education sciences

MDPI

Article

Implementing Mastery Grading in Large Enrollment General Chemistry: Improving Outcomes and Reducing Equity Gaps

Joshua D. Hartman * and Jack F. Eichler *

Department of Chemistry, University of California, Riverside, CA 92521, USA
* Correspondence: joshua.hartman@ucr.edu (J.D.H.); jack.eichler@ucr.edu (J.F.E.)

Abstract: Specifications and mastery grading schemes have been growing in popularity in higher education over the past several years, and reports of specifications grading and other alternative grading systems are emerging in the chemistry education literature. The general goal of these alternative grading approaches is to reduce the reliance on high-stakes exams and give students a more transparent pathway to achieving the course learning outcomes. More importantly, relying less on infrequent high-stakes exams may help reduce historical equity gaps in introductory gateway STEM courses. Herein, we describe the implementation of two versions of mastery grading systems in large enrollment general chemistry courses at a public R1 institution. Class-wide course outcomes, equity gaps in performance on a common final exam, and student feedback on their experience navigating these grading schemes are presented. We show that combining mastery grading with interactive courseware tools improved the average performance on a common final assessment for under-represented minority (URM) students by 7.1 percentage points relative to an active control course that used infrequent high-stakes exams.

Keywords: mastery grading; testing effect; second-change testing; alternative grading

check for updates

Citation: Hartman, J.D.; Eichler, J.F. Implementing Mastery Grading in Large Enrollment General Chemistry: Improving Outcomes and Reducing Equity Gaps. *Educ. Sci.* **2024**, *14*, 1224. https://doi.org/10.3390/educsci14111224

Academic Editor: Daniel Muijs

Received: 29 September 2024
Revised: 25 October 2024
Accepted: 25 October 2024
Published: 7 November 2024

1. Introduction

Mastery grading has been discussed in the education research literature since the 1960s, notably in Bloom's proposal to develop broader mastery learning curricula [1]. Though several meta-analyses corroborate the general efficacy of mastery learning approaches, broader adoption of this assessment approach has not been observed in higher education, and the American higher educational system generally relies on high-stakes exams [2]. Research studies that specifically examine the impact of the mastery outcomes approach in higher education STEM remain limited but are beginning to emerge in the literature [3]. A mastery outcomes structure that utilized second-chance testing in an undergraduate engineering course resulted in significantly improved final exam performance relative to a course that used traditional high-stakes exams, and students in the mastery outcomes course earned twice as many As and half the number of failing grades. Most importantly, traditionally underrepresented students performed on par with non-URM students [3].

Alternative grading and assessment models have been explored across numerous STEM disciplines, including chemistry education [4]. For example, a learner-centered grading method using a standards-based assessment structure for general chemistry has been shown to improve grading transparency. This implementation did not quantify the impact on student learning outcomes, focusing instead on observational data related to the generally positive student learning experience [5]. In another study involving high school chemistry students, mastery learning improved performance and attitude toward learning [6]. Although the research is limited, these studies highlight the generally positive shift in student perspective when moving from a traditional high-stakes grading system to a mastery approach [7,8].

Another outcomes-based assessment approach that has gained recent attention in the chemistry education community is specifications grading [9]. The specifications grading system differs from the mastery outcomes (second chance testing) approach in that specifications grading allows students to demonstrate mastery by completing bundles of assignments or tasks (e.g., a letter grade of A can be earned by completing 9/10 components in the bundle, a letter grade of B can be earned by completing 8/10 components in the bundle, etc.). Within the chemistry education literature, specifications grading has been implemented predominantly in laboratory courses [7,8]. We speculate that this is due to this type of grading structure being well suited to lab courses that focus on skills and completion of tasks [7], and specifications systems can lead to mixed results with respect to student satisfaction [8]. Nevertheless, implementing the specification grading system in a large enrollment organic chemistry laboratory setting notably improved final letter grades [8].

Implementing a specifications grading system in a lecture setting represents a dramatic overhaul in course design from traditional points-based systems, where grades are largely determined by a single attempt, to high-stakes summative assessments [2]. As a result, traditional points-based grading remains the norm for lecture courses despite mounting evidence that points-based models increase student stress levels, decrease equity, and de-emphasize the acquisition of content knowledge [10]. Several innovative strategies have been examined for implementing specifications grading in general chemistry [11], organic chemistry [12–14], analytical chemistry [15], and upper-division chemical biology [16] lecture courses. Many of these studies primarily focus on qualitative aspects of the specifications grading implementations, highlighting reduced self-reported anxiety and generally positive feedback from professors. Hollinsed et al. did report an increase in the conversion of B students to A students. However, the specifications grading model did not significantly impact the number of lower-performing students [11].

The promise of improving the learning environment for students and professors while creating a more equitable grading system continues to motivate the development of alternative grading models. Recently, Noell et al. implemented a hybrid-specs grading system introducing an element of second-chance testing to shift the emphasis toward content mastery without a full course redesign [17]. The hybrid-specs system increased the conversion of B to A grades, but there was a small increase in the DFW rates using this hybrid-specs model. The success of the hybrid-specs model is likely tied to the testing effect. The testing effect is a framework based on research linked to retrieval practice, which is a critical component of the learning process [18]. However, for students to benefit from the second-chance testing model, they must have the skills, resources, and metacognitive strategies required to fill knowledge gaps, thereby improving scores on subsequent tests [19]. For this reason, recent literature suggests that the testing effect can decrease or even disappear as the complexity of the learning materials increases [20]. Yet, when regular testing is coupled with additional tools and resources, student performance gains have been realized for advanced topics directly related to chemistry [21].

