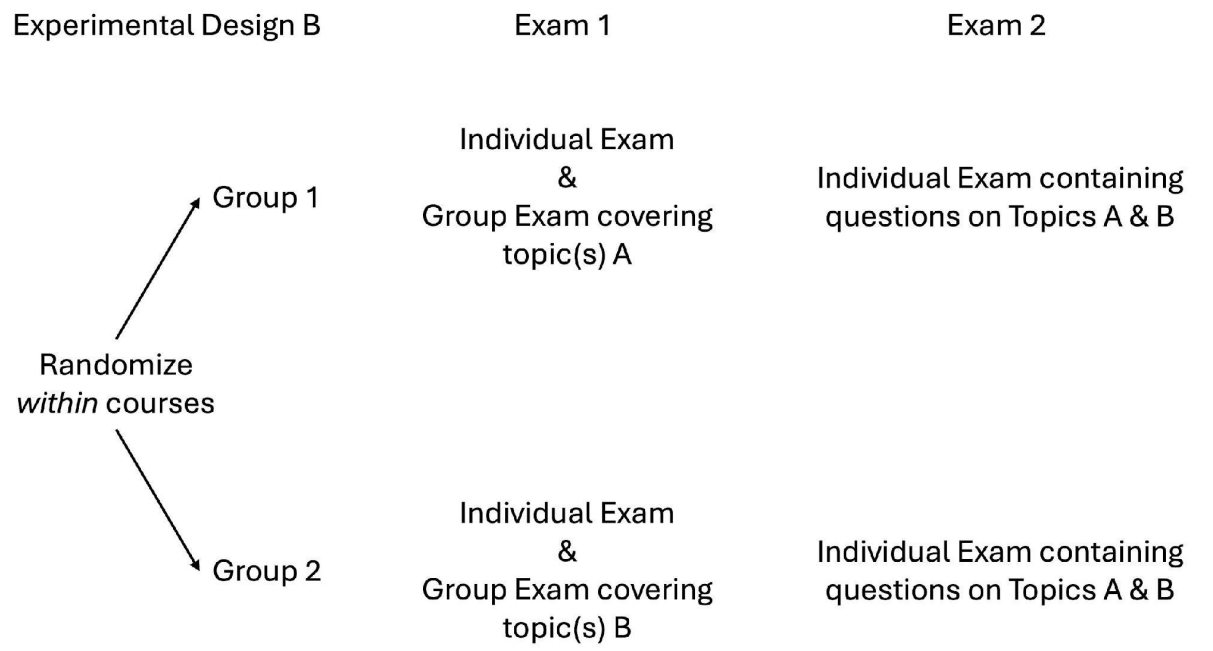
Appendix

Figure 1: Experimental Design A



Key Assumption: $A \cap B = \emptyset$

Figure 2: Experimental Design B



Key Assumption: $A \cap B = \emptyset$

Figure 3. Google survey for instructors – to be administered as a follow-up to the survey sent by EENE in June/ July of 2026 asking whether people are planning on using two-stage exams in their AY26-27 courses.

You are receiving this survey because you indicated on the most recent EENE survey that you are planning on using two-stage exams during the 2026-27 academic year. We would appreciate it if you would answer the following questions about your course.

1. Are you interested in participating in a study of the effectiveness of two-stage exams at improving learning?
[Yes, No, Maybe]
If No, exit from survey.

2. Is your course in-person? [Yes, No] *If No:*

2a. Do you teach an in-person course in which you would like to study two-stage exams? [Yes, No] *If No, send to exit message (see below).*

3. Is your course being taught at a U.S.-based university? [Yes, No] *If No, send to exit message (see below) .*

4. Are you willing to have students take the exams for your course in person (even if the exams are currently administered on a computer)? ? [Yes, No] *If No, send to exit message. .*

5. What resources do you allow students to use while taking the exam? Please check all that apply.
[drop down menu]
-Calculator
-Instructor-generated formula sheet
-Student-generated formula sheet
-Student-generated handwritten notes sheet
-Student-generated typed notes sheet
-My exams are fully open book and open notes
-Other [open text box]

[*If fully open book and open notes is checked*]
5a. The study design does not permit the exams to be fully open book and open notes. Are you willing to modify this portion of your course to participate in the study? [Yes, No]
If No, send to exit language.

6. Are you willing to require your students to take both the individual and group exams? [Yes, No]
If No, send to exit language.

7. Are you willing to have both a student's individual and group exam scores count towards their grade on the exam? (Note that under this grading scheme, if a student's individual score exceeds their combined individual and group score, then their exam score is their individual score.)
[Yes, No]
If No, send to exit language.

