bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Public exams may decrease anxiety and deepen critical thinking

Wiggins, Benjamin L.[12*]; Lily, Leah S.[23¬]; Busch, Carly A.[4¬]; Landys, Mėta M.[5¬]; Shlichta, J. Gwen[6¬]; Shi, Tianhong[7¬]; Ngwenyama, Tandi R.[8¬]

[1] Department of Biology, Shoreline Community College, Shoreline, WA (USA)

[2] Department of Pharmacology, University of Washington, Seattle, WA (USA)

[3] Department of Biology, Western Washington University, Bellingham, WA (USA)

[4] School of Life Sciences, Arizona State University, Tempe, AZ (USA)

[5] Department of Integrative Biology, Oregon State University, Corvallis, OR (USA)

[6] Department of Biology, Edmonds College, Edmonds, WA (USA)

[7] Ecampus Division, Oregon State University, Corvallis, OR (USA)

[8] Carlson College of Veterinary Medicine, Oregon State University, Corvallis, OR (USA)

*Corresponding author email: benlwiggins@gmail.com [updated email will be available soon]

¬These authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract:

Assessment methods across post-secondary education are traditionally constrained by logistics, built on prior practice instead of evidence, and contribute to the widespread inequities in education outcomes. As part of attempts to improve and diversify the methods used in assessment, the authors have developed a flexible and low-tech style known as ‘public exams’ based in best practices. Public exams attempt to bring students authentically into the process of assessment through the use of pre-released portions of the exam. Through an initial mixed-methods research endeavor at a closely-matched pair of a research-intensive university and a community college classroom, we observe signals of positive impact from the public exam on student experiences. Public exams appear to prompt deeper thought, direct students more efficiently to the core concepts in the discipline, and may decrease anxiety in and around exams. The public exam experience does not show evidence for exacerbating gaps in exam outcomes for students from minoritized backgrounds. This evidence suggests that public exams are worth deeper investigation as an evidence-based and effective assessment style.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction:

High-stakes examination-based assessments (hereafter, exams) are a common and widespread feature of postsecondary education (Stobart & Eggen, 2012). Whether used to give formative feedback to students, to summatively assess students' knowledge, to create selection barriers for capacity-constrained programs or careers, or simply to assign grades for external use, these exams are complex structural elements that students must grapple with (Wideen et al., 1997). Problematically, the educational practices used widely in college and universities are often based in traditional routines and logistical concerns instead of evidence-based, student-centered practices (Ambrose, 2010; Handelsman, 2006). Improving the practices in giving and taking exams has the potential to improve the educational experience for a more diverse, deeper, and more talented pool of future students (Intemann, 2009; Ralph et al., 2022).

The choices that college faculty make around course assessment methods have profound impacts on students. Within a highly unequal power relationship, students have little to no voice about the ways in which they should be assessed. Students for whom college practices are new (to them, or to their communities) are figuring out the rules of the game on the fly; those rules change between classrooms. The same challenges that multilingual learners experience in monolingual classrooms play out (with higher stakes) during an exam. Anxiety around education can be exacerbated by exams and this anxiety tends to impact groups of students unjustly. Students from a wide array of diverse backgrounds find their progress metered by exam challenges that are designed by a professoriat that is rarely as diverse as they are (Hurtado, 2007). Because strategies and tactics change in meaningful ways even between closely matched practitioners, there is a wide range of experiences that a student might encounter even within a single institution or unit. Faculty are under constant pressure to use time effectively, and many evidence-based practices require significant investments of time, energy and training that are rarely valued at the same level as research achievements or ratings of other aspects of teaching (Rossing & Lavitt, 2016). The traditional style for postsecondary education is to reveal assessment tasks to students only during the exam itself. While a dynamic mix of active learning principles have become more widespread, similar best-practices in giving college exams are less-well defined and relatively difficult to adopt even for the most conscientious of instructors.

There are many ongoing attempts to improve the practices around exams, though largely at the practitioner level and less often codified in research literature. Our contribution is an interrelated set of evidence-based practices collectively described as the public exam system. While public exams are based in best practices well-known in education, here we describe the implementation and research findings that result. In this work, we take a lens of educative assessment: a theoretical framework summarizing that assessments have many purposes but the primary among them should be as a tool for facilitating student learning (Buxton et al., 2013; Fink, 2003; Jönsson, 2008; G. Wiggins, 1998, 2011). Specifically, educative assessment suggests that educators can create challenging exams for students that are useful practice for their careers and lives such that teaching directly to these exams will be beneficial. Our methodology follows a design-based tradition in which education interventions are implemented and researched dynamically and iteratively, and that each of our model organisms is a human

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

being in a crucial, formative part of their life. To explore our research questions rigorously, we apply mixed quantitative and qualitative methods and attend to signals in the data that triangulate similarly across multiple types of investigation. Our goal in this work is to demonstrate how public exams impact college students.

What is a public exam?

Public exams have three elements that attempt to address three common problems:

- Partial exam content is pre-released to students prior to the exam to deepen the thinking that students can accomplish during the actual assessment. This allows students to read meta-information about their tasks beforehand as well as to engage with content that might take more time to comprehend than is available in a traditional exam. Traditionally, exam content is often encountered all at once in the context of the exam, and this rapid transmission of large amounts of relevant information constrains the asking of interesting and higher-order cognitive exam questions due to the high cognitive load (Sweller, 2010). Throughout the manuscript, we use the term "deepening thought" to refer to this aspect of public exams.
- Pre-released exam content provides opportunities for students to edit much of the exam. Language barriers around exam content are hard to disassociate from true struggles with content. By allowing students an opportunity to give feedback on exam formats and wording, we leverage a larger group of motivated editors to address challenges that are separate from conceptual knowledge. These same developing experts can also contribute to the writing of the exam itself. Traditional, surprise-based exams cannot be co-created and the experience of power relationships and secret information can detract from positive student-teacher relationships that are crucial to maximizing learning. Whether by improving language, increasing transparency, or by utilizing students as exam question creators, we hope to draw students authentically into the creation of their own assessments. Throughout the manuscript, we use the term "language barriers" to refer to this aspect of public exams.
- Lastly, the pre-released material gives a direct conduit for instructors to amplify the parts of course material that are most important. Instead of indirectly indicating core concepts through study guides or practice exams or review sessions, students are given strong cues in the actual exam about the concepts and skills that are core to the discipline and that they are expected to master. We use the term ‘core concepts’ here to broadly describe the content that instructors believe is more central to the practice of their discipline. Traditional surprise-based exams can only do this after the fact, at which point the opportunity to direct optimal study is generally lost. Throughout the manuscript, we use the term "directing to core concepts" to refer to this aspect of public exams.

As a simplified example, imagine an exam question in which the student is directed “For ten points, explain in three sentences or less how detoxification of human blood is performed by the cells in the liver.” By pre-releasing the exam question for students but withholding only the word ‘liver’, the possible variants of the exam question are increased to include at least several organs. While providing the meta-information for the task as well as the framing of the topic area

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

itself, this question maintains enough surprise to deeply examine student understanding. A further variant of a pre-released exam question might be: “For ten points, explain in three sentences or less how [withheld] of human blood is performed by the cells in the [withheld].” By withholding just a single additional word, students are now given direct information about both the method/scope of written assessment as well as tangible evidence that their understanding of processes impacting human blood will be crucial for demonstrating mastery of the topic.

A timeline comparison of a public exam and a traditional exam is shown in Figure 1.

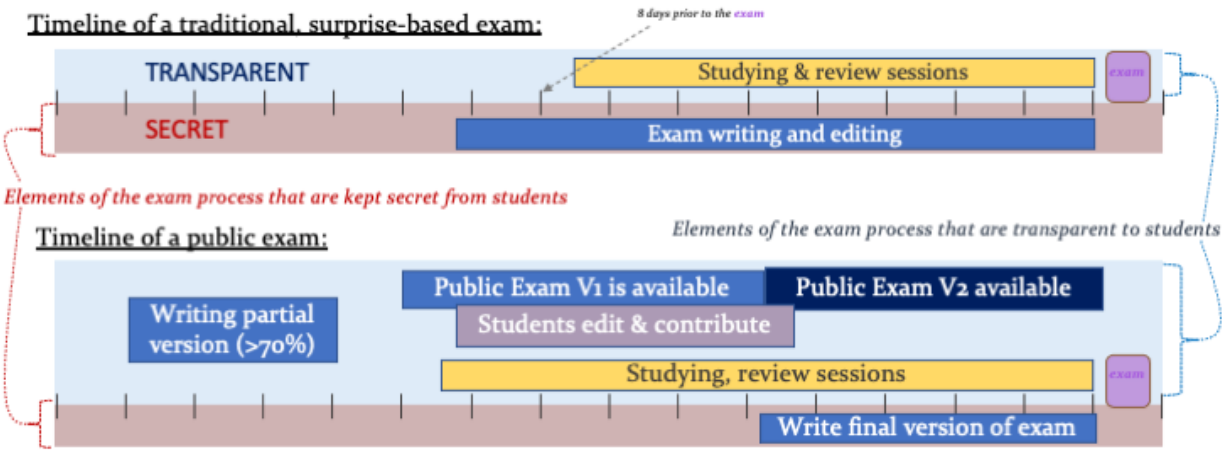

Fig 1. ***Comparative timeline of traditional and public exams.***
Tasks to be completed are separated into those that are transparent to students and those that must necessarily be kept secret from students at the risk of giving away exam answers. For readers unfamiliar with traditional exams, the top timeline is offered as an approximation. The bottom timeline is an approximation of a public exam structure. The purpose of this figure is to illustrate the differences in increased transparency and opportunities to study from exam material in public exams.

