International Journal of Engineering Education Vol. 40, No. 2, pp. 303–321, 2024
Printed in Great Britain

0949-149X/91 $3.00+0.00

Effects of Standards-Based Testing via Multiple-Chance Testing on Cognitive and Affective Outcomes in an Engineering Course*

AUTAR KAW**
Professor of the Department of Mechanical Engineering, University of South Florida, 4202 E Fowler Ave ENG030, Tampa, FL 33620-5350, USA. E-mail: kaw@usf.edu

RENEE CLARK
Associate Professor of the Department of Industrial Engineering and Director of Assessment for the Engineering Education Research Center, University of Pittsburgh, 1025 Benedum Hall, Pittsburgh, PA 15261, USA. E-mail: rmclark@pitt.edu

Multiple-chance testing was used to conduct standards-based testing in a blended-format numerical methods course for engineering undergraduates. The process involved giving multiple chances on tests and post-class learning management system quizzes. The effectiveness of standards-based testing was evaluated through various forms of assessment, including an analysis of cognitive and affective outcomes, and compared to a blended classroom that did not use standards-based testing. Based on a two-part final exam, a concept inventory, final course grades, a classroom environment inventory, and focus groups, the results showed that standards-based testing had overall positive effects. Standards-based testing was associated with a more significant percentage of students (15% vs. 3%) earning a high final exam score, a higher proportion of A grades (36% vs. 27%), and a better classroom environment on dimensions of involvement, cohesiveness, and satisfaction. Focus group discussions revealed that students appreciated the benefits of enhanced learning, second chances, and reduced stress with standards-based testing. The study also included an analysis of the impact of standards-based testing on underrepresented minorities, Pell Grant recipients (low socioeconomic groups), and low-GPA students, as well as an examination of test-retaking behaviors. The methodology and comprehensive results of the study are presented in this paper.

Keywords: standards-based grading; multiple-chance tests; standards-based testing; mastery grading; traditional grading; numerical methods

1. Introduction

Many consider traditional grading practices [1, 2] in postsecondary courses are rigid and not always a good measure of a student's understanding of the course material. For example, if students fail a high-stakes test, they have a limited chance of recovering in the course. Hence, their situation is not characterized by much flexibility. Consider a course grading scheme with four tests weighing 25% each, and a student scores 25% on the first test. If the student works hard for the rest of the semester and scores 90% on the other three tests, the highest possible grade they could achieve would be a C. Would that reasonably represent the student's understanding of the course?

One way to overcome this rigidity of traditional grading and improve student outlook and persistence is by introducing standards-based grading (SBG), where a course is broken down into standards, required objectives, topics, or competencies. Each standard is evaluated for mastery via several assessments, including homework assignments, quizzes, projects, etc. Each standard is graded on a proficiency scale (e.g., exceeds expectations, meets expectations, partial mastery, little or no mastery) rather than a percentage or points scale. This approach is taken so students can concentrate on the feedback, not the points deducted. As such, feedback emphasizes what students did wrong, showing them the correct answer with significant steps and hopefully leading them to corrective actions. Students can retake an assessment multiple times to achieve higher proficiency on a standard. Final letter grades are assigned based on the percentage of standards the student was highly proficient in.

When SBG is used in medium – and large-sized college classes, the burden on the instructor can become exceptionally high [3]. One way to reduce the burden is by using standards-based testing (SBT), where only tests are used to measure if a standard has been met. However, multiple test attempts for a standard require creating new tests, increasing time spent formulating and grading them. Also, keeping track of a student's current grade may not be accommodated directly by a learning management system. Further, it can be

** Corresponding author.

* Accepted 17 January 2024.

challenging for the student to stay on top of the number of standards they need to fulfill to get a desired grade. Intricate rules surrounding retakes can get overwhelming. This uncertainty can cause stress and confusion, making it difficult for students to succeed. Some students might procrastinate, leading to an end-of-semester rush to meet the standards [4]. The additional testing time also leads to reduced class time. If the tests are given outside of class time, there is a risk of academic dishonesty. This consequence could occur through unsupervised tests or students sharing test information on a standard with one another. Additionally, conducting tests outside class time may be inequitable for students with work schedules, other courses, commuting, and family responsibilities. It is essential to consider that not all students live near campus or are enrolled full-time. However, these concerns should not discourage using SBG or its variations, such as SBT.

In light of these concerns of keeping the grading manageable, maintaining academic integrity, discouraging student procrastination [5], encouraging spaced practice, limiting retesting to class time, being equitable to all students, and reducing the cognitive burden of tracking standards, the first author adopted multiple-chance testing (MCT) as a way to incorporate SBT [6–8]. The basic premise is the same as conventional SBT – provide students with multiple opportunities to show proficiency in the content and aim to enhance learning in a reduced-stress environment. However, several standards are tested in a single session via proctored in-class tests, and the number of opportunities given to the students for retakes is limited.

Taught in a blended modality, the MCT was implemented in a Numerical Methods engineering course in the spring 2023 semester. The multiple chances were limited for the three unit tests and the weekly post-class LMS quizzes, representing 45% and 15% of the course grade, respectively. The standards to be met were the eight topics of the course. Each student could take a retake test on topics of the three unit tests at a scheduled class meeting time. The post-class LMS quizzes were open for unlimited attempts until the last day of class. The original attempt scores on tests and quizzes were included when calculating the topic score to avoid procrastination [5, 9]. The final exam stood as a separate grading item (25%) but was also used as a proxy for a second retake attempt of the three unit tests. The remainder of the grade was for computer projects (10%) and an end-of-semester concept inventory (5%).

In this paper, we seek to assess and compare cognitive and affective outcomes for a junior-level, postsecondary course in Numerical Methods in a blended classroom with SBT (experimental group) vs. a blended classroom without SBT (control group). We used a concept inventory and a final examination to measure the cognitive outcomes and a classroom environment inventory and student focus groups to assess the affective outcomes.

2. Literature Review

SBG has its roots in the 1963 paper by John Caroll, in which they argued that different students often need different amounts of time to learn the same content [10, 11]. Then, in 1971, work by Bloom showed that mastery-based learning (now called SBG) coupled with tutoring could improve a typical student's performance by two standard deviations [12]. In SBG, students are graded based on proficiency instead of points and have multiple opportunities to demonstrate meeting the course standards. Feedback given by the instructor on failed attempts becomes an opportunity for growth rather than a measure of incompetence. It is assumed to foster a growth mindset [13] and improve self-efficacy [14]. Students can work beyond their current proficiency level while the emphasis on grading is shifted from points-based to criteria-based.

SBG in engineering courses has garnered renewed attention, prompting several studies to explore its implementation and assess its impact. Carberry et al. [15] conducted a comprehensive investigation involving six instructors across different institutions. Their work established best practices for seamlessly integrating SBG into courses and identified barriers to its effective implementation. In a Mechanics of Materials course, Siniwaski et al. [16] employed SBG and discovered that students not only considered it superior for learning but also expressed a preference for it over traditional grading methods. However, it was noted that students still desired numerical values akin to traditional grading to monitor their progress throughout the course. Post and Agritech [17] adopted SBG in a Thermodynamics course, utilizing a pass-fail proficiency scale across 11 standards. Notably, the pass rates on these standards ranged from 61% to 100%. Nevertheless, student feedback indicated dissatisfaction with the binary nature of the proficiency scale. Averill et al. [18] analyzed SBG and traditional grading sections in a Mechanics of Materials course. Results revealed that students in the mastery learning sections consistently outperformed their counterparts in conventional sections, scoring 1.5 to 3 letter grades higher in a common final examination. Implementing mastery learning in dynamics and thermodynamics engineering courses, Moore [19] conducted surveys indicating

that SBG was perceived as fairer than the conventional points system by students. In a more recent development, a 2023 paper conducted a systematic review of using SBG in engineering [20]. The findings in the review paper suggested that the efficacy of SBG was evident in transcript course and homework grades but not consistently observed in final exams. Survey results indicated a mixed sentiment towards SBG, emphasizing the absence of conclusive evidence regarding its buy-in by students.

One of the grading systems under SBG is called standards-based testing (SBT), which uses only testing to check for proficiency achieved in standards. Harsy [21] used SBT in four mathematics courses. They used limited opportunities for retesting to get students to take them seriously. 'Retesting weeks' occurred, in which students could be retested on any objective covered thus far in the course. The retesting was conducted during office hours and at proctored Math Study Tables, which are similar to peer tutoring sessions. Harsy and Hoofnagle [22] compared SBT with traditional testing in an Integral Calculus course by using surveys, final exams, and transcript grades. They found that SBT was associated with students receiving higher transcript grades. Also, students reported a better understanding of course concepts and not having to spend as much time studying for the course. However, students may not necessarily be the best judges of assessing their learning, which these authors noted in their work.

