8-18-2025

Difference in High School Geometry End-of-Year State Exam Scores Between Three Different Retake Time Frames

Christina Noel Walton
Walden University

<u>Walden University</u>

College of Education and Human Sciences

This is to certify that the doctoral study by

Christina N. Walton

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. Jennifer Seymour, Committee Chairperson, Education Faculty
Dr. Christopher Cale, Committee Member, Education Faculty

Chief Academic Officer and Provost
Sue Subocz, Ph.D.

Walden University
2025

Abstract

Difference in High School Geometry End-of-Year State Exam Scores Between Three Different Retake Time Frames

by

Christina N. Walton

MS, Air Command and Staff College, 2000

MS, Air Force Institute of Technology, 1987

MS, Air Force Institute of Technology, 1987

BS, University of Southern California, 1986

Project Study Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Education

Walden University

August 2025

Abstract

A grading policy was implemented in a Northeastern suburban public school district that left it to curriculum teams to define the amount of time between the original test and its retake resulting in a slight decrease in high school students' geometry end-of-year state exam scores. Guided by Bloom's theory of mastery learning, the purpose of this quantitative causal-comparative study was to investigate if there were statistically significant differences between high school students' geometry end-of-year state exam scores or between high school students' geometry math growth scores when controlling for the preassessment score differentiated by three schemes of retake timing: *short-term*, *end of quarter*, and *anytime*. The Welch ANOVA analyzed the difference in high school students' (N = 1197) geometry end-of-year exam scores differentiated by retake timing and determined how much time was given between the original test and its retake was statistically significant, (p = .009). A Games-Howell post hoc test determined end-of-year scores decreased when the amount of time between the original exam and its retake was increased from the 2-3 week to anytime. However, there were validity issues with these results. An ANCOVA analyzed the difference in high school students' math growth scores when controlling for the preassessment score differentiated by retake timing and determined the amount of time given between the original test and its retake was not statistically significant. A policy paper was created recommending modification of how assessments are tracked so an improved data set can be created for future analysis. By optimizing data collection, positive social change, including the ability to improve processes and evaluate changes in practice, may result in improved student learning.

Difference in High School Geometry End-of-Year State Exam Scores Between Three Different Retake Time Frames

by

Christina N. Walton

MS, Air Command and Staff College, 2000

MS, Air Force Institute of Technology, 1987

MS, Air Force Institute of Technology, 1987

BS, University of Southern California, 1986

Project Study Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Education

Walden University

August 2025

Dedication

To my husband, Frank, and my children, Casey, Courtney, and Justin, I dedicate this doctoral achievement to you. Your love, support, and patience throughout this journey have been my inspiration. Frank, your belief in me and your support throughout this journey were my constant guide. Casey, Courtney, and Justin, you bring me immeasurable joy and purpose to my life. Your passion and persistence have always motivated me to chase my dreams and embrace the joy in my pursuits - may we teach you both to do the same. This accomplishment would not have happened without you.

Acknowledgments

I wish to thank my doctoral committee Dr. Jennifer Seymour and Dr. Chris Cale for offering their advice, guidance, and patience throughout the process. Their sage words and mentorship helped me to traverse the many obstacles I faced to complete this research. I also want to thank my school sponsor, the assistant superintendent of Instructional Services, my principal, whose many conversations and mentorship were a support on both the research study as well as how to navigate through the district hierarchy, and the district's superintendent, who helped clear the hurdles so I could finish this study. Finally, to my best friend, advisor, and husband. Thank you for supporting me as I worked towards accomplishing this goal.

Table of Contents

List of Tables

List of Figures

Section 1: The Problem

The Local Problem

Mastery grading attempts to address the issue of students passing a class through partial credit without ever being able to correctly solve a problem (Dudley, 2010), addressing this issue by focusing on whether students can master a clear list of student academic objectives rather than awarding points or partial credit for incomplete work. A retake policy in a Northeastern suburban public school district was implemented in 2016 leaving to curriculum teams the discretion to determine the amount of time between the original test and retake. This has resulted in many different practices being used across the district and has ultimately resulted in minimal to no improvement in high school students' geometry end-of-year state exam scores.

In April of 2015, a Northeastern suburban public school district convened a task force to develop an improved system of grading and reporting based on Guskey and Bailey's (2001) five imperatives. One result of the task force was district officials implemented a retake process for every major assignment in summer of 2016, providing students additional time to work towards mastery of the material and an additional opportunity to demonstrate their knowledge (District policy, 2021). The retake policy as implemented by the school district addressed the requirement for a retake but did not determine the specific practices necessary to implement the policy, leaving it to the curriculum teams to determine the implementation, including appropriate interventions and timeline for reassessment. In the 7 years since initial implementation, multiple practices have been implemented by curriculum teams across the district, resulting in many different choices on the amount of time between the original test and its retake.

The policy leaving it to the curriculum teams to define the amount of time between the original test and its retake has resulted in many different practices implemented within the district and a slight decrease in high school students' geometry end-of-year state exam scores. This discrepancy between accepted theory and practical application is the genesis for the identified gap in practice. It is not known if any of the practices currently being used in the district are producing different results, or if any one practice is better than the rest at improving high school students' geometry end-of-year state exam scores. Schools are currently executing a variety of practices without the data to determine which is most effective. As a result, it is unclear if it is acceptable that different practices are being used or if a singular one should be adopted that provides better access to mastery learning. The proposed research attempts to fill this gap in practice by investigating what the differences in high school students' geometry end-of-year state exam scores are, given the implementation of different retake time frames and which of these practices currently in use by schools in the district have the highest influence on student learning.

Rationale

Evidence of the Problem at the Local Level

Evidence of the problem at the local level includes both evidence of the different practices that have resulted from the policy and evidence of the minimal improvement of high school students' geometry end-of-year exam scores.

Different Practices

A data discovery was performed by requesting syllabi from secondary and high school's Geometry and Geometry Honors curriculum teams for the 23-24 school year (N = 48). The syllabi were used to determine how much time individual curriculum teams across the district used between the original school-based test(s) and their retakes at their respective schools. Syllabi were not requested from either the alternative high schools or the governor's school as they are outside the scope of the study. The time frames identified in the research problem were chosen from those identified as being in use by curriculum teams within the syllabi.

Table 1

Summary of Time Frames Used in a Northeastern Suburban School District

School year	Amount of time provided	Number of curriculum teams	Details
2023-2024	End of quarter	8	Retake of multiple units taken at end of quarter
	2-3 weeks	14	Retake 2-3 weeks after original exam
	Anytime	6	Retake available anytime during the school year
	Not enough information	20	Syllabi provided did not identify amount of time provided or used a different method

Instructional staff from across the district have noted unintended consequences with the implementation of the retake policy (Instructional staff, personal communication, September 15, 2021 & data discovery, Mar 15, 2022). Instructional staff noted that retakes are effective for the student who had a bad day or made minor mistakes on the original test but were not effective at closing the gap in understanding. Students who struggle the whole year normally do just as poorly or even worse on the retake. General mathematics skills have weakened over time as students rely on the retake to recover their grade as opposed to focusing on mastering content (Instructional staff, personal communication, December 2, 2021). Instructional staff have also noted that students are no longer proactively preparing for assessments, coming into assessments asking when the retake will be given or attempting to use the retake as a practice test for the second exam opportunity. Across the district the observation was students generally do not prepare properly for the retake, and many students who take the first retake tend to get into a cycle of needing to take multiple retakes, never fully recovering from the effects of the additional workload associated with retaking exams (department meeting, January 19, 2022 & data discovery, Mar 15, 2022).

High School Students' Geometry End-Of-Year Exam Scores

The retake policy was intended to provide students with additional time to work towards mastery of material and an additional opportunity to demonstrate this knowledge (District policy, 2021). However, the policy resulted in a slight decrease in the percentage of high school students who passed the geometry end-of-year state exam scores. Using the state end-of-year geometry exam as a measure of course material mastered during the school year, and therefore effectiveness of a student's use of the retake to master material, Table 2 below demonstrates that the percentage of high school students who have passed the geometry exam across the district declined slightly since the implementation of the retake process. Although it appears that there was a jump in percentage during the 2018-2019 school year, there was a change of policy on who took end-of-year exams beginning with this year. Instead of requiring state exams for all students who were taking Algebra I, Geometry, and Algebra II, the requirement changed to only one mathematics exam was required at the high school level. The table notes a 6% decline in the number of students who sat the exam this year. This resulted in an increased percentage of honors students in the pool of students who took the exam, as the general education students who took the Algebra I test at the high school level were no longer required to take the geometry exam, resulting in an arbitrary increase in overall score that year.

Table 2

Percentage of Students Meeting or Exceeding the High School Geometry End-of-Year State Exam

School year	Number of students	Percentage of students who passed
2014-2015	13,752	85
2015-2016	14,019	85
2016-2017	14,506	83
2017-2018	14,607	82
2018-2019*	13,704	85

Note. From Northeastern State Department of Education (n.d.).

* Increased percentage resulted from new policy requiring only one math exam per student increasing the proportion of honors level students in comparison to all students who took the exam.

Evidence of the Problem in the Professional Literature

Mastery learning is the overarching theory from which mastery-based assessments, mastery-based grading, and standards-based grading were derived. All these practices include the ability to retest to demonstrate mastery of material as part of their implementation. The district's retake policy stems from standards-based grading which is based on Bloom’s mastery learning theory and Dweck’s implicit theories or mindsets (Bloom, 1968; Dweck, 2000). However, current research does not provide clear guidance on practice, demonstrating both positive and negative effects resulting from individual implementation schemes (Cundiff et al., 2020; Harsy & Hoofnagle, 2020; Heubach & Krinsky, 2020). Research also demonstrates the unintended consequences of implementing a retake process identified by instructional staff (Olsen & Buchanan, 2019; Winget & Persky, 2022; Wisch et al., 2018).

No grading system by itself improves student learning; however, better grading can provide more accurate information on student learning and achievement. Knight and Cooper (2019) and Selbach-Allen et al. (2020) both showed that standards-based grading generally yields a stronger relationship between grades and external measures of student achievement, but neither study provided evidence that standards-based grading improves student achievement. Although proponents stress retakes, or allowing students multiple opportunities to reassess, improve student self-concept, student-teacher relationships, and improved communication on what students have learned, there is little empirical evidence confirming these benefits (Link & Guskey, 2022). Additionally, reassessment policies that allow for unlimited retakes generate teacher push back as vast curriculum coverage requirements with innumerable retakes can cause a burdensome workload for teachers (Scarlett, 2018; Townsley, 2019; Wisch et al., 2018). How teachers and students use the assessment results is what matters, following assessments with high-quality corrective instruction to help students remedy whatever learning errors the assessment identified (Bloom, 1968; Guskey, 2005).

Although Bloom (1968) allows for multiple corrective opportunities and re-assessments to determine mastery, the school district being evaluated only provides one opportunity to re-assess material. The problem involves the inherent trade-off between the certainty of mastery and the efficiency of practice, minimizing false positives where the system declares mastery but the learner does not fully understand the content material. Pelánek and Řihák (2018), using both simulated and real data, showed there was not a significant difference between simple mastery criteria such as consecutive correct or exponential moving averages, and more complete methods. The choice of data sources used for the mastery decision and setting of thresholds, driven by the trade-off between the risk of premature mastery and over-practice, are more important than the choice of learner modelling technique. Bälter et al. (2018) performed a study (N = 3813) to determine how many opportunities were necessary, on average, for a university level learner to master a skill. Inputs into an on-line learning system were evaluated across 74 different course offerings in four topic areas (statistics, biology, engineering statics, and psychology). ANOVA and two-sample t-tests determined the mean number of attempts to reach mastery is between three to six attempts. Analysis revealed a minimum of seven interactions with a specific academic standard is required for all students to reach mastery of a particular standard. Nelson et al. (2018) evaluated the number of attempts required to mastery specific mathematics subskills among fourth through eighth graders receiving supplemental mathematics support. They determined that for most subskills, students mastered the material at a 50% level after one try; however, for 71% of the subskills between two and four attempts were necessary to master at a 90% level, with many students failing to reach mastery even after direct, supplemental intervention. Although the choice of data sources and mastery thresholds determines whether a student achieves premature mastery or over-practice (Pelánek & Řihák, 2018), it requires between two and four attempts at the middle school level or three to six attempts at the university level to achieve mastery of material (Bälter et al., 2018; Nelson et al., 2018).

Providing students multiple opportunities to demonstrate their knowledge using a retake is a best practice, but there is no consensus on the best implementation strategy for a retake process (Brookhart et al., 2016; Olsen & Buchanan, 2019; Wisch et al., 2018). Wisch et al. (2018) identified that there is a discrepancy between theory and practice, highlighting that there are differences of opinion on what constitutes best practice. Alternatively, there is also research that demonstrates that retakes have no or negative effect on students. In Guskey's (2010) discussion of Bloom's theory of mastery learning, he acknowledged students did not consistently improve on subsequent retakes, possibly resulting from trying to simultaneously keep up with new content while trying to remediate previous content. Brookhart et al.'s (2016) literature review on assessment identified that teachers are not fully embracing standards-based grading, and standards-based grading did not improve student knowledge as assessed by high-stakes end-of-year testing. Brookhart et al. (2016) also highlighted that there were few empirical studies on the effectiveness of the standards-based grading pedagogy. The purpose of this quantitative study is in a Northeastern suburban public school district, to investigate if there are differences between high school students' geometry end-of-year state exam scores or between high school students' geometry math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*.

Definition of Terms

A grading system that uses mastery grading identifies clear objectives or specifications, assesses for mastery, and focuses on eventual mastery of material (Campbell et al., 2020). Mastery grading is used in three assessment implementations discussed in this literature review.

Mastery-based testing (MBT): MBT is a "mastery grading style that focuses on tests or exams," typically in conjunction with traditional homework or other point-based assignment (Campbell et al., 2020, p. 839; Halperin, 2020).

Standards-based grading: Standards-based grading "refers to a broad category of mastery grading systems with the common theme that the student's final grade is based on demonstrating mastery of a list of objectives" where each objective can be re-tested until the student demonstrates mastery (Campbell et al., 2020, p. 839; Guskey et al., 2020).

Specifications grading: Specifications grading is like standards-based grading in that all grades are determined by the number of specifications successfully completed. The primary difference is the focus on clearly stating what a successful submission should look like, typically using rubrics. Assignments may be resubmitted until specifications have been met or mastered (Campbell et al., 2020; Katzman et al., 2021).

Significance of the Study

The district has implemented a policy resulting in many different implementations of the amount of time given between the original school-based test and its retake. This has ultimately resulted in a slight decrease in high school students' geometry end-of-year state exam scores. It is unclear if the effects of amount of time between the original school-based test and its retake influence high school students' geometry end-of-year state exam scores, or if there is one time frame that is better. This research will determine the amount of time between the original test and the retake that has the highest influence on high school students' geometry end-of-year state exam scores. Identifying if the recommended amount of time between the test and retake that maximizes student acquisition and demonstration of content knowledge is significant as it will be used to inform district leaders leading to the development of an updated policy improving student learning and mastery of content material.

Research Question(s) and Hypotheses

It is unclear if there are differences between high school students' geometry end-of-year state exam scores or between high school students' geometry math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter* at a Northeastern suburban public school district. The study will be guided by the following research questions and null and alternate hypotheses:

RQ1: In a Northeastern suburban public school district, what are the differences between high school students' geometry end-of-year state exam scores differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*?

$H_0 1$: There are no differences between high school students' geometry end-of-year state exam scores differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*.

$H_1 1$: There are differences between high school students' geometry end-of-year state exam scores differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*.

RQ2: In a Northeastern suburban public school district, what are the differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*?

$H_0 2$: There are no differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*.