The underlying goal in these three elements of public exams is to engender trust and authentic engagement between students and instructors. Throughout the manuscript, we use the term "authentic involvement" to refer to aspects of public exams that address this overarching theme relating to trust (Brown, 2017). The four evidence-based practices described above are frequently addressed throughout K-12 education and are useful in convincing students more often that the assessment process can work for them (Darling-Hammond & Bransford, 2005; R. Keith. Sawyer, 2005; Zeichner et al., 2000). A few types of examples of public exam questions are presented in Figure 2. Because students and classrooms differ so greatly, the use of the public exam style is not intended to be narrowly prescriptive. Instead, we offer this stylistic definition of public exams in order to a) help guide instructors incrementally closer to more engaging assessments and b) provide a basis for exploratory research to identify impacts on and benefits for postsecondary students.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

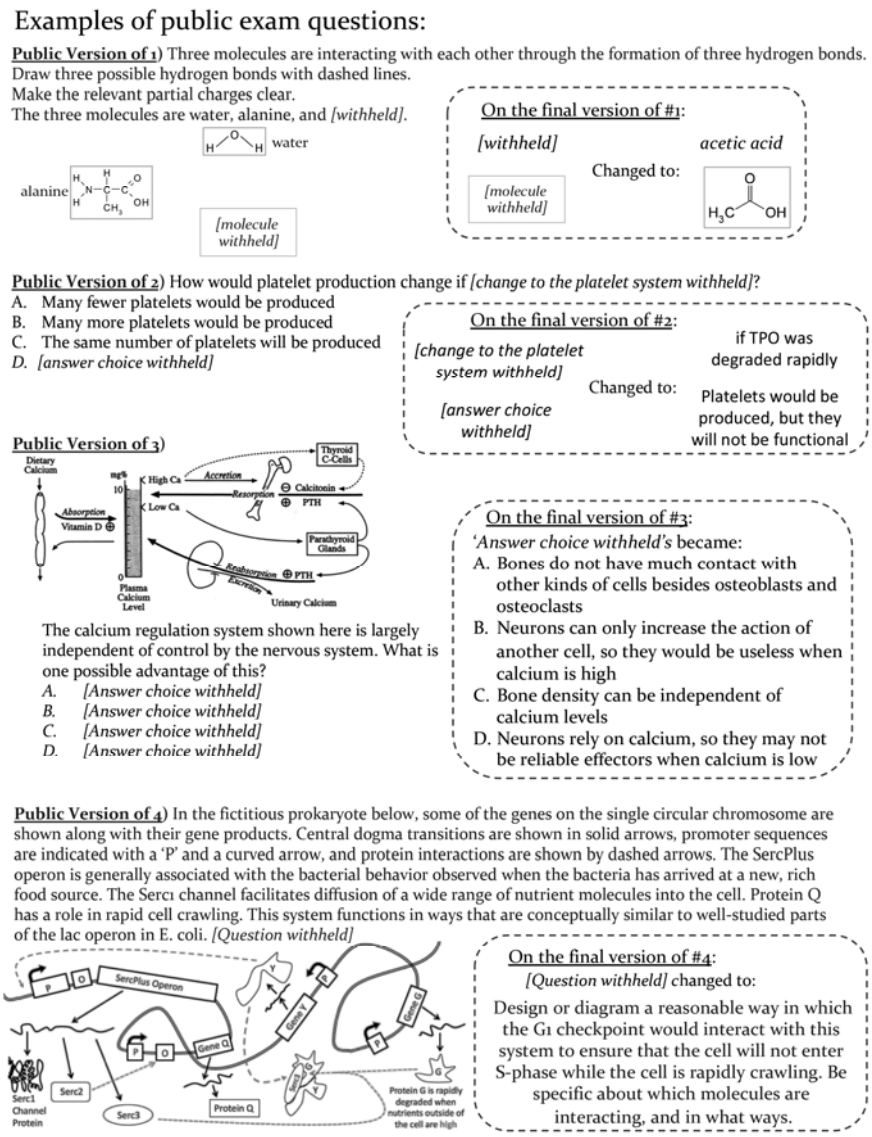

Fig 2. ***Examples of public-style exam questions.***
For each of 4 exam questions, the pre-released version provided to students well before the exam is shown. In dashed insets are the changes made to the question for the actual version that students complete for course points. The purpose of this figure is to give examples of a few of the types of exam questions that can be used in public exams.

Pedagogical Framework

Pedagogical frameworks that support the practice of public exams as described above include cognitive load theory (Sweller, 2010), retrieval practices (Moreira et al., 2019), active learning (Bonwell & Eison, 1991), pedagogy of care (McMullen et al., 2015), and inclusive pedagogy (Florian & Black□Hawkins, 2011). When students engage with an exam, they are retrieving information from long-term memory into working memory in order to answer a test question. If the exam questions are unfamiliar to students, do not match what was taught, or have unclear instructions, students are likely to experience cognitive overload (McMullen et al., 2015), and ultimately a negative impact on their academic performance. In public exams, giving students

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

opportunities to practice exam-relevant skills in similar formats and on similar content is a solution aiming to reduce cognitive load and, ultimately, test anxiety. Another effect of giving students opportunities to practice exam-relevant skills is utilizing the benefits of retrieval practice to achieve mastery learning. Researchers define active learning as “instructional activities involving students in doing things and thinking about what they are doing” (Ambrose, 2010; Bonwell & Eison, 1991; Moreira et al., 2019). Allowing students to engage with and edit the pre-released exam applies the principles of active learning. Public exams give students opportunities to lessen potential cultural barriers or linguistic barriers to a full understanding of the exam questions, aiming to create an inclusive learning environment for all students. Pedagogy of care is defined as “a teaching practice based on reciprocity where teachers take on the role of caregiving and students receive care on the basis of the teachers’ concern for their overall well-being” (Obuaku-Igwe, 2021). Public exams apply pedagogy of care by attending to students’ emotional stress related to test anxiety. Inclusive pedagogy is the application of the diversity and inclusive social movement into education, and is a student-centered approach to teaching and learning that supports learners of all backgrounds (Shi & Blau, 2020). The public exam style is designed to align with evidence-based research on best practices in assessment.

Research Questions:

Our research questions are the following:

- In what ways do public exams impact the student experience?
  - Are these impacts negative or positive?
  - Are these experiences impacted by Language issues, Directing to core concepts, Deepening thought, and/or Authentic engagement?
- Do public exams impact grade equity?
- Are public exams likely to be applicable across postsecondary education contexts?

In summary, exams are a widespread and problematically complex aspect of the college experience. Public exams are designed around best practices in education, but the combined application of these methods has not been rigorously studied. We apply mixed-method design research to understand how and for which students public exams can impact their educative experiences in college courses.

Methods:

Research environments:

Research was conducted at a research university (R1) and a community college (CC) in the Pacific Northwest of the United States. Students were enrolled in lower-division courses in Biology departments during Quarter 2 of 2021. The R1 course was taught for 300 students and the CC course was taught for 48 students from which populations of 292 and 32 participants, respectively, were included through IRB-approved consent processes (under protocol #s

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

STUDY00012237, ECIRB-20210512 and IRB-2020-0813). These courses were chosen for consistency of general topic and level, for the large population in the R1 course which allowed quantitative analysis of subgroups, and for institutional access to research. Students in the R1/CC courses were 63%/59% non-white, 77%/66% registrar-identified female, 24%/20% first-generation attending college, 12%/24% international and (at the R1) 31% identified as being from historically underserved populations by the R1 university. Students in both courses typically have interest in a wide range of career goals around healthcare, science, research and business. Public exam techniques were used in both courses. Both the CC and R1 courses were using public exams for the first time in those environments. In the large R1 course, students were graded largely on the basis of 5 total exams given every 2 weeks throughout the 10-week quarter. In the smaller CC course, students completed a total of two exams that were written in the public exam style.

Research flow:

This work was conducted using a design-based research methodology, which allows for preliminary research findings to be used to guide the collection and analysis of subsequent data in an iterative fashion (Collins et al., 2004). Examining human experiences in this methodology is intended to be more rigorous than simple, self-reported data while allowing a greater breadth of possible findings than quantitative experiments alone would observe. This methodology is a good fit for education systems where iterative redesign and incremental improvement of human experiences are the primary goals of research and implementation work (Sandoval, 2004).