Henriksen et al. [23] studied student anxiety in an Ordinary Differential Equations course. The survey requested students to rate their anxiety, motivation, and time management during weeks 4, 5, 12, and 14 of a 14-week course [23]. Anxiety monotonically decreased during the semester with SBT but monotonically increased during the semester with traditional grading. Lewis used SBT and found that it reduced test anxiety and promoted mastery. However, they could not show a change in the student's growth mindset [24]. Chamberlain used SBT in an entry-level College Algebra class for asynchronous, mastery-based learning with 500+ students [25]. They found that students needed multiple attempts to succeed but improved their thinking. Lenarz and Pelatt [26] used SBT to increase student persistence. Their implementation was at a liberal arts university for women, where one-third of the students were first-generation college students. SBT was preferred by students over traditional grading and was also found to be linked with higher letter grades.

As mentioned in Section 1, although less, SBT has similar drawbacks as SBG when used in mid to large-size classes. An alternative is the use of multiple-chance testing (MCT). MCT is simply used as second-chance testing to replace a lower grade on an original test. Herman et al. [8] studied MCT (second-chance testing) in four large-size classes in Computer Organization, Dynamics, Solid Mechanics, and Introduction to Electronics. They used three different test score replacement policies – partial grade replacement, completing a zero-credit homework assignment before being allowed to take the replacement test, and full grade replacement. They found that these three policies and the second-chance exams had no effect on student performance or study habits for the first-chance exam. Still, it increased the study time by 60% in between exams. Noell et al. [27] used second chance testing (MCT) in a general chemistry course. Although it increased the DFW rate from 17% to 24%, the MCT helped lower-achieving students and doubled the number of A-grade students compared to the point-based sections. MCT also decreased the grading effort by eliminating partial credit. Emeka et al. studied student perceptions and behavior related to second-chance testing [6]. Their main conclusions were that second-chance testing promoted fairness, alleviated stress, and improved learning.

As outlined in Section 1, in this paper, we are also using MCT as a bridge to conduct and gain advantages of SBT while reducing the resources and time expended by the student and instructor. Proficiency in each standard gets measured, and students get a chance to show improvement in meeting standards in the second-chance testing and the final examination.

3. Significance of the Study and Research Questions

Through the use of MCT, this paper studies the effect of SBT on the cognitive and affective outcomes of university engineering students in a numerical methods course. Our research provides additional understanding of the association between SBT and student outcomes, including the effect on individual standards (i.e., topics) and overall course knowledge. We investigated the effects associated with SBT for underrepresented minority (URM) and low-socioeconomic (Pell grant recipients) students as well as students of low and high prerequisite achievement (i.e., via the prerequisite GPA). To better understand students' perspectives towards SBT, we conducted a classroom environment inventory and focus groups. Our study addresses the following research questions for SBT compared to traditional assessment and grading approaches in a blended engineering classroom.

RQ1: What effect is associated with SBT for a final exam that assesses both lower- and higher-order

skills in an engineering course? What are the effects for URM, Pell Grant recipients, and students of varying prerequisite achievement?
RQ2: What effect is associated with SBT for a concept inventory in an engineering course? What are the effects for URM, Pell Grant recipients, and students of varying prerequisite achievement?
RQ3: What effect is associated with SBT on the student's final letter grade in an engineering course?
RQ4: What effect is associated with SBT for the classroom environment in an engineering course? What are students' perceptions of SBT?
RQ5: What participation and performance characteristics are associated with test and quiz retaking?

4. Methods

4.1 Instructional and Grading Methods

The undergraduate course within this study is Numerical Methods, which is taught in the Mechanical Engineering department at the University of South Florida. The course is taken by third-year students, and they learn numerical methods related to the following eight topics: Introduction to Scientific Computing, Differentiation, Nonlinear Equations, Simultaneous Linear Equations, Interpolation, Regression, Integration, and Ordinary Differential Equations. Errors and their relationship to the accuracy of the numerical solutions are underlined throughout the course. Programming is used to model numerical methods and solve intractable and application problems.

The first author has taught the course in a blended modality since 2003, using web-based open educational resources [28]. It has also been taught using flipped instruction [29] while including adaptive learning for pre-class preparation [30]. For this study, the control group semester was Spring 2017, when the course was last taught in a face-to-face blended modality by the first author. The modalities used after Spring 2017 included flipped, remote instruction during COVID-19, and flipped with adaptive learning.

In the control group, the course was taught in a blended modality with traditional grading components mentioned in Section 1 for the experimental group but without multiple test attempts. About 25–40% of class time was spent on active learning conducted mainly via personal response systems and in-class exercises. Pre-class learning was expected only for prerequisite course materials, and class lectures were used to introduce new content. Because of the use of class time for active learning, some content was delivered after class via video lectures available through the learning management system (LMS).

In Spring 2023, the first author taught the course in the same blended format as the control group but with MCT. Students had the choice to retake the post-class LMS quizzes and tests on the eight topics of the first-chance unit tests 1–3. The course consisted of eight topics, which are delineated as the eight chapters of the textbook. By using chapters as the standards or topics, students were clear about the meaning of each topic or standard, and this kept the bookkeeping (of points) reasonable for both instructor and student. More importantly, it enabled the instructor to ask questions on the retakes that could be objective, procedural, evaluative, and comparative.

The SBT was applied to first-chance unit tests 1–3 (45% of the course grade) and the weekly post-class LMS quizzes (15%). The retakes for each first-chance test 1–3 were given one to three weeks afterward so students could get their graded tests back, decide whether to retest, act on the feedback, and prepare for the test. The score and proficiency level were provided for each topic covered on a given first-chance test. The proficiency levels were as follows: Highly Proficient (A) 90–100%; Proficient (B) 80–89%; Progressing (C) 70–79%; Beginning (D) 60–69%; and Not Yet (F) 0–59%. A student who received 'Highly Proficient' (90% or more) on a standard did not need to retake the test on that standard, as their grade would remain unchanged. However, they could take the retake test if they wished. Also, the final score for retakers who made less than 90% on a standard was capped at 90%. This cap was used less to reduce the grading load and more so to discourage students from spending time obtaining just a few extra points. Instead, these students would be better served by learning and practicing new topics. To avoid students retaking a test just to have access to the questions for their future preparation, the retake tests were made available to all students on the LMS after they were administered.

The retakes of the first-chance unit tests were conducted as follows. Using unit test 2 as an example, two topics (Topics 4 and 5) were covered. Students took the first-chance test on both topics in a 75-minute class session. The test was returned to students with grades and proficiency levels for each of the two topics. Two weeks later, students took the retest during regular class time for both topics (i.e., 25 minutes per topic) during a single class session. A student could retake the test on none, one, or both topics, but the start and end time for each topic test was fixed. To maintain academic integrity, if any student left a retake test early, a student coming in late was not allowed to take the

test. However, this rule never needed to be applied. Also, the retake tests consisted of new questions, and they were not simply algorithmic equivalents of the first-chance test questions. Topic scores of each student were updated by one-half of the difference between the retake and the original topic score. This adjustment was only applied if the student's retake score exceeded the original test score.

Although the final exam was a separate component of the grading scheme, it was also used as a proxy for a *second* retake of all topics. Adjustments were made to the topic scores similar to the retake tests. Points for each topic corresponded to the questions asked on the final exam. Although final exams are not generally given in SBG classes, the final exam was kept as a separate component of the grading because students must recognize the integrated connections between the topics and course prerequisites and that it improves long-term retention [31, 32]. The final exam is also critical since Numerical Methods is a prerequisite for several other courses in the Mechanical Engineering curriculum.

Thirty-one (31) post-class LMS quizzes on the eight topics were also assigned, and they had regular weekly deadlines to encourage spaced practice and ensure adequate preparation for in-class active learning exercises. Each quiz consisted of two short objective questions and one algorithmic question, and these questions were chosen randomly from question banks. The retake quiz was released soon after the due date of the original quiz. The due date for all retake quizzes was the last day of classes for the semester. The adjustment for each quiz was calculated in the same way as for the retake tests; that is, the quiz score was increased by one-half of the difference between the retake and the corresponding original quiz. Of course, this adjustment was only applied if the retake score was higher than the original quiz score. The grading scheme for the two implementations of blended modality without and with SBT is summarized in Table 1.