Successful engagement with the mastery grading model is closely linked to a student's background, particularly through metacognitive development and familiarity with effective learning strategies. Therefore, comparing student performance across different demographics, such as familial education history, ethnicity, and income, can provide valuable insight into the design and implementation of alternative grading models. Previous reports did not study the specific impacts of these alternative grading strategies on traditionally underrepresented students [3,7,8,12,15,16]. A previous study implementing mastery learning in high school general chemistry did find a particularly pronounced positive effect on learning outcomes and attitudes from students who are struggling with the chemistry content [6]. However, the results were not disaggregated by ethnicity or familial education level.

To the authors' knowledge, the present work represents the first study that assesses the impact of a mastery grading system in large enrollment general chemistry courses relative to an active control course that used traditional high-stakes exams. This work highlights the importance of coupling second-chance testing with interactive courseware designed to promote asynchronous active learning and metacognitive development [22]. Particular emphasis is placed on exploring correlations between a student's familial education history, ethnicity, and socioeconomic status by looking at disaggregated student performance data. This study shows that supporting a mastery grading model with robust interactive courseware improved the average student performance for the entire class population on a common final assessment by 6.9 percentage points relative to a control using infrequent high-stakes exams. The improvement was more pronounced (11.6 percentage points) for first-generation college students with an URM background receiving financial aid.

1.1. Theoretical Frameworks

In this project, we implement a mastery outcomes assessment approach rooted in the theoretical frameworks of the testing effect [18] and mindset theory [23]. The testing effect is linked to the phenomenon of retrieval practice, in which it has been found that the act of retrieving information is, in some cases, a more impactful learning event than an information coding event, where information encoding refers to learning new knowledge [24]. The mastery outcomes approach implemented in this study naturally leverages the positive impacts of the testing effect by providing more frequent self-assessment scenarios. Furthermore, the mastery grading approach naturally facilitates the incorporation of metacognitive strategies.

Student feedback on mastery assessments is directly linked to interactive courseware content with built-in tools for helping students monitor the learning process by identifying gaps in their understanding and addressing them through targeted practice. The assessment, reflection, and practice cycle is designed to provide a pattern of engagement that promotes a growth mindset. Mindset theory is based on research indicating students who believe that intelligence is malleable often experience more positive learning outcomes. Students with a growth mindset tend to view initial failure as an opportunity for improvement rather than a predictor of future negative outcomes [23].

1.2. Research Questions and the Current Study

1. How do student performance outcomes differ between courses incorporating a mastery grading/test-retake system and a course using infrequent high-stakes exams?
2. How do courses incorporating a mastery grading/test-retake system affect equity gaps compared to a course using infrequent high-stakes exams?
3. What is the general qualitative student affective response to courses using mastery grading/test-retake system?

The present work employs a second-chance testing strategy through weekly unit mastery assessments (denoted as Mastery). However, we have chosen a mastery-focused model that dispenses with the token economy and allows every student a fixed number of scheduled retakes for a given mastery assessment [7,9]. This mastery grading approach directly fosters a growth mindset in students and demonstrates a commitment from instructors that students possess the capacity to improve through persistent effort. The cues hypothesis states that instructors who espouse a fixed mindset create threatening situational cues that can demotivate students, especially traditionally under-represented students [25]. By adopting a mastery outcomes assessment structure, instructors will show a commitment to a student growth mindset that should lead to improved cognitive and affective outcomes.

Coupling the mastery grading model with metacognitive coaching and interactive courseware designed to promote asynchronous active learning (Mastery+OLI) is a crucial complement to the mastery learning model. Suppose that students are not given guidance on evaluating and reflecting upon their learning. In that case, it is unlikely that having multiple attempts on the various content assessments will lead to meaningful learning gains [20]. The General Chemistry curriculum available through the Open Learning Initiative (OLI) at Carnegie Mellon University was selected for the present work. OLI General Chemistry provides a rich, interactive learning environment built upon the OpenStax Chemistry textbook with embedded problems that provide extensive hints and feedback [26,27]. The structure of OLI General Chemistry is based on the literature findings, which suggest students learn more by doing interactive problems rather than reading text or watching videos [28].

2. Materials and Methods

2.1. Implementation Details

The study involved the second course in a three-quarter sequence for general chemistry, which is the required introductory chemistry sequence for all students in the UCR College of Natural and Agricultural Sciences (CNAS). The topics are separated into six units: gases, thermochemistry, liquids and solids, solutions chemistry, thermodynamics, and kinetics. A quasi-experimental study compared two versions of mastery outcomes grading system to a teaching-as-usual course that used traditional high-stakes exams. Performance on a common final assessment administered to the three sections was used to compare student learning outcomes.

The study took place during the winter 10-week quarter at a large public research university in Southern California, federally designated as a Hispanic Serving Institution. All three sections were taught at the same time by three different instructors. Student demographic data were separated based on ethnicity, financial aid status, and first-generation status (see Table 1). Ethnicity and first-generation status were determined by self-reporting in admission files. First-generation status was defined as neither parent completing a 4-year degree (i.e., the highest level of education being "some college", "high school", or "some high school"). Student data were separated into two groups based on ethnicity. Students who self-reported as white or Asian were classified as not belonging to an underrepresented minority group (not URM). Students belonging to all other ethnic backgrounds were classified as URMs.

Table 1. Descriptive statistics for the performance on the common final exam across the three groups. The control (*N* = 239), mastery (*N* = 242), and mastery with OLI (*N* = 244) sections had similar total enrollment. YES corresponds to students belonging to the corresponding demographic population listed in the left-hand column.