Here, we used *qualitative interviews* to broadly assess the experiences of students taking public exams around our three main research questions. Interviews were also used to assess differences in the student experience between institutions. These interview findings refined our analysis of a larger data set by *coding open-ended survey items*. In parallel to this qualitative and mixed-method work, students in the R1 course took exams that used both public and traditional questions to experimentally observe signals of inequity in exam outcomes. This *within-exam experimental system* controls for student identity, instructor impact, classroom environment, and content material in comparing data from two types of assessment questions. This *quantitative data collection and analysis* is intended to cast a wide net for possible negative impacts or inequities of public exams on student experiences. Any positive impacts of the public exam system that are observed are likely to be conservative because of issues with first-time implementation fidelity (in both R1 and CC courses) and incomplete application of the public exam system (in the R1 course). *Student self-reported preferences* for exam style were collected for triangulating with other types of data; this self-reported data may help to illuminate the presence of unknown negative impacts of the intervention, but is not in itself convincing of positive impacts of the intervention. The overall process of data collection is described below in Figure 3. The purpose of collecting a wide range of types of data is to broadly investigate the possible outcomes from this intervention and better understand the possible avenues for future, deeper research investigations. Here we present the results of this initial design-based research study.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

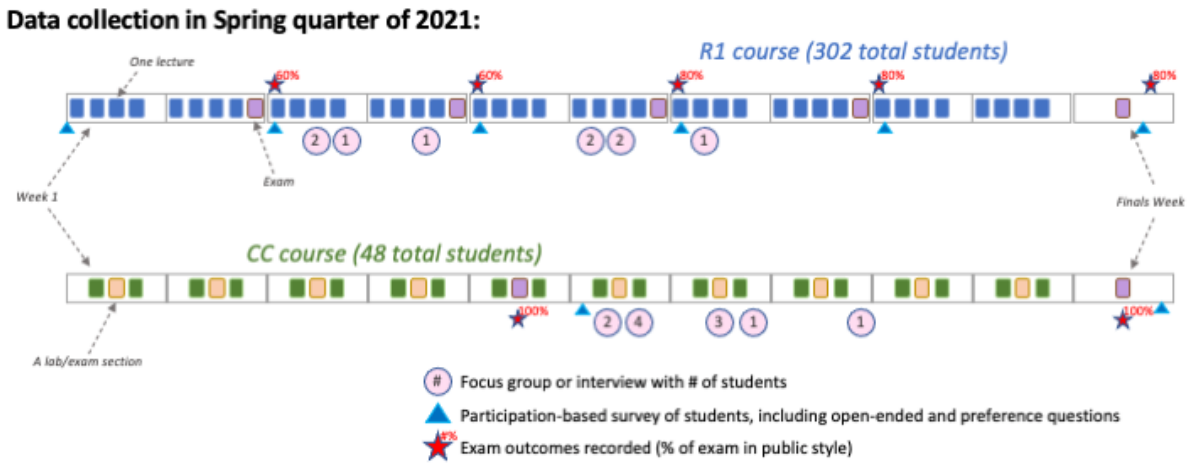

Fig 3. ***Data collection scheme.***
The purpose of this figure is to make clear when and in which class environment the data were being collected.

Qualitative interviews:

Facilitation of interviews and transcription were completed by an experienced qualitative researcher (LSL) who has an M.Ed. in Curriculum and Instruction, was a Research Assistant on the project, has experience in clinical psychology, and has prior publications using qualitative coding and interview research in education (Dahlberg et al., 2019; B. L. Wiggins et al., 2017, 2021).

Group and individual interviews were used to hear student experiences using grounded ethnographic principles (Glaser & Strauss, 1968; Rubin, 2012) and with subject-centered and -driven methodology from dialectical behavioral therapy (Linehan, 2018). Eleven interviews totalled 488 minutes of recorded discourse with 19 participants. Student participants were recruited to participate via random email to course lists. Interviewer non-affiliation with the courses was communicated and students were given a small Starbucks card for participating in the Zoom interview. Participants did not know the interviewer prior to the interview process. Data around the interviews at both sites as well as all transcripts are available in Supplement 1.

During qualitative interviews, broad experiential opening questions were used (e.g. “How is [course] treating you?”) to elicit a broad spectrum of conversations around students’ experiences (Cameron, 2005). Rather than bringing in specific questions or prompts, the facilitator followed up with probing questions on student-raised topics pertaining to our research questions. Opportunities to segue organically into these discussions were taken using light reinforcement and broad questioning (Rubin, 2012). This method enabled us to influence the focus of discussion without disclosing our specific research methods or interests, which permits students to consider their impressions of the course and content within their own framework of values, memories, and needs. Anonymized transcripts of recorded conversations were analyzed afterwards, and participants did not give feedback on the findings. Thematic representation saturated (Saunders et al., 2018) at the R1 site after 6 interviews, so interviews at this site were discontinued.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transcripts of qualitative interviews were analyzed by coding of statements. Initially, we prioritized the following four themes drawn from our original research question: 1) Language issues, 2) Directing to core concepts, 3) Deepening thought, and 4) Authentic engagement (coded as numbered here). We decided to investigate these four themes due to anecdotal discussions with students in prior courses, and these themes were prevalent in research interviews. While several original themes appeared to be less frequently encountered and were dropped from the final analysis, other themes emerged from qualitative analysis. The new codes 5) Anxiety or Confidence and 6) Collaboration emerged during iterative qualitative analysis of interviews with students at the CC and R1 institutions, where students strongly expressed the importance of these themes. Lastly, a final code 7) Not Related to the Public Exam System was designed to capture student experiences that were not part of the public exam system. The descriptive language found in the coding tables was iteratively improved for clarity and to better match student language. Transcripts were subsequently re-coded by BW and LSL using this improved set of seven codes. The research team discussed coding with a lens towards observing possible differences in experience between the institution types throughout the design-based process. The final consensus coding table for interviews with exemplary quotes is available in Supplement 2.

Coding of open-ended survey items:

Open-ended survey items were used as a quantifiable source of qualitative data at scale. In a participation-only study, all students at both sites were asked to answer the question: *"Did the style of exams in [this course] work for you? Why or Why not?"*. Cognitive testing for validity of this question was performed with a separate group of students that were of the same age and progression as students at the CC and R1 sites. Four student participants were involved, and all were of sophomore/junior standing. Student participants were randomly recruited by email from a large list of similarly matching biology students. The goal of this cognitive testing was to identify any problematic interpretations of this question that had been previously vetted through iterative writing and editing. Cognitive testing was facilitated by BW. Students read the question in paper form and then afterwards worked as a group to read the question aloud, come to a consensus meaning, and then discuss any possible alternative interpretations. No significant confusions or alternative meanings were described. BW and LSL discussed all interpretations and concluded that the final wording of this question was understood by all participants and likely to be valid for the research population. Responses to the final version of this open-ended survey question were collected and anonymized from 242 participants at the R1 site and 32 participants at the CC site.

Open-ended survey responses were qualitatively coded. The final seven thematic codes used were developed and improved as described above. The seven final codes were iteratively coded and discussed and then coded for presence in the larger survey-based set of 242 open-ended responses by LSL and BW. Each code was also sub-coded as positive or negative with regards to literature-based learning outcomes for students. This was not opinion-based coding on the part of students, but rather researcher-based assessment of whether the practices or experiences presented were positive or negative based on educational best practices as

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

described in national best-practices texts including How People Learn II (National Academies of Sciences, 2018) and the biology-focused AAAS document Vision and Change (AAAS, 2011a). In other words, these results were not coded for what students enjoyed or appreciated (see examples in-text below) but rather for conditions in which learning is likely to be supported. Two researchers (BW and LSL) independently coded 15% of the responses and achieved an acceptable interrater reliability score of kappa = 0.88 (McHugh, 2012). One researcher (LSL) coded the remaining responses.

For examples of the positive or negative coding, a student who indicated "*The public exam made it harder to know what I needed to know*" would be coded into the category of 'Directing to core concepts' and as a 'Negative' impact, since confusion about core concepts is a problematic distractor for learning across fields (Meyer, 2004; National Academies of Sciences, 2018; R. Keith. Sawyer, 2005). If a student indicated that they "*hate public exams because they force me to think more deeply*," then this would be coded as a 'Positive' impact within the theme of 'Deeper thought', even though the student may not have enjoyed that aspect of the learning challenge. Further examples and final codes are available in Supplement 3.

To determine if the prevalence of any codes was significantly impactful on student experiences, we calculated the percent of students who provided feedback on each qualitative theme of the public exam system in the open-ended survey items and whether that feedback was positive or negative. To determine if there was a relationship between the type of feedback students provided (i.e., about the public exam system or not) and the nature of that feedback (i.e., positive or negative), we conducted a series of Pearson's chi-square tests of independence for each of the six factors of the public exam system as well as an aggregate of all six factors. This approach used Code 7 ('Not related to the Public Exam System') as a control group, which is more conservative than a simple control ratio like 1:1 and controls for the likely general tendency for participants to report positive experiences more often than negative experiences. When a given count in the contingency table was too small (i.e., less than five) to conduct a chi-square test, we used a Fisher's exact test (Bower, 2003; McCrum-Gardner, 2008).