4.2 Outcomes Assessment

To assess both the affective and cognitive-based effects of applying SBT in the Numerical Methods course, we used multiple instruments and protocols, which included a demographics survey, a classroom environment inventory, focus groups, a numerical methods concept inventory, and a final examination having both multiple-choice and free-response questions. Similar triangulated approaches were

Table 1. Grading scheme for blended Numerical Methods classroom with and without SBT

Grading component	Weight	Without SBT	With SBT
Post-class LMS quizzes	15%	• 31 quizzes due weekly to keep students accountable	
			• Quizzes could be retaken until the last day of class • Quiz scores updated by one-half of the difference between retake and corresponding original quiz
Personal response system quizzes	0%	• Declarative and conceptual questions asked in class to conduct think-pair active learning	
In-class exercises	0%	• Scaffolded exercises requiring higher-level thinking with help from the instructor and two learning assistants • Exercises submitted for ungraded extensive feedback	
Problem sets	0%	• End-of-chapter problems (not graded), as doing most problems improves performance on tests	
Unit-tests 1-3 (first chance tests)	45%	• Three-unit-tests (non-cumulative) • Cover specific topics • Unit Test 1 – three topics; Unit Test 2 – two topics; Unit Test 3 – three topics	
			• Retake of unit-tests by topic given within 1–3 weeks of tests • The topic score increased by one-half of the difference between the retake and the corresponding original score
Concept inventory	5%	• An 18 multiple-choice question test • Given at the end of the semester to assess conceptual understanding	
Programming project	10%	• Requires modeling a physical problem, collecting experimental data, identifying required mathematical procedures, and writing code for solution	
Final exam	25%	• Consists of multiple-choice (50%) and free-response (50%) questions to assess lower-order and higher-order skills, respectively • Free-response questions are comprehensive and may require the application of several topics and prerequisites	
			• The final exam was also considered as a retake of all eight topics

used in our other NSF-funded studies [29, 33, 34]. In addition, we analyzed the behaviors and performance characteristics associated with retaking the first-chance tests and post-class LMS quizzes.

4.2.1 Demographics Survey

A demographics survey [29] was administered to students who opted to participate in the study so that stratified statistical analyses could be conducted for particular groups of students and students as a whole. For example, this survey was used to collect data on race and ethnicity to identify underrepresented minority (URM) students. URM students were defined as Black/African American, Hispanic, American Indian, and/or Hawaiian/Pacific Islander. This survey was also used to collect Pell Grant status, which was used to identify students with low socioeconomic status [35]. The Pell Grant is given to US students with exceptional financial needs [35]. The survey was also used to collect data on letter grades received in prerequisite courses, which included Calculus 1–3 (i.e., differential, integral, and multivariable calculus), ordinary differential equations, physics, and programming. The prerequisite GPA, calculated from the letter grades, served as a covariate (control variable) for the statistical analyses of the final exam and concept inventory scores. The prerequisite GPA was also used to investigate the effects of SBT for students of differing achievement levels as measured by this GPA.

4.2.2 College and University Classroom Environment

The CUCEI, or College and University Classroom Environment Inventory [36], is a 49-item validated instrument that measures seven psychosocial dimensions of the classroom, as given in Table 2. Each item is rated by students on a 1 to 5 scale, with 5 being the most desirable. The CUCEI was administered to students during the last two weeks of the semester. We had previously used [29] this instrument in our research on blended classrooms because it is well-suited for the dynamic environment of such classrooms. Therefore, we had comparison data readily available.

Since seven dependent variables are associated with the CUCEI (i.e., the seven dimensions), a MANOVA (i.e., multivariate analysis of variance) was used to compare the blended classroom with and without SBT. Bonferroni's highly conservative adjustment for multiple comparisons was applied to each univariate p-value by multiplying it by seven before comparison to $\alpha = 0.05$ [37]. In addition, Cohen's d-effect size was used to measure the practical significance of the seven differences [38]. The following were used for Cohen's d effect sizes: small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) [39].

4.2.3 Student Focus Groups

Two semi-structured focus groups were conducted via Zoom by the second author toward the end of the semester to collect student perspectives and feedback on the SBT. Each focus group was approximately one hour in length. The instructor (i.e., the first author) recruited student volunteers from the course. Two focus groups were conducted to accommodate the number of interested students. The following questions given in Table 3 were posed.

A total of 17 students participated in the focus groups, with 8 and 9 in each of the two focus

Table 2. Seven psychosocial dimensions of the CUCEI [36]

Psychosocial Dimension	Extent to Which
Cohesiveness	Students know and help one another.
Individualization	Students are treated individually and differentially.
Innovation	New class activities or teaching techniques are used.
Task orientation	Class activities are well-organized.
Involvement	Students participate in class activities.
Personalization	Interaction takes place with the instructor, and there is a concern for students.
Satisfaction	Classes are enjoyed by students.

Table 3. Focus group questions for blended with SBT modality

1. What do you think about standards-based grading, where you get to retake tests and online LMS (Canvas) quizzes?
2. Do you read and use the extensive feedback given to you on the tests and the retakes?
3. Did retaking the tests take time away from learning the new content?
4. How can standards-based grading (or any element of it) be improved in this course?
5. To what degree did standards-based grading enhance your learning compared to traditional grading?
6. Since LMS (Canvas) cannot handle standards-based grading, did you use the standards-based grading Excel program that the instructor posted to calculate your grade in the class? What would you suggest for improving the Excel sheet?

groups. After collecting the focus group data, the second author read all responses line-by-line and listed the frequent, interesting, and/or relevant themes [40]. This reading led to the development of several emergent coding schemes based on the data, which were then used to conduct a content analysis [41]. One of the coding schemes was used to code focus group questions 1, 4, and 5 (Table 3). The responses to these three questions overlapped; therefore, the same coding scheme was applied to all three questions, and the results were combined. This coding scheme contained positive (i.e., benefits) and negative (i.e., drawbacks/suggestions) themes related to our implementation of SBT and is shown in Table 4. The positive themes, or benefits, with SBT identified by students, pertained to the enhancement of their learning (*LEARN*), the ability to focus on only particular topics (*TOPIC FOCUS*), and the opportunity to use the feedback obtained from the first-chance tests (*APPLY FEEDBACK*). In addition, with SBT, students felt their stress was reduced (*LESS STRESS*), their grades could be boosted (*POINTS OR GRADE BOOST*), second chances were possible (*CHANCES*), and fairness was promoted (*FAIRNESS*).

The themes or categories in Table 4 are found in the recent literature, as discussed in section 2. They are, therefore, grounded in the research of others in standards-based grading and testing. For example, Lewis found that students experienced significantly less test anxiety as measured by the TAI-5 inventory in their standards-based grading (SBG) math course versus their other non-SBG courses [24]. The theme of reduced stress was frequently found in interview data from students who had taken STEM courses with second-chance testing, in addition to the themes of improved learning, better grades, promotion of fairness, and the opportunity for focused and targeted study [6]. In survey results from students who had experienced mastery-based instruction with retesting in a Calculus 2 course, the themes of extra chances enhanced conceptual understanding and less stress were found [22]. The theme of reduced stress is quite prevalent in literature. One article is actually entitled *Specifications-Based Grading Reduces Anxiety for Students of Ordinary Differential Equations* [23].

Conversely, there were negative themes (i.e., drawbacks and suggestions) identified by students with SBT, including an upper limit on the final score with a retake (*CAP*), the need to manage other work and commitments alongside the retakes (*OTHER WORK*), the need for points accounting (*TRACKING*), the desire for richer feedback (*MORE FEEDBACK*), and a possible reduction in first-chance seriousness or motivation (*LESS MOTIVATION*). These themes were likewise present in the recent SBT literature. In Emeka et al.'s interviews, students admitted to reduced studying for the first-chance test, knowing they would have a second chance [6]. Their students mentioned they had to consider other commitments and time constraints in conjunction with the test retakes [6]. Student survey data from math courses [21] included complaints about points that could be earned with mastery-based assessment, specifically a lack of partial credit, in addition to the time required for retesting, considering their other classes or work.

Table 4. Coding Scheme for focus group questions 1, 4, and 5

Code	Description	
LEARN	Retake helps promote understanding or learning. Retake promotes the reinforcement of knowledge or practice. Retake allows more time to learn or absorb.	Benefit
LESS STRESS	Less stress with retakes. Less worry with retakes.	
POINTS OR GRADE BOOST	Grade can be enhanced with a retake. Points can be retrieved with a retake. Retake allows multiple ways to earn points.	
TOPIC FOCUS	Retake allows focusing on particular topics.	
CHANCES	Retake provides a second (or more) chance.	
APPLY FEEDBACK	Can apply feedback to the retakes or tests.	
FAIRNESS	Retake promotes fairness or equity.	
CAP	Cap set at 90%; grade cannot be improved beyond 90%. Points limited (that can be earned).	Drawback or Suggestion
TRACKING	Retake creates special needs in tracking points.	
OTHER WORK	Retake must be handled alongside other work or commitments.	
MORE FEEDBACK	Want more feedback. Want fuller solutions.	
LESS MOTIVATION	Retake may reduce motivation or serious approach initially.	

Separate coding schemes were developed for focus group questions 2 and 3 (Table 3) and consisted of simple, mutually exclusive yes/no/other categories. Focus group question 6 (Table 3) was not coded but served solely as formative feedback for the instructor. After developing the coding schemes, the second author and a second analyst independently coded the focus group responses using these coding schemes [41]. Each distinct statement was coded by both analysts, where a distinct statement is defined as a response made by one participant to a question before another participant subsequently responded, and so on. The analysts later discussed their codes and agreed on any differences; thus, the focus group data were double-coded by two analysts. Their interrater reliability associated with applying the coding scheme in Table 4 was Cohen's kappa = 0.91, indicating strong agreement beyond chance [42]. The coders were in 100% agreement in coding questions 2 and 3 with the yes/no responses.