$H_1 2$: There are differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*.

Review of the Literature

This section begins with a description of the theoretical framework based on the writing and research of Bloom (1968, 1974). Because of the different implementations discussed in the literature and the different timeframes used by the participating schools in the study, the review of the broader problem was separated into two major sections: results of mastery implementation and the impact of different timeframes for retaking assessments. The results of mastery implementation section include research on implementations factors and considerations, effect on student readiness for follow-on learning, and the testing effect. The impact of different timeframes for retaking assessments includes research on the *short-term* timeframe, the *anytime* timeframe, *end-of-year* high stakes testing, and comparing the amount of time between tests. The literature review ends with a critical summary.

The literature review encapsulates a comprehensive evaluation and assessment of relevant research from scholarly, peer-reviewed journals and dissertations that were identified through online searches of databases including ERIC, Education Research Complete, and Google Scholar. Relevant articles between the years 2017 and as possible 2022 were included. The Walden University Library provided access to a wide variety of information from peer-reviewed articles. The following keywords were used: mastery learning, standards-based grading, mastery grading, resits, reassessments, retakes, mastery assessments, timing, and retention.

Theoretical Foundation

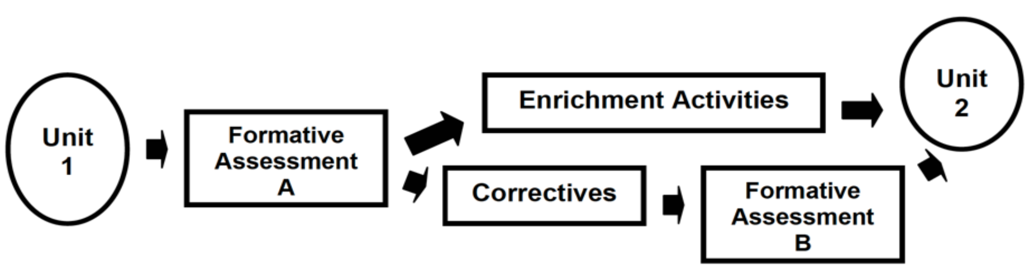
The theoretical foundation for this study was Bloom's theory of mastery learning which was informed by Carroll's (1963) work on school learning. Carroll posited that, if the quality of instruction were constant, a student's ability to learn new material was not dependent on student motivation but on the amount of time given (Carroll, 1963).

Building upon Carroll's work, Bloom's theory of mastery learning posited that all students could attain mastery of any content if provided sufficient time and academic support (Bloom, 1968). Additional academic support provides the opportunity to further diversify and differentiate instruction, increasing student acquisition of content knowledge. According to the theory, the amount of time necessary to acquire new material would vary from student to student based upon the student's aptitude for the curricula, the quality of instruction, and the student's ability to understand the instruction (Guskey, 2007). The theoretical framework differs from the practice of retakes being conducted in this study as additional academic support and the opportunity to take a retake offered to students are not provided prior to the beginning of new instruction, but in parallel with new instruction. Depending on the curriculum team execution at different school sites within the district, there are varying amounts of time given between the original test and the retake. Therefore, the school district is using a modified version of the theoretical framework described below.

Bloom (1968) believed that instructors should utilize a variety of instructional methods. The core of Bloom's mastery learning approach was assessment of small units of learning, individualized feedback, collaborative learning, and providing students with the time needed for them to learn (Guskey, 2007). Mastery learning is a core element of many instructional models and interventions, including mastery-based assessments, mastery-based grading, standards-based grading, and in Europe, resits (Campbell et al., 2020). Figure 1 below provides the key features of Bloom's mastery learning instructional practice, the use of frequent, targeted assessments to guide both learning and instruction (Guskey, 2005). Within Bloom's (1968) mastery learning theory, a unit of instruction, Unit 1, is provided followed by a formative assessment (A) to identify where students are having problems with the content. These relatively brief formative assessments provide both students and teachers with feedback about whether a particular instructional goal has been mastered. Students who do not meet the pre-determined mastery criteria follow the lower path and are given individually defined interventions followed by an opportunity to take a second formative assessment (B), or retake, to demonstrate their subsequent mastery of the material. If they fail to pass this assessment as well, they are given additional opportunities to study and retake the assessments until virtually all students achieve mastery. In parallel, students who have successfully mastered the material follow the upper path and are provided enrichment activities to further engage with the content. Once all students have mastered the material than the instructor begins a new unit of instruction, Unit 2.

Figure 1

Bloom's Mastery Learning Instructional Process



(Guskey, 2005)

Bloom (1968) made two predictions regarding children's learning in his discussion of mastery learning. His first prediction was individual differences in students' achievement would diminish over time, with 90% of the students in a class achieving a level at which only the top 10% of the students previously reached under traditional instructional practices. Well-documented research on the implementation of mastery learning in the classroom has demonstrated higher levels of achievement among students and increased student confidence in their ability to learn (Adesope et al., 2017; Çevik & Bakioğlu, 2022; Kulik et al., 1990). Bloom also predicted that weaker students who initially need longer time periods to learn will need less time as they master the fundamental material and become familiar with the mastery learning approach. The seminal study performed by Livingston and Gentile (1996) determined there was no demonstrated decrease in time necessary for learners to acquire mastery of material over successive units. Over time, faster learners consistently maintained the same relative advantage over slower learners.

The current study takes place in the context of a mastery learning grading policy implemented by a Northeastern suburban public school district. The study is focused on the amount of time [*short-term* (2-3 weeks), *anytime* (during the course), and *end of the quarter/year*], provided between the original test and the retest, however, none of these configurations follow the original path of Bloom's (1968) mastery theory. Within Bloom's theory, all members of a class would master content, with formative assessment B in Figure 1 as the retake, before moving on to new content. These just-in-time corrections prevent minor learning difficulties from accumulating and becoming major difficulties (Guskey, 2007). Instead, in this district enrichment activities are not performed, and all school-selected amounts of time to retake happen after moving on to new content, requiring students to complete correctives outside of class or during special study sessions. This implementation of Bloom's theory rarely experiences success as the students who could benefit most from the corrective process are the least likely to take part (Guskey, 2007).

Although the subject school district's policy does not fully align with all aspects of Bloom's mastery learning instructional process, the policy does specify that interventions will be defined and executed by the school curriculum teams with a second formative assessment to be provided to each student. In this study, a unit of instruction will occur, the original formative assessment will be given, and additional supports and correctives will be provided to students who have not demonstrated mastery of material. Although outlined in Bloom's (1968) theory above, the district does not provide specific time to allow the correctives and second formative assessment to be accomplished prior to the beginning of new content, resulting in these steps being executed in parallel to new content learning. This study evaluates the use of differing amounts of additional time between the original formative assessment and the second formative assessment, or retake, increasing the student's opportunity to access correctives and determining the difference in student mastery of content knowledge as defined as increases in end-of-year state testing scores and math growth scores.

Review of the Broader Problem

The review of the broader problem was separated into two major sections: results of mastery implementation and the impact of different timeframes for retaking assessments. The results of mastery implementation section includes research on implementation factors and considerations, effect on student readiness for follow-on learning, and the testing effect. The impact of different timeframes for retaking assessments includes research on the *short-term* timeframe, the *anytime* timeframe, end-of-year high stakes testing, and comparing the amount of time between tests.

Results of Mastery Implementation

Implementation Factors and Considerations. This section of the literature review includes research on implementation factors and considerations for mastery learning resulting from qualitative research. Several studies were performed on the implementation of mastery learning within a school environment and identified themes in the research were mostly convergent (Emery et al., 2018; Knight & Cooper, 2019; Olsen & Buchanan, 2019; Scheopner Torres et al., 2018; Wisch et al., 2018).

The following two articles identified a common theme that an increase in instructional workload adversely affected teachers' ability to implement retakes. Olsen and Buchanan (2019) performed a case study with 15 teachers, two principals, and one professional developer to study changes in grading practices after a one-year professional development program on rethinking grading that included retakes as part of the grading design. Two themes were identified, a lack of clarity regarding the purpose and practice of grading and all teachers modified their grading practices based on their personal perspectives as opposed to what they learned from the professional development. Teachers found that if given a chance to retake, students began arriving unprepared for assessments and retake the assessment until they received the score they wanted, raising concerns that scores were no longer a measure of knowledge attainment but of the practice effect. Teachers also noted that creating and giving multiple assessments was time-consuming, resulting in teachers abandoning the practices after a year. Wisch et al. (2018) evaluated survey results from 429 teachers to determine how redo's/retakes are implemented within the umbrella of standards-based grading and how teachers perceived these retake policies. School policies, teacher beliefs about student ownership and accountability, and teacher beliefs about the importance of deadlines were determined to effect implementation. Respondents noted that redo's/retakes negatively affected student achievement. Standards-based grading requires students to manage their time and put their best effort on the first assessment. These skills are already an issue for students who are struggling before implementation of standards-based grading, and further complicate student success after implementation. Research additionally noted issues with teacher workload and the amount of time required to implement the policy. There was a clear difference of opinion across the respondents on what defines best practice on redo's/retakes. Although both articles identified the common theme that an increase in instructional workload adversely affected teachers' ability to implement retakes, Olsen and Buchanan (2019) also identified themes of an increase in students arriving to their first test unprepared and an increased student focus on grades rather than learning.

The following three articles identified common themes of students arriving to their first test unprepared and an increase in student procrastination leading to student stress or increased course failures. Emery et al. (2018) executed semi-structured interviews with administrators (n = 3), teachers (n = 4), and middle school students (n = 9) to examine the impact of mastery learning on mastery goal structures. Concerns raised included expecting students to take the initiative in their education when they can't remember to bring a pencil to class, the time it takes to study and remediate to get to a mastery level, and the effects of student procrastination. Students felt increased pressure from the setting of mastery standards as they were constantly being compared to their peers. Additionally, the remediation policy for all student work prevented teachers from enforcing deadlines, raised concerns regarding students' self-regulation, and increased the number of students who gave up rather than facing failure multiple times. Time was discussed as a burden for students who are required to spend additional time in remediation, with the subsequent student perception that remediation was a penalty. Knight and Cooper (2019) used a phenomenological design to investigate teachers' perceptions (N = 7) of how standards-based grading effects teacher planning, instruction, assessment, classroom management, and student behavior. Standards-based grading was found to make instruction more purposeful, assessments clearer, and classroom environments that were more conducive to learning. Student accountability was shown to decrease, with students taking the first test to see it prior to taking a second test for grade, leading to a reassessment spiral where students were trying to learn two sets of content simultaneously. Additionally, student procrastination created situations where students were trying to recover large portions of the material at the end of the class. Scheopner Torres et al. (2018) used purposive, reputational case sampling to explore the implementation of competency education, a system that requires students to demonstrate mastery of specific competencies to earn credit and graduate, in secondary schools in three New Hampshire school districts three years after the reform implementation.

Teachers and administrators noted that students who took advantage of reassessment were often the high-performing students looking to increase their grade rather than students who had failed to demonstrate mastery. To attempt to prevent this, one district revised the reassessment policy so students could only earn up to a prespecified maximum grade. Parents noted high-achieving students were blowing off the original assessments because they knew they could reassess. Teachers noted the time requirement to create and implement recovery plans and reassessments were not logistically feasible for the instructional staff. Reassessments had to be limited to reduce the impact on teachers. To fully implement, students needed the metacognitive and self-regulation skills to take ownership of their education, something that had not been accomplished after 3 years. Findings determined that although progress has been made towards implementation, the reform effort has been met by many challenges that still need to be addressed. All three research teams identified common themes of students arriving to their first test unprepared and an increase in student procrastination leading to student stress or increased course failures. Both Emery et al. (2018) and Scheopner Torres et al. (2018) also identified a common theme that there was an adverse effect in instructional workload. In addition, Knight and Cooper (2019) found that standards-based grading made instruction more purposeful, assessments clearer, and classroom environments more conducive to learning

A divergent finding was identified by Buckmiller et al. (2017), who found that student ownership of learning increased over the course of the initial implementation of mastery learning. Buckmiller et al. performed a case study in a college-level Educational

Technology class (N =21) to examine student attitudes and experiences with a standards-based grading system. Data were collected via two rounds of focus groups and from student writing prompts given at three different points in the course. Themes identified were student apprehension to change, increased rigor and student critical thinking, student ownership of their own learning, and increased clarity on grades and their meaning, supporting the correlation that increased time and access to feedback improves student learning. Although Buckmiller et al. found that student ownership of learning increased over the course of the initial implementation of mastery learning, unlike the rest of the research identified above, it should be noted that the participants in this study were university-level students rather than high school students that are the subject of this investigation.

Synthesis. Themes identified in the qualitative research are mostly convergent. Researchers identified that students begin to arrive to their first test unprepared (Emery et al., 2018; Knight & Cooper, 2019; Olsen & Buchanan, 2019; Scheopner Torres et al., 2018), there was an increase in instructional workload adversely affected teachers' ability to implement retakes (Emery et al., 2018; Olsen & Buchanan, 2019; Scheopner Torres et al., 2018; Wisch et al., 2018), an increased focus on grades rather than on learning (Olsen & Buchanan, 2019), and an increase in student procrastination leading to student stress or increased course failures (Emery et al., 2018; Knight & Cooper, 2019; Scheopner Torres et al., 2018). In contrast, Buckmiller et al. (2017) found divergent positive findings that although university students were initially apprehensive, student ownership of their learning increased over the course of the initial implementation of mastery learning, and

Knight and Cooper (2019) found that standards-based grading made instruction more purposeful, assessments clearer, and classroom environments more conducive to learning.

Effect on Student Readiness for Follow-On Learning. This section of the literature review includes research on how mastery learning in a course affects student success in subsequent learning and is broken into three sections: retention of knowledge within a course, retention of knowledge in follow-on courses, and student perceptions on retention.

Retention of Knowledge Within a Course. In each of these studies, the results reflect how much students have retained content knowledge within a course or year. These studies demonstrated the effect on retention of content knowledge within a course implemented using mastery assessments is mixed when compared to courses taught without mastery assessments, and there is a clear change in the distribution of grades within a course that uses mastery assessments (Harsy & Hoofnagle, 2020; Jones, 2020; Katzman et al., 2021; Lenarz & Pelatt, 2020; Weir, 2020).

Two articles provided similar results supported by statistical analyses. Katzman et al. (2021) compared student pass rates and grade distributions in classes using traditional (n = 236) and standards-based grading (n = 252) teaching methods in a university level cell biology class. A chi-square test was performed to compare pass rates and grade distributions. Students were assessed on 20 individual quizzes, with each quiz having two retake opportunities. In addition, a traditional midterm and final were given. No significant difference was observed on pass rate of students; however, a significant difference with a medium effect size was found in the overall grade distribution. Among passing students, a higher percentage of students earned an A, but in the lower grades, a larger fraction of students withdrew from the course or earned a D. As the student data did not follow a normal distribution, a Kruskal-Wallis test was performed showing that student performance on content questions demonstrated standards-based grading sections performed significantly better but with a small effect size. Harsy and Hoofnagle (2020) compared 53 students in a mastery-based testing classroom versus 43 students in a traditional assessment in six Calculus II classes over a 2-year period. A final mastery assessment was used in all classes to assess retention of concepts. Data collected included results of the final mastery assessment and final course grades. Grades were grouped by letter grade for the statistical analysis of the final assessment and final grade in the course. There were no statistically significant differences on the number of concepts mastered or the score on the final mastery assessment. There was a statistically significant result for end-of-course grades with a higher percentage of students earning an A and a higher fail rate among the mastery-based testing group. Both Katzman et al. (2021) and Harsy and Hoofnagle (2020) demonstrated there was no significant difference in pass rates on the course final exam but identified a significant difference in final course grade distributions. Although Katzman et al. (2021) demonstrated that on content questions standards-based grading sections performed significantly better with a small effect size, this was based on content questions that were included in an end-of-course survey for students in traditional ($n = 29$) and standards-based grading ($n = 112$) courses.