Within-exam experimentation:

Within the large R1 course, students completed five summative exams in which assessment questions were a mix of traditional 'surprise' style exam questions for which they had no pre-knowledge and other 'public' style exam questions for which students had exposure to much of the information in the individual question well before the time of assessment. For this course, all exam questions were written in multiple choice format. The relative amounts of traditional or public exam questions changed throughout the course. Students began the quarter with two exams that used the same distribution of multiple choice exam questions: 15 public-style exam questions and 10 traditional, surprise-style exam questions. Subsequent exams (in response to student survey responses, see discussion) included 20 public-style exam questions and 5 traditional, surprise-style exam questions. The purpose of this within-exam experimentation is to collect well-controlled data that might lead to the observation of any inequities in this style of assessment, should they exist, rather than to prove any particular value of an assessment style. Because the variation between assessment styles happens within each

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

exam, data collected about exam outcomes can be compared on the same scientific topic areas, for the same student identities, and with the same instructor among other variables that are otherwise difficult to control.

Quantitative data collection and analysis:

Within the large R1 course, the following discrete data were collected for each participant: College GPA, course grade, exam results for each question on each exam, scores for participation-based assignments, completion or not of an exam editing activity, and (via the university registrar) race/ethnicity, gender, international student status, first-generation in college status, and inclusion in the university-assigned Education Opportunity Program (EOP). This last categorization is particularly important to this work: the R1 institution defines "under-advantaged" students as students identified as part of the EOP and these students hail from educationally or economically disadvantaged backgrounds. Because this EOP categorization is based on family income and other variables not typically represented in simpler demographic statistics, we chose this measure as the single variable on which we would pre-build models for analysis as has been used in other, similar work (Freeman et al., 2017; Wright et al., 2016). These data were collect in order to analyze quantitative data in a way that would make observation of any inequitable outcomes of the intervention more likely. All data collected in these ways are available in anonymized form in Supplement 4.

A primary purpose of quantitative data was to observe any inequities in the assessments that might exist. In order to determine if students performed differently on public or traditional exams, we used a two-sample t-test to compare the total percentage of points students earned on all public exam questions and all traditional exam questions throughout the term.

In order to determine whether there were differences in exam performance on each type of exam question based on students' demographic characteristics, we used linear regression models and included gender (male/female), EOP group of interest (yes/no), and overall GPA (from the registrar on a 4-point scale) as predictors. (Example model: percent score on public exam questions ~ gender + interest group + GPA.) Gender has been shown to affect student exam performance (Odom et al., 2021) and students in our EOP group of interest have been found to do worse than their peers on exams at this institution (Cooper et al., 2020). We acknowledge that registrar data for gender that includes only male/female do not best represent all individuals' gender identity and that not every person identifies in the gender binary (Cooper et al., 2020), but we did not ask students to self-report their gender.

To examine potential demographic differences in students' self-reported preferences for the proportion of each question type on an exam, after the second and third exams, we asked students if they would prefer to have more public questions, fewer public questions, or keep the same ratio of public to traditional questions for future exams. After the fourth exam, we asked students if they would prefer more or fewer public questions with no neutral option. We calculated the percentage of students who selected each option and assessed potential demographic differences of students' preferences after the second and third exams using

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

multinomial regressions and using logistic regression for preferences after the fourth exam. We again included gender (male/female), EOP group of interest (yes/no), and overall GPA (based on registrar data on a 4-point scale) in our models. (Model for post-exam two and three preferences: exam preference (more public/fewer public/same) ~ gender + interest group + GPA; model for post-exam four preferences: exam preference (more public/fewer public) ~ gender + interest group + GPA.)

Preceding each exam, students were given the opportunity to provide edits on the public portion of the exam. This was an optional part of a required online assignment which students were able to bypass and still receive full participation points. To investigate the extent to which a student providing edits on the exams might have impacted their overall course grade, we used a linear regression with the total number of exams for which the student provided edits, EOP group of interest (yes/no), and overall GPA as the predictors in our model. (Model: course grade ~ total edits + interest group + GPA.)

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results:

Qualitative interviews:

Interview-based methods were used to guide the overall flow of research. Interview transcripts were used to allow for the emergence or loss of code themes in response to student experiences, and Codes 5 and 6 ('anxiety or confidence'and 'collaboration') were added to the coding table. Interviews were also used to iteratively improve research questions and codes, including cognitive testing of an open-ended survey question.

To better understand whether public exams might be applicable to community college courses, which are generally smaller and less available to quantitative research, we undertook qualitative interviews in a closely-matched community college course. This CC course closely matched the R1 course in terms of topic, location, timeline, and the first-time use of the public exam style for the course. Comparing two environments through qualitative interviews is an inexact method, but it is a rigorous way to explore broadly for signals that there are substantial important differences in either the environment or the intervention. In this case, analysis through iterative coding of interview transcripts brought us to the conclusion that students in the two courses had similar experiences with public exams. Our primary codes were evident in similar proportions, and student comments to interviewers brought up similar challenges and gains. No thematic signals appeared to us in one environment and not the other. This is an initial attempt to explore the possible broad application of public exams, and clearly more research will be required on a greater scale to make similar conclusions. In the meantime, the outcomes of these analyses are consistent with public exams being similarly applicable across these two institution types.

Coding of open-ended responses:

Students in the large R1 course answered a survey item: "*Did the style of exams in [the R1 course] work for you? Why or Why not?"*. All coding data for open-ended survey items is available in Supplement 5. When compared with a conservative control group using Code 7 ('Not part of the public exam system'), we observed a strongly significant statistical signal for the overall positive impacts of public exams (Table 1, Row 1). No significance (positive or negative) was observed for student mentions of language barriers, authentic involvement in the process of assessment, or collaboration. Student experiences with 'Directing to core concepts' were significantly positive (p value = 0.0002). Student experiences with 'Deeper thought' were also significantly positive (p value = 0.004). Student experiences with 'Anxiety' were strongly, significantly positive (p value = 0.0101). Positive or negative experiential impact showed no statistical difference for students in the EOP group. These data suggest that students' unprompted experiences with public exams are predominantly positive, which correlates well with preference data described below. These data also triangulate well with interview results noting that deeper cognitive work, decreased anxiety, and more efficient directing to core concepts are likely outcomes of public exams. These results of the quantitative analysis of open-ended coding are presented in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Signal Pos:Neg	*Null Pos:Neg*	χ^2 *Test Statistic*	***P value***
Overall Impacts of Public Exams	*97:22*	*74:38*	*7.1547*	*0.0075*
Language barriers	*9:9*	*74:38*	*1.7353*	*0.1877*
Authentic involvement	*11:4*	*74:38*	*0.3152*	*0.5745*
Collaboration	*13:0*	*74:38*	*Cannot run test with a zero result. Does not approach significance.*	
Directing to core concepts	*55:5*	*74:38*	*13.6508*	*0.0002*
Deeper thought	*27:2*	*74:38*	*8.2834*	*0.0040*
Anxiety	*31:4*	*74:38*	*6.6150*	*0.0101*
Results are different for students in minoritized groups	*24:9*	*19:15*	*2.0669*	*0.1505*

Table 1. ***Results of coding of open-ended survey items.***
Instances of codes are tabulated from open-ended survey item responses from 242 students in the R1 environment. In each entry for Signal (Column 2) and Null (Column 3) the occurrences are presented as 'PositiveInstances:NegativeInstances'. The Null ratio of codes used as a control is taken from all codes not related to features of the public exam for the same population of students. Significance tests compare Signal ratios to Null ratios (which are themselves conservatively more positive than 1:1) using a Chi-squared test statistic. The purpose of this table is to show which codes were found to have statistically significant presence in students' unprompted self-reported experiences, and whether those codes had an impact that is likely to be positive or negative on learning.

Within-exam experimentation:

Comparisons of exam outcomes on mixed traditional/public exams were used to quantitatively assess possible issues of equity. Student exam outcomes on public and traditional exam questions were analyzed for two groups of students: a university-identified diverse group of students in the Educational Opportunity Project (EOP), and the rest of the student population. As shown in Figure 4, we observed in our model that all students performed better on public exam questions compared to traditional exam questions (blue lines). Because of the differences in learning processes between public and traditional exam questions, this difference in performance is not evidence of learning differences between contents assessed in a given method. We also observed the expected decrease in high-stakes exam scores across question types for students from EOP minoritized groups (red lines). The combination of these trends was consistent for students in both EOP and non-EOP groups, giving no indication that public exam questions resulted in increasing inequity.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