4.2.4 Direct Assessment – Concept Inventory and Final Exam

A final exam and concept inventory were given to the students to assess learning in the course directly. The concept inventory [43, 44] was designed to measure conceptual understanding of the course material. It was developed by several numerical methods instructors and consists of 18 multiple-choice questions. There are three questions per each of the six concepts the instructors agreed upon as most important using a Delphi method [45].

The final examination consists of two parts and is a two-hour test during the semester's last week. The first part consists of 14 multiple-choice questions based on the lower-order skills (i.e., remember, understand, apply) of the revised Bloom's taxonomy. The second part consists of four free-response questions based on the higher levels of the revised Bloom's taxonomy. A rubric is used to grade the free-response questions, with a score of 0 indicating no understanding, a blank response, or using irrelevant formulas. A score of 4 indicates a complete understanding of the problem in which all task requirements were included in the response.

An analysis of covariance was used to compare the blended classroom with and without SBT, using the GPA of prerequisite courses as the covariate [42]. Given the smaller sample sizes for some demographic categories and the attendant uncertainty of the data being normally distributed, results from a non-parametric analysis of covariance (i.e., Quade's test) were reported for this study [46, 47]. Bonferroni's adjustment for multiple comparisons was applied to the multiple tests conducted for the various demographic categories of interest [37]. Hedges' g effect size, used for smaller sample sizes, was calculated to measure the practical significance of the differences using the same ranges used for Cohen's d [38].

In analyzing categorical data associated with the final exam, concept inventory, and course performance, such as the proportion of students within a score range or earning a final course grade, a z-test of proportions or Fisher's Exact test was used [48, 49]. Fisher's Exact test was used when the sample size was too small to meet the assumptions for a z-test of proportions [49]. The odds ratio (OR) was calculated as the effect size measure when comparing two proportions, with the ranges of small (1.5), medium (2.0), and large (3.0) [50].

5. Results

5.1 Student Participants

The Numerical Methods course, which is taught at a large R1 institution in the southeastern US, is taken by junior-level undergraduates in mechanical engineering as a required course. For the control group, we had the final exam and demographics data from 62 students with which to conduct statistical analysis. For the experimental group, we had this data from 47 students. These numbers represented 57% and 82% of class enrollment, respectively. Of the 62 students in the spring 2017 semester, 87% were male, 13% were female, and 27% were underrepresented minority students. Of the 47 students in the spring 2023 semester, 81% were male, 17% were female, 2% represented another gender, and 38% were underrepresented minority students.

5.2 Direct Assessment of Learning with SBT

To directly assess learning and academic performance with and without SBT in the blended classroom, we used a two-part final exam and a numerical methods concept inventory (CI). The multiple-choice portion of the final exam assessed lower-order skills and the free-response portion assessed high-order skills. We present stratified results from each portion separately in answering the first research question:

RQ1: What effect is associated with SBT for a final exam that assesses both lower and higher-order skills in an engineering course? What are the effects for URM, Pell Grant recipients, and students of varying prerequisite achievement?

Fig. 1 shows the final exam score distribution with and without SBT. For this culminating exam, the differences were visually notable for the 80–100% and 20–39% bins. A significantly higher percentage of students (15% vs. 3%) earned a high

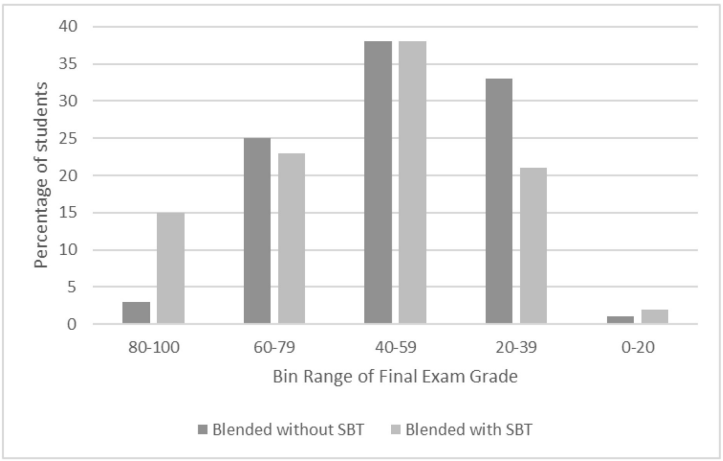

Fig. 1. Distribution of final exam percentage scores for two groups: blended without SBT and blended with SBT.

final exam score (80–100%) in the SBT group compared to the non-SBT group. The effect size associated with this was considerable at $OR = 6.5$, and the difference in the proportions was significant based on Fisher's Exact test ($p = 0.016$). For the combined 0–20% and 20–39% bins associated with low final exam scores, a significantly higher percentage of students (34% vs. 23%) were from the non-SBT versus SBT group. However, the effect size for this was small at $OR = 1.7$, and the difference in the proportions was not significant based on a z-test of proportions ($p = 0.18$). Thus, the effect associated with SBT on the final exam was particularly favorable for higher performers.

Positive results for SBT were also evident in the prerequisite GPA-adjusted averages on the final (culminating) exam and, more notably, for the free-response questions. For the multiple-choice questions (Table 5), which assess lower-order skills, the adjusted mean for all students was higher with SBT but not significantly so based on Quade's test (unadjusted $p = 0.67$), including a small effect size ($g = 0.06$). The most significant effect associated with SBT was for students in the low prerequisite GPA group. The effect size for this group fell between small and medium, with a value of $g = 0.38$. For the URM and Pell Grant groups, a trivial effect was associated with SBT for the multiple-choice questions, with $g = 0.11$ and $g = 0.14$, respectively. For students in the high prerequisite GPA group, the effect associated with SBT was not positive, although small ($g = -0.10$).

The positive effect associated with SBT was higher for the free-response (versus multiple-choice) questions on the final exam for all students combined, with $g = 0.22$, albeit not statistically significant (unadjusted $p = 0.46$) based on Quade's test. As shown in Table 6, the positive effect associated with SBT was medium for students in the high prerequisite GPA group ($g = 0.53$). How-

Table 5. Final exam multiple-choice percentage score comparison: blended without SBT vs. blended with SBT

Dem Group	**Adjusted Mean Percentage % (*s*) n**		**Quade's Test *p unadj***	**Quade's Test *p adj***	**Effect Size *g***
	Blended without SBT	**Blended with SBT**			
All	**58.2** (*15.3*) 62	**59.1** (*15.3*) 47	0.67	1.00	0.06
URM	**53.9** (*13.3*) 17	**55.4** (*13.3*) 18	0.73	1.00	0.11
Pell	**55.7** (*14.8*) 21	**57.8** (*14.8*) 14	0.39	1.00	0.14
Low prereq GPA	**50.2** (*12.3*) 36	**55.0** (*12.4*) 19	0.11	0.55	0.38
High prereq GPA	**66.3** (*17.1*) 26	**64.7** (*17.1*) 28	0.77	1.00	–0.10

Table 6. Final exam free-response percentage score comparison: blended without SBT vs. blended with SBT

Dem Group	Adjusted Mean Percentage % (*s*) n		Quade's Test *p unadj*	Quade's Test *p adj*	Effect Size *g*
	Blended without SBT	Blended with SBT			
All	**42.6** (*19.5*) 62	**46.8** (*19.6*) 47	0.46	1.00	0.22
URM	**36.8** (*18.4*) 17	**38.5** (*18.4*) 18	0.84	1.00	0.09
Pell	**44.7** (*19.9*) 21	**33.4** (*19.9*) 14	0.07	0.35	–0.56
Low prereq GPA	**33.3** (*15.9*) 36	**32.3** (*16.0*) 19	0.89	1.00	–0.06
High prereq GPA	**50.6** (*19.6*) 26	**61.2** (*19.6*) 28	0.08	0.41	0.53

ever, the difference was not quite statistically significant (unadjusted $p = 0.08$). The effect associated with SBT for the free-response questions was highest for this group of students (i.e., high prerequisite GPA). Interestingly, this was the opposite of the effect observed for this group with the multiple-choice questions. For URM students, the effect associated with SBT was again small ($g = 0.09$). For Pell Grant recipients and low prerequisite GPA students, the effect associated with SBT on the free-response questions was negative, with a medium effect size for the Pell Grant recipients of $g = -0.56$.

Secondly, a concept inventory (CI) was used to assess the conceptual understanding of numerical methods directly. We present stratified results from the CI in answering the second direct-assessment research question:

RQ2: What effect is associated with SBT for a concept inventory in an engineering course? What are the effects on URM, Pell Grant recipients, and students of varying prerequisite achievement?

Fig. 2 shows the concept inventory score distribution with and without SBT. The concept inventory is a second culminating assessment for students. The difference between the two groups was most visually notable for the 60–79% group. For the two top-scoring groups combined (i.e., 60–79% and 80–100%), SBT was associated with a significantly higher proportion of students (58% vs. 33%). The associated effect size was medium at OR = 2.8, and the difference in proportions was significant based on a z-test of proportions ($p = 0.006$).