Three other articles described similar results without providing statistical analyses to support their conclusions. Weir (2020) described an implementation of mastery grading in a university level Precalculus and Calculus I course that included testing objectives every 2 weeks with the opportunity to retest until students had mastered the content, which required showing correct solutions on two separate assessments. It was noted anecdotally that the pass rates demonstrated mixed results. Jones (2020) discussed the implementation of specifications grading in a post-graduate public administration classroom where assignments were bundled, with different levels of work accomplishment leading to different final grades in the course. Finally, Lenarz and Pelatt (2020) discussed the impact of transitioning to mastery-based testing on student persistence in university level Pre-Calculus, Calculus I, and Calculus II courses comparing student performance by analyzing course grades in the four years prior to switching to mastery-based testing with the two years after implementation. Within these studies, the authors noted an increase in the number of As and Bs, but also a clear increase in the number of Fs and drop-outs (Lenarz & Pelatt, 2020; Weir, 2020), or students choosing to accept lower grades in the course with a subsequent drop in the average course grade point average over time (Jones, 2020) and though anecdotal, pass rates demonstrated mixed results (Weir, 2020)

Synthesis. Recent literature on retention of knowledge within a course resulted in two primary findings. Specifically, the effect on retention of content knowledge within a course implemented using mastery assessments demonstrated no difference in pass rates or a mixed result when compared to courses taught without mastery assessments (Harsy & Hoofnagle, 2020; Katzman et al., 2021; Weir, 2020). In addition, there is a clear change in the distribution of grades within a course that uses mastery assessments, with a higher percentage of students earning an A but also a greater percentage of students who either withdraw from courses or earn a D or F (Harsy & Hoofnagle, 2020; Jones, 2020; Katzman et al., 2021; Lenarz & Pelatt, 2020; Weir, 2020). There was one study with a significant limitation. Although Katzman et al. (2021) demonstrated that standards-based grading sections performed significantly better on content questions with a small effect size, this was based on content questions that were included in an end-of-course survey for students and not on a final assessment in traditional (n = 29) and standards-based grading (n = 112) courses, with no analysis provided to justify a comparison of this limited data. Overall, there were two primary findings each with several articles supporting each finding as indicated above: (a) no difference in pass rates for courses with and without mastery assessments and (b) more students both earn A or F grades in mastery courses.

Retention of Knowledge in Follow-On Courses. In each of these studies, the results reflect how much students have retained content knowledge and transferred this knowledge into the next course or year. These studies either demonstrated mixed results or a reduction in content retention in follow-on courses (Carr et al., 2017; Childers & Lu, 2017; Lyle et al., 2020; Zimmerman, 2020).

In the following two articles, a reduction in content retention was demonstrated in follow-on courses. Carr et al. (2017) evaluated the retention of mathematics knowledge at the beginning of their fourth year for 42 students who took an on-line mathematics skill knowledge test during their third year at university. During their third year, students could retake the exam as many times as necessary until they passed, requiring an average of three to four times. Although no analysis of differences for students who took the test once versus multiple times was performed, a one-sided paired t-test noted a statistically significant reduction in test average results between third year (51.9%) and fourth year (44.3%) students, demonstrating students did not retain knowledge between the third and fourth year. Childers and Lu (2017) performed an ANOVA analysis using 753 students divided into 4 cohorts to determine if the implementation of mastery learning in a computer-based, college level developmental mathematics course resulted in changes in completion rates, length of completion time, success in subsequent college-level mathematics courses. The course was broken into 8-10 modules, each of which had to be mastered to a level of 80% before entry into the next module. Only 33.3% of students completed the program, and of those who completed the program, only 58.1% went on to pass a college-level mathematics course. There was no comparison to comparable data prior to the implementation of mastery learning. The program did not reduce the amount of time students spend in developmental math courses. Both Carr et al. (2017) and Childers and Lu demonstrated a reduction in content retention in follow-on courses, though Childers and Lu did not provide a comparison to data for students prior to the implementation of mastery learning.

In contrast to the above articles that found there was a reduction in content retention, the following two articles demonstrated *mixed results* on how well students retained knowledge into the follow-on course. Zimmerman's (2020) study on the retention of college algebra content knowledge into either trigonometry or calculus courses resulted in mixed results. Two classes of students executed concurrently by the same instructor with the same materials were compared, a traditional class with 140 students and a treatment class with 50 students. However, the two classes could not be compared as the standards-based grading group was pulled from STEM majors who had low scores on a math placement test and were in a program that provided additional supports whereas the traditional class used students with similar math placement scores as the standards-based grading group but were in a cohort of students who shared multiple classes. Although the author did not provide a direct comparison by evaluating the common final exam, significant differences were found in the students' follow-on courses, with trigonometry students mean grades improving demonstrating a higher retention of college algebra while the calculus students mean grades declining demonstrating a reduction in retention of college algebra. Lyle et al. (2020) studied how increasing the amount and spacing of retrieval practice effected mathematics knowledge retention on the final, and four weeks after the final at the beginning of a follow-on course. Amount and spacing of retrieval were manipulated using quizzes, with spacing determined by whether a learning objective appeared on a single quiz or was distributed across three quizzes, and practice determined by answering either three or six questions on a learning objective. A within-subjects design (N =111) used a single class split into four groups: control, increased spacing, increased practice, increased spacing and practice. Analysis was performed on the 62 subjects who completed all quizzes and the final exam, and of this a subset of 51 subjects who also took the exam in the follow-on course. The main effect of practice and practice with spacing was not significant, and though increased spacing was statistically significant with a moderate effect size, the mean difference was numerically small. Within-semester retention was higher under the condition of increased practice but did not have statistically significant effect on long term retention. The effect of spacing was similar on retention in both the within-semester and long-term conditions. The increased-both condition resulted in the highest level of within-semester performance. The effect of the amount of practice was less robust than the effect of spacing, but across semester retention benefited exclusively from increasing spacing. However, though increased spacing benefitted retention, it significantly reduced quiz performance within the course. Both Zimmerman (2020) and Lyle et al. demonstrated mixed results on how well students retained knowledge into the follow-on course, though Zimmerman's study was limited as it compared two different cohorts of students with different academic strengths, and Lyle et al.'s study split a class into four cohorts with no analysis presented on their equivalence.

Synthesis. Evidence provided by these researchers provide mixed results on how effective the use of retakes are as an instructional practice on the long-term retention of content knowledge as evidenced by the impact on subsequent learning. Carr et al.'s (2017) study evaluated participants with a common background and showed a significant reduction in retention after a 3-month timeframe, but Childers and Lu's (2017), which also demonstrated poor retention of content knowledge, did not provide a comparison to previous data prior to the implementation of the retake process. In contrast Zimmerman's (2020) research on students with mixed backgrounds provided mixed results, with students entering trigonometry classes demonstrating retention but students entering calculus classes demonstrating a lack of retention. It is unclear if student background had an impact on retention based on these results as the Zimmerman study was limited in that it compared two groups of students from very different backgrounds with no systematic accounting. Additionally, Lyle et al.'s (2020) research presented conflicting results, with increased practice demonstrating higher retention within a course but having no effect on long-term retention and increased spacing having a negative effect within course but a positive effect on long-term retention.

Student Perceptions on Retention. In each of these studies, the results reflect student perceptions on how effectively mastery assessments support retention of content knowledge. These studies demonstrated mixed results on the effectiveness of the use of mastery assessments (Erbes et al., 2021; Guskey et al., 2020; Harsy & Hoofnagle, 2020; McKell & Danowitz, 2020).

Two studies focused on university student perception on the effectiveness of the use of standards-based grading in high school settings towards preparing students for follow-on learning. To evaluate how college students felt about how traditional and standards-based grading at the high school affected student learning at the university level, Erbes et al. (2021) performed a mixed-method inquiry using survey data from 72 college sophomores. Of the original 75 respondents, two had participated in a school using standards-based grading and one had been home-schooled. The resultant evaluation used the 72 participants, all of whom had graduated from a public or private high school that used a traditional grading system. Students believed that traditional grading inadequately measures student learning, but they still prefer traditional letter grades over a standards-based grading model as they felt that they were better prepared for college given these institutions generally use a traditional grading system as well. However, 97% of the respondents noted they were more concerned with the letter grades than with their learning (Erbes et al., 2021). Guskey et al. (2020) also performed a mixed methods study to investigate if the implementation of standards-based grading at the high school level effected student performance at the university level. Thirteen students were surveyed to determine the validity of the standards-based grading implementation at their high school, and of these seven students participated in a one-on-one structured interview to investigate how standards-based grading had prepared them for transition to the university level. High schools were determined to not implement standards-based grading with fidelity, however the degree of implementation did not affect student performance at the university level. Although the research determined that standards-based grading did not have a negative influence on performance at the university level, in the response to open ended items on the survey, three of seven participants noted that the access to retakes assessments at the high school level did not prepare them properly and was a potential failure of standards-based grading. The results of both studies were consistent, student perceptions were that standards-based grading was a better measure of student learning, but since university courses primarily used traditional grading, the use of standards-based grading inadequately prepares students for follow-on learning (Erbes et al., 2021; Guskey et al., 2020).

Two additional studies focused on university student perception on the effectiveness of the use of standards-based grading in university settings. McKell and Danowitz (2020) explored the implementation of standards-based grading within a university level electrical engineering course. Data showed student learning of course topics was comparable between the control (n = 34) and experimental (n = 33) as measured on a pretest/posttest measure. Students were assessed on two midterms and a cumulative final in the traditional course, and in the standards-based grading course quizzes were taken on specific objectives, which could be retaken any time, but a standard could only be retaken three times. The normalized gain in the traditional section was 0.44 and the normalized gain in the standards-based grading section was 0.35. Student opinion of standards-based grading was mixed on end-of-course surveys, where students judged standards-based grading to be less educationally effective than traditional methods. standards-based grading was seen as a burden by some students who would rather have set deadlines and structure. Students in the standards-based grading section evidenced poor time management, with many students attempting quizzes well after the topics were covered in lecture (McKell & Danowitz, 2020). It was noted that the workload on the instructor was daunting, managing the individual standards, writing multiple versions of quizzes, grading, and review of student's corrective work (McKell & Danowitz, 2020). Harsy and Hoofnagle (2020) compared 53 students in a mastery-based testing classroom versus 43 students in a traditional assessment in six Calculus II classes over a 2-year period. Data collected on an end-of-course survey noted a statistically significant result of whether in-class assessments deepened understanding of course concepts with mastery-based testing students agreeing more strongly than traditional assessment students. Student comments included issues with procrastination and mastery-based testing allowing students to study just enough to master a concept without thoroughly learning it by having to work and demonstrate the concept on multiple occasions and in multiple ways or in different implementations. These two studies demonstrated mixed results with one study stating that standards-based grading was less educationally effective and the other noting that the use of standards-based grading deepened student understanding of course content (Harsy & Hoofnagle, 2020; McKell & Danowitz, 2020)

Synthesis. Students' perceptions on the effectiveness of mastery assessments on course retention was mixed. Erbes et al. (2021) evaluated how students from schools who used traditional grading schemes impacted student learning, whereas Guskey et al. (2020) interviewed students who attended high schools who used standards-based grading to understand their perspectives after their first semester at university. Students from traditionally graded courses believed using these methods better prepared them for subsequent learning at the university level which generally implements traditional grading methods, whereas students from schools using standards-based grading stated that access to retake assessments did not properly prepare them for subsequent learning at the university level. Students evaluating the use of mastery assessments within a university level course also provided mixed results with standards-based grading being seen as both deepening the understanding of course content (Harsy & Hoofnagle, 2020) and less educationally effective (McKell & Danowitz, 2020). Of special note, students were concerned with the burden and time management issues associated with mastery assessments (Harsy & Hoofnagle, 2020; McKell & Danowitz, 2020) and instructional staff were concerned with the increase in workload to implement (McKell & Danowitz, 2020).

The Testing Effect. The testing effect refers to the retention benefits resulting from using assessments as a learning activity. Trumbo et al. (2021) provided a review of current research on the factors affecting the testing effect, to determine the degree to which test feedback, test format, the number of tests, level of test questions, timing of tests, and retention duration transitions from laboratory settings to the classroom. A notable gap in testing effect research exists, as most of the laboratory studies have been performed with undergraduate students. Evidence of the testing effect has been found following even a single test (Bangert-Drowns et al., 1991). If students are given the opportunity to restudy material and are provided multiple testing attempts, short-answer and multiple-choice formats generally produce equivalent learning outcomes, however the magnitude of the benefit decreases as the complexity of the curriculum increases (Adesope et al., 2017; Foss & Pirozzolo, 2017; Rawson et al., 2020). The timing of quizzes, between study sessions, has been shown to effect retention as well as facilitate learning of new material (Chan et al., 2018; Lyle et al., 2020). The testing effect appears to benefit learning of underlying concepts, though variables such as test format and question type influence how beneficial the effect is (Adesope et al., 2017; Trumbo et al., 2021). The test benefit may persist as long as 9 months but is stronger with the implementation of refresher quizzes. However, there are few studies with long retention intervals (Trumbo et al., 2021). The following section investigates three aspects of the testing effect: test format, distributed practice, and test-enhanced transfer.

Test Format. This section reviews research on how the choice of test format impacts the results of mastery assessments. The research provided converging evidence. that test format impacts the testing effect (Adesope et al., 2017; Bangert-Drowns et al., 1991; Choi & Lee, 2020; Foss & Pirozzolo, 2017; Morphew et al., 2020).

A meta-analysis performed to extend the seminal work by (Bangert-Drowns et al., 1991) confirmed the effectiveness of learning with practice tests versus restudy. Adesope et al. (2017) performed a meta-analysis of 118 articles, 272 independent studies, and 15,427 participants investigating the effects of practice tests versus non-testing learning conditions, the features of practice tests that foster learning effects on the final test, and the impact of feedback between the original and final test. Of the studies (k = 216), the majority had participants read passages, word lists, or word pairs. Overall weighted mean effect size was moderately large (g = 0.61) and statistically significant supporting the effectiveness of learning with practice tests. Most studies compared practice tests with restudying (k = 195) resulting in a moderately statistically significant mean effect size (g = 0.51). Researchers mainly used recall practice tests (k = 134) within their studies, but of those that did include other forms of tests, multiple-choice practice tests (g = 0.70) were associated with a much higher weighted mean effect size than short-answer tests (g = 0.48). The testing effect was stronger if the practice and final test formats were identical (g = 0.63) then when dissimilar (g = 0.53). The effects of practice tests were larger when used only once (g = 0.70) then when used two or more times (g = 0.51). Testing effects were robust, whether feedback on practice tests was given or not, with weighted means not differing significantly. Most studies (k = 117) examined retention intervals of 1 day, however the studies using a lag of time of 1 to 6 days demonstrated a stronger effect size (g = 0.82) then those with a lag time of 7 to 42 days (g = 0.69). This meta-analysis confirmed the effectiveness of learning with practice tests versus restudy, but the effect size was larger for multiple-choice versus short-answer test items, was stronger if the tests were identical, and demonstrated a reduced effect size with longer lag times between tests.