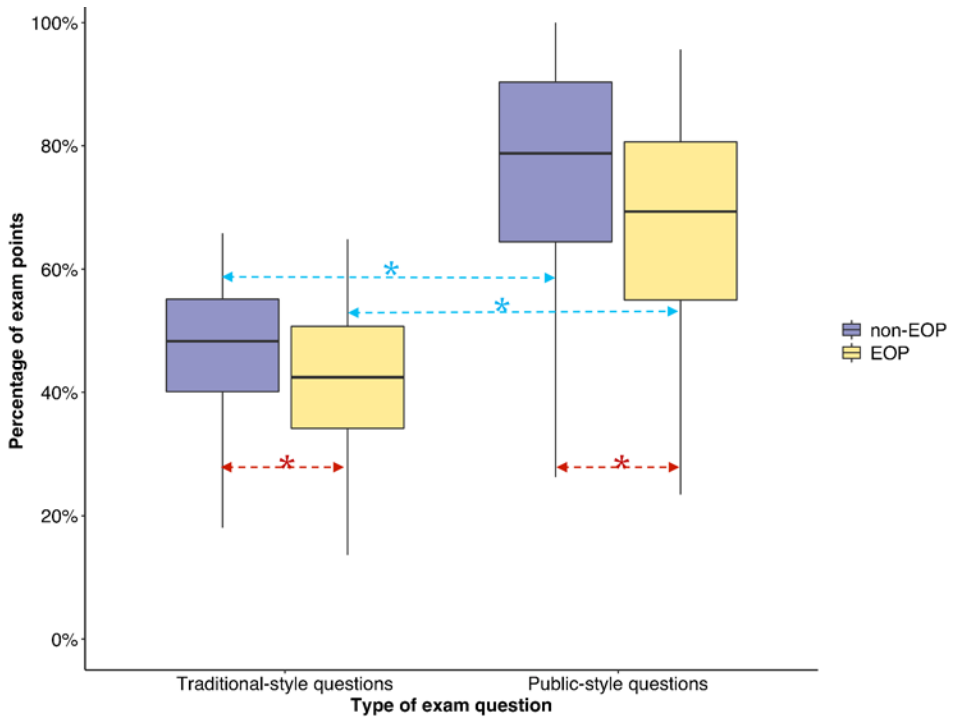

Fig 4. ***Exam outcomes for traditional- and public-style exams.***
Color plots are separated by underserved EOP group in yellow and non-EOP (majority) group in purple. Significant differences were found in the higher scores for students on public style exam questions as compared to traditional exam questions (indicated with blue asterisks), although the difficulty or achievement on these exam questions cannot be directly compared as the learning structures were different. Significant differences were found in exam scores between groups of students, which is consistent with pernicious gaps in outcomes in postsecondary education (indicated with red asterisks). No differences in the patterns of outcomes for traditional/public exam questions were found in either group of students, which is consistent with public exams being similarly equitable compared to traditional exams. The purpose of this figure is to display the outcomes of this experiment intended to observe any differences in equitable treatment of students if they exist.

Student self-reported preferences:

While student self-reported data is insufficient for establishing research findings, it can be useful in broad or initial research investigations to help observe experiential problems in an intervention. Our self-reported data is presented here for completeness. In the large R1 course, students were asked about their preferences for public or traditional exam questions. After experiencing two mixed exams with 15 public and 10 traditional questions each, 41% of students preferred to keep the same distribution for future exams, 3% of students wanted more traditional questions, and 56% of students wanted future exams to have a greater proportion of public-style questions. After listening to this student voice and increasing the proportion of public questions for the following exam, students were surveyed with the same options. After this exam with 20 public and 5 traditional questions, 67% of students wanted to keep the increased 20:5 distribution while 6% wanted more traditional questions and 24% wanted more than 20 of the 25 questions to be public. Course instructors kept the 20:5 ratio for the next exam, and students after this exam were given only two options so as to better understand the preferences of the majority of students. In this final survey prior to the final exam, 15% of students wanted to

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

decrease the number of public questions and 85% wanted to increase it. Throughout these exams and the overall self-reported desire for more public exam questions than traditional questions, there was no significant signal for a demographic basis on which these preferences were made, nor was preference correlated with course grade outcomes.

Does editing of the exam impact students?

As part of the public exam, students were given the opportunity to suggest edits or contributions to the public exam document itself. Three examples of the kinds of edits suggested by students were:

- Highlighting a grammatical error in the exam: The initial public exam had a question that ended with "...is likely to experience which of the following symptoms effects." A student responded via survey by writing "What do you mean by "symptoms effects"? Is this asking which symptoms the patient will experience?". This made clear to the exam authors that the word 'effects' was confusing and could be removed.
- Suggesting an improvement to the grammar in the exam: An initial public question used the word 'reasonable', and a student noted *"... 'reasonable' is a subjective and vague descriptor here, leaving it open to different interpretations."* The student went on to suggest that the exam writers should *"...either including a more precise definition of what you mean by 'reasonable' in the question or using a different word that more clearly gets at what you are looking for in this question would make it easier to understand. For example, by reasonable do you mean 'could possibly happen' or 'is likely to happen'?*" The authors used one of these suggestions in later versions of the exam.
- Suggesting creative text to complete a question: A public question asked students to assess the conclusions that could be drawn from a given graph on clinical outcomes for patients with diabetes. A student suggested that one of the possible answers could be "*Based on these graphs, should we be optimistic about the progress of diabetes care in the United States?*". This answer choice was not taken up as written by exam authors, but did catalyze the use of a similar incorrect answer choice for a later version: "*Based on these graphs, should we be pessimistic about the progress of diabetes care*?".

Students who undertook these optional, non-credit opportunities, when controlling for course grades and demographic backgrounds, were significantly more likely to perform better in their overall course grade (p value = 0.000402). This result suggests that the act of being engaged and legitimately contributing to the exam, even for non-content contributions, may help students learn the concepts. We present this interesting but relatively less-important finding for completeness.

Discussion:

We have described above an initial mixed methods research investigation on the impacts of public exams for students. Here we discuss the results in light of our research questions, as well as future research questions and limitations of this work.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In what ways do public exams impact the student experience?

We observed impacts of the public exam system on the student experience through analysis of three types of data: open-ended survey responses, self-reported preference surveys, and modeling of student outcomes based on a feature of student behavior around exam editing.

Analysis of students' open-ended survey responses showed an overall significant and positive impact of public exams on student experiences in a large STEM course. The positive impact of public exams on student experiences was significant even when controlled against other student responses in the same environment. They also triangulate well with themes from interviews, self-reported preference surveys, and with anecdotal narratives from public exam practitioners more widely. The aspects of the student experience that were significantly positive were in 1) Directing students to core concepts, 2) Deepening thought in the exam experience, and 3) Helping students to address problems around anxiety or confidence.

1) Directing students to core concepts speaks directly to a consistent challenge for novice learners. While accepting the deluge of information present in any fast-paced course, novice learners struggle to develop mental models to organize incoming information (26). Modern courses typically offer an array of learning materials to assist students in developing understanding of which pieces of information are core to the discipline and which pieces of information are facts or ideas that simply reinforce the concepts that an instructor feels are core to mastering the material in their course. Within the public exam structure, students have early access to exam materials that are directly connected to the reinforcement scheme of the course (typically, in course points). Instead of deducing core concepts from a string of lectures, assignments, study guides and other sources, students in a public exam course have the opportunity to infer value by placement (or not) on the actual assessment itself. Meta-contextual clues like the amount of exam points that can be earned can be a powerful reminder for students to study THIS skill and not THAT one. In contrast, traditional exams hide these valuable assessments until the moment of the exam itself. For students in multiple courses or studying while maintaining employment or families, efficiency in deciding which parts of the course to study can help learning and keep college work manageable. The significant, positive impact of 'Directing to core concepts' on public exams may be a reflection of these benefits to learning. In open ended responses in which students were asked "*Did the style of exams in [the R1 course] work for you? Why or Why not?"* students reported that having access to some part of the exam ahead of time allowed them to focus on what was important instead of feeling overwhelmed by all the content. As one R1 participant said:

> *"...they provide me with some direction on what to study a lot for. I think that there's a lot of material that's covered in this course throughout the lectures, and it would be hard to remember every single detail from the textbook, so I think the guidance of the public questions really helps you to look back at that specific part in your notes and/or the lecture to refresh your memory on what you learned."*

Many instructors are frequently asked by students before exams, "What do I need to know for the exam?" Perhaps similar to some types of practice exams given before an exam, public exams were seen to provide a similar type of focus on important content.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2) Deepening thought for students was an original motivating factor in early development and implementation of the public exam style. For instructors, the 'flattening' of thought required by the logistical constraints in many types of assessments has been a constant source of dismay. Further research should assess directly for creativity and critical thinking, but evaluation of responses for those characteristics is daunting, especially at scale. It is possible that benefits from public exams come from the increase in higher-order exam question (Anderson & Krathwohl, 2001; Barnett & Francis, 2012; Lemons & Lemons, 2013), which was the intent of the designers but not rigorously assessed in this study. The significant, positive benefits from the public exam style may be due to shifting exam-provoked thought from a one-time performance into a longer and more collegial set of learning cycles (Schwartz et al., 1999). Because students are less limited by the time needed to read and comprehend a complex exam scenario, more interesting scenarios can be approached by the instructor. Assessment materials transmit the values of the instructor into real terms (G. Wiggins, 1998, 2011). Moreover, students can spend their valuable study time working on intriguing, layered problems instead of re-hashing simple factual information. Students reported being challenged by the public exam format to more in-depth learning of a concept. In interviews, students realized that with the extra time to think about and discuss exam questions, there was an expectation of exam responses that demonstrated deeper thought and synthesis. For example, a CC student said:

> *"Personally I liked this type of exam a lot more. I didn't feel like I had to memorize anything. More like I understood the concept and could be asked questions about [it] from multiple angles. It helped learning with others as well because when explaining to other people a certain topic, and they begin to understand tells me that I understand the concept exceptionally well."*

As more disciplines make calls for deeper critical thinking skills (AAAS, 2011b; Halpern, 2001; McConnell et al., 2019; *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics*, 2012), it is possible the pre-release of exam material (as in (Crowther et al., 2020)) is a motivating factor in pushing students to do, share, and enjoy this deeper thought.