Among the culminating assessments, the positive effect associated with SBT for all students combined was highest for the concept inventory. This effect was close to medium at $g = 0.45$, although not

Fig. 2. Distribution of concept inventory percentage scores for two groups: blended without SBT and blended with SBT.

Table 7. Concept inventory percentage score comparison: blended without SBT vs. blended with SBT

Dem Group	Adjusted Mean Percentage % (*s*) n		Quade's Test *p unadj*	Quade's Test *p adj*	Effect Size *g*
	Blended without SBT	Blended with SBT			
All	**52.3** (*16.9*) 62	**59.8** (*17.0*) 48	0.03	0.15	0.45
URM	**46.9** (*17.2*) 17	**54.8** (*17.2*) 18	0.13	0.65	0.45
Pell	**54.2** (*13.6*) 21	**57.6** (*13.6*) 14	0.49	1.00	0.24
Low prereq GPA	**46.3** (*13.3*) 36	**53.1** (*13.4*) 19	0.12	0.60	0.50
High prereq GPA	**57.8** (*19.6*) 26	**66.6** (*19.6*) 29	0.11	0.55	0.44

statistically significant after the Bonferroni correction. This correction was obtained by multiplying the unadjusted $p = 0.03$ by 5 since tests were run for five groups, as shown in Table 7. The effects were of a similar positive size for the URM and the lower and high prerequisite GPA groups ($0.44 \leq g \leq 0.50$). Thus, the effect associated with SBT for the concept inventory was similarly positive for the low and high prerequisite GPA groups, which was different than observed with the final exam. Although the differences were not statistically significant in Table 7, the direct assessment results were most promising for the concept inventory.

The third direct-assessment analysis investigates students' final letter grade in the course with and without SBT in addressing the following research question:

RQ3: What effect is associated with SBT for the final letter grade in an engineering course?

Fig. 3 shows the final grade distribution with and without SBT. SBT was associated with an increased proportion of A grades (36% vs. 27%). However, the proportions were not significantly different based on a z-test of proportions ($p = 0.28$), and the effect size was small, with the odds ratio $OR = 1.52$. The percentage of students who achieved a C or lower was also not statistically different with SBT

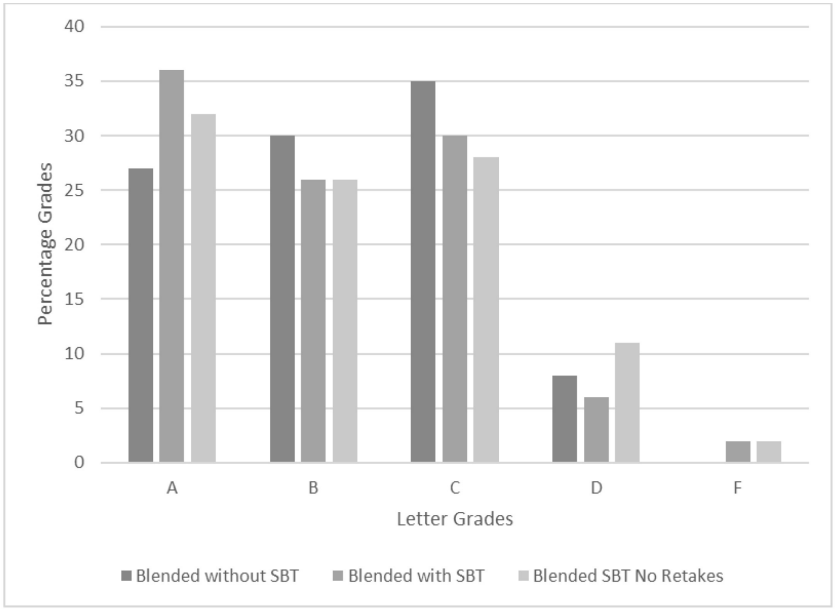

Fig. 3. Distribution of transcript letter grades for three groups: blended without SBT, blended with SBT, and blended with SBT if no retakes were allowed.

based on a z-test of proportions ($p = 0.51$). Without SBT, 44% achieved a C or less; with SBT, the percentage was somewhat lower at 38%, resulting in a small effect ($OR = 1.28$). The proportion who had to repeat the course (D or F grades) remained the same at 8% with SBT vs. without SBT. The bars on the right in each bin in Fig. 3 are the percentage of students who would have earned the course grade had the retakes not been offered. Thus, the effects associated with SBT for the final course letter grade were small, albeit positive.

5.3 *Affective Assessment with SBT*

To investigate students' feelings and perspectives associated with SBT, we utilized an inventory measuring the classroom environment and conducted two focus groups. In this section, we discuss the results of these assessments and address the following research question:

RQ4: What effect is associated with SBT on the classroom environment in an engineering course? What are students' perceptions of SBT?

5.3.1 *Classroom Environment Assessment*

From a general perspective, the classroom environment was preferable in the blended classroom with SBT. As shown in Table 8, each of the seven CUCEI dimensions (enumerated in Table 2) had higher mean values with SBT (versus without). The CUCEI dimension that was significantly higher with SBT (including after adjustment with the Bonferroni correction) was the *Involvement* dimension, which pertains to active student participation in class ($p < 0.007$). The effect size was also medium at $d = 0.71$. The *Cohesiveness* and *Satisfaction* dimensions also had approximately medium effect sizes ($d = 0.43$, $d = 0.47$, respectively) and were statistically significant before adjustment with the highly conservative Bonferroni correction. Although the CUCEI questions were not specific to SBT or its goals, there may have been an indirect connection between SBT and the classroom environment.

5.3.2 *Focus Group Assessment of Student Perspectives*

Focus Questions 1, 4, and 5 (Table 3)

The frequencies for the final codes assigned to focus group questions 1, 4, and 5 (Table 3) combined are given in Table 9, which were determined after coding and discussion by both analysts to reach a consensus. The responses to these three questions overlapped; therefore, the same coding scheme was applied to all three questions. To refresh, these questions were as follows:

1. What do you think about standards-based grading, where you get to retake tests and online LMS (Canvas) quizzes?
4. How can standards-based grading (or any element of it) be improved in this course?
5. To what degree did standards-based grading enhance your learning compared to traditional grading?

The most frequently occurring category in Table 9 was LEARN (i.e., promotion of learning by retakes), which was mentioned in 50% of the distinct statements in response to questions 1, 4, and 5. The second and third most frequently occurring categories were also positive – CHANCES and LESS STRESS, each mentioned in 22% and 19% of the distinct statements, respectively. In fact, 63% of the distinct statements contained one or more positive (i.e., benefits) categories. Out of all the

Table 8. Classroom environment comparison for seven dimensions: blended without SBT vs. blended with SBT.

Dim	Mean (s)		Univar p *unadj*	Univar p *adj*	Effect Size d
	Blended without SBT	Blended with SBT			
Cohesiveness	2.62 (*0.69*)	2.94 (*0.85*)	0.029	0.20	0.43
Individualization	2.48 (*0.62*)	2.63 (*0.64*)	0.21	1.00	0.24
Innovation	2.97 (*0.69*)	3.02 (*0.59*)	0.68	1.00	0.08
Involvement	3.02 (*0.58*)	3.46 (*0.66*)	<0.001	<0.007	0.71
Personalization	3.88 (*0.70*)	4.09 (*0.67*)	0.11	0.77	0.31
Satisfaction	3.08 (*0.99*)	3.54 (*0.96*)	0.015	0.11	0.47
Task orientation	3.94 (*0.63*)	4.09 (*0.63*)	0.25	1.00	0.23
n	63	47			

Table 9. Frequency of codes for focus group questions 1, 4, and 5

Code	Description	% of Distinct Statements	
LEARN	Retake promotes understanding or learning. Retake promotes the reinforcement of knowledge or practice. Retake allows more time to learn or absorb.	50%	Benefit
CHANCES	Retake provides a second (or more) chances.	22%	
LESS STRESS	Less stress with retakes. Less worry with retakes.	19%	
POINTS OR GRADE BOOST	Grade can be enhanced with retakes. Points can be retrieved with retakes. Retake allows multiple ways to earn points.	16%	
TOPIC FOCUS	Retake allows focusing on particular topics.	13%	
APPLY FEEDBACK	Can apply feedback to the retakes or tests.	6%	
FAIR	Retake promotes fairness or equity.	3%	
CAP	Cap set at 90%; grade cannot be improved beyond 90%. Points limited (that can be earned).	16%	Drawback or Suggestion
MORE FEEDBACK	Want more feedback. Want fuller solutions.	16%	
OTHER WORK	Retake must be handled alongside other work or commitments.	6%	
TRACKING	Retake creates special needs in tracking points.	3%	
LESS MOTIVATION	Retake may reduce motivation or serious approach initially.	3%	

statements, only 41% had drawbacks or suggestions, like CAP or MORE FEEDBACK categories. This shows that students generally have a positive perspective on the SBT.

The following are sample student responses associated with various coding categories in Table 9.

LEARN. This benefits code is assigned to responses that indicate that SBT and its retakes enhanced student learning and understanding, possibly due to reinforcement or extended time for content absorption.