The following two articles demonstrated converging evidence that test format impacts the testing effect. Foss and Pirozzolo (2017) performed a study (N = 588) using repeated measures ANCOVA to evaluate the effect of test frequency and repeating versus varying test items on the testing effect. Students were given either two midterms or four or eight smaller midterms exams, with the total test items being the same. The effect of increasing testing frequency from two to either four or eight exams was not significant, however there was a significant effect of student aptitude. A statistically significant improvement was demonstrated for test items that were repeated (M = .72) vice new items (M = .63). Students were found to perform better on fact-based test items (least squares M = .56) than application items (least squares M = .52). Choi and Lee (2020) investigated whether the forward testing effect was due to students' expectation about the test format or the actual test experience itself, manipulating both the timing and how participants were informed on the test format. Three experiments were performed that determined: 1) if informing students (n = 61) about the test format via either a pretest (m = 48.95, SD = 15.38) or an interim test (m = 81.05, SD = 14.03) enhanced student learning, 2) if the effect was seen (n = 40) when paired with an actual test-taking opportunity (m = 60.50, SD = 17.44) versus restudy (m = 41.80, SD = 19.16), and 3) if retesting enhanced subsequent learning of new material when students were not provided feedback on the original test (n = 76). The results were not significant, suggesting the benefits of interim testing are independent of providing feedback. Although neither increasing test frequency nor providing feedback were statistically significant, repeating test items was statistically significant with higher performance on fact-based items vice application items.

Synthesis. The research provided converging evidence that test format positively impacts the testing effect, with higher performance on recall and multiple-choice formats vice short answer or application questions (Adesope et al., 2017; Choi & Lee, 2020; Foss & Pirozzolo, 2017; Morphew et al., 2020). The effects of the practice test were larger when performed once vice more than once and when the test items were identical (Adesope et al., 2017; Bangert-Drowns et al., 1991; Foss & Pirozzolo, 2017), and neither increasing test frequency nor providing feedback changed the effect of test format (Choi & Lee, 2020; Foss & Pirozzolo, 2017).

Distributed Practice. The following research evaluates the impact of distributed practice versus testing. The effect of distributed practice on the test effect provided mixed results (Barzagar Nazari & Ebersbach, 2019; Ebersbach & Barzagar Nazari, 2020; Guo, 2021; Latimier et al., 2021; Racsmány et al., 2020; Rawson et al., 2020).

A meta-analysis was performed to evaluate the effectiveness of distributed practice. Latimier et al. (2021) performed a meta-analytic review of 29 articles evaluating the benefits of spaced and massed retrieval practice (n = 39) and expanding and a uniform spacing schedule (n = 54) on content retention. Most of the studies were in laboratory settings and included a population derived from undergraduate students. Results determined a strong benefit of spaced retrieval practice in comparison with massed retrieval practice (*g* = 0.74) but no significant difference between expanding and uniform spacing schedules of retrieval practice (*g* = .034). Additionally, the more students are tested, the more beneficial the expanding schedule is to the uniform one. The results of this meta-analysis support the advantage of spacing out the retrieval practice episodes on the same content, but do not support that inter-retrieval internals should be progressively increased until a retention test.

Three articles were performed by the same research team to investigate distributed learning at the university level. Prior work across three experiments and 533 students demonstrated positive results with verbal learning skills when practice and test questions were identical, resulting in a recommendation to practice to three correct recalls over three widely spaced intervals (Rawson & Dunlosky, 2011). Rawson et al. (2020) performed three experiments with a pre-test/post-test design to determine if single session learning or mastery learning combined with distributed practice affected retention as defined as scores on the final exam in university level statistics course. In all three experiments, the mastery learning combined with distributed practice group showed a statistically significant but weak effect over the single session group on identification questions and performed modestly better at problem solving. Although mastery learning with distributed practice has been shown as an effective technique in language acquisition

(Rawson et al., 2018), it provided minimal benefit for learning and executing mathematical procedures (Rawson et al., 2020).

Two articles were performed by the same research team to investigate distributed practice on procedural knowledge. Barzagar Nazari and Ebersbach (2019) examined the effect of distributed mathematics practice on procedural knowledge in seventh graders (N = 118) on two follow-up tests, using practice sheets, 1- and 6-weeks after the practice periods were completed. Both tests included a transfer task to determine if the effect of distributed practice could be generalized to a problem that had not been previously solved. Results were so poor on the transfer tasks that the data was not analyzed. Bayesian linear regression identified strong evidence of a positive effect of distributed practice of approximately one point within a 0- to 12-point score range. However, the practice condition did not influence the performance development between the first and second tests. Ebersbach and Barzagar Nazari (2020) followed this investigation by focusing on how the retention of mathematical procedures were affected by distributed practice and lag effect, or how increasing inter-study intervals increased retention. University students (N = 235) solved permutation tasks with study intervals of 0-, 1-, or 11-days, and tested after 1- or 5-weeks. A 3 by 2 between-subjects design evaluated with a repeated measures ANOVA with the practice set as the between-subjects variable revealed no effect of distributed practice, and a significant effect on test performance, lower at five weeks ($M = 1.63$, $SD = 1.82$) than at one week ($M = 2.53$, $SD = 2.01$), independent of lag time. This research provided mixed results. Although there was strong evidence of a strong positive effect of distributed practice with 7th graders, at the undergraduate level, there was no effect from distributed practice.

Two additional articles demonstrated mixed results on the effectiveness of distributed practice. Guo (2021) studied the effect of the initial test interval (short versus long) and feedback timing (immediate and delayed) on vocabulary retention over 2-weeks on undergraduate students (N = 60) in a lower-level ESL course. Intervals were chosen to accommodate the real-world classroom setting with immediate feedback provided right after the test and delayed feedback provided the next day. Test items included definitions, cued recall, and short answer questions. The study used one initial learning session, three test sessions (Days 2 or 4, Day 9, and Day 14), and one new learning session before the intermediate test. A 2 by 2 repeated measures ANOVA determined that the optimal schedule to promote retention was a short-interval initial test followed by delayed feedback. A two-way ANOVA with the initial test interval and feedback timing as the between-subject factors demonstrated the delayed feedback condition produced better final test performance than groups in the intermediate feedback condition, regardless of study-test lag. Racsmány et al. (2020) performed three experiments with university level students (N =84) investigating whether the effect of feedback on the testing effect is modified by the level of retrieval success during practice. The investigation on the level of success at retrieval practice was manipulated with multiple pre-practice learning trials of 36 Swahili-Hungarian word pairs, and multiple tests with feedback applied after a 1-week retention period. A repeated-measures ANOVA determined the level of retrieval success was the key factor in reversing the testing effect.

After the first round of final recall, previously tested items were recalled better than restudied items. However, after one re-learning session with feedback, the testing effect was reversed. These two articles demonstrated mixed results, suggesting the optimal schedule to promote retention is a short-interval initial test followed by delayed feedback and repeated retrievals, not re-testing, may insulate memory strength (Guo, 2021) and determining that the key factor in reversing the testing effect was one relearning session with feedback (Racsmány et al., 2020).

Synthesis. The effect of distributed practice on the test effect provided mixed results. There was a consistent advantage of spacing out retrieval practice demonstrated in the literature (Barzagar Nazari & Ebersbach, 2019; Guo, 2021; Latimier et al., 2021; Racsmány et al., 2020; Rawson et al., 2020). However the testing effect was demonstrated to be reduced when feedback was introduced after the first assessment (Racsmány et al., 2020) and provided a minimum effect on higher order test items such as in mathematics (Ebersbach & Barzagar Nazari, 2020; Rawson et al., 2020). Finally divergent evidence was found that it is repeated retrievals and not repeated testing that supports increased memory (Guo, 2021).

Test-Enhanced Transfer. The following research focuses on if the act of testing itself can help students apply or transfer their knowledge to different contexts. The research is convergent, the effect of test-enhanced transfer is stronger with memorization type questions but is less effective with questions that require higher order thinking like mathematics (Adesope et al., 2017; Barzagar Nazari & Ebersbach, 2019; Foss & Pirozzolo, 2017; Pan & Rickard, 2018; Peterson & Wissman, 2018).

In the following meta-analysis, it was demonstrated that test-induced transfer was not statistically significant or significant with a weak effect size. Pan and Rickard (2018) performed a meta-analytic review of 122 experiments and 67 articles to investigate if testing yielded test-enhanced learning that transfers to different contexts ($d = 0.40$, $p < .00001$). Testing can yield transferrable learning relative to a non-exposure control group, however this transfer is greatest across test formats, application and inference questions, and related word cues. The transfer was statistically not significant for stimulus-response items ($d = 0.22$, $p = 0.66$), and previously untested materials seen during initial study ($d = 0.16$, $p = 0.20$). A weak transfer effect was found for problems involving worked examples ($d = 0.29$, $p = 0.10$). Transfer is higher when the initial and final assessments are the same, and less when they are dissimilar and yields different degrees of transfer across different problem types. To improve transfer, providing correct answer feedback emerged as having a significant negative effect, whereas feedback that includes some postretrieval restudy or extended or explanatory feedback positively affected transfer. For assessments typically used within the mathematics curriculum, this meta-analysis demonstrated that, test-induced transfer was not statistically significant for stimulus-response items or previously untested materials, or significant with a weak effect size for problems involving worked examples (Pan & Rickard, 2018).

One new article and several articles already reviewed included converging evidence on test-enhanced transfer. Peterson and Wissman (2018) performed a study to evaluate the testing effect on learning transfer for problem-solving test items. A 2 (restudy versus retrieval) x 2 (2-min delay versus 48-hr) design ($N = 280$) was used.

ANOVA analysis yielded no numerical retrieval practice advantage, and the effect was not impacted by delay length. The results of the study revealed there is no transfer effect from retrieval practice on analogical problem-solving test items. Barzagar Nazari and Ebersbach (2019) examined the effect of distributed mathematics practice on procedural knowledge in seventh graders ($N = 118$) on two follow-up tests that included a transfer task to determine if the effect of distributed practice could be generalized to a problem that had not been previously solved. Results were so poor on the transfer tasks that the data was not analyzed. Foss and Pirozzolo (2017) investigated the testing effect by evaluating the effect of frequency of tests, and repeating and varying test items, on student performance on the final exam score. Results do not support the transfer effect as students were found to perform better on fact-based test items (least squares $M = .56$) than application items (least squares $M = .52$). In Adesope et al.'s (2017) meta-analysis, the transfer effect is not supported as multiple-choice practice tests ($g = 0.70$) were associated with a much higher weighted mean effect size than short-answer tests ($g = 0.48$) and the testing effect was stronger if the practice and final test formats were identical ($g = 0.63$) rather than when dissimilar ($g = 0.53$). All four articles demonstrated converging results, that there is minimal evidence for the transfer effect, but when present, it is stronger for identical and multiple-choice items and much reduced for short-answer test items or analogical problem-solving test items (Adesope et al., 2017; Barzagar Nazari & Ebersbach, 2019; Foss & Pirozzolo, 2017; Peterson & Wissman, 2018).

Synthesis. There is converging evidence that the transfer effect is stronger with memorization type questions but is less effective with questions that require higher order thinking like mathematics. The transfer effect was strongest in memorization type questions or when the practice and final test formats are identical (Adesope et al., 2017; Pan & Rickard, 2018). There was not a statistically significant transfer for stimulus-response items (Pan & Rickard, 2018), and there was no transfer effect for problem-solving test items (Barzagar Nazari & Ebersbach, 2019; Peterson & Wissman, 2018). There was a weak effect for problems involving worked solutions (Foss & Pirozzolo, 2017; Pan & Rickard, 2018).

The Impact of Different Timeframes for Retaking Assessments.

The following sections review the research on the impact of different timeframes for retaking assessments organized by research on the *short-term* timeframe, the *anytime* timeframe, end-of-year high stakes testing, and comparing the amount of time between tests. The practices evaluated in the literature are those that are predominantly in use by teachers at the study site.

Short-Term Timeframe. The *short-term* time frame allows retakes during a 2-3 week timeframe after the initial test administration. Short-term implementation of mastery learning showed mixed results with a majority demonstrating improvements in exam grades and a smaller number demonstrating no difference in exam grades (Adeniji et al., 2018; Blackstone & Oldmixon, 2019; Early et al., 2016; Ee et al., 2018; Fernandez, 2021; Halperin, 2020; Heubach & Krinsky, 2020; Katzman et al., 2021; Morphew et al., 2020).

The following four articles demonstrated an improvement in exam results. Morphew et al. (2020) performed a quasi-experimental study (N = 480) to investigate the effects of transitioning from two midterm exams to seven shorter exams that incorporated mastery testing by comparing identical final exam results over two consecutive semesters in a solid mechanics course. Students in the treatment group performed seven percentage points higher on the final exam with a medium effect size and performed better on both the multiple-choice and free-response questions. Halperin (2020) described a mastery-based testing implementation in a calculus class with a final exam to evaluated mastery and retention of material. A linear correlation between cumulative objective grades and the final exam score was performed resulting in a moderate correlation ($r^2 = 0.76$) in Calculus I and a strong correlation ($r^2 = 0.82$) in Calculus II. Implementation resulted in increased preparation than a typical course, additional time necessary to create multiple cumulative assessments, a significant increase in grading time, and additional time in office hours supporting students. Adeniji et al. (2018) performed a 2-week, quasi-experimental, pre-test, post-test control group design to determine the effect of mastery learning on high school students' retention of geometry content knowledge (N = 172). Analysis determined there was a statistically significant increase in post-test scores for students in the experimental group, however, retention data collected on the experimental group demonstrated a significant decrease in test scores after two weeks. The analysis did not provide a comparison of retention scores between the control and experimental groups. The author noted a limitation of mastery learning as it requires more instructional time to implement than traditional methods. Ee et al. (2018) investigated the impact of including a 2-week time constraint with an online mastery learning approach on the final exam results in a post-graduate finance course taught with personalized system of instruction ($n = 265$) and mastery learning ($n = 395$). A Wilcoxon two-sample t-test on the final examination scores found that both the mean and median scores for mastery learning ($m = 44.10$) were higher than with no time constraint ($m = 34.96$). There was a statistically significant association between the length of time taken to complete a quiz and the final examination score, with scores decreasing with an increase in the total time taken to complete a quiz. All four articles provided convergent evidence that demonstrated an improvement in exam results using the two-week timeframe between the original assessment and its retake (Adeniji et al., 2018; Ee et al., 2018; Halperin, 2020; Morphew et al., 2020), however, it was noted that two weeks after the final assessment, retention data demonstrated a significant decrease (Adeniji et al., 2018).

Two articles demonstrated convergent evidence that there is no difference in exam results. Katzman et al. (2021) used a chi-square analysis to evaluate the implementation of standards-based grading in a university level cell biology class, comparing student pass rates and grade distributions using traditional ($n = 236$) versus standards-based grading ($n = 252$) teaching methods. Students were assessed on 20 individual quizzes, with each quiz having two retake opportunities. In addition, a traditional midterm and final were given. No significant difference was observed on pass rate of students; however, a significant difference with a medium effect size was found in the overall grade distribution. Among passing students, a higher percentage of students earned an A, but a larger fraction of students withdrew from the course or earned a D. As the student data did not follow a normal distribution, a Kruskal-Wallis test was performed showing that students in standards-based grading sections performed significantly better on content questions with a small effect size. Early et al. (2016) evaluated an instructional improvement program called Every Classroom, Every Day (ECED) which included a mastery assessment component, by performing a two-year long school-randomized trial of 20 high schools split evenly into treatment and control groups to determine if there were increases in ninth and 10^{th} grade math and English scores on standardized tests, and to determine if improved fidelity of implementation resulted in higher scores. Math scores were not statistically significant in the first year but were statistically significant in the second year with a small effect size. English scores showed no statistically significant improvement. Fidelity of implementation was shown to have a positive effect on math scores. There are issues with validity as different tests were used at different schools, therefore the researchers were making comparisons on different baselines within the analysis. The results of both studies demonstrated no difference in exam results between traditional and mastery assessments (Early et al., 2016; Katzman et al., 2021), though Katzman et al. (2021) found a significant difference in grade distribution, and Early et al. (2016) had issues with validity of result.