	Control		**Mastery**		**Mastery+OLI**	
	YES	**NO**	**YES**	**NO**	**YES**	**NO**
URM:						
Mean Final Exam %	60.2	67.8	59.3	70.7	67.3	75.0
Stnd. Dev.	21.5	19.2	20.1	17.4	18.5	14.3
% of Population	38.5%	61.5%	28.5%	71.5%	39.4%	60.6%
First-Generation:						
Mean Final Exam %	54.6	69.5	58.9	71.0	67.0	74.6
Stnd. Dev.	20.7	18.6	17.8	18.3	19.4	14.1
% of Population	31.0%	69.0%	28.9%	71.1%	34.4%	65.6%
Financial Aid Status:						
Mean Final Exam %	64.2	67.5	66.4	72.9	71.6	73.4
Stnd. Dev.	20.7	19.2	19.2	16.7	16.8	15.3
% of Population	79.9%	20.1%	83.9%	16.1%	80.9%	19.1%

The control group used traditional publisher textbook resources, while the Mastery group used the Atoms First General Chemistry text available through OpenStax [26]. To ensure equitable access, students enrolled in the Mastery+OLI group were given free access to the Open Learning Initiative general chemistry resources. Care was taken to ensure content coverage was the same throughout the three sections. The instructor for the control group wrote a set of 20 common questions to be administered on the final exam for the Control, Mastery, and Mastery+OLI courses. The common questions covered various topics throughout the course and were given to the instructors running the mastery grading sections after the last day of instruction to minimize bias. The common test questions were evaluated for content validity by the Mastery and Mastery+OLI instructors, and though these test items were not evaluated for internal reliability prior to being administered on the final exam in the three courses, post hoc item analyses suggest these items were reliable measures of content knowledge (see questions, item means, and item discrimination indexes in Supplementary Materials (Section S4)). The two mastery grading sections were coordinated in terms of structure, content, and instructional approach. The instructor running the Mastery grading section is a Distinguished Professor of Teaching, and the instructor running the Mastery+OLI section is an Assistant Professor of Teaching.

2.2. Design of the Mastery Grading System

The instructors for the Mastery and Mastery+OLI sections coordinated designing the mastery test retake system. The mastery grading system was based on a second-chance testing system in which the course content was divided into six units. Mastery unit assessments aligned with each unit's learning objectives were developed and administered in bi-weekly proctored assessment sessions. New versions of the unit mastery assessment were released two times a week, and students had the opportunity to demonstrate mastery of both new and previous content according to a predetermined testing schedule outlined in the syllabus (see Supporting Information). The testing schedule was designed to provide three attempts for each unit, with the highest score counting toward the final course grade. If students were satisfied with their score on the first mastery exam, retakes were optional. All mastery assessments were administered during the discussion sections and proctored by graduate teaching assistants (discussion sections are mandatory one-hour weekly meetings in which the class size is approximately 30–40 students).

The unit mastery exams were composed of ten multiple-choice and numerical response questions administered through the learning management system. The questions were selected from extensive question banks separated by course learning objectives. A mastery grading scheme was employed when scoring the mastery assessment. According to the mastery grading scheme, a score of 9/10 was assigned full points, scores at or below 5/10 were assigned zero, and scores in between were assigned the corresponding percentage. The mastery grading scheme was applied to the highest score of their multiple attempts, and each unit mastery exam accounted for 10% of the final course grade. The unit mastery assessments accounted for 60% of the final course grade. The threshold for mastery was defined as a score of 60% on the unit mastery assessments (historically, 60% has been defined as a 'C-' grade in the fixed grading scale used in the department). See Table 2 for a complete breakdown of the grading for each section and the syllabus language, along with a tentative course schedule in Section S1 in the Supporting Information.

Table 2. Breakdown of the grading scheme for the three groups. [a] Homework for the control was administered through a standard publisher online homework system, the mastery course assigned instructor-created assignments in Canvas, and homework was assigned within the OLI platform for the Mastery+OLI section. [b] For all three courses, participation and discussion points were assigned based on completing in-class poll questions during the main lecture and attendance at weekly recitation/discussion sections.

	Control	**Mastery**	**Mastery+OLI**
Midterm 1	12%	10%	10%
Midterm 2	12%	–	–
Final Exam	24%	10%	10%
Homework [a]	24%	10%	15%
Discussion/Participation [b]	28%	10%	5%
Mastery Exams	–	60%	60%

Traditional comprehensive midterm and final exams were administered at the midpoint and end of the quarter, respectively—the midterm and final exams each account for 10% of the final course grade. However, the midterm exam covered content from the first two units, and if the score on the midterm was higher, it was used to replace the mastery assessment score for either or both of the first two units. Similarly, the final exam covered content from all six units, and the final exam score could replace the mastery exam scores and midterm scores. The remaining 20% of the final course grade was allocated to homework and participation. The mastery-only section assigned homework using the pre-recorded lecture videos with embedded questions through Canvas Studio. Engagement with the online courseware was incentivized for the mastery+OLI section, with 15% of the final course grade allocated to homework assigned within the OLI platform. We considered this a pure mastery (second-chance) testing system as opposed to a specifications model. We were not building groups of tasks or student products. Instead, the mastery grading model assessed blocks of content through multiple attempts on individual assessments. Furthermore, there was no token economy; all students were provided multiple attempts on every mastery assessment [7].

2.3. Design of the Interactive Courseware Platform

A combination of Canvas practice exams, recorded lecture videos, and the OLI interactive courseware formed the foundation for asynchronous content delivery in the Mastery+OLI group. The interactive online chemistry courseware was developed through the Open Learning Initiative (OLI) at Carnegie Mellon University. OLI General Chemistry is a comprehensive, data-driven, and evidence-based general chemistry curriculum. It should be noted that the OLI General Chemistry textbook is based on the atoms-first OpenStax Chemistry text, which covers the same curriculum as the textbook used by the control sections [26]. The OLI platform is designed to promote asynchronous active learning and provides rich student user data [29]. The chemistry content consists of units equivalent to a textbook chapter. Each unit is separated into modules, which include 5 to 10 content pages. Each content page includes didactic instruction in the form of text, images, and videos. In addition, “Learn by Doing” and “Did I Get This” activities are interspersed throughout each content page to promote active engagement with the content. The “Learn by Doing” activities typically break the problem-solving process into steps and provide extensive hints and feedback to guide the student to the correct answer. “Did I Get This” activities typically provide less extensive feedback and scaffolding. Each module concludes with a checkpoint activity, which can serve as homework, where feedback is provided only at the end of the quiz. The checkpoint activities were assigned as homework, accounting for 15% of the final course grade. Interaction with the OLI courseware was incentivized through participation points, accounting for 2.5% of the final course grade. A detailed analysis of student interaction within the OLI courseware and mastery grading system was provided in previous work [22].