3) Anxiety around education (and more specifically exams) is a constant and increasingly-pressing concern (Disability, 2017; Health, 2020). While this is well-studied in STEM courses (Cooper et al., 2018; Downing et al., 2020; Schussler et al., 2021), it may be more relevant instead to courses for which high-stakes exams are a primary feature (Brady et al., 2018; Culler & Holahan, 1980; Harris et al., 2019). STEM courses (among many others) generally meet this description (Momsen et al., 2010). Learning is maximized at moderate levels of stress (Rudland et al., 2020), but greater stress hampers learning and motivation and disproportionately impacts students from groups traditionally underrepresented in the holders of college degrees (Lee et al., 2021; Medina, 2011; Misra & McKean, 2000; Vaidya & Mulgaonkar, 2007). There is some indication that this current most-diverse and most-economically challenged generation of students in college are also understandably the most over-stressed that have ever enrolled (Lederer & Hoban, 2020)l. With less anxiety associated with the surprise of the exam, they were able to feel more confident and prepared. A R1 student noted:

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

"... with the availability of the public exam I am able to study the possible directions the questions might take. It reduces the amount of stress and anxiety I usually get when I take exams, I feel more prepared."

Students reported a decrease in anxiety, albeit not always initially. Student experiences suggest that the positive perception of these exams takes time and that students need to get used to the new exam style. A CC student described this evolution of mindset:

"At first It was a bit of an adjustment because I had never taken a public exam, but the second time around I enjoyed it."

This sentiment was reiterated by a R1 student:

"During the first exam of the quarter, the style of the exams did not work for me because the format was new and I barely knew what to do to prepare for it. As of now, the style of the exams is working for me because even though I second guess myself…"

Public exams may help students to alleviate some of their stress through some familiarity with the assessment itself. The non-content information like formatting can be comprehended at relative leisure. Strategic points like where to focus effort and time can be usefully discussed and digested at home. Shifting non-content mental effort out of the exam performance time may explain why coding analysis shows better outcomes in public exams and would be in line with prior research (Hacker et al., 2008; Pate et al., 2019). It is also possible that the steps made towards exam transparency have a role to play, as signals of equitable behavior on the part of powerful authorities may suggest to students that they need not worry about being caught in a negative power-dynamic over some other disputed element within assessment (Bang & Medin, 2010; Bell et al., 2012; Fredricks et al., 2004).

Beyond these three emergent aspects of the student experience, student self-reported preferences for exam style were strongly in favor of public exams. While it appears that students prefer public exam questions in this context, and that those preferences are not explainable by demographics or class success, these data are presented only as a triangulation of other data sources. If these preference surveys can be taken at face value, then student preferences for public exam questions are relatively strong and in accordance with findings from open-ended coding and qualitative interview analysis.

Public exams include opportunities for students to authentically engage in the creation of the assessment through edits and suggestions. Students who took advantage of these opportunities also performed better in the class. Those edits are sparse among many exam questions, and the changes suggested rarely alter content, so this trend is unlikely to be explainable by gains on the specific question edited by the student. The statistical model used controlled for demographics and for student course grade, so it is less likely that this is a self-selection of which students choose to take on this extra task. If the correlation observed (p value = 0.000402) indicates a causative relationship, then it may be explainable in one of three ways. It might be that students who engage with the exam in this editorial mode are finding a new way to engage with the material. By seeing the content from a different angle, one more closely aligned with the perspective of the faculty instructor, they may find their own perspective on the content to be broadened in useful ways. This is in line with learning theory about critical thinking skills (Halpern, 2001). A second possibility is that engaging with assessment as a partner, even

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in a temporary way, may help students to feel authentically involved in the process of assessment. Affective impacts can improve learning (Dweck, 1986), so this specific observation would be in line with learning theory. Lastly, it is possible that this result conflates students who did not provide edits with students who never accessed the public exam materials (even after frequent instructor guidance), which might contribute to their lower course grade. In the first two models, the benefit to student learning would be valuable and further research will be required to better understand how, for which students, and under what conditions this benefit is generated.

Do public exams impact grade inequity?

Prerequisite to understanding more about the specific impacts of public exams, and as part of feminist and anti-racist drives within education research, we want to ensure that public exams do not contribute to the extant inequities in student outcomes within postsecondary education (Museus et al., 2015). Those concerns are most pressing for assessments, which often represent a gateway for student success at which inequities are both created and revealed. The primary goal of our quantitative within-exam experimental design in a large R1 course was to help understand if public exams are creating or exacerbating inequities for students from groups historically marginalized in postsecondary education. Analysis of question-by-question exam outcomes in a large course is our most likely opportunity to observe a signal of inequitable outcomes. Close analysis of question-by-question outcomes make clear that these pernicious gaps in outcomes exist beyond our research environment: Students from minoritized groups are associated with lower scores on both public and traditional exam questions. Clearly, improving outcomes for all students will take much more than the use of public exams. Of particular importance for our study is that outcome gaps are not exacerbated by public exams. In other words, the gaps between public and traditional question outcomes are not different between groups of students. While we could imagine a hypothetical situation where some benefits from an intervention might be so positive as to be worth some negative impact on equity, it is relieving to know that this choice does not appear to be necessary and that public exams appear to be as inequitable or equitable as existing traditional exams.

Are public exams likely to be applicable across postsecondary contexts?

Our analysis is largely based on data collected in an R1 institution. While R1 institutions are frequently the site for postsecondary education research projects, they account for a relatively small proportion of postsecondary students. Crucially, interventions must be useful in larger contexts like primarily undergraduate institutions, comprehensive colleges, and (perhaps most importantly) the vast community college system. To better understand whether public exams might be applicable to community college courses, which are generally smaller and less available to quantitative research, we undertook a similar qualitative study in a community college course.

For example, a CC student noted that:

> "*we were able to sit down and start bouncing information off of each other and asking different questions about the questions...just kinda sharing information right before the*

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

exam and that just gave me so much confidence as to how much I know going into the exam so"

This student suggests a deeper questioning style beyond memorization, and notes the affective impact of this practice as well. A second CC participant mentioned:

"*it helps more with like understanding but sometimes when you're panicking about an exam you're like 'I don't want understanding ; I just wanna know' but at the same time you do have to understand things...if we hadn't had the public exam I would have studied all five of the chapters and had like less knowledge on each of the things and I don't feel like I would have remembered the exact definition of phenotypic plasticity as well as like when I saw the question and was like, I really do need to know this for the exam.*"

These three themes of Anxiety, Directing to core concepts and Deepening thought are evident here and were strongly present in both environments. Weaker themes of collaboration, language issues on exams, and authentic engagement with assessment were evident in both environments but less so. While we did identify emergent themes in this work, no thematic signals appeared to us in one environment and not the other. This is an initial attempt to explore the possible broad application of public exams, and clearly more research will be required on a greater scale to make similar conclusions. In the meantime, the outcomes of these analyses are consistent with public exams being similarly applicable across these two institution types.

Limitations of this study:

As an initial foray into research on public exams, this study has many limitations. The design-based research model used in this study is likely to unearth important features of the student educational experience. However, this model is not intended to prove that a particular feature is more or less important than another, or to compare overall impacts of the student experience on learning or career success. Future research, using longitudinal analysis and topical challenges, will be important for assessing the overall impacts of the public exam intervention beyond these initial analyses. Constructs like anxiety are treated as emergent themes; future research should apply established theoretical frameworks around anxiety to make use of established survey instruments that may be a good fit to better understand the ways in which and extent to which public exams affect student anxiety. The core features of public exams are examined as a unit, and more work will be required to understand if benefits can be achieved modularly. Largely a single-course study, this analysis may be conflated by the specific instructors or the environment of early 2021 (in itself, a unique time to be working in postsecondary education during a pandemic). Education impacts tend to be relatively weak in comparison to impact size of other fields, so it is possible that other important features have gone unexamined for lack of analytic power in a single course of 300 students. This is especially true for particular groups of students of historic importance, for whom numbers are smaller and backgrounds unique to this particular study environment. Furthermore, the newness of the public exam style in post-secondary classrooms means that existing instruments for investigating assessments on more traditional models was not appropriate. Future research should involve validation of specific instruments for assessing these learning cycles, such as those seen in (Arikan et al., 2022; Chang et al., 2021; Hicks et al., 2017; Johnson et al., 2022;

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Reynders et al., 2020). Perhaps most importantly, this study did not directly assess student learning but rather the student experience. We hope that the benefits demonstrated, combined with positive anecdotal reports on the strengthened student/instructor relationships in similar courses, motivate future research to better understand how varied assessment styles can better serve the next generation of students and improve on this work.

One salient criticism of public exams is that the process can be summarily characterized as 'teaching to the test'. This pejorative has a long and well-deserved history in K-12 education, especially in situations where externally-created assessments are linked to a motivation to maximize scores for the purposes of accumulating outcome-linked resources (Jensen et al., 2014; Ravitch, 2020). We propose that many college and university exams are fundamentally different in that the instructors have wide purview to create exactly the kinds of assessments that reflect the values, skills and content needed in modern pursuits. In other words, professors can create the kinds of exams for which 'teaching to the exam' is a great thing for students. Creating worthwhile assessments that help students to develop relevant and high-level skills is a core principle of educative assessment (Jensen et al., 2014; G. Wiggins, 1998). We hope that public exams are a useful way to do this.