8. Are you willing to require students to take the follow-on exam (which could be the second midterm or the final exam) to count towards their final grade?
[Yes, No]
If No, send to exit language, exit from survey.

If exit before end of survey: Your answers indicate that your course is not a good candidate for this study. If you are teaching a course in future semesters that you think might satisfy the criteria for the study, please check the box below.

[ ] I am interested in potentially participating in this study in a future semester.

THANK YOU for taking the time to complete this survey.

If complete the survey: Your answers indicate that one or more of your courses might be a good candidate for this study. You can expect to be contacted by the study organizers regarding your possible participation. THANK YOU for taking the time to complete this survey.

Figure 4. Google survey for instructors who successfully complete the screening survey. The purpose of this survey is to collect data about how eligible instructors typically run their two-stage exams to inform some details of the study design.

1) For each study condition requirement listed below, please indicate whether or not you think you can structure your course to satisfy this requirement. If you respond NO, please provide an explanation for why you do not think you can structure your course to meet this requirement in the space provided.

	YES	NO
Include only question types that can be scored as correct or incorrect (e.g. multiple choice or short answer questions with unambiguously correct answers) on the portion of the first stage exam and the second-stage exam used for the study. (Note: Questions on the second stage should be identical to a subset of questions on the stage 1 exam.)		
Use questions on the study portion of the first stage exam, second-stage exam and study portion of the follow-on exam that are commonly missed by students.		
Create questions for the individual portion of the follow-on exam that are as similar as possible to the treatment questions, but use (for example) different numbers so that the questions are not identical.		
Use group sizes of four students to the maximum extent possible.		
If the course uses fixed student groups for group work, use the same pre-established group for the two stage exam. OR		
If groups are formed only for the two-stage exam, randomly assign students to groups.		
Hold the group exam immediately after the individual exam. OR		
Report the length of the time gap between the individual and group exams.		
Provide data about groups and students' group membership. (Whether groups are test-only or semester-long; method of assigning students to groups; number of students per group)		
Provide test question and performance data (Individual and group exam scores for each question in the study; students' individual exam scores; individual and group responses to multiple choice questions)		
Provide student responses to the demographic survey		

2) How do you typically form groups for the second stage of the two-stage exam?

- Pre-established groups used in group work, groups are randomly assigned by instructor at the beginning of the term
- Pre-established groups used in group work, groups are intentionally assigned by instructor at the beginning of the term using student characteristics
- Pre-established groups used in group work, groups are selected by students at the beginning of the term
- Groups created only for the purpose of the two-stage exam, randomly assigned by the instructor
- Groups created only for the purpose of the two-stage exam
- Other [open text box]

3) What weight do you typically give the group exam in a student's two-stage exam grade? If you have never used two-stage exams, what weight would you give the group exam in a student's overall exam score? (Assuming that their individual score is not higher than the weighted score.)
[Slider with a range from 0-100%]

4) What fraction of the individual exam content do you or would you plan to include on the group exam?
[Slider with a range from 0-100%]

5) How many questions from the individual exam do you/would you include on the group exam?
[Response is a positive integer]

6) What percentage of the time for the individual exam do you/would you allow for the group exam? For example, if students have 50 minutes to complete the individual exam and 25 minutes to complete the group exam, this percentage is 50%.
[Slider with a range from 0-100%]

7) As a rule, do you find that students have difficulty completing the group exam within the allotted time? [Yes/No]

8) Do you/would you allow students to retain the individual exam as a reference while they complete the group exam? [Yes/No]

9) Please describe whether and how you return exams to students and go over the answers. (Examples: Graded exams are returned to students electronically; Graded exams are returned to students in class but no class time is devoted to reviewing the correct answers.)
[Open-ended answer]

10) Please include any other relevant information about your course that we did not ask about in the space below.
[open text box]

Figure 5. Proposed Demographic Survey for Students

If you consent to allow your information to be included in the study of effectiveness of two-stage exams, please answer the questions below. When you are finished, place the completed form in the envelope provided.

If you choose not to allow your information to be included in the study, please put the blank form in the envelope provided.

1. To which gender do you most identify?

□ Female
□ Male
□ Gender Variant/Non-Conforming
□ Prefer Not to Answer

2. Are you of Hispanic, Latino, or of Spanish origin?

□ Yes
□ No

3. How would you describe yourself? (Check all that apply.)

□ American Indian or Alaska Native
□ Asian
□ Black or African American
□ Native Hawaiian or Other Pacific Islander
□ White
□ Prefer Not to Answer

4. Are you in your first year as a student at this institution? [Yes/No]

5. Do you consider yourself a first-generation college student? [Yes/No/Prefer not to answer]