2.4. Statistical Analyses

The final exam data was analyzed as part of a post hoc observational study approved by the UCR Institutional Review Board (IRB) under protocol number 30202. The approved protocol included obtaining student demographic information from the UCR Office of Institutional Research, and because this was carried out as a post hoc observational study, students were not required to complete an informed consent. Analysis of variance (ANOVA) was used to compare final exam scores between the three study groups; these were carried out for the entire class populations and for specific demographic groups (see Figures 1 and 2). The assumptions for ANOVA were evaluated (e.g., the dependent variable was normally distributed, and homogeneity of variance was observed across the three study groups), and post hoc pair-wise tests were used to determine where significant differences in final exam scores were observed for cases when the omnibus ANOVA was found to be statistically significant. Effect sizes were calculated using the partial eta-squared statistic from the omnibus ANOVA. The omnibus ANOVA and post hoc pair-wise tests were carried out using the IBM SPSS software program version 28.0.0.0 [30].

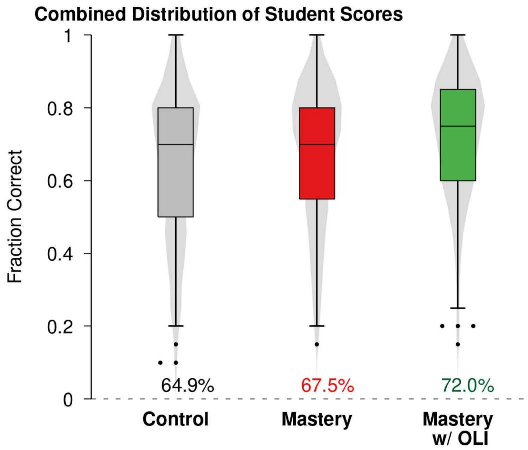

Figure 1. The distribution of student scores on the set of 20 common final exam questions is represented by the fraction of questions answered correctly. A box plot within each violin shows the median error (black line), middle 50th percentile (colored box), the range of errors (black lines), and outliers—represented by dots. The corresponding average scores, expressed as a percent correct, are provided below each distribution. Omnibus ANOVA comparing mean score on common final exam items ($F = 8.847$; $p < 0.001$); post hoc pair-wise comparisons with Bonferroni correction (Mastery+OLI vs. Control mean difference = 7.08, $p < 0.001$; Mastery+OLI vs. Mastery mean difference = 4.48, $p = 0.026$; see Supplementary Tables S2 and S3).

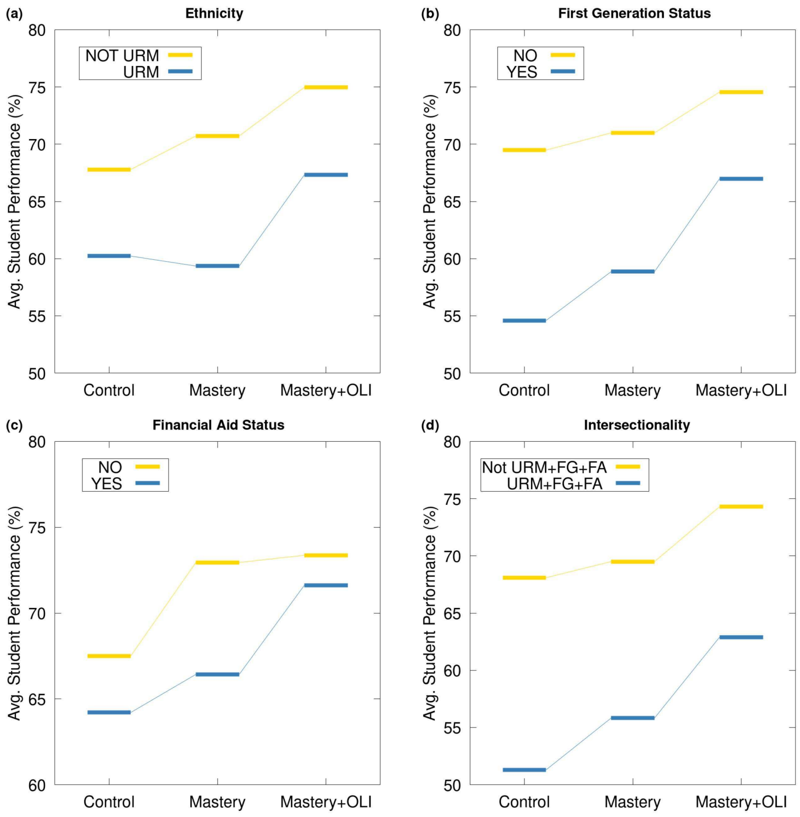

Figure 2. Average performance on common final exam questions disaggregated by (**a**) ethnicity: see Supplementary Tables S4 and S5, (**b**) first-generation status: see Supplementary Tables S6 and S7, and (**c**) financial aid status: see Supplementary Tables S8 and S9. Panel (**d**) compares the average student performance of all students to that of students belonging to all three groups: first-generation (FG), URM, and receiving financial aid (FA); see Supplementary Tables S10 and S11. (URM = underrepresented minority students; non-URM = white/Asian students).

3. Results

We begin by comparing student performance on the common final assessment questions. Figure 1 plots the distribution of student scores for the control (gray), Mastery (red), and Mastery+OLI (green) courses. The Mastery and Mastery+OLI courses led to higher scores on the common final exam questions than the control. Specifically, the mean performance expressed as the percent of correct responses on the common final exam questions for the Mastery+OLI group was 7.1 percentage points greater than the control ($p < 0.001$). Similarly, the mean performance for the Mastery group was 2.6 percentage points higher relative to the control; however, this improvement was not statistically significant at the 0.05 significance level. Figure 1 shows a narrower distribution in scores for the Mastery+OLI section, and the standard deviation for the Mastery+OLI group (16.4) is less than both the Mastery (18.9) and Control (20.4) groups.