The following three articles discussed the implementation of mastery assessments but included minimal analyses on their effectiveness. Heubach and Krinsky (2020) presented a large-scale implementation of mastery-based grading in a college level statistics class across 37 instructors and 96 sections. Grades are determined by a decaying average. A simple comparison of course pass-rates pre- and post- implementation showed an improvement from 68.2% to 78%. The biggest challenge identified with the course design was excessive grading workload. Fernandez (2021) discussed the design of a standards-based grading system that allowed students to retake material at the 4-week, 8-week, and 12-week point within a semester long university mathematics course. Data from 139 end-of-course evaluations over 10 courses was provided as evidence of success. The author stated that all predictions from research had been evidenced in the classroom, although no evidence was provided to support this statement. The author acknowledged the downside of offering unlimited or on-demand re-takes as they are time consuming for the instructor and they de-incentivize students to put in their best effort on the initial assessment. Finally, Blackstone and Oldmixon (2019) explored the implementation of specifications grading in an undergraduate political science classroom. A token system was used to control the number of retakes used in the classroom. Noted that specifications grading did not reduce the total time spent on grading. Although average scores were higher on multiple choice portion of the first two exams, there were no significant differences between the means of classes with and without specifications grading on the third unit exam nor was there a significant difference in average course grades or the distribution of grades. Although there were minimal analyses provided on the effectiveness of the mastery assessment implementations, the evidence on their effectiveness was mixed. Heubach and Krinsky (2020) and Fernandez (2021) both stated they saw test score improvements in the classroom while (Blackstone & Oldmixon, 2019) found no significant difference in average course grades.

Synthesis. *Short-term* implementation of mastery learning showed mixed results with a majority demonstrating improvements in exam grades (Adeniji et al., 2018; Ee et al., 2018; Halperin, 2020; Heubach & Krinsky, 2020; Morphew et al., 2020), a smaller number demonstrating no difference in exam grades (Blackstone & Oldmixon, 2019; Early et al., 2016; Katzman et al., 2021) or an overall reduction in retention on follow-on assessments (Adeniji et al., 2018). There were several common themes across the literature. Multiple articles noted the increase in instructional workload associated with creating and grading assessments (Adeniji et al., 2018; Blackstone & Oldmixon, 2019; Fernandez, 2021; Halperin, 2020; Heubach & Krinsky, 2020) and an increase in student procrastination (Ee et al., 2018; Fernandez, 2021). In addition, there was mention of an increase in course drop-outs (Katzman et al., 2021) or schools dropping out of a study and no longer implementing mastery learning (Early et al., 2016).

Anytime Timeframe. The *anytime* time frame allows retakes at any point after the initial test administration but prior to the end of the course. Anytime implementation of mastery learning demonstrated mixed results (Barth & Higginbottom, 2021; Birhan, 2018; Carlisle, 2020; Childers & Lu, 2017; Collins et al., 2019; Cundiff et al., 2020; Elsinger & Lewis, 2020; Harsy & Hoofnagle, 2020; Jones, 2020; Mangum, 2020; McKell & Danowitz, 2020; Nevid & Gordon, 2018; Stange, 2018; Weir, 2020; Williams, 2018).

Two articles demonstrated an improvement in grades. Nevid and Gordon (2018) evaluated the effect on final exam scores (N=123) from the use of an integrated learning system (ILS) to implement online homework and quizzes to a 100% mastery level with unlimited retakes in comparison to a control class that had access to the ILS but were not required to use it. To balance the workload, the non-ILS students were provided with additional homework on content not assessed on course exams. A mixed factorial ANOVA demonstrated a statistically significant effect for the ILS class with a medium size effect, demonstrating a 6% improvement on exams with a medium effect. Students in the non-ILY class failed to use the ILS resource consistently with a near zero usage by the end of the course, suggesting the requirement of a grade to incentive student effort. Birhan (2018) performed an 8-week, time series quasi-experimental research design to study the effects of mastery learning on first-year university engineering students' (N = 107) academic writing skills. Pre- and post-test data was collected and analyzed using repeated measure ANOVA, independent t-tests, and descriptive statistics. Although the experimental group had a slightly higher mean score on the pre-test (36.35 versus 32.94), the post-test score demonstrated a greater improvement by the experimental group (58.17 versus 45.05). The t-test demonstrated a statistically significantly result, that students who engage in continuous assessment as well as peer and teacher feedback improved their writing skills. Both articles together provided convergent evidence that an improvement in grades though it was also shown that with continuous assessment, writing skills improved, but when students were provided the opportunity to retake to improve their understanding without a grade benefit, students chose not to put in the effort (Birhan, 2018; Nevid & Gordon, 2018).

Two articles demonstrated improvements in grades but had issues with validity. Barth and Higginbottom (2021) discussed the implementation of a mastery exam using an on-line exam generator with a pre- and post-test gain comparison of two sections of

Calculus I students, one that included the mastery exam (n = 19) and a control section (n = 20). A statistically significant mean gain of 11.6 was demonstrated by the mastery exam section, but there was not a significantly significant difference demonstrated on the final exam. It was noted that there is a large time commitment required by the instructor to implement mastery testing, estimated at 10 hours for 20 students, times three exams over the normal planning requirements, as well as increased opportunities for students to present the work of others as their own on the mastery assessments. The number of students in each section was too small to support generalizing results, and there was no analysis to determine if there were any statistical differences between the two groups to ensure that the significant increase was due to student ability versus the mastery exams. Cundiff et al. (2020) performed a mixed-methods study of the mastery learning implementation in a set of social science courses, with an experimental group (n = 282) and control group (n = 165) from a previous section of the course. Students in the experimental group were given a pre-test and final exam which demonstrated all students learned content material over the course. In relation to the student's cumulative GPA, over half of the students received a score equal to or higher in the course, but no statistical analysis was performed to identify effect size. An ordinary least squares regression was calculated using a subsample of the experimental group which showed a statistically significant improvement of 5 percent on selected exam questions over the control group, however, no definition of how the subsample was selected was provided. Acknowledged teacher time management challenges associated with scheduling mastery quiz retakes and assessment development. There was no assessment of long-term retention. These two articles provided convergent evidence that improvements in grades but had issues with validity of results, either by not including enough subjects to be able to generalize results, providing a comparison of cohorts to determine if they are statistically similar, or by not including a definition on how subjects were selected for including in analyses (Barth & Higginbottom, 2021; Cundiff et al., 2020).

In contrast, the following two articles were convergent in that they either found no measurable difference in grades or mixed results. As previously reviewed, Harsy and Hoofnagle (2020) demonstrated there was no statistically significant difference on the number of concepts mastered or the score on the final mastery assessment in a Calculus II course. There was a statistically significant result for end-of-course grades ($m = 2.918$ vs $m = 2.537$) with a higher percentage of students earning an A with mastery-based testing. Carlisle (2020) described a mastery-grading design for university-level discrete mathematics and discrete and combinatorial algebra courses. Students were provided three opportunities to show mastery on topics, at a quiz, test, or final exam, with the instructor providing an opportunity for minor re-writes to pass a topic with a token system to allow for late work, opportunities to re-write homework or projects, or to access one-on-one oral quizzes to demonstrate mastery of a topic. Analysis was simple average of scores on each of five classes, two using a traditional point system and three using the new grading system. Although there appears to be improvement in one course, the other courses had mixed results with a concerning increase in course failures. Concerns raised included grade inflation, less flexibility within the course to delve into other advanced topics, and not reducing instructor workload as predicted.

The following three articles discussed the implementation of mastery-based testing or mastery grading but included minimal analyses on their effectiveness. Collins et al. (2019) described a mastery-based testing implementation in undergraduate mathematics courses. Multiple attempts were provided to demonstrate mastery of concepts, during either a quiz, each of four tests, or the final. Student end-of-course survey results were provided but no statistical analysis was performed. Mangum (2020) discussed the implementation of mastery-based testing where concepts could be retaken either during retake opportunities given once between exams or on any subsequent exam, with all students taking a common final exam. Due to workload associated with retakes and students not properly preparing for retakes, student access to retakes was limited to once between tests which did not appear to affect the overall grade average of students taking the course. Results of student end-of-course data and a comparison of the average of course grades were provided as evidence. A comparison could not be made on the common final average as the author noted that courses were provided by different instructors, with different emphasizes placed during the course and with different textbooks. In addition, a comparison of student responses on end-of-course surveys could not be compared as the survey instruments were different. Finally, Weir (2020) described an implementation of a mastery grading in a university level Precalculus and Calculus I course. Components were divided into objectives which were tested every 2 weeks and could be retested until students had mastered the content, defined as correct solutions on two separate assessments. Anecdotal data was mixed with a higher-than-average pass rate one semester, and an average pass rate the following semester. Of concern, a higher number of students failed the course under this system. These three articles discussed the implementation of mastery-based testing or mastery grading but included minimal analyses on their effectiveness.

Three additional articles discussed the implementation of standards-based grading but included minimal analyses on their effectiveness. Elsinger and Lewis (2020) described a standards-based framework implemented from 2015-2019 in lower-level university mathematics courses, including the structure of reassessments and the use of oral reassessments. A modification made to traditional standards-based grading was to require core standards, content determined to be required knowledge for follow-on courses, be demonstrated twice for credit. Aligning content into standards left little room for discussion of or testing synthesis of topics. There was no evidence comparing student attainment of knowledge under this process versus the traditional weighted grading process. Williams (2018) presented a methodology to integrate standards-based grading into a college level proofs course including 100 learning objectives and 22 proof-learning objectives. Learning objectives can be retaken anytime during the course, but proof-learning objectives are retaken during the original midterm or final exam. Williams acknowledged that the proposed methodology would create a feedback workload and tracking challenge when scaled to a typical student load of 160 students at the secondary level. Stange (2018) discussed the implementation of standards-based grading in a university level abstract mathematics course. Quizzes were provided at the end of every class and students could choose and take the quizzes based on individual mastery levels. Scores could be raised but never lowered, and a traditional midterm and final were provided. Although the averages on the quizzes ranged from 81 – 85%, the average grade on the final exam was 58%. These three articles discussed the implementation of standards-based grading but included minimal analyses on their effectiveness.

Synthesis. *Anytime* implementation of mastery learning revealed mixed results with some researchers demonstrating improvements in grades (Barth & Higginbottom, 2021; Birhan, 2018; Cundiff et al., 2020; Nevid & Gordon, 2018) but more demonstrating mixed results for grades (Carlisle, 2020; Childers & Lu, 2017; Jones, 2020; Mangum, 2020; McKell & Danowitz, 2020; Weir, 2020). Several articles noted either an overall reduction or minimal change to the level of retention demonstrated on end of course assessments (Barth & Higginbottom, 2021; Childers & Lu, 2017; Harsy & Hoofnagle, 2020; Stange, 2018), and others noted either grade inflation (Carlisle, 2020; Stange, 2018) or active measures used to ensure retention and accurate grading by requiring students to verify mastery of material on more than one assessment (Elsinger & Lewis, 2020; Weir, 2020). There were several common themes across the literature. Multiple articles noted the increase in instructional workload associated with creating and grading assessments or active measures to limit availability due to the increased workload (Barth & Higginbottom, 2021; Carlisle, 2020; Collins et al., 2019; Cundiff et al., 2020; Elsinger & Lewis, 2020; Mangum, 2020; McKell & Danowitz, 2020; Stange, 2018; Williams, 2018) and an increase in student procrastination (Collins et al., 2019; Harsy & Hoofnagle, 2020; Mangum, 2020; McKell & Danowitz, 2020; Weir, 2020). In addition, several articles mentioned an increase in course drop-outs (Carlisle, 2020; Harsy & Hoofnagle, 2020; Jones, 2020; Weir, 2020).

End-of-Year High Stakes Testing. The following articles focus on how the use of mastery learning affects end-of-year high stakes testing. The following research demonstrated mixed results (Brookhart et al., 2016; Early et al., 2016; Kulik et al., 1990; Lehman et al., 2018; Pollio & Hochbein, 2015; Townsley & Varga, 2018).

Two seminal works discuss the impact of mastery assessments on end-of-year high stakes testing. C.-L. C. Kulik et al. (1990) performed a meta-analysis of 108 studies evaluated mastery learning programs, of which 36 studies focused on Bloom's mastery learning approach. A one-way ANOVA determined that studies with the biggest positive effects shared the following characteristics: locally developed versus nationally standardized tests, a mastery requirement of 90-100%, teacher-paced units, and course content. Overall, mastery programs were shown to have a statistically significant, moderate size, positive effect on student learning as demonstrated by school-based unit exam scores. The effect of mastery learning on standardized tests was consistently small across the studies. When quiz feedback was given to both control and treatment groups, the treatment effects were smaller. Mastery learning programs in social sciences were more effective than in mathematics and the natural sciences. Of the studies that noted instructional time, all studies indicated an increased requirement for instructional time and a higher demand on student time for mastery learning compared to conventional classes. At the university level, a lower completion rate was determined in the treatment groups, though the effect was small. In the second seminal work, Brookhart et al. (2016) performed a meta-analysis of 114 articles on grading focusing on: 1) early studies on grade reliability, 2) quantitative studies comparing grades to educational outcomes, 3) teacher's perceptions of grades and grading practices, 4) standards-based grading practices, and 5) grading practices in higher education. Noted that there is extensive literature on standards-based grading, but few empirical studies. Literature on standards-based grading has focused on how it is implemented, how it compares to state achievement tests designed to measure the same or similar standards, and what are student, parent, and teacher perceptions. The articles (N = 8) evaluating the consistency between standards-based grades and state achievement test scores all demonstrated similar results, with a modest correlation of 0.5. These two seminal works provided convergent evidence that although there was a modest correlation between grades and nationally standardized tests, mastery learning showed the biggest positive effects for locally developed versus nationally standardized tests (Brookhart et al., 2016; Kulik et al., 1990). Additionally, it was noted that, feedback reduced the effect size, mastery learning requires an increased demand for both instructors and students, mastery learning is not as effective when used within the mathematics curriculum, and although there is extensive literature there were few empirical studies. (Brookhart et al., 2016; Kulik et al., 1990)

The following four articles provide mixed results on the impact to end-of-year high stakes testing. Pollio and Hochbein (2015) performed a quasi-experimental non-equivalent control group design on algebra 2 students to examine the association of different grading approaches with the end-of-year state exams for students from 11 high school in Louisville, Kentucky. The control group (n = 1,163) was selected from the previous year and the experimental group (n = 1,256) was selected from the first-year group of standards-based grading implementation. A Pearson correlation coefficient on the relationship between student grades and state exams demonstrated a weak positive correlation for both the control (r = .38) and experimental groups (r = .45). A multiple linear regression indicated both prior academic achievement and grades within the content area were the strongest predictors of state-exam achievement scores. A stronger association existed between course grades and state-exam scores for students in the experimental group. There was no analysis to determine if there were any differences in implementation across the 11 schools or classrooms. Lehman et al. (2018) performed a study to understand the relationship between classroom grades in traditional-graded classrooms (n =1,900, 4 middle schools) and standards-based graded classrooms (n = 500, 1 middle school) using Scholastic Math Inventory exam scores for sixth-, seventh-, and eighth-grade mathematics students. The Pearson correlation for both grading schemes was moderate, with traditionally assessed end-of-course mathematics grades for all students (r = 0.355, R^2 = 0.126) and standards-based grading assessed end-of-course mathematics grades for all students (r = .392, R^2 = 0.154), demonstrating similar results. As previously reviewed, Early et al. (2016) performed a 2-year long randomized trial of high schools (N = 20) to determine if there was an increase in ninth and 10th grade math and English scores on end-of-year standardized tests and if improved fidelity of implementation impacted these scores. Math scores were not statistically significant in the first year but were statistically significant in the second year with a small effect size (0.14), English scores showed no statistically significant improvement. Townsley and Varga (2018) implemented a quasi-experimental research design with a stepwise linear regression to determine if standards-based grading affected high school grade point averages (GPAs) or ACT exam scores. Existing data from two comparable high schools (N =327) were evaluated, one implementing standards-based grading and one using traditional grading practices. Neither math nor English GPAs were found to be statistically significant, however the analysis of both math and English ACT scores were both found to be statistically significant, with students in traditional grading scoring higher on ACT exams. The results of these four studies were all different, finding either mixed, a weak positive correlation, similar results, or that traditional grading resulted in higher ACT scores (Early et al., 2016; Lehman et al., 2018; Pollio & Hochbein, 2015; Townsley & Varga, 2018). In two studies, there were no comparisons across schools to determine if they had common baselines affecting the validity of the results (Early et al., 2016; Pollio & Hochbein, 2015).