Considerations for interested practitioners:

Transitioning from traditional exams to a public exam style is a low-tech strategy to employ many of the practices identified in education literature to improve student learning. Instructors found that they could make simple changes to the exams or exam blueprints that they were already using by withholding some of the information. In many cases these adjustments shorten the exam by augmenting the higher cognitive exam questions and allowing students to discuss core concepts in more detail because students had more time to reflect on the question. Additionally, instructors were receiving meaningful feedback from students during the editing process of their new public exam that improved their exam questions. Importantly, instructors do not need to adjust the entire exam to the public method. Instructors can slowly transition to a greater percentage of the exam being publicly available over the quarter or semester or academic year. Anecdotally, students were excited to be part of the public exam process and a new assessment strategy that they participated in. This is the first research that we know of that has examined the impact of public exams on R1 and CC students. Our research suggests that public exams do not appear to create additional inequity, work similarly for R1 and CC students and, perhaps most importantly, are valued by students themselves. More research is going to be important to understand the impacts this type of exam has on student learning, particularly with respect to anxiety and impacts on students from minoritized groups.

Postsecondary instructors have numerous choices when designing exams (Gezer-Templeton et al., 2017; Hodges, 2004; Knierim et al., 2015; Wieman et al., 2014)**.** For those who want to take up public exams as a classroom practice, we suggest adjusting a small number of questions on an upcoming exam into a public, pre-released style. This helps create a positive feedback loop for instructor design and feedback from students, and it also helps to

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

avoid taking on an unsustainable overhaul of all assessment in one course. In our experience, instructors who take up a few challenging pre-released exam questions a) quickly develop the communication needed for students to understand how and why to access the materials, and b) invariably lead to greater use of these methods in future assessments. Discussing an exam draft with someone experienced in public exams is especially useful; please do write to the corresponding author if this would be useful for you. A few examples of public exams (both pre-released and final versions) are available here in supplemental materials. An earlier, deeper, non-peer-reviewed logistical discussion of public exams within the field of molecular biology may be of interest to practitioners (B. L. Wiggins, 2019).

As already discussed, anxiety around education, particularly associated with exams, does not impact all groups of students equally. We have proposed that public exams may be a strategy to address some of the anxiety associated with taking exams. It is important to note that this student adjustment period as instructors move away from a more traditional exam may be longer for some students compared to others. Instructors may need to provide guidance and support during this adjustment period into the exam process. Some strategies that could facilitate a smoother transition are starting off with lower stakes quizzes or exams, practice assignments or quizzes, or setting up student groups where students can support each other. Although we did not find support for “Collaboration” in the quantitative coding analysis, at least some students recognized the advantage in collaboration when preparing for the exam. A R1 student described this by saying:

> *“I have noticed that it only works for me when I work with other people in study sessions. I try to study on my own. I have a more difficult time understanding the material, which is something quite new to me since I am used to studying on my own. But overall I like it.”*

Students may not have recognized that collaboration was not only acceptable but highly encouraged, often not utilizing that strategy until later exams. As a CC participant explained:

> *“I loved the second exam because I was able to meet up with others outside of the classroom to go over a couple different concepts before the exam.“*

Emphasizing and encouraging collaboration as a strategy for student success on the exam, may be another way the instructor can facilitate the transition from a more traditional exam model.

Conclusion:

In an initial study, we analyzed the impacts of public exams on student class experiences. Our mixed-methods design research shows that students find significant positive impacts on their experiences. Those impacts are largely focused on improving the direction of students to core concepts, the deepening of thought in the assessment process, and structural assistance for students in managing negative stress and anxiety. The public exam method is likely to be similarly equitable to traditional methods and potentially applicable across institutional contexts without exacerbating issues of educational equity. We present this work in the spirit of improving assessment for all students as a core feature of critical, high-quality education.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments:
This work was funded by the generosity of a private, independent research gift from CourseHero, without which none of this research would have been possible. Participants and their data are protected by Institutional Review Boards under #s STUDY00012237 (University of Washington), ECIRB-20210512 (Edmonds Community College) and IRB-2020-0813 (Oregon State University). Thank you to the volunteers and staff at Seattle Public Libraries for maintaining access to research-quality facilities even throughout a global pandemic. We appreciate the efforts and contributions of Greg Crowther, Deb Donovan, Kelly Hennessey, Lori Kayes, Devon Quick, Christine Savolainen, Mandy Schivell, Katie Simons, Shelley Stromholt, Jeannette Takashima, Liz Warfield, and Seth Wiggins. C.A.B was supported by an NSF Graduate Research Fellowship Program Grant No. 026257-001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bibliography:

1. Stobart G, Eggen T. High-stakes testing–value, fairness and consequences. Assess Educ Princ Policy Pract. 2012;19(1):1–6.
2. Wideen MF, O'Shea T, Pye I, Ivany G. High-stakes testing and the teaching of science. Can J Educ Can Léducation. 1997;428–44.
3. Ambrose SA, editor. How learning works: seven research-based principles for smart teaching. 1st ed. San Francisco, CA: Jossey-Bass; 2010. 301 p. (The Jossey-Bass higher and adult education series).
4. Handelsman J. Scientific teaching. New York: W.H. Freeman and Co.; 2006.
5. Intemann K. Why diversity matters: Understanding and applying the diversity component of the National Science Foundation's broader impacts criterion. Soc Epistemol. 2009;23(3–4):249–66.
6. Hurtado S. Linking diversity with the educational and civic missions of higher education. Rev High Educ. 2007;30(2):185–96.
7. Rossing JP, Lavitt MR. The neglected learner: A call to support integrative learning for faculty. Lib Educ. 2016;102(2):34–41.
8. Wiggins G. Educative Assessment. Designing Assessments To Inform and Improve Student Performance. ERIC; 1998.
9. Wiggins G. A true test: Toward more authentic and equitable assessment. Phi Delta Kappan. 2011;92(7):81–93.
10. Buxton CA, Allexsaht-Snider M, Suriel R, Kayumova S, Choi Y, Bouton B, et al. Using educative assessments to support science teaching for middle school English-language learners. J Sci Teach Educ. 2013;24(2):347–66.
11. Jönsson A. Educative assessment for/of teacher competency. A study of assessment and learning in the" Interactive examination" for student teachers. Malmö University; 2008.
12. Fink LD. A self-directed guide to designing courses for significant learning. Univ Okla. 2003;27(11):1–33.
13. Sweller J. Cognitive load theory: Recent theoretical advances. 2010;
14. Brown B. Braving the wilderness: The quest for true belonging and the courage to stand alone. Random House; 2017.
15. Zeichner KM, Miller L, Silvernail DL, Darling-Hammond L, American Association of Colleges for Teacher Education, National Commission on Teaching & America's Future (U. S.). Studies of excellence in teacher education□: preparation in the undergraduate years. Washington, DC: AACTE Publications; 2000. xi, 109 p. p.
16. Darling-Hammond L, Bransford J. Preparing teachers for a changing world□: what teachers should learn and be able to do. 1st ed. San Francisco, CA: Jossey-Bass; 2005. xxx, 593 p. p.
17. Sawyer RK. The Cambridge Handbook of the Learning Sciences. 2nd ed. Cambridge University Press; 2006. 784 p. (Cambridge Handbooks in Psychology).
18. Kirschner PA. Cognitive load theory: Implications of cognitive load theory on the design of learning. Elsevier; 2002.
19. Moreira BFT, Pinto TSS, Starling DSV, Jaeger A. Retrieval practice in classroom settings: a review of applied research. In: Frontiers in Education. Frontiers; 2019. p. 5.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20. Bonwell CC, Eison JA. Active Learning: Creating Excitement in the Classroom. Washington, D.C.: The George Washington University, School of Education and Human Development; 1991.
21. Obuaku-Igwe C. Teaching and Re-Imagining the Role of Medical Sociology in South Africa During COVID-19: A Reflection. In: Strategies for Student Support During a Global Crisis. IGI Global; 2021. p. 175–94.
22. Shi T, Blau E. Contemporary Theories of Learning and Pedagogical Approaches for All Students to Achieve Success. In: Optimizing Higher Education Learning Through Activities and Assessments. IGI Global; 2020. p. 20–37.
23. Collins A, Joseph D, Bielaczyc K. Design Research: Theoretical and Methodological Issues. J Learn Sci. 2004 Jan 1;13:15–42.
24. Sandoval W& B. Design-based Research Methods for Studying Learning in Context. Educ Psychol. 2004;39:3.
25. Meyer H. Novice and expert teachers' conceptions of learners' prior knowledge. Sci Educ. 2004;88(6):970–83.
26. Bransford JD, Brown AL, Cocking RR. How people learn: Brain, mind, experience, and school. National Academy Press; 1999.
27. Dahlberg C, Lee S, Leaf D, Lily L, Wiggins B, Jordt H, et al. A Short, Course-Based Research Module Provides Metacognitive Benefits in the Form of More Sophisticated Problem Solving. J Coll Sci Teach [Internet]. 2019 [cited 2019 Jul 29];048(04). Available from: https://www.nsta.org/store/product_detail.aspx?id=10.2505/4/jcst19_048_04_22
28. Wiggins BL, Sefi-Cyr H, Lily LS, Dahlberg CL. Repetition Is Important to Students and Their Understanding during Laboratory Courses That Include Research. Vol. 22, Journal of Microbiology & Biology Education. 2021. p. e00158-21.
29. Wiggins BL, Eddy SL, Wener-Fligner L, Freisem K, Grunspan DZ, Theobald EJ, et al. ASPECT: A Survey to Assess Student Perspective of Engagement in an Active-Learning Classroom. Cbe-Life Sci Educ. 2017 Jun 20;16.
30. Rubin HJ. Qualitative interviewing: the art of hearing data. 3rd ed. Thousand Oaks, Calif.: SAGE; 2012.
31. Glaser BG, Strauss AL. The discovery of grounded theory: strategies for qualitative research. London,: Weidenfeld and Nicolson; 1968. xiii, 271 p. p. (Observations).
32. Linehan MM. Cognitive-behavioral treatment of borderline personality disorder. Guilford Publications; 2018.
33. Cameron J. Focusing on the focus group. In: Qualitative Research Methods in Human Geography. Oxford: Oxford University Press; 2005. p. 116–32.
34. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant. 2018;52(4):1893–907.
35. Wright CD, Eddy SL, Wenderoth MP, Abshire E, Blankenbiller M, Brownell SE. Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. CBE—Life Sci Educ. 2016;15(2):ar23.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