Despite the fact the Mastery+OLI course had the highest percentage of URM, first-generation, and students receiving financial aid (see Table 1), the mean common exam scores for the Mastery+OLI course were significantly higher when the entire class population was analyzed (Figure 1). Disaggregating student performance data by ethnicity, first-generation status, and financial aid status reveals striking trends. In particular, Figure 2a shows that mastery grading alone did not significantly impact the average student performance for URM students (blue) relative to the control. However, when the mastery grading model was coupled with the OLI interactive courseware tools, the average performance for URM students improved by 7.1 percentage points relative to the control ($p < 0.001$), with an effect size of 0.181 (see Table 3). Even more pronounced improvements are observed for students with first-generation status, as seen in Figure 2b). Relative to the control, first-generation students enrolled in the Mastery+OLI course demonstrated a mean difference of 12.5 percentage points, with a moderate effect size of 0.272 ($p < 0.001$) compared to only a 5.1 percentage point improvement for students who did not identify as first-generation (see Table 3). These findings suggest mastery grading provides improved student learning for students with sufficient scaffolding for addressing gaps in content knowledge. The rich interactive tools provided through OLI in the Mastery+OLI course are likely crucial for assisting URM and first-generation college students in addressing gaps in their content knowledge. These findings are consistent with our recent work examining the link between engagement with the OLI courseware and performance on mastery assessments [22].

Table 3. Omnibus analysis of variance (ANOVA) comparing mean score on common final exam items. Results are reported for URM students (Table S4), first-generation college students (Tables S6 and S7), students who are receiving financial aid (Tables S8 and S9), and students belonging to all three categories (Tables S10 and S11).

		Omnibus ANOVA		**Mastery+OLI vs. Control**		**Mastery+OLI vs. Mastery**	
	F	*p*	*Cohen's f*	**Mean Diff.**	*p*	**Mean Diff.**	*p*
URM	4.19	0.016	0.181	7.10	0.048	7.97	0.038
First-Generation	8.36	<0.001	0.272	12.40	<0.001	8.13	0.031
Financial Aid	7.80	<0.001	0.163	7.40	<0.001	5.19	0.019
Intersectionality	3.97	0.021	0.248	11.60	0.018	7.07	0.343

Financial aid status was used as a proxy for identifying students more likely to have experienced financial hardship. Similar to the results found for ethnicity and first-generation status, students who received financial aid disproportionately benefited from the Mastery+OLI implementation relative to the control group. In particular, the Mastery+OLI group showed a 7.4 percentage point improvement in the mean relative to the control. Interestingly, the mean performance of the students who had not received financial aid in the Mastery and Mastery+OLI courses was the same. These results further support the hypothesis that students from more affluent backgrounds are more likely to have family members who have attended college and are, therefore, more likely to have developed habits and practices conducive to mastering chemistry content.

Finally, we considered intersectionality in the data by comparing the average performance of all students with those who simultaneously identify as first-generation college students, members of a URM group, and receive financial aid. Figure 2d and Table 3 show that mastery grading alone yielded a mean difference of 4.5 relative to the control ($p = 0.956$), and Mastery+OLI yielded a mean difference of 11.6 percentage points relative to the control ($p < 0.018$). The Mastery+OLI design showed a moderate effect, with a Cohen's f value of 0.248 (see Table 3) [31]. It should be noted that financial aid status was highly correlated with URM and first-generation status. Specifically, there were only two students who were both URM and first-generation status and not receiving financial aid.

Student Feedback Results

Recent work suggests a poor correlation between student evaluations and student learning, and the efficacy of student evaluations has been questioned [32]. However, this lack of correlation has been observed for numerical ranking systems, and evidence suggests student evaluations remain a crucial tool for providing instructors with feedback [33]. In particular, the comments section can provide valuable insights from the student's perspective, and such insights are particularly useful when evaluating novel instructional tools and course design. Nevertheless, analyzing hundreds of student responses for sentiment and relevance to a specific intervention while minimizing the introduction of bias is a challenging task. Recently, Hoar et al. suggested using natural language processing tools to facilitate this process [34]. Here, we apply natural language processing tools to student course evaluation data to analyze student feedback on the mastery grading system.

Student feedback was collected as anonymous course evaluation data for both mastery grading sections. The comments section in the course evaluation data was separated by sentence, providing 397 responses for the mastery with OLI section and 95 responses for the mastery grading section that did not use OLI. Each sentence was processed using Google's Natural Language API to identify keywords with a corresponding salience ranking. We selected keywords related to mastery grading and collected the associated sentiment scores (see the Supporting Information for details). This analysis resulted in 103 student comments related to the mastery grading model across both sections. Finally, each relevant student comment was subject to sentiment analysis using Google's Natural Language Processing tools. Google Cloud Sentiment Analysis (GCSA) provides a sentiment score between -1 and $+1$, with larger scores corresponding to more positive sentiment. Sentiment analysis was carried out using two additional algorithms and the results were similar to those reported in Figure 3 (see Supporting Information for details). Readers interested in the details of implementation and the comparative performance of alternative sentiment analysis algorithms are directed to the following review [35].

The histogram in Figure 3 illustrates the sentiment analysis results. For ease of interpretation, scores below -0.25 are classified as negative (red), between -0.25 and $+0.25$ as neutral (yellow), and above $+0.25$ as positive (green). This analysis was replicated with similar results using the VADER and TextBlob algorithms (see the Supporting Information for details) [36,37]. The mastery grading model resulted in generally positive or neutral student feedback. In particular, numerous students noted decreased stress and anxiety from the second-chance testing. On the other hand, numerous students expressed concern about sacrificing small-group discussion time in favor of repeated mastery exam attempts. Additionally, several students remarked on the lack of flexibility regarding testing times.