Synthesis. The research on the impact of mastery assessments on end-of-year high stakes testing demonstrates there are no clear results on the impact of mastery learning on end-of-year high stakes exams. Two seminal works provided convergent evidence that there is a modest correlation between grades and nationally standardized tests, but mastery learning showed the biggest positive effects for locally developed versus nationally standardized tests. Although feedback reduced the effect size, mastery learning requires an increased demand for both instructors and students and is not as effective when used within the mathematics curriculum (Brookhart et al., 2016; Kulik et al., 1990). Compared to traditional learning, mastery learning was demonstrated to have similar results (Brookhart et al., 2016; Lehman et al., 2018), mixed results (Early et al., 2016), weak positive results (Kulik et al., 1990; Pollio & Hochbein, 2015), and negative results (Townsley & Varga, 2018). The strongest predictors on end-of-year high stakes testing were prior academic achievement and grades (Early et al., 2016; Pollio & Hochbein, 2015).

Comparing the Amount of Time Between Tests. The following articles focus on comparing how different timeframes used between the original and retake tests affect retention as evidenced by test scores on assessments. The research focuses on the impacts of spaced retrieval, increasing testing, and the length of time between testing intervals.

Spaced Retrieval. Three articles discussed the impacts of spaced retrieval on retention. Latimier et al.'s (2021) meta-analytic review demonstrated that spaced retrieval practice outperformed massed retrieval practice with a high effect size, and the more students are tested, the more beneficial expanding the schedule is. Alternately, Lyle et al. (2020) determined that increasing the spacing between retrievals increased long-term retention for mathematics learning but decreased retention within a class. Guo (2021) used 2 by 2 repeated measures ANOVA comparing initial test interval (short versus long) and feedback timing (immediate and delayed) on vocabulary retention over 2 weeks on undergraduate students (N = 60) in a lower-level ESL course. The analysis determined a short-interval initial test followed by delayed feedback was the optimal schedule to promote retention. The delayed feedback condition producing better final test performance than groups in the intermediate feedback condition regardless of study-test lag, suggesting that repeated retrievals and not re-testing may insulate memory strength.

These articles demonstrate a consensus that increasing spaced retrieval improves long-term retention.

Increased Testing. The following articles discussed the impact of increasing the number of tests on retention. Bangert-Drowns et al.'s (1991) seminal work performed a meta-analysis on 40 studies to determine the impact of frequency of classroom testing on student achievement, determining students who took one test during the 15-week term outperformed students who took no tests by 0.5 standard deviations, but the amount of improvement diminished as the number of tests were increased**.** Although increased performance was shown when test items were distributed into a larger number of short tests, it did not translate into better posttest performance on criterion tests, with an average effect size of zero. Adesope et al. (2017) extended the work performed by Bangert-Drowns et al., reconfirming that the effects of practice tests were larger when used only once than when used two or more times. Recent articles provided additional contributions to the conversation on the frequency of testing on retention. Lyle et al. (2020) determined that increasing the number of retrievals from three to six increased retention for mathematics learning within a class, but this increase had no effect on long-term retention and Foss and Pirozzolo (2017) determined that the effect of increasing the frequency of tests from two per semester to four or eight exams per semester on student final exam performance was not significant. Although both meta-analyses were consistent, showing that the effect of practice tests were larger when used once, in the study that focused on mathematics content the use of frequent short assessment demonstrated higher retention within the course (Lyle et al., 2020).

Length of Time Between Testing Intervals. The following articles discussed the impact of the length of time between testing intervals on retention. The following research demonstrated mixed results (Chan et al., 2018; Cho & Blood, 2020; Ee et al., 2018; Kliegl et al., 2019; Lin & Chen, 2020; Wyse, 2021; Wyse & Anderson, 2020; Zhou & Cao, 2020)

The following article investigated the length of time required to avoid test recall. Wyse (2021) explored how the number of days between taking reading and math computerized adaptive tests to find the optimum time between tests that is long enough to avoid recall effects but short enough that student ability does not change dramatically. The time between the first and second tests was between 0 and 150 days for reading tests (n = 6,221,329) and math tests (n = 4,048,494). The test instrument were computerized adaptive tests used to measure student progress and growth in first through eighth grade students. Highest alternate form reliability coefficients were obtained for both reading and math at least 2-3 weeks after the first test, but if keeping student ability similar is a concern then the best time to retest is shortly after 3 weeks after the first test.

The following articles focused on the impact on retention with increased time between test and retest. Ee et al. (2018) determined that the inclusion of an instructional time constraint of 2 weeks to complete a quiz resulted in a statistically significant association between the length of time taken to complete a quiz and the final examination score, with increased time to take a quiz resulting in decreased scores on the final exam. Chan et al. (2018) performed a study using four experiments to test the effects on new learning in both restudying and re-testing conditions across various retention intervals.

The study determined that increasing the amount of time between the original learning and the effects of test induced learning substantially reduced the effectiveness for both re-testing and restudying participants. Kliegl et al. (2019) performed experiments to determine if delay-induced forgetting was driven by interference or contextual drift. Three experiments were performed manipulating delay conditions of 1-, 10- and 30-minutes, with an interference task (n = 81), delay of 1- minute and 30-minutes with or without a reinstatement task (n = 90), and a delay of 1-minute and 10-minutes with or without an interference task (n = 30) on university level students. All three experiments demonstrated that prolonged delay fixed with distractor tasks reduces both recall rates and response latencies. In contrast, Cho and Blood (2020) performed a multi-level modelling analysis on Japanese (n = 1681) and Turkish (n = 3532) students, ages 8-13 years old, who took the TOEFL test two or three times, to determine the effect of time interval between test administration and age of the test taker on TOEFL test scores. The age of the test taker had the highest correlation with test score. The correlations between time and test score were statistically significant but with a low effect size. Results showed there was a positive effect on test score with increasing time, but it only explained 1-2% of the variability in test scores, with the range of average score increases of 0.08 and 1.5 points for test retake timeframes between 1 to 6 months. Where Chan et al. (2018), Ee et al. (2018), and Kliegl et al. (2019) demonstrated a reduction in retention with increased time between test and retest, Cho and Blood demonstrated a slight positive increase in test scores with increasing time.

The following articles looked at the effect of time on large scale examinations. Zhou and Cao (2020) investigated if the time interval between repeaters' initial and second test attempt on a large-scale computer-based licensure test for entry-level mathematics teachers was associated with the repeaters' test score change. Data were evaluated from 9,124 test takers who repeated the exam, of which 1,081 took the same exam, 4,827 took a different form, and 3,216 took a test that had common questions as the original exam. The time interval between exams varied from 22 to 1443 days, with a mean of 129 days and an *SD* of 156 days. The data were positively skewed with approximately 60% of the test takers retaking the first time within 3 months and more than 80% within 6 months. The mean score and the interquartile range differed little based on length of time between exams, with a mean score change varying from 5.4 to 6.8 points. Same-form repeaters had an average increase of 8.0 points (effect size of 0.74) and different form repeaters had an average increase of 5.9 points (effect size of 0.50). Wyse and Anderson (2020) performed a study on the different kinds of pass rates for 14 medical credentialing programs to recommend which pass rate statistics should be reported to stakeholders. In 13 of 14 credentialing programs (N = 25,009), the first attempts pass rate was the highest (*M* = 77.7%), and the repeat attempts pass rate the lowest (*M* = 51.7%). Lin and Y. Chen (2020) performed a longitudinal study to evaluate the writing score and linguistic feature changes on the Canadian English Language Proficiency exam (N = 562) for repeat test takers between 2014 and 2018. A short interval between the first and second attempts (30-40 days) and a longer interval between the first and third attempts (90-180 days) were chosen based on the statistical modes within the data. Score distributions remained stable across attempts, with most test takers falling in the middle range. Proficiency groups were determined based on the score received on the initial attempt. The low proficiency group performed better between the first and second attempt with a marginal improvement on the third attempt. The medium proficiency group scores remained stable across all attempts, and the high proficiency group scores went down across all attempts. All mean differences were statistically significant, but effects sizes were not provided. For all proficiency groups, lexical features were more likely to improve with the longer timeframe, but cohesion and syntactic sophistication did not change significantly. These articles provide convergent evidence that there are mixed results, with either the mean score differing little based on length of time between tests (Zhou & Cao, 2020), the first pass attempts rate being highest and the repeat attempts pass rate the lowest (Wyse & Anderson, 2020), or the pass rate of the repeat exam depending on how the student did on the original exam, with students who performed poorly the first time showing improvement, but those who performed better either doing the same or worse on the second attempt (Lin & Chen, 2020).

Synthesis. The above articles focused on comparing how different timeframes used between the original and retake tests affect retention as evidenced by test scores on assessments. The research focused on the impacts of spaced retrieval, increasing testing, and the length of time between testing intervals. The research demonstrated a consensus that increasing spaced retrieval improves long-term retention (Guo, 2021; Latimier et al., 2021; Lyle et al., 2020). Increasing testing demonstrated that the effect of practice tests was larger when used once (Adesope et al., 2017; Bangert-Drowns et al., 1991; Foss & Pirozzolo, 2017), however, the use of frequent short assessments demonstrated higher retention for mathematics content (Lyle et al., 2020). The research on the effect of length of time between testing intervals was not consistent. Research determined students demonstrated similar abilities if tested 2-3 weeks after the original test (Wyse, 2021), no change in retest scores with increased time (Cho & Blood, 2020; Zhou & Cao, 2020), and a reduction in retention when time is increased (Chan et al., 2018; Ee et al., 2018; Kliegl et al., 2019; Wyse & Anderson, 2020). Finally, one researcher determined that the differences in score between the original and retake were determined by the proficiency on the original exam, with low proficiency students performing better, but medium and high proficiency students performing the same or worse (Lin & Chen, 2020).

Critical Summary

Current literature on mastery-based assessments does not provide a clear picture on the practice of mastery-based assessment, evidencing multiple practices associated with the amount of time between the original test and retake, and demonstrating mixed results from the practice. The review of the broader problem was separated into two major sections: (a) results of mastery implementation and (b) the impact of different timeframes for retaking assessments. The results of mastery implementation section included research on implementations factors and considerations, effect on student readiness for follow-on learning, and the testing effect.

Although the themes from research on the factors and considerations for mastery learning are mostly convergent demonstrating increased student ownership of learning, making learning more purposeful, and assessment practices clearer (Harsy & Hoofnagle, 2020; Knight & Cooper, 2019; Weir, 2020), as with any practice there were unintended consequences. Themes in the current literature provide convergent evidence of the following unintended consequences: 1) student's arriving unprepared for initial exams, 2) student procrastination leading to increased course failures or drop-outs, 3) grade inflation, 4) increased instructional workload, 5) increased burden and time management required of students with the resultant increase in student stress, 6) increased student focus on grades, not learning, and 7) reduced ability to modify instruction to follow student interests or to synthesize topics (Carlisle, 2020; Elsinger & Lewis, 2020; Emery et al., 2018; Fernandez, 2021; Harsy & Hoofnagle, 2020; McKell & Danowitz, 2020; Olsen & Buchanan, 2019; Scheopner Torres et al., 2018; Stange, 2018). In addition, there is a clear change in the distribution of grades within a course that uses mastery assessments, with a higher percentage of students earning an A but also a greater percentage of students who either withdraw from courses or earn a D or F (Harsy & Hoofnagle, 2020; Jones, 2020; Katzman et al., 2021; Lenarz & Pelatt, 2020; Weir, 2020). The themes evidenced in the literature are convergent, demonstrating that the mastery learning process has many positive impacts on student learning, but has resulted in significant unintended consequences that affect its implementation.

The results of the mastery implementation section continued with literature on retention with three primary foci: retention within a course, retention in follow-on courses and student perceptions on retention. This study focuses on the impacts of different time implementations within mastery learning on end-of-year high stakes exams. End-of-year exams were chosen as they are a common exam given to all students and because these exams are a measure of student's retention of content which is critical for follow-on learning. Mastery learning both within course and in follow-on courses demonstrated either mixed results or a reduction in content retention (Carr et al., 2017; Childers & Lu, 2017; Harsy & Hoofnagle, 2020; Katzman et al., 2021; Weir, 2020; Zimmerman, 2020). In particular, Lyle et al.'s (2020) research presented diverging results, with increased practice demonstrating higher retention within a course but having no effect on long-term retention and increased spacing having a negative effect within course but a positive effect on long-term retention. Finally, students' perceptions on the effectiveness of mastery assessments on course retention was also mixed. As traditional grading methods are generally implemented at the university level, students from schools using traditional grading believed these methods better prepared them whereas students from schools using standards-based felt they were not properly prepared due to their previous reliance on retake assessments (Erbes et al., 2021; Guskey et al., 2020). Students evaluating the use of mastery assessments within a university level course provided mixed results with the practice seen as both deepening the understanding of course content (Harsy & Hoofnagle, 2020) and being less educationally effective (McKell & Danowitz, 2020). Mastery learning both within and across courses demonstrated mixed results or a reduction in content retention, with student perceptions on the implementation of mastery learning also mixed.

The results of the mastery implementation section finished with literature on the three aspects of the testing effect to determine their impact on retention of content material: test format, distributed practice, and test-enhanced transfer. There is converging evidence that both test format and test-enhanced transfer positively impacts the testing effect, with higher performance using recall and multiple-choice formats (Adesope et al., 2017; Choi & Lee, 2020; Ebersbach & Barzagar Nazari, 2020; Foss & Pirozzolo, 2017; Morphew et al., 2020; Racsmány et al., 2020; Rawson et al., 2020). There was also convergent evidence from all three aspects of the testing effect that the magnitude of the benefit decreases with the complexity of the curriculum or for higher order test items (Ebersbach & Barzagar Nazari, 2020; Foss & Pirozzolo, 2017; Peterson & Wissman, 2018; Racsmány et al., 2020; Rawson et al., 2020). The effects of practice tests were larger when performed once and the transfer effect is strongest in memorization type questions or when the practice and final test formats are identical (Adesope et al., 2017; Bangert-Drowns et al., 1991; Foss & Pirozzolo, 2017; Pan & Rickard, 2018). There was no transfer effect for problem-solving test items (Barzagar Nazari & Ebersbach, 2019; Peterson & Wissman, 2018) and a weak effect for problems involving worked solutions (Foss & Pirozzolo, 2017; Pan & Rickard, 2018). Distributed practice provided mixed results, with a consistent advantage of spacing out retrieval practice demonstrated in the literature (Barzagar Nazari & Ebersbach, 2019; Guo, 2021; Latimier et al., 2021; Racsmány et al., 2020; Rawson et al., 2020). Divergent results were found about the impact of feedback, with results obtained that stated neither increasing test frequency nor providing feedback changed the effect of test format (Choi & Lee, 2020; Foss & Pirozzolo, 2017) and the testing effect was reduced when feedback was introduced after the first assessment (Racsmány et al., 2020). Finally divergent evidence was found that it is repeated retrievals and not repeated testing that supports increased memory (Guo, 2021). There are few studies with long retention intervals and there is a notable gap in testing effect research as most of the studies are performed in the laboratory with undergraduate students (Trumbo et al., 2021). The research shows a convergent connection between the testing effect and the complexity of the curriculum and test format, with curriculum such as mathematics demonstrating reduced effect.