> "My learning has improved due to the retakes because it forces me to look over material. I certainly go over topics I didn't understand, which doesn't happen with other traditional grading."

> "The SBG creates more of a learning atmosphere versus a pass/fail atmosphere. I won't review things I did wrong otherwise. It can be engrained into my brain this way."

LESS STRESS. This benefits code applies to responses that discuss a decrease in students' stress or worry with the retakes.

> "I like the test retakes. It takes pressure off, especially when studying. If I don't understand something when studying, I tense up."

> "I can focus on learning versus being worried about doing well and what my grade will be. It is more comfortable, and I am more ok with being wrong."

TOPIC FOCUS. This benefits theme pertains to students' ability to focus their study in certain areas with the retakes.

> "I like the opportunity to retake, as I don't worry about my grade. Also, if I am wrong on something, I can focus on just that topic for the retake. I get a second chance."

LESS MOTIVATION. This drawbacks category relates to a possible reduction in student motivation or a serious approach to their first-chance test, knowing that a second-chance test was also possible.

> "I feel that subconsciously I may not take the original test as seriously if I know I can retake it."

Focus Question 2

2. Do you read and use the extensive feedback given to you on the tests and the retakes?

Out of the 12 statements made in response to the question in the two focus groups, 10 (83%) of them answered Yes, while 2 (17%) answered No. This suggests that a substantial number of students probably read and utilized the feedback from the tests and retests, which is a positive outcome for instructors. See below for an example statement for each response.

Yes

> "I see the notes from Dr. K after the original test, which point to examples in the book. I go to the book and practice these, including after the retake, because that might help me on the final exam."

No

> "I have never used the feedback. I simply look at the questions I got wrong and practice those topics more."

Focus Question 3

3. Did retaking the tests take time away from learning the new content?

Out of the 12 responses given to this question during the focus groups, 8 (67%) were answers of

'No', and 4 (33%) were categorized as 'Other Impacts'.' This suggests that, in general, the students did not consider the retakes to be an additional burden on top of the new course material they were learning. Examples of statements for each category are provided below, and there were no responses categorized as 'Yes'.

No

> "No, it doesn't take time away. It just involves reviewing concepts. I go over what I missed and not the whole unit."
>
> "A little bit, but it gives you peace of mind that you can retake the test."
>
> "It does not take time away. I should have studied more for the first test, so I just made it up with the retake."
>
> "I only need an hour or two to be more prepared because I have already studied a lot."

Other Impacts

> "Retakes are not convenient if they are on the same day as a quiz is due."

5.4 Analysis of Test and Quiz Retaking Behaviors and Outcomes

In this section, we review the course's test and quiz retaking characteristics, specifically participation levels, the student groups who retook tests and quizzes with the most significant relative frequency, and the retesting outcomes. Specifically, we address the following research question in this section:

RQ5: What participation and performance characteristics are associated with test and quiz retaking?

Fig. 4 displays three different percentages per bin along the x-axis, with each bin corresponding to the first-chance test topic score. Since 53 students were enrolled in the study, and eight topics were covered in the course, there were a total of 53*8, or 424, topic scores for first-chance testing. Thus, 424 retake-topic scores (i.e., second-chance) were possible. The left bar in each bin corresponds to the percentage of first-chance topic scores. For example, in Fig. 4, 46% of the first-chance topic scores were between 80–100%. The middle bar displays the percentage of first-chance topic scores students tried to improve through a retake of the topic test. For example, only 15% of the retake opportunities were pursued in the 80–100% bin, compared to 76% in the 60–79% bin. The right bar in each bin shows the percentage of topic retakes that resulted in a higher updated topic score. The trend for the right bar was similar to the middle bar.

Fig. 5 is a scatter plot of first chance versus retake topic test scores. Of the possible 424 opportunities for retaking a topic test throughout the semester (i.e., 53 students * 8 topics), only 209 opportunities (49%) were pursued. That is why many points along the abscissa in Fig. 5 have ordinate values of zero. The low 49% participation rate may have occurred due to students being satisfied with their first-chance scores, as 46% of the first-chance topic scores were in the 80–100% bin. Encouragingly, of the total retake opportunities pursued, 60% (125/209) resulted in a higher score on the second-chance topic test, resulting in an increased topic test score used to calculate the course grade.

Fig. 6 is analogous to Fig. 4 and displays information about the post-class LMS quiz scores administered through the LMS. Similar to Fig. 4, Fig. 6 shows three different percentages per bin, with each bin corresponding to the first-chance online quiz

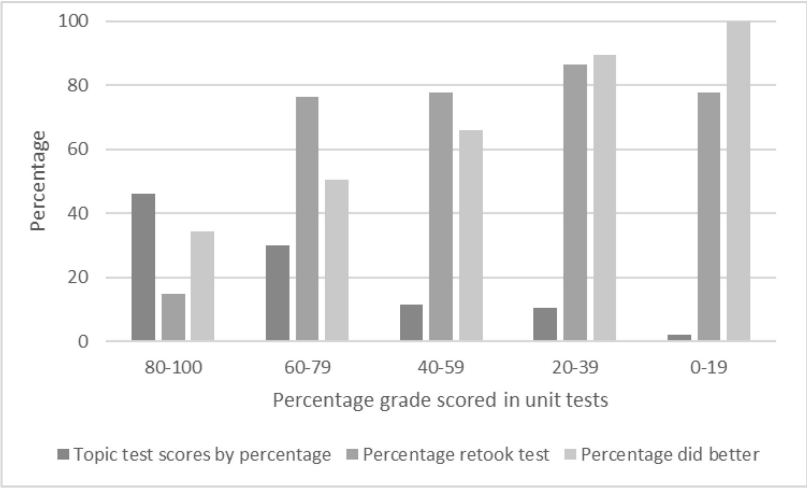

Fig. 4. Percentage of students given by first-chance topic test scores, percentage of students who retook the topic tests, and percentage of students who improved their score via the retake tests.

Fig. 5. Retake topic test percentage score vs first-chance topic test percentage score.

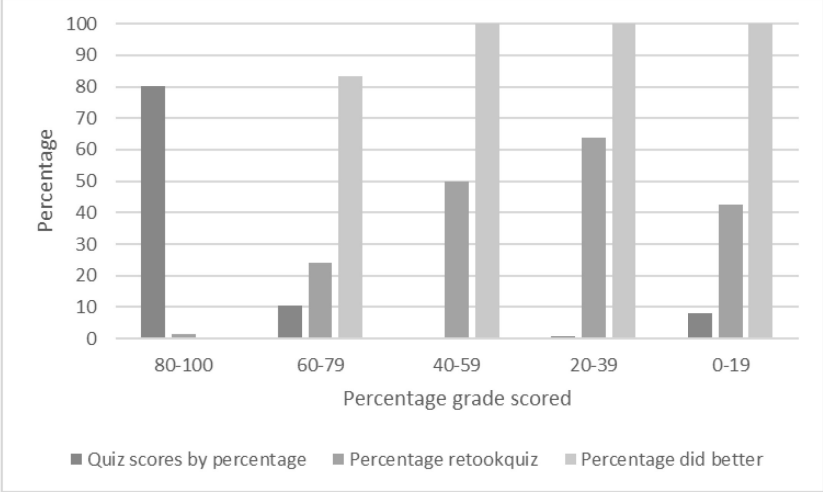

Fig. 6. Percentage of students given by first-chance LMS quiz scores, percentage of students who retook the LMS quizzes, and percentage of students who improved their score via the retake quizzes.

score. Since 53 students were enrolled, and 31 post-class LMS quizzes were given throughout the semester, there were a total of 53*31 = 1,643 first-chance quiz scores. Thus, 1,643 retake (i.e., second chance) quiz scores were possible. The left bar in each bin corresponds to the percentage of first-chance online quiz scores in that performance bin. For example, in Fig. 6, 80% of the first-chance quiz scores were between 80–100%, which is higher than the 46% for the topic tests (see Fig. 4). The middle bar displays the percentage of first-chance quiz scores that students tried to improve through a retake of the quiz. For example, only 1.4% of the possible retake opportunities were pursued in this same bin, compared to 24% in the 60–79% bin. Thus, as in Fig. 4, the percentage of quiz retakes pursued generally increased with decreasing first-chance scores. The right bar shows the percentage of quiz retakes that resulted in a higher updated quiz score.

Throughout the semester, there were 1,643 chances for students to retake online quizzes (53 students multiplied by 31 quizzes). However, only 127 opportunities (8%) were taken advantage of, which is significantly lower than the 49% of retaken topic tests. The low participation rate of 8% could be attributed to the fact that 80% of the first-chance quiz scores were already in the 80–100% range. Additionally, students may have perceived the post-class LMS quizzes as having lower stakes. Students who did retake quizzes generally did so because they missed them the first time and could recover half the points. Notably, 80% of the retaken quizzes resulted in a higher score on the second attempt, which led to an increase in the quiz score used to calculate the course grade.