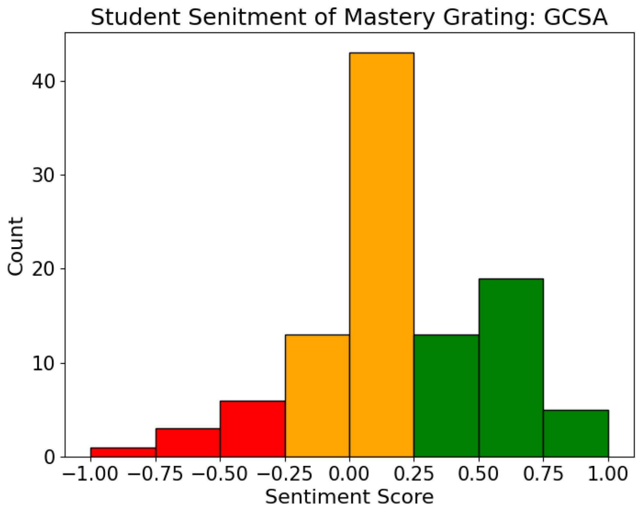

Figure 3. Student sentiment toward mastery grading based on 103 student comments from sections employing mastery grading. Each comment is classified based on the sentiment score, with scores below −0.25 classified as negative (red), between −0.25 and +0.25 as neutral (yellow), and above +0.25 as positive (green).

4. Discussion

The preliminary implementation of mastery grading in large-enrollment chemistry courses appears to have improved overall student learning outcomes and reduced equity gaps. This study contributes to the growing body of literature on alternative grading systems by demonstrating the efficacy of a mastery-focused approach, particularly when supplemented with interactive courseware like the General Chemistry curriculum provided through the Open Learning Initiative (OLI). The overall performance improvements we observe with the Mastery+OLI course are comparable to those seen in a pharmacokinetics (PK) and pharmacodynamics (PD) course that used a weekly quizzing model. Specifically, Henning et al. reported a 7.93 percentage point improvement in the average scores on the PK/PD component of the final exam [21].

Although several studies have explored alternative grading systems in chemistry, these studies do not directly compare performance across different demographics [6,11,12,14]. However, one study involving high school chemistry students found that a mastery grading model improves learning outcomes for students having difficulty with the content [6]. These previous results reported for high school chemistry are broadly corroborated by our finding that URM, first-generation, and students receiving financial aid showed the greatest improvement when using the Mastery+OLI model (Figure 2). Another study implementing a mastery grading model in an undergraduate engineering course found that women and URM students benefited from the alternative grading model to the same extent as the general population [3]. The present work did not consider gender; however, Figure 2a does show similar improvement for both URM and non-URM students when using the Mastery+OLI model.

Examining the disaggregated statistics for learning outcomes highlights the importance of incorporating metacognitive tools within the mastery grading model. Mastery grading alone did not improve student learning relative to the control for URM students (Figure 2a). These findings are not necessarily surprising when considered in the context of the recent literature. The role of retrieval practice in consolidating learning is well established [18,24]. However, consolidating learning through repeated testing may be of little value for students who are struggling to grasp the challenging and complex concepts in general chemistry [20]. Casselman et al. have shown that providing responsive online content designed to promote the development of metacognitive skills improved ACS exam performance by 4% relative to the control [19]. The OLI interactive courseware provides an accessible platform where students can regularly assess their abilities, receive detailed feedback regarding progress toward learning goals, and create a future study plan. Developing metacognitive skills is expected to be particularly impactful for students with minimal previous training of this nature (e.g., URM and first-generation students in Figure 2).

Chemsitry-specific growth mindset interventions in first-year general chemistry have been shown to improve the student learning experience and even eliminate the ethnicity achievement gap [38]. The mastery grading system reinforces the belief that chemistry content knowledge and problem-solving skills can be developed and improved over time. A closer examination of the sentiment analysis data presented in Figure 3 shows that the majority of the positive student feedback references a reduction in stress surrounding testing and a shift toward viewing mistakes as opportunities for growth, resulting in improved performance on subsequent exams. Though this was not directly explored in the study, we speculate that the absence of a curve in the mastery grading model promoted peer-to-peer engagement because the course grade was no longer tied to a student's performance relative to their peers.

Limitations and Future Work

Although we do not have incoming knowledge data, all students were placed into the course on the same track based on math placement. The distribution of students is likely equal across all three sections with respect to those math placement scores. However, as with any observational study, we ultimately could not account for the various confounding variables that might have negatively impacted student performance (prior chemistry knowledge, differences in co-curricular demands among the class populations, etc.) Additionally, though instructor bias could not be accounted for, it is noted the Mastery course was taught by a Distinguished professor of teaching and the control instructor provided the common questions. These factors suggest there was no bias favoring higher student exam performance in the Mastery+OLI group. Further studies are currently being designed to include initial knowledge assessment and the deployment of the Mastery+OLI model at scale across multiple institutions.

The common assessment items were generally emphasizing more traditional skills and knowledge. There is an emerging emphasis in the chemistry education community to move beyond a procedural and skill-based focus and promote learning objectives associated with conceptual understanding and more meaningful learning [39,40]. Therefore, future work will investigate how a mastery grading approach can improve this type of higher-order learning. However, multiple-choice questions administered through an online testing system better accommodate the volume of testing inherent to the Mastery+OLI model, and building a test retake system with more open-ended conceptual assessment items will be a challenge that needs to be overcome. Finally, the end of the term presents a logistical limitation, wherein students must complete multiple mastery attempts in rapid succession. This did lead to some student anxiety, and future implementations will focus on strategies for providing a more flexible testing schedule.