The second major section of the current literature focused on the impact of different timeframes for retaking assessments which included research on the *short-term* and *anytime* timeframes, end-of-year high stakes testing, and comparing the amount of time between tests. At the study site there are varying amounts of time given between the original test and the retake. This was used to focus the literature review on the different implementations on the length of time given between the original test and its retake. Although the anytime implementation revealed some research demonstrating improvements on locally developed exams (Barth & Higginbottom, 2021; Birhan, 2018; Cundiff et al., 2020; Nevid & Gordon, 2018), the majority of the research for both *short-term* and *anytime* implementations demonstrated mixed results on locally developed exams (Adeniji et al., 2018; Carlisle, 2020; Childers & Lu, 2017; Ee et al., 2018; Halperin, 2020; Heubach & Krinsky, 2020; Jones, 2020; Mangum, 2020; McKell & Danowitz, 2020; Morphew et al., 2020; Weir, 2020), or either no difference or an overall reduction on locally developed exams (Adeniji et al., 2018; Blackstone & Oldmixon, 2019; Early et al., 2016; Katzman et al., 2021). There were divergent results from the research on the impact of mastery assessments on end-of-year high stakes testing

(Brookhart et al., 2016; Early et al., 2016; Lehman et al., 2018; Pollio & Hochbein, 2015; Townsley & Varga, 2018). Two seminal works demonstrated mastery learning showed the biggest positive effects for locally developed versus nationally standardized tests and was shown to be not as effective when used within the mathematics curriculum (Brookhart et al., 2016; Kulik et al., 1990). The strongest predictors on end-of-year high stakes testing were prior academic achievement and grades (Early et al., 2016; Pollio & Hochbein, 2015). The impact of individual timeframes, whether *short-term*, *anytime*, or on *end-of-year* testing on locally developed exams was mixed.

Finally, research on comparing how different timeframes used between the original and retake tests affect retention as evidenced by test scores on assessments focused on the impacts of spaced retrieval, increasing testing, and the length of time between testing intervals. Research was convergent for the impacts of both spaced retrieval which demonstrated that increasing spaced retrieval improves long-term retention (Guo, 2021; Latimier et al., 2021; Lyle et al., 2020) and increased testing which demonstrated the effect of practice tests was larger when used once (Adesope et al., 2017; Bangert-Drowns et al., 1991; Foss & Pirozzolo, 2017; Lyle et al., 2020). However, it was noted that within mathematics content area, the use of frequent short assessment demonstrated higher retention within a course (Lyle et al., 2020). There were divergent results from the research on the effect of length of time between testing intervals. Research determined students demonstrated similar abilities if tested 2-3 weeks after the original test (Wyse, 2021), no change in retest scores with increased time (Cho & Blood, 2020; Zhou & Cao, 2020), and a reduction in retention when time is increased (Chan et al., 2018; Ee et al., 2018; Kliegl et al., 2019; Wyse & Anderson, 2020). Finally, the differences in score between the original and retake were determined to be driven by the proficiency on the original exam, with low proficiency students performing better, but medium and high proficiency students performing the same or worse (Lin & Chen, 2020). Although there is convergent research for both spaced retrieval and increased testing, there were divergent results on the effect of length of time between testing intervals.

Although mastery-based assessments are the context of the study, the specific practices related to the amount of time teachers provide between the original test and the retake is the focus of the study. Current literature on mastery-based assessments does not provide a clear picture on the practice of master-based assessment, evidencing multiple practices associated with the amount of time between the original test and retake, and demonstrating mixed results from the practice. The school district is using a modified version of the theoretical framework as proposed by Bloom (1968) as additional academic support and the opportunity to take a retake offered to students are not provided prior to the beginning of new instruction, but in parallel with new instruction. Although there are many recent qualitative reviews or discussions on the implementation of mastery learning in the classroom there are fewer that provide statistical analyses and fewer still that compare across implementation choices to determine the factors that have an impact on student retention of content. Complicating the current review, most research is performed on read passages, word lists, or word pairs. The working memory demand is stronger for a mathematical procedure than simpler tasks like recalling word lists as mathematical procedures are more complex requiring both recalling underlying concepts such as formulas and the ability to apply these concepts within a solving task (Ebersbach & Barzagar Nazari, 2020). The proposed research attempts to address this gap in practice by identifying which practices demonstrate the greatest effect on student learning within a high-school mathematics course.

Implications

The purpose of this quantitative study is in a Northeastern suburban public school district, to investigate if there are differences between high school students' geometry end-of-year state exam scores or between high school students' geometry math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*. Identifying if the recommended amount of time between the test and retake that maximizes student acquisition and demonstration of content knowledge is significant as it will be used to inform district leaders leading to the development of an updated policy improving student learning and mastery of content material. The district is currently performing a review of the grading policy (Superintendent Town Hall, 2023). If there is an optimized amount of time between test and retake, then this can be implemented in future district assessment policy. A subsequent project based on the findings of this study is a position paper outlining a recommended policy modification. A potential PowerPoint presentation either in person or delivered by video recording may be a productive avenue to pursue. A curriculum plan or a professional development are not appropriate for this project study.

Summary

Current literature on mastery-based assessments does not provide a clear picture on the practice of master-based assessment, evidencing multiple practices associated with the amount of time between the original test and retake, and demonstrating mixed results from the practice. The review of the broader problem was separated into two major sections: results of mastery implementation and the impact of different timeframes for retaking assessments. Although there are many recent qualitative reviews or discussions on the implementation of mastery learning in the classroom there are fewer that provide statistical analyses and fewer still that compare across implementation choices to determine the factors that have an impact on student retention of content or that focused on mathematics as the content knowledge. The proposed research attempts to address this gap in practice by identifying which practices demonstrate the greatest impact on student learning within a high-school mathematics course. The following sections will cover the methodology (Section 2), the project (Section 3), and reflections and conclusions (Section 4).

Section 2: The Methodology

Research Design and Approach

The purpose of this quasi-experimental quantitative study was in a Northeastern suburban public school district, to investigate if there were differences between high school students' geometry end-of-year state exam scores or between high school students' geometry math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*. A quantitative causal-comparative approach was used to analyze the difference in high school students' geometry end-of-year exam scores for students who retook a school-based exam differentiated by three schemes of retake timing: *short-term*, *anytime*, or *end of the quarter*. An ANOVA was appropriate for this study as the goal was to determine if there was a statistically significant difference between the three timeframes.

This causal-comparative design derived logically from this problem because the research question naturally lent itself to the need to compare the effectiveness of the three timeframes for retakes. It was a clear question to ask which of the retake methods being used were most effective so that best practices could be employed. The timeframes were determined by curriculum teams at individual schools within the target school district. In the 2023-2024 school year, fourteen curriculum teams implemented retakes within two to three weeks of the original test (*short-term*), six curriculum teams implemented retakes anytime during the course (*anytime*), and eight curriculum teams implemented at the end of the quarter (*end of the quarter*). The different practices for retake timing were the independent variable, and the end-of-year state exam score was the dependent variables in the analysis. Archival data was requested from schools from these pools to evaluate in this project. A one-way ANOVA analysis was used to determine if there were statistically significant differences between the means of the three independent practices currently in use. A post hoc test was used to determine which of the groups differ from the others on the high school geometry end-of-year state exam scores based on the different practices for retake timing (Dr. Liu, personal communication, Mar 1, 2022 & Laird Statistics, n.d.a).

Setting and Sample

The local setting was a large, public-school division in the northeastern United States. The district has approximately 200 pre-K through grade 12 schools and centers serving a diverse student population of 181,000 students, speaking over 200 languages. Approximately 38.5% of the total student population is economically disadvantaged, 16.3% are students with disabilities, and 26.5% of students are English learners. Demographically 36.1% of students are White, 28.7% are Hispanic, 18.7% are Asian, 9.8% are Black, 6.3% are two or more races, 0.3% are American Indian, and 0.1% are Native Hawaiian (state Department of Education School Quality Profile, 2025).

The sampling strategy used was non-random convenience sampling. As already existing archival data was used for this study, it was not possible to randomly sample students, and the researcher used the data that was available. A g*power for ANOVA analysis was performed to determine the minimum sample size for this research study. With a power of .8, an effect size of .25, an alpha level of .05 and three groups the total sample size was calculated to be 159, or 53 for each group (UCLA Advanced Research Computing, 2021).

The eligibility criteria for study participants was high school Geometry and Geometry Honors students who took the course in the 2023-2024 school year in the 28 sites identified in Table 1: 14 curriculum teams implement retakes within two to three weeks of the original test (*short-term*), six curriculum teams implement retakes anytime during the course (*anytime*), and eight curriculum teams implement at the end of the quarter (*end of the quarter*). Student data was included if the student retook a locally developed exam during the identified school years. Student data was excluded if the student did not retake a locally developed exam during the identified school years. Participants were not recruited for this study, instead deidentified archival data from locally developed exam scores and end-of-year high stakes testing scores was used. The sample was expected to represent the characteristics of the population of the district as every student within the targeted population could retake an exam if their original score on the exam was less than 80% (District policy, 2021).

Instrumentation and Materials

The data collection required to address the research question consisted of student scores on the state end-of-year geometry exam which was provided as an actual total score out of 600 points and students who had retaken a locally developed retake exam either in the *short-term* timeframe, the *anytime* timeframe, or the *end of quarter* timeframe. Experimental data was not used in this study. A request was made to the local district for deidentified archival data consisting of state end-of-year geometry exam scores. The state end-of-year geometry test was published in 2016 and is administered by Pearson through the online testing application, TestNav™, which includes the Desmos online calculator (state Department of Education website, 2023). The state end-of-year exam is not a published instrument, and permission from the publisher is not required as the instrument itself will not be used in this research study, just the results of previous administrations of the instrument by the district. This instrument was appropriate and sufficient for the study as it is a common and independent exam provided to all geometry students across the district whose results could be compared across the different retake methodologies. A researcher developed instrument was not used in this study.

The state has determined a minimum set of expectations that each student must reach within a curriculum. The end-of-year exam measures content knowledge, mathematical processes, and reasoning and critical thinking skills using both multiple-choice and technology-enhanced items (state Department of Education website, 2023). The concepts specifically measured by the geometry end-of-year state exam encompass the state standards for the geometry curriculum, which include reasoning, lines, and transformations; triangles; polygons and circles; and three-dimensional figures (state Department of Education, 2016). All items on the end-of-year exam are reviewed by classroom teachers for accuracy and fairness.

End of year exam scores are calculated using scaled scores. There are 55 test items on the geometry end of year state exam, of which 16 cover reasoning, lines, and transformations, 13 cover triangles, and 16 cover polygons, circles, and three-dimensional figures. In addition to the abovementioned 45 operational items, there are also ten field-test items included in each test administrations, but these items are not included in the computation of a student's score (state Department of Education, 2023). Each student receives 45 test items from a bank of approved test items; therefore, no two tests are guaranteed to be the same either within a year or across year groups. To ensure that the end-of-year test score has the same meaning across student and year, the raw score is converted to a scaled score. The scaled score is used to adjust for slight variations in individual item difficulty from one version of a test to another and provides a score that can be used to compare student achievement across test versions and year of implementation. Scaled scores range from a 0 – 600, with a score of 400 or greater set as pass/proficient and a score of 500 of greater set as pass/advanced.

Reliability answers whether the same student taking the end-of-year state exam multiple times, would earn the same score and validity describes how accurately the score received on the end-of-year exam measures a student's understanding of the geometry course content (Babbie, 2017). According to the state's Standards of Learning Assessments Technical Report 2021-2022 Administration Cycle, the end-of-year exam demonstrated an overall reliability of 0.90. The reliability analysis included computations on the internal consistency, measurement error, and decision accuracy and consistency, with an addition coefficient alpha analysis by gender as well as ethnic and status subgroups. According to the state's Standards of Learning Assessments Technical Report 2021-2022 Administration Cycle validity is evidenced through the development processes for test content which include field testing, student response processes, and internal structure.

To evaluate the learning of students in geometry, an online exam is administered every May. The process to complete this instrument begins in May, when every geometry student who has not already passed a mathematics end-of-year exam at the high school level takes the exam. The test is administered by local instructional staff using the online testing application, TestNav™, which includes the Desmos online graphing calculator (state Department of Education website, 2023). The test is not timed; students have as long as needed to finish the exam. Exams are scored by Pearson, and scores are provided back to the district. A passing score is any score of 400 points or higher out of 600. If a student earns between 375 and 400 points on the exam, they are provided with a summary of the exam results which includes what content the student performed well on, and what content the student needs to improve upon. Students are provided with a second opportunity to pass the exam 2 weeks later (state Department of Education website, 2023). Although the state end-of-year exam is published, it has not been released to the public and therefore cannot be attached to the study. Example problems are available on the state department of education website, 2023.

The district collects and maintains this data in their student information system. The district has maintained their rights to the data and has not approved its general release. Request for the data may be made directly to the site.

Data Collection and Analysis

The data collection required to address the research question consisted of student scores on the state end of year exam which were provided as an actual total score out of 600 points and students who have retaken a locally developed retake exam either in the *short-term* timeframe, the *anytime* timeframe, or the *end of quarter* timeframe. The data collection process began with a request for deidentified archival data from the district to retrieve locally developed exam scores and end-of-year test scores for students taking Geometry or Geometry Honors in the previously identified fourteen curriculum teams who implemented retakes within two to three weeks of the original test (*short-term*), the six curriculum teams who implemented retakes anytime during the course (*anytime*), and the eight curriculum teams who implemented retakes at the end of the quarter (*end of the quarter*).

There were two steps to prepare the data for analysis. The first step was to identify the students who had retaken a locally developed exam during the identified school year. To ensure only students who had retaken tests were included in the data set, information from both the centralized district electronic gradebook and individual teacher-maintained gradebooks were obtained and used to identify students to be included in the sample population. Specifically, retakes are captured within the electronic gradebook in one of two ways, either directly by having their own separate grade, or within teacher comments that are associated to the original test grade. Teacher-maintained gradebooks will be obtained as a secondary source to confirm the students who should be included in the sample population. Only students who had retaken locally developed tests will have their end of year scores included in the analysis.

The second step to prepare the data for analysis using the IBM Statistical Package for the Social Sciences (SPSS), Version 28, was the creation of an excel spreadsheet with two columns. The first column was coded with either a '1', '2', or '3' to represent which implementation was used. Within the first column, students who retook a locally developed exam at a school implementing the *short-term* timeframe were coded with a '1', students who retook a locally-developed exam at a school implementing the *end of quarter* timeframe were coded with a '2', and students who retook a locally developed exam at a school implementing the *anytime* timeframe were coded with a '3'. The second column was the student's state end-of-year exam score. For example, Happy school implements retakes *anytime* so all students that were identified in the teacher comments as having retaken a test at Happy school were coded with a '3' in the first column. In summary, the gradebooks determined who had retaken the tests, and school attendance determined what code was used in the first column. The ANOVA then compared the end of year exam scores by implementation.

The procedure for gaining access to the data set began by making a request for deidentified archival data from the district. An application to access the data was submitted to the district's division of research screening committee. The application included an overview of the research design, the identification of a district sponsor, and the approval form from the Institutional Review Board. The application asked for the deidentified locally developed exam scores with any associated teacher comments and state end-of-year exam scores for students taking Geometry or Geometry Honors in the selected schools during the specified school year.