6. Summary and Discussion

This study has demonstrated positive, desirable outcomes (both cognitive and affective) associated with standards-based testing (SBT) in an undergraduate numerical methods course for engineers. In doing this, a blended classroom with SBT was compared to a blended classroom without SBT, with the instructor being the same for both classrooms. With our implementation of SBT, students could retake topic tests for the eight topics (i.e., standards or objectives). They received feedback that they could apply to the retests and the final exam. If a student earned a higher retake score for a given topic, an adjustment was made to their original test score for the topic, which was equal to half the difference. Students could also retake post-class LMS quizzes.

On the culminating direct assessments for students as a whole, all effects associated with SBT were positive. These effects ranged in size from small for the multiple-choice ($g = 0.06$) and free-response ($g = 0.22$) parts of the final exam to near-medium for the concept inventory ($g = 0.45$). None of the differences were statistically significant after applying the Bonferroni correction; however, the sample sizes were small within the demographic strata. These positive effects may be attributable to revisiting and reinforcement of content, spaced and increased practice, focus on true learning versus grades, and application of extensive feedback provided on the first- and second-chance tests. The effects associated with SBT on the final course grade were also positive, albeit small. This effect was supported by the series of second chances made available to students, in which they could boost their scores on the topic tests.

Likewise, for all seven dimensions of the classroom environment survey, the effects associated with SBT were positive. Although the CUCEI questions were not necessarily specific to SBT or its goals, there may have been an indirect connection between SBT and the classroom environment, with positive feelings about SBT translating to positive in-class experiences or positive student feelings. The dimension that was significantly higher with a medium effect with SBT was Involvement ($p < 0.007$). The Cohesiveness and Satisfaction dimensions also had near-medium effect sizes with SBT. The focus group responses also demonstrated positive student perspectives towards SBT, with 63% of distinct statements containing one or more 'benefits' categories, such as promotion of learning, decreased stress, and second chances. A lower percentage (i.e., 41%) of distinct statements contained one or more 'drawbacks or suggestions' categories, and students generally did not find the retakes burdensome.

Taking a second viewpoint on the direct assessments, more students earned a high final exam score (80–100%) in the SBT versus the non-SBT group. The effect for this was substantial ($OR = 6.5$), and the difference was significant ($p = 0.016$). Likewise, with the concept inventory, the two top-scoring groups combined (i.e., 60–79% and 80–100%) contained a significantly higher proportion of students who had taken the course with SBT. This effect was medium, and the difference was statistically significant ($p = 0.006$). The effects of the direct assessments associated with SBT were mainly positive for the highest performers.

With the stratified analyses, the positive effects associated with SBT were the largest for the strata defined by the academic variable (i.e., prerequisite GPA). This effect is in contrast to the strata based on race/ethnicity and socioeconomic status. For example, with the multiple-choice questions, the effect for SBT for students with a low prerequisite GPA was $g = 0.38$, while the effects for URM and Pell Grant recipients were only $g = 0.11$ and $g = 0.14$, respectively. Thus, the SBT may help low-achieving students. On the free-response questions, the effect associated with SBT for students with a high prerequisite GPA was $g = 0.53$, while the URM stratum was only $g = 0.09$. On the concept inventory, the effect associated with SBT for students having a low prerequisite GPA was $g = 0.50$, and for the high GPA students, $g = 0.44$. The effect was only $g = 0.24$ for the Pell Grant recipients. Thus, from a general perspective, the effects associated with SBT were most prominent for the strata defined by the academic variable versus the non-academic variables, with SBT potentially supporting students with low GPAs.

Relative to assessment retake behaviors, we found that students' percentage of retake opportunities generally increased with decreasing first-chance scores. Others have noted this result as well. Emeka et al. [6] concluded that scores on the first-chance exam were the primary factor in students' decisions to take the second-chance exam. With the topic tests, 49% of all possible retake opportunities were pursued, compared to only 8% for the online quizzes. However, 80% of first-chance quiz scores, versus 46% of first-chance test scores, were already high (i.e., between 80–100%). Encouragingly, of the test retakes, 60% resulted in a higher score and, therefore, a higher final topic test score. The corresponding percentage for the online quizzes was 80%.

A limitation of this study was that it was conducted for one course at one university during two different semesters. Therefore, the sample sizes

within the demographic strata were small, which we accommodated by applying non-parametric statistical techniques. The smaller sample sizes may have been an issue in demonstrating statistical significance.

The instructor plans to use SBT in his courses in the future. He values the multiple chances whereby students can show they know the topic. He believes this reduces test anxiety and automatically creates spaced practice and likely enhanced retention, particularly for lower-performing students. The instructor had expected an increase in final letter grades with SBT. Still, with SBT, students showed an improvement on average in their conceptual and comprehensive understanding through their end-of-term concept inventory and final exam scores.

Modern LMSs unfortunately do not have a suitable system for reporting standards-based testing results. Students, therefore, had to keep track of their current grade status via an instructor-programmed Excel sheet, although they knew their lowest possible grade via the LMS. Also, on the flip side, the instructor estimated that with the second-chance tests, his grading time increased by approximately 50%, and his test development time increased by 60%, although class time decreased by about 5%. Thus, instructors need to be prepared for more grading and assessment-development time.

Given the out-of-class time needed for grading and creating retake examinations, and the in-class challenges due to reduced lecture time, we recommend that the instructor be relatively experienced in teaching the course before implementing SBT. Experience will assist the instructor in handling these challenges, including coverage of the necessary content in less time. Also, given the 'advanced' level of accounting needed for the revised scores, we recommend the development of spreadsheets that will automate this for students and instructors alike.

7. Conclusions

This study compared a blended classroom with SBT (experimental group) to one without SBT (control group), presenting a comprehensive analysis of cognitive and affective outcomes. The use of multiple-chance testing (MCT) in implementing standards-based testing (SBT) in a blended-format numerical methods course has yielded significant and positive outcomes, both quantitatively and qualitatively.

The results demonstrate a clear advantage for standards-based testing, with a substantial increase in the percentage of students earning high final exam scores (15% vs. 3%) and a higher proportion of A grades (36% vs. 27%). The classroom environment, measured through the dimensions of involvement, cohesiveness, and satisfaction, exhibited significant improvement with SBT, particularly in involvement ($p < 0.007$) and near-medium effect sizes for cohesiveness and satisfaction.

Quantitative direct assessments, including concept inventory and final exam results, revealed positive effects associated with SBT. Although not statistically significant after Bonferroni's correction, the effects were consistently positive, ranging from small to near-medium effect sizes. The study also uncovered that SBT had several more pronounced positive effects for students with lower prerequisite GPAs, emphasizing its potential to support lower-achieving students.

Analysis of assessment retake behaviors indicated a higher percentage of opportunities taken for students with lower first-chance scores, aligning with previous observations in the literature. While acknowledging the study's limitation regarding sample size and single-course focus, the findings contribute valuable insights into the potential benefits of SBT.

This study advocates for the thoughtful adoption of standards-based testing, specifically through the multiple-chance testing approach, emphasizing its positive impact on learning outcomes, student experiences, and the overall classroom environment. The findings encourage further exploration and implementation of SBT, recognizing its potential to enhance education practices in postsecondary courses. The instructor's intention to continue using SBT in future courses underscores the perceived advantages, including multiple chances for students to demonstrate proficiency, reduced test anxiety, and automatic creation of spaced practice. However, grading time and test development challenges must be considered. Therefore, it is recommended that instructors be experienced and well-prepared.

Acknowledgments – This research was made possible by the Engineering Education Research Center (EERC) in the Swanson School of Engineering at the University of Pittsburgh and the College of Engineering at the University of South Florida. The work would not have been possible without the data collected under the work supported partially by the National Science Foundation under Grant Numbers 1609637 and 2013271. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. S. M. Brookhart, Graded achievement, tested achievement, and validity, *Educational Assessment*, **20**(4), pp. 268–296, 2015.
2. J. Schneider and E. L. Hutt, *Off the Mark: How Grades, Ratings, and Rankings Undermine Learning (but Don't Have To)*, Harvard University Press, 2023.