We have demonstrated that this test retake system can be implemented in large enrollment intro courses at an R1 institution with TA support and built-in recitations/discussions. The absence of either TA support or recitations/discussion sections would make it difficult to replicate this approach. Furthermore, sacrificing small-group discussion time in favor of testing presents a limitation, evidenced by student comments that expressed concern over the loss of small-group interaction time in the discussion/recitation sections. Future work will include exploring alternative second-chance testing models using a testing center, collecting more data on the student experience, and carrying out a more detailed study on the affective outcomes and/or outcomes related to student mindset.

5. Conclusions

Despite these limitations, the results described herein provide compelling evidence that a mastery outcomes/test retake system can significantly improve equity gaps in gateway STEM courses. There has been an ongoing long-term effort to improve retention of underserved students in higher education STEM [19,38], yet these equity gaps persist. It is proposed here that a contributing factor to this problem is the fact that many recruitment and retention programs are implemented in parallel to introductory STEM courses that continue to employ infrequent high-stakes exams. It is argued here that traditional assessment structures will ultimately limit the impact of these co-curricular programs and likely negatively impact outcomes for underserved students. It is hoped this current study provides a template for a new way forward.

Supplementary Materials: The following Supporting Information can be downloaded at: https://www.mdpi.com/article/10.3390/educsci14111224/s1, Section S1. Syllabus Language; Section S2. Pair-wise Comparison Statistics; Section S3. Student Feedback; Section S4. Common Test Questions.

Author Contributions: Conceptualization, J.D.H. and J.F.E.; methodology, J.D.H. and J.F.E.; validation, J.D.H. and J.F.E.; formal analysis, J.D.H. and J.F.E.; writing—original draft preparation, J.D.H.; writing—review and editing, J.D.H. and J.F.E.; funding acquisition, J.D.H. and J.F.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the NSF grant number DUE-2421279. This project was also funded in part by the California Education Learning Lab, a state-funded grantmaking organization with a mission to improve learning outcomes and close equity gaps for students attending public higher institutions in California.

Institutional Review Board Statement: The study was approved by the Institutional Review Board of the University of California, Riverside (protocol code 30202 approved 6 May 2024).

Informed Consent Statement: Because all data collection and data analysis were carried out as part of a post hoc observational study subsequent to the course delivery and final grade submissions, no informed consent was administered to students.

Data Availability Statement: The raw datasets used to carry out the statistical analyses in this article are not available due to restrictions in the IRB human subjects research protocol.

Acknowledgments: We wish to thank our colleagues in the Chemistry Department at the University of California, Riverside for their continued support. In particular, we thank Christopher Bardeen and Linlin Zhao for numerous helpful discussions. Access to the interactive courseware was provided free of charge through the Open Learning Initiative at Carnegie Mellon University. We wish to thank the developers and support staff at OLI. In particular, we thank Mark Blaser, Sandra Rasor, David Yaron, Norman Bier, and Ariel Anbar.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

OLI	Open Learning Initiative
GCSA	Google Cloud Sentiment Analysis
URM	Under Represented Minority
FG	First-Generation
DFW	Drop, Fail, and Withdrawal

References

1. Bloom, B.S. Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. *Eval. Comment* **1968**, *1*, n2.
2. Zimmerman, B.J.; Dibenedetto, M.K. Mastery learning and assessment: Implications for students and teachers in an era of high-stakes testing. *Psychol. Sch.* **2008**, *45*, 206–216. [CrossRef]
3. Morphew, J.W.; Silva, M.; Herman, G.; West, M. Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering. *Appl. Cogn. Psychol.* **2020**, *34*, 168–181. [CrossRef]
4. Kulik, C.L.C.; Kulik, J.A.; Bangert-Drowns, R.L. Effectiveness of mastery learning programs: A meta-analysis. *Rev. Educ. Res.* **1990**, *60*, 265–299. [CrossRef]