36. Freeman S, Theobald R, Crowe AJ, Wenderoth MP. Likes attract: Students self-sort in a classroom by gender, demography, and academic characteristics. Act Learn High Educ. 2017;18(2):115–26.
37. Stanich CA, Pelch MA, Theobald EJ, Freeman S. A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women. Chem Educ Res Pract. 2018;19(3):846–66.
38. Odom S, Boso H, Bowling S, Brownell S, Cotner S, Creech C, et al. Meta-analysis of Gender Performance Gaps in Undergraduate Natural Science Courses. CBE—Life Sci Educ. 2021;20(3):ar40.
39. Cooper KM, Auerbach AJJ, Bader JD, Beadles-Bohling AS, Brashears JA, Cline E, et al. Fourteen Recommendations to Create a More Inclusive Environment for LGBTQ+ Individuals in Academic Biology. CBE—Life Sci Educ. 2020 Jul 14;19(3):es6.
40. McCrum-Gardner E. Which is the correct statistical test to use? Br J Oral Maxillofac Surg. 2008;46(1):38–41.
41. Bower KM. When to use Fisher's exact test. In: American Society for Quality, Six Sigma Forum Magazine. 2003. p. 35–7.
42. Anderson LW, Krathwohl DR, editors. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Complete ed. New York: Longman; 2001. 352 p.
43. Lemons PP, Lemons JD. Questions for assessing higher-order cognitive skills: It's not just Bloom's. CBE—Life Sci Educ. 2013;12(1):47–58.
44. Barnett JE, Francis AL. Using higher order thinking questions to foster critical thinking: A classroom study. Educ Psychol. 2012;32(2):201–11.
45. Schwartz DL, Lin X, Brophy S, Bransford JD. Toward the development of flexibly adaptive instructional designs. Instr-Des Theor Models New Paradigm Instr Theory. 1999;2:183–213.
46. AAAS. Vision and Change in Undergraduate Biology Education□» Final Report [Internet]. 2011 [cited 2015 Oct 11]. Available from: http://visionandchange.org/finalreport/
47. Presidents Council of Advisors on Science and Technology. Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. 2012.
48. McConnell KD, Horan EM, Zimmerman B, Rhodes TL. We Have a Rubric for That: The VALUE Approach to Assessment. ERIC; 2019.
49. Halpern DF. Assessing the effectiveness of critical thinking instruction. J Gen Educ. 2001;50(4):270–86.
50. Crowther G, Wiggins B, Jenkins L. Testing in the Age of Active Learning: Test Question Templates Help to Align Activities and Assessments. HAPS Educ. 2020 May 6;24:74–81.
51. Disability NC on. Mental health on college campuses: Investments, accommodations needed to address student needs. 2017;
52. Health C for CM. 2020 Annual report (Publication No. STA 21-045). 2020;
53. Schussler EE, Weatherton M, Chen Musgrove MM, Brigati JR, England BJ. Student Perceptions of Instructor Supportiveness: What Characteristics Make a Difference? CBE—Life Sci Educ. 2021;20(2):ar29.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

54. Cooper KM, Downing VR, Brownell SE. The influence of active learning practices on student anxiety in large-enrollment college science classrooms. Int J STEM Educ. 2018;5(1):1–18.
55. Downing VR, Cooper KM, Cala JM, Gin LE, Brownell SE. Fear of Negative Evaluation and Student Anxiety in Community College Active-Learning Science Courses. Vol. 19, CBE—Life Sciences Education. 2020. p. ar20.
56. Brady ST, Hard BM, Gross JJ. Reappraising test anxiety increases academic performance of first-year college students. J Educ Psychol. 2018;110(3):395.
57. Culler RE, Holahan CJ. Test anxiety and academic performance: The effects of study-related behaviors. J Educ Psychol. 1980;72(1):16.
58. Harris RB, Grunspan DZ, Pelch MA, Fernandes G, Ramirez G, Freeman S. Can test anxiety interventions alleviate a gender gap in an undergraduate STEM course? CBE—Life Sci Educ. 2019;18(3):ar35.
59. Momsen JL, Long TM, Wyse SA, Ebert-May D. Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. CBE—Life Sci Educ. 2010;9(4):435–40.
60. Rudland JR, Golding C, Wilkinson TJ. The stress paradox: how stress can be good for learning. Med Educ. 2020;54(1):40–5.
61. Lee J, Jeong HJ, Kim S. Stress, anxiety, and depression among undergraduate students during the COVID-19 pandemic and their use of mental health services. Innov High Educ. 2021;1–20.
62. Misra R, McKean M. College students' academic stress and its relation to their anxiety, time management, and leisure satisfaction. Am J Health Stud. 2000;16(1):41.
63. Vaidya PM, Mulgaonkar KP. Prevalence of depression anxiety and stress in undergraduate medical students and its co-relation with their academic performance. Indian J Occup Ther Indian J Occup Ther. 2007;39(1).
64. Medina J. Brain rules: 12 principles for surviving and thriving at work, home, and school. ReadHowYouWant. com; 2011.
65. Lederer AM, Hoban MT. The development of the American College Health Association-National College Health Assessment III: An improved tool to assess and enhance the health and well-being of college students. J Am Coll Health. 2020;1–5.
66. Pate A, Lafitte EM, Ramachandran S, Caldwell DJ. The use of exam wrappers to promote metacognition. Curr Pharm Teach Learn. 2019;11(5):492–8.
67. Hacker DJ, Bol L, Keener MC. Metacognition in education: A focus on calibration. Handb Metamemory Mem. 2008;429455.
68. Bang M, Medin D. Cultural processes in science education: Supporting the navigation of multiple epistemologies. Sci Educ. 2010;94:1008–26.
69. Bell P, Tzou C, Bricker L, Baines A. Learning in Diversities of Structures of Social Practice: Accounting for How, Why and Where People Learn Science. Hum Dev. 2012 Jan 1;55:269–84.
70. Fredricks JA, Blumenfeld PC, Paris AH. School engagement: Potential of the concept, state of the evidence. Rev Educ Res. 2004;74(1):59–109.
71. Dweck CS. Motivational processes affecting learning. Am Psychol. 1986;41:1040.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.15.488479; this version posted August 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

72. Museus SD, Ledesma MC, Parker TL. Racism and Racial Equity in Higher Education. ASHE High Educ Rep. 2015 Nov 1;42(1):1–112.

73. Johnson C. Teaching to the test: How schools discourage phronesis. In: Vice Epistemology. Routledge; 2020. p. 225–38.

74. Ravitch D. How, and How Not, to Improve Schools. N Y. 2020;

75. Jensen JL, McDaniel MA, Woodard SM, Kummer TA. Teaching to the test… or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. Educ Psychol Rev. 2014;26(2):307–29.

76. Gezer-Templeton PG, Mayhew EJ, Korte DS, Schmidt SJ. Use of exam wrappers to enhance students' metacognitive skills in a large introductory food science and human nutrition course. J Food Sci Educ. 2017;16(1):28–36.

77. Knierim K, Turner H, Davis RK. Two-stage exams improve student learning in an introductory geology course: Logistics, attendance, and grades. J Geosci Educ. 2015;63(2):157–64.

78. Wieman CE, Rieger GW, Heiner CE. Physics exams that promote collaborative learning. Phys Teach. 2014;52(1):51–3.

79. Hodges LC. Group exams in science courses. New Dir Teach Learn. 2004;2004(100):89–93.

80. Wiggins BL. The Public Exam System: Simple Steps to More Effective Tests. CourseHero Faculty Club [Internet]. 2019 Aug 6; Available from: https://www.coursehero.com/faculty-club/classroom-tips/benjamin-wiggins/