3. E. Tuson and T. Hickey, Mastery learning and specs grading in discrete math, in *Proceedings of the 27th ACM Conference on Innovation and Technology in Computer Science Education*, **1**, pp. 19–25, 2022.
4. J. Cotter and R. Guldiken, Remote versus in-class active learning exercises for an undergraduate course in fluid mechanics, in *Proceedings of the Virtual ASEE Conference & Exposition*, 2021.
5. C. Ott, B. McCane and N. Meek, Mastery learning in cs1-an invitation to procrastinate?: Reflecting on six years of mastery learning, in *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education*, **1**, pp. 18–24, 2021.
6. C. Emeka, T. Bretl, G. Herman, M. West and C. Zilles, Students' perceptions and behavior related to second-chance testing, in *IEEE Frontiers in Education Conference (FIE)*, pp. 1–8, 2021.
7. O. E. Fernandez, Second chance grading: An equitable, meaningful, and easy-to-implement grading system that synergizes the research on testing for learning, mastery grading, and growth mindsets, *PRIMUS*, **31**(8), pp. 855–868, 2021.
8. G. Herman, K. Varghese and C. Zilles, Second-chance testing course policies and student behavior, in *IEEE Frontiers in Education Conference (FIE)*, pp. 1–7, 2019.
9. D. Ariely and K. Wertenbroch, Procrastination, deadlines, and performance: Self-control by precommitment, *Psychological Science*, **13**(3), pp. 219–224, 2002.
10. C. A. Tomlinson and J. McTighe, *Integrating differentiated instruction & understanding by design: Connecting content and kids*. ASCD, 2006.
11. J. B. Carroll, A model of school learning, *Teachers College Record*, **64**(8), pp. 1–9, 1963.
12. B. S. Bloom, The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring, *Educational researcher*, **13**(6), pp. 4–16, 1984.
13. C. S. Dweck, *Self-Theories: Their Role in Motivation, Personality, and Development*. Psychology Press, 2000.
14. A. Bandura, W. H. Freeman and R. Lightsey, *Self-Efficacy: The Exercise of Control*, ed: Springer, 1999.
15. A. R. Carberry, M. Siniawski, S. A. Atwood and H. A. Diefes-Dux, Best practices for using standards-based grading in engineering courses, in *Proceedings of the ASEE Annual Conference & Exposition*, 2016.
16. M. T. Siniawski, A. Carberry and J. D. N. Dionisio, Standards-based grading: An alternative to score-based assessment, in *Proceedings of the ASEE PSW Section Conference*, 2012.
17. LS Post. Standards-based grading in a thermodynamics course, International Journal of Engineering Pedagogy, **7**, pp. 173–181, 2017.
18. R. Averill, S. Roccabianca and G. Recktenwald, A new assessment model in mechanics of materials, in *Proceedings of the ASEE-NCE Annual Conference & Exposition*, 2019.
19. J. P. Moore and J. Ranalli, A mastery learning approach to engineering homework assignments, in *Proceedings of the ASEE Annual Conference & Exposition*, 2015.
20. C. Perez and D. Verdin, Mastery learning in undergraduate engineering courses: A systematic review, in *Proceedings of the ASEE Annual Conference & Exposition*, 2022.
21. A. Harsy, Variations in mastery-based testing, *PRIMUS*, **30(**8–10), pp. 849–868, 2020.
22. A. Harsy and A. Hoofnagle, Comparing mastery-based testing with traditional testing in calculus II, *International Journal for the Scholarship of Teaching and Learning*, **14**(2), Article 10, 2020.
23. M. Henriksen, J. Kotas and M. Wentworth, Specifications-based grading reduces anxiety for students of ordinary differential equations, *Community of Ordinary Differential Equations Educators Journal*, **13**(1), pp. 1, 2020.
24. D. Lewis, Impacts of Standards-Based Grading on Students' Mindset and Test Anxiety, *Journal of the Scholarship of Teaching and Learning*, **22**(2), pp. 67–77, 2022.
25. D. Chamberlain Jr, How one instructor can teach a large-scale, mastery-based College Algebra course online, *PRIMUS*, **33**(8), pp. 867–888, 2023.
26. J. Lenarz and K. E. Pelatt, A transition to mastery-based testing with the hope of increasing student persistence, *PRIMUS*, **33**(2), pp. 97–106, 2023.
27. S. L. Noell, M. Rios Buza, E. B. Roth, J. L. Young and M. J. Drummond, A bridge to specifications grading in second-semester general chemistry, *Journal of Chemical Education*, **100**(6), pp. 2159–2165, 2023.
28. A. Kaw, G. Besterfield and J. Eison, Assessment of a web-enhanced course in numerical methods, *International Journal of Engineering Education*, **21**(4), pp. 712–722, 2005.
29. R. Clark, A. Kaw, Y. Lou, A. Scott and M. Besterfield-Sacre, Evaluating blended and flipped instruction in numerical methods at multiple engineering schools, *International Journal for the Scholarship of Teaching & Learning*, **12**(1), Article 11, 2018.
30. R. Clark, A. Kaw and R. Braga Gomes, Adaptive Learning: Helpful to The Flipped Classroom in The Online Environment of Covid?, *Computer Applications in Engineering Education*, **30**(2), pp. 517–531, 2022.
31. A. L. Glass, M. Ingate and N. Sinha, The effect of a final exam on long-term retention, *The Journal of General Psychology*, **140**(3), pp. 224–241, 2013.
32. K. K. Szpunar, K. B. McDermott and H. L. Roediger, Expectation of a final cumulative test enhances long-term retention, *Memory & Cognition*, **35**(5), pp. 1007–1013, 2007.
33. R. M. Clark, A. Kaw and M. Besterfield-Sacre, Comparing the effectiveness of blended, semi-flipped, and flipped formats in an engineering numerical methods course, *Advances in Engineering Education*, **5**(3), 2016.
34. A. Yalcin, A. Kaw and R. Clark, On learning platform metrics as markers for student success in a course, *Computer Applications in Engineering Education*, **31**(5), pp. 1412–1432**,** 2023.
35. Federal Pell Grant Program. https://studentaid.gov/understand-aid/types/grants/pell (accessed January 11, 2024).
36. B. J. Fraser and D. F. Treagust, Validity and use of an instrument for assessing classroom psychosocial environment in higher education, *Higher Education*, **15**(1–2), pp. 37–57, 1986.
37. T. V. Perneger, What's wrong with Bonferroni adjustments, *BMJ*, **316**(7139), pp. 1236–1238, 1998.
38. D. Lakens, Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs, *Frontiers in Psychology*, **4**, Article 863, 2013.
39. J. Cohen, *Statistical Power Analysis for the Social Sciences*, Lawrence Erlbaum Associates, 1988.
40. D. Howitt and D. Cramer, *Research Methods in Psychology*, Pearson, 2020.
41. K. Neuendorf, *The Content Analysis Guidebook*, Thousand Oaks, CA: Sage Publications, 2002.

42. M. Norusis, *SPSS 14.0 Statistical Procedures Companion*, Upper Saddle River, NJ: Prentice Hall, 2005.
43. A. Kaw, Y. Lou, A. Scott and R. Miller, Building a concept inventory for numerical methods: a chronology, in *Proceedings of the ASEE Annual Conference and Exposition*, 2016.
44. A. Kaw and A. Yalcin, Measuring student learning using initial and final concept test in a STEM course, *International Journal of Mathematical Education in Science and Technology*, **43**(4), pp. 435–448, 2012.
45. M. Adler and E. Ziglio, *Gazing Into the Oracle: The Delphi Method and its Application to Social Policy and Public Health*, Jessica Kingsley Publishers, 1996.
46. D. Quade, Rank analysis of covariance, *Journal of the American Statistical Association*, **62**(320), pp. 1187–1200, 1967.
47. A. Lawson, Rank analysis of covariance: Alternative approaches, *Journal of the Royal Statistical Society: Series D (The Statistician)*, **32**(3), pp. 331–337, 1983.
48. R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye, *Probability and Statistics for Engineers and Scientists*, Macmillan New York, 1993.
49. A. Agresti, *Statistical Methods for the Social Sciences*, 5th ed. New York: Pearson Education, 2017.
50. G. M. Sullivan and R. Feinn, Using effect size – or why the p value is not enough, *Journal of Graduate Medical Education*, **4**(3), pp. 279–282, 2012.

Autar Kaw is a professor of mechanical engineering at the University of South Florida whose scholarly interests include engineering education research, adaptive, blended, and flipped learning, open courseware development, composite materials mechanics, and bascule bridge design. The National Science Foundation, Air Force Office of Scientific Research, Florida Department of Transportation, and Wright Patterson Air Force Base have funded him. Under his leadership and funding from NSF, he and his colleagues from around the nation developed, implemented, refined, and assessed online resources for open courseware in Numerical Methods. This courseware receives over 1 million page views, 1.6 million views of the YouTube lectures, and 90,000 visitors to the ''numerical methods guy'' blog annually. This courseware is also used to measure the impact of flipped, blended, and adaptive settings on how well engineering students learn content, develop group-work skills, and perceive the learning environment. He has written over 120 refereed technical papers, and his opinion editorials have been featured in the Tampa Bay Times, the Tampa Tribune, and the Chronicle of Higher Education. He has earned several teaching awards at the national level, including the 2012 US Professor of the Year Award (doctoral and research universities) from the Council for Advancement and Support of Education and the Carnegie Foundation for Advancement of Teaching.

Renee Clark is an Associate Professor of Industrial Engineering, Data Engineer for the Swanson School of Engineering, and Director of Assessment for the Engineering Education Research Center (EERC). She uses data analytics to study techniques and approaches in engineering education, focusing on active learning techniques and the professional formation of engineers. Current NSF-funded research includes the use of adaptive learning in the flipped classroom and systematic reflection and metacognitive activities in the mechanical engineering classroom. She also serves as Associate Editor for *Advances in Engineering Education*. She has 30 years of experience as an engineer and IT analyst in industry and academia. She completed her post-doctoral studies in engineering education at the University of Pittsburgh.