5. Toledo, S.; Dubas, J.M. A learner-centered grading method focused on reaching proficiency with course learning outcomes. *J. Chem. Educ.* **2017**, *94*, 1043–1050. [CrossRef]
6. Damavandi, M.E.; Kashani, Z.S. Effect of mastery learning method on performance, attitude of the weak students in chemistry. *Procedia-Soc. Behav. Sci.* **2010**, *5*, 1574–1579. [CrossRef]
7. Howitz, W.J.; McKnelly, K.J.; Link, R.D. Developing and implementing a specifications grading system in an organic chemistry laboratory course. *J. Chem. Educ.* **2020**, *98*, 385–394. [CrossRef]
8. McKnelly, K.J.; Howitz, W.J.; Thane, T.A.; Link, R.D. Specifications grading at scale: Improved letter grades and grading-related interactions in a course with over 1000 students. *J. Chem. Educ.* **2023**, *100*, 3179–3193. [CrossRef]
9. Nilson, L.B.; Stanny, C.J. *Specifications Grading: Restoring Rigor, Motivating Students, and Saving Faculty Time*; Routledge: London, UK, 2015.
10. Saucier, D.; Schiffer, A.; Renken, N. Five Reasons to Stop Giving Exams in Class. *Fac. Focus.* 2022. Available online: https://www.facultyfocus.com/articles/educational-assessment/five-reasons-to-stop-giving-exams-in-class/ (accessed on 1 January 2023).
11. Hollinsed, W.C. Applying innovations in teaching to general chemistry. In *Increasing Retention of Under-Represented Students in STEM Through Affective and Cognitive Interventions*; ACS Publications: Washington, DC, USA, 2018; pp. 145–152.
12. Ring, J. ConfChem conference on select 2016 BCCE presentations: Specifications grading in the flipped organic classroom. *J. Chem. Educ.* **2017**, *94*, 2005–2006. [CrossRef]
13. Houseknecht, J.B.; Bates, L.K. Transition to remote instruction using hybrid just-in-time teaching, collaborative learning, and specifications grading for organic chemistry 2. *J. Chem. Educ.* **2020**, *97*, 3230–3234. [CrossRef]
14. Ahlberg, L. Organic chemistry core competencies: Helping students engage using specifications. In *Engaging Students in Organic Chemistry*; ACS Publications: Washington, DC, USA, 2021; pp. 25–36.
15. Hunter, R.A.; Pompano, R.R.; Tuchler, M.F. Alternative assessment of active learning. In *Active Learning in the Analytical Chemistry Curriculum*; ACS Publications: Washington, DC, USA, 2022; pp. 269–295.
16. Kelz, J.I.; Uribe, J.L.; Rasekh, M.; Link, R.D.; McKnelly, K.J.; Martin, R.W. Implementation of specifications grading in an upper-division chemical biology course. *Biophys. J.* **2023**, *122*, 298a. [CrossRef]
17. Noell, S.L.; Rios Buza, M.; Roth, E.B.; Young, J.L.; Drummond, M.J. A bridge to specifications grading in second semester general chemistry. *J. Chem. Educ.* **2023**, *100*, 2159–2165. [CrossRef]
18. Karpicke, J.D.; Blunt, J.R. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* **2011**, *331*, 772–775. [CrossRef] [PubMed]
19. Casselman, B.L.; Atwood, C.H. Improving general chemistry course performance through online homework-based metacognitive training. *J. Chem. Educ.* **2017**, *94*, 1811–1821. [CrossRef]
20. Van Gog, T.; Sweller, J. Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educ. Psychol. Rev.* **2015**, *27*, 247–264. [CrossRef]
21. Hennig, S.; Staatz, C.E.; Bond, J.A.; Leung, D.; Singleton, J. Quizzing for success: Evaluation of the impact of feedback quizzes on the experiences and academic performance of undergraduate students in two clinical pharmacokinetics courses. *Curr. Pharm. Teach. Learn.* **2019**, *11*, 742–749. [CrossRef]
22. Asher, M.W.; Hartman, J.D.; Blaser, M.; Eichler, J.F.; Carvalho, P.F. Test, Review, Repeat: Mastery-Based Testing and its Benefits for Student Engagement and Performance in a General Chemistry Course. *OSFPreprints* **2024**. *preprint*. [CrossRef]
23. Kapasi, A.; Pei, J. Mindset theory and school psychology. *Can. J. Sch. Psychol.* **2022**, *37*, 57–74. [CrossRef]
24. Karpicke, J.D.; Roediger, H.L., III. The critical importance of retrieval for learning. *Science* **2008**, *319*, 966–968. [CrossRef]
25. Canning, E.A.; Muenks, K.; Green, D.J.; Murphy, M.C. STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. *Sci. Adv.* **2019**, *5*, eaau4734. [CrossRef]
26. Flowers, P.; Theopold, K.; Langley, R.; Robinson, W.R. Chemistry (OpenStax), 2015. Available online: https://openstax.org/details/books/chemistry-2e/ (accessed on 1 January 2023).
27. Bier, N.; Moore, S.; Van Velsen, M. Instrumenting courseware and leveraging data with the Open Learning Initiative (OLI). In Proceedings of the Companion Proceedings 9th International Learning Analytics & Knowledge Conference, Tempe, AZ, USA, 4–8 March 2019.
28. Koedinger, K.R.; Kim, J.; Jia, J.Z.; McLaughlin, E.A.; Bier, N.L. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In Proceedings of the Second (2015) ACM Conference on Learning@ Scale, Vancouver, BC, Canada, 14–18 March 2015; pp. 111–120.
29. Lovett, M.; Meyer, O.; Thille, C. JIME-The open learning initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *J. Interact. Media Educ.* **2008**, *2008*, 13. [CrossRef]
30. IBM Corp. *IBM SPSS Statistics for Windows*; Version 28.0; IBM Corp.: Armonk, NY, USA, 2021.
31. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Routledge: London, UK, 2013.
32. Uttl, B.; White, C.A.; Gonzalez, D.W. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Stud. Educ. Eval.* **2017**, *54*, 22–42. [CrossRef]
33. Benton, S.L.; Ryalls, K.R. *Challenging Misconceptions About Student Ratings of Instruction*; IDEA Paper# 58; IDEA Center, Inc.: Buffalo, NY, USA, 2016.
34. Hoar, B.B.; Ramachandran, R.; Levis-Fitzgerald, M.; Sparck, E.M.; Wu, K.; Liu, C. Enhancing the Value of Large-Enrollment Course Evaluation Data Using Sentiment Analysis. *J. Chem. Educ.* **2023**, *100*, 4085–4091. [CrossRef]

35. Al-Otaibi, S.T.; Al-Rasheed, A.A. A review and comparative analysis of sentiment analysis techniques. *Informatica* **2022**, *46*. [CrossRef]
36. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 216–225.
37. Loria, S. textblob Documentation. *Release 0.15* **2018**, *2*, 269.
38. Fink, A.; Cahill, M.J.; McDaniel, M.A.; Hoffman, A.; Frey, R.F. Improving general chemistry performance through a growth mindset intervention: Selective effects on underrepresented minorities. *Chem. Educ. Res. Pract.* **2018**, *19*, 783–806. [CrossRef]
39. Stowe, R.L.; Scharlott, L.J.; Ralph, V.R.; Becker, N.M.; Cooper, M.M. You are what you assess: The case for emphasizing chemistry on chemistry assessments. *J. Chem. Educ.* **2021**, *98*, 2490–2495. [CrossRef]
40. Holloway, L.R.; Miller, T.F.; da Camara, B.; Bogie, P.M.; Hickey, B.L.; Lopez, A.L.; Ahn, J.; Dao, E.; Naibert, N.; Barbera, J.; et al. Using Flipped Classroom Modules to Facilitate Higher Order Learning in Undergraduate Organic Chemistry. *J. Chem. Educ.* **2024**, *101*, 490–500. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.