The data was analyzed using the IBM Statistical Package for the Social Sciences (SPSS), Version 28. The independent variable was a nominal variable and was defined by the different practices for retake timing. The dependent variable was student end-of-year state exam scores and was measured on a continuous scale from 0 – 600. Table 3 summarizes the variables used in the analysis.

Table 3

ANOVA Variables

	Research question	Variables	Data analysis
RQ1	What are the differences between high school students' geometry end-of-year state exam scores differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*?	IV: Retake Timing *Short-term* 1 *End of Quarter* 2 *Anytime* 3 DV: End-of-Year Exam Scores Continuous from 0 - 600	One-way ANOVA

A one-way ANOVA will be used to determine if there were statistically significant differences between the means of the three independent practices currently in use. A post hoc test was used to determine which of the groups differed from the others on the high school geometry end-of-year state exam scores based on the different practices for retake timing (Dr. Liu, personal communication, Mar 1, 2022 & Laird Statistics, n.d.a).

Assumptions, Limitations, Scope, and Delimitations

Assumptions

The first three of six statistical assumptions for using ANOVA were met within the design: a) the dependent variable was measured on a continuous scale, b) the independent variable was from two or more independent groups, and c) there were independence of observations as there is no relationship between the groups since they were from different schools within the district. The other three of six assumptions, no outliers, an approximately normal distribution of the dependent variable, and homogeneity of variances, were assessed after data collection and during the data analysis (Laird Statistics, n.d.a).

The following additional assumptions are inherent in this analysis. End-of-year state assessments used by the district were administered with fidelity and were reliable. All students performed their best on every assessment. The data sets provided by the district were all-inclusive and accurate. All geometry instructors across the district had the same skill level and assessed students equivalently. Finally, all curriculum teams using the same amount of time were internally self-consistent, meaning all members of the curriculum teams used the same amounts of time and had the same skill level.

Limitations

The data analysis did not include a statistical procedure to determine if groups were equal. However, as the group sizes were large, this limitation is negligible. Random assignment was also not possible. Additionally, the process to determine if a student took a school-based retake exam was complicated as teachers denote this information in different ways in the district's student information system. If possible, an annotated class list was requested from teachers to determine which students took a school-based retake exam. Both test grades and test comments were collected as there are multiple ways teachers entered data into the student information system used by the district.

A potential limitation was meeting the sample size requirements for each group as not every student in a class takes the retake. The minimum sample size was met for each group to analyze the different amounts of time [*short-term* (2-3 weeks), *anytime* (during the course), and *end of the quarter*]. Another potential limitation was teachers within the district vary how they listed their test scores and retake scores in the electronic gradebooks used by the district, causing issues when collecting and creating the data set for SPSS analysis. This was mitigated by requesting an annotated student list if available.

Scope and Delimitations

The scope of this study was Geometry and Geometry Honors students who retook a locally developed exam during the 2023-2024 school year. The study itself looked at these students' end of year exam scores and compared them. It did not compare locally developed exams from the different teachers. Student social-economic status, prior academic achievement, and student demographic data were not included in this analysis because IRB no longer allows researchers to collect this information unless it is necessary to answer the research questions.

Protection of Participants' Rights

Deidentified archival data provided by the district was used in this research study. Therefore, there was minimal risk to high school students and consent was unnecessary. The Walden University Office of Research Ethics and Compliance granted permission on December 2, 2024 (IRB approval number 11-08-23-1047586), and there was no attempt to collect data until after approval was received.

Data Analysis Results

Upon receipt of the original request for data as outlined above, it was determined that though the research met the technical merits and needs of the site, the request was denied because the data was not readily available, and the site could not justify spending public funds to access and provide the requested data. In conjunction with the committee, school administration, and the site, it was determined that two data sources, student's end-of-year exam scores and student's math growth scores were available and could be used to address the research problem. This section will first outline the changes to the above sub-sections of Section 2: The Methodology, to address the necessary changes in the analysis plan resulting from the change in data provided by the site. Finally, the results will be presented for the two analyses performed to support the original and new research questions.

Research Design and Approach Changes

As stated above, a quantitative causal-comparative approach was used to analyze the difference in high school students' geometry end-of-year exam scores differentiated by three schemes of retake timing: *short-term*, *anytime*, or *end of the quarter*. An ANOVA was appropriate for this study as the goal was to determine if there was a statistically significant difference between the three timeframes.

A second research question was added to evaluate the differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter* using the math growth data. A quantitative causal-comparative design derived logically from this problem to compare the effectiveness of the three timeframes for retakes. An ANCOVA was appropriate for this research question as the goal was to determine if there was a statistically significant difference in the math growth scores between the three timeframes when controlling for the effects of the pre-test.

For the ANCOVA analysis, the different practices for retake timing were the independent variable, the spring math growth score (post-test) was the dependent variable, and the fall math growth score (pre-test) was the covariate in the analysis. A one-way ANCOVA was used to determine if there were statistically significant differences between the means of the three independent practices currently in use. A post hoc test was used to determine which of the groups differ from the others on the math growth scores based on the different practices for retake timing (Dr. Liu, personal communication, Oct 15, 2024 & Laird Statistics, n.d.b).

Setting and Sample Changes

The local setting, sampling strategy, g*power analysis, and use of deidentified archival data did not change. The primary change was the eligibility criteria for study participants. The eligibility criteria initially included only students who had completed a retake, but the schools were not able to provide that data. So, the criteria changed to include all students' data in the identified schools, including those who had not retaken any tests. The eligibility criteria for study participants was high school Geometry and Geometry Honors students who took the course in the 2023-2024 school year in the 28 sites identified in Table 1: fourteen curriculum teams implement retakes within two to three weeks of the original test (*short-term*), six curriculum teams implement retakes anytime during the course (*anytime*), and curriculum teams implement at the end of the quarter (*end of the quarter*). Student data were included if the student had an end-of-year state exam score, a fall math growth score (pre-test) and a spring math growth score (post-test). Student data was excluded if the student did not have all three data sources (scores).

Instrumentation and Materials Changes

The instrumentation and materials discussion for the end-of-year state exam did not change. In addition to this data, the data collection required to address the second research question consisted of student fall and spring math growth scores for students who had access to a locally developed retake exam either in the *short-term* timeframe, the *anytime* timeframe, or the *end of quarter* timeframe. A request was made to the local district for deidentified archival data consisting of state end-of-year geometry exam scores, fall math growth scores and spring math growth scores. Math growth scores were assessed using the Scholastic Math Inventory assessment. Math Inventory is a computer-adaptive screener that assesses student understanding across the five strands of math aligned with the state standards for college and career readiness: number and number sense; computation and estimation; measurement and geometry; probability and statistics; and patterns, functions, and algebra (state Department of Education website, 2022).

The Math Inventory software was released in 2019 and is taken using an online testing platform (Houghton Mifflin Harcourt, 2020). The Math Inventory assessment is not a published instrument, and permission from the publisher is not required as the instrument itself will not be used in the conduct of this research study. This study used the results of previous administrations of the instrument by the district. This instrument was appropriate and sufficient for the study as it is a common and independent exam provided to all geometry students across the district, allowing the comparison of results across the different retake methodologies.

The Math Inventory assesses understanding across the five strands of mathematics aligned to state standards: number and number sense; computation and estimation; measurement and geometry; probability and statistics; and patterns, functions, and algebra (Houghton Mifflin Harcourt, 2020). Math Inventory provides a quantile score for the math skills successfully accomplished during the assessment. Both content and student achievement are measured along a single scale. The quantile measure is determined when the content and the student scores align. The quantile measure of a skill or concept is constant as the skill or concept does not change. The math growth score is determined by the change in Math Inventory quantile scores between the Fall and Spring assessments.

The quantile framework includes more than 500 skills and concepts aligned to both the Common Core Standards and individual state standards. Math Inventory is a computer-adaptive assessment that includes an item bank of more than 6,000 items across the five strands of mathematics. The Math Inventory program demonstrated an overall reliability of 0.97 and a test-retest correlation of 0.78 (Houghton Mifflin Harcourt, 2020). The Math Inventory was determined to provide stable and consistent results, to measure what it was supposed to measure, and to have a sufficiently robust item bank producing comparable results for similar students (Houghton Mifflin Harcourt, 2020) The quantile framework used in Math Inventory was based on the analysis of over 40,000 K-12 students from 22 states during two national field studies. The items in the item bank were developed using the same protocol as was used for the development of items for the quantile scale. The content-description validity is based on the explicit connections to concepts and skills described by state-level mathematics standards documents. The Math Inventory indicated expected age-related, special education, and gifted and talented related differences, and the Math Inventory results satisfactorily correlate with the results of state assessments (Houghton Mifflin Harcourt, 2020).

To evaluate student's math growth scores, the online Math Inventory assessment was administered in the fall, winter, and spring. Students who passed the assessment in the fall or winter were not required by the district to take the exam in subsequent seatings within the school year. However, multiple school sites determined that all students would take all three exams regardless of previous score. The test was administered by local instructional staff using the online Math Inventory testing application, which included an online calculator imbedded in questions where calculators were allowed (district website, 2023). Math Inventory is provided in an adaptive environment, where the next question is determined by what questions the student had previously answered correctly. The test is 30 questions long and is not timed, though it normally took between 30 and 60 minutes to complete. If a student completed 25 or more questions, a quantile score could be generated for the student. Assessments are scored by Math Inventory, and quantile scores are provided back to the district, schools, teachers, students, and parents. Quantile scores are determined by grade level and are provided for 9th and 10th graders in the chart below (these grades were chosen as they represent the majority of the high school students who took geometry within the district). For all students in all grades, the range of growth scores are between 0 and 1600 (Houghton Mifflin Harcourt, 2020).

Table 4

Math Inventory Quantile Score Ranges

Grade	Below basic	Basic	Proficient	Advanced
9	680-977	978-1132	1133-1214	1215-1459+
10	705-1003	1004-1215	1216-1248	1249-1509+

The math growth score is determined by the change in Math Inventory quantile scores between the Fall and Spring assessment opportunities. Although the math inventory software is published, it has not been released to the public and therefore cannot be attached to the study. Deidentified results of the math inventory, in the form of math growth scores, were provided by the district for this study.

Data Collection and Analysis Changes

The data collection required to address the research questions consisted of student scores on the state end of year exam which are provided as an actual total score out of 600 points and student fall and spring math growth scores for school curriculum teams who implemented retakes using either the *short-term* timeframe, the *anytime* timeframe, or the *end of quarter* timeframe. The data collection process began with a request for deidentified archival data from the district to retrieve end-of-year test scores and fall and spring math growth scores for students taking Geometry or Geometry Honors in the previously identified fourteen curriculum teams who implemented retakes within two to three weeks of the original test (*short-term*), the six curriculum teams who implemented retakes anytime during the course (*anytime*), and the eight curriculum teams who implemented retakes at the end of the quarter (*end of the quarter*).

There were two steps to prepare the data for analysis. The first step was to identify the students from the previously identified schools who had taken both the end-of-year state exam and both the fall and spring math growth tests. Only students who had scores from all three tests were included in the analysis. The second step to prepare the data for analysis was the creation of an excel spreadsheet with four columns. The first column was coded with either a '1', '2', or '3' to represent which implementation was used. Within the first column, students who were from a school implementing the *short-term* timeframe were coded with a '1', students who were from a school implementing the *end of quarter* timeframe were coded with a '2', and students who were from a school implementing the *anytime* timeframe were coded with a '3'. The second column was the student's state end-of-year exam score and the third and fourth columns were the fall and spring math growth scores. The ANOVA then compared the end of year exam scores by implementation and the ANCOVA compared the spring math growth score by implementation when controlling for the pre-assessment (fall math growth) score.

The data for research question two was analyzed using the IBM Statistical Package for the Social Sciences (SPSS), Version 28. The independent variable was a nominal variable and was defined by the different practices for retake timing. The dependent variable was student post-test (spring) math growth scores and was measured on a continuous scale from 0 – 1600, and the covariate was the pre-test (fall) math growth scores and was measured on a continuous scale from 0 – 1600. Table 4 summarizes the variables used in the ANCOVA analysis.

Table 5

ANCOVA Variables

	Research question	Variables	Data analysis
RQ2	What are the differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*?	IV: Retake Timing *Short-term* 1 *End of Quarter* 2 *Anytime* 3 DV: Post-test (spring) math growth scores Continuous from 0 – 1600 Covariate: Pre-test (fall) math growth scores Continuous from 0 – 1600	One-way ANCOVA

A one-way ANCOVA was used to determine if there were statistically significant differences between the means of the three independent practices currently in use. A post hoc test was used to determine which of the groups differed from the others on the high school geometry end-of-year state exam scores based on the different practices for retake timing (Dr. Liu, personal communication, Oct 15, 2024 & Laird Statistics, n.d.b).

Assumptions, Limitations, Scope, and Delimitations Changes

Assumptions Changes

The following assumptions are in addition to the assumptions listed above. The first four of ten statistical assumptions for using ANCOVA were met within the design: a) the dependent variable was measured on a continuous scale, b) the independent variable was from two or more independent groups, c) there is one covariate variable measured at the continuous level, and d) there were independence of observations as there is no relationship between the groups since they were from different schools within the district. The other six of ten assumptions, the covariate should be linearly related to the dependent variable at each level of the independent variable, homogeneity of regression slopes, there is normality of within-group residuals, homoscedasticity, homogeneity of variances, and no outliers were assessed after data collection and during the data analysis (Laird Statistics, n.d.b). In addition, it is assumed the math growth assessments used by the district were administered with fidelity and were reliable.

Limitations, Scope and Delimitations Changes

The scope of this study is Geometry and Geometry Honors students who took the end-of-year state exam, and fall and spring math growth assessments, during the 2023-2024 school year from the identified schools that implemented retakes on locally developed tests using the three different time frames. A limitation of this study was the inability to identify specifically which students had accessed a retake exam or the inability to exclude those students who did not take a locally developed retake exam. Additionally, this study assumed that the remediation provided to students between the original exam and its retake was consistent across the district.

End-of-Year State Exam ANOVA Results

To answer the first research question, what are the differences between high school students' geometry end-of-year state exam scores differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*, an ANOVA analysis was executed on the deidentified data provided by the site. The resulting sample sizes after reducing the original data set to only those students who had an end-of-year state score and a pre-test (fall) and post-test (spring) math score were *short-term* (n = 500), *end of the quarter* (n = 510), and *anytime* (n = 225) for a total of 1235 students across 19 curriculum teams from 11 high or secondary schools.

Three assumptions, no outliers, an approximately normal distribution of the dependent variable, and homogeneity of variances, were assessed after data collection and during the data analysis (Laird Statistics, n.d.a). The results of testing the assumption that there were no significant outliers in the groups of the independent variables resulted in a finding that there were outliers in the original data set. The outliers constituted 3% of the data set, or 38 out of the original 1,235 student samples (14 from Method 1, 19 from Method 2, and 5 from Method 3). An inspection of a boxplot for values greater than 1.5 box-lengths from the box's edge confirmed the removal of the outliers, and the subsequent analysis confirmed the ANOVA results were not materially affected by the removal of the outliers. The resulting sample sizes after reducing the original data set without outliers were *short-term* (n = 486), *end of the quarter* (n = 491), and *anytime* (n = 220) for a total of 1197 students across 19 curriculum teams within 11 high and secondary schools.

The assumption that the dependent variable should be approximately normally distributed for each group of the independent was assessed using the Shapiro Wilk's test. The Shapiro-Wilk's test determined that the data were not normally distributed for

Method 1 (p = .01) and Method 2 (p = .031) but were normally distributed for Method 3 (p = .171). If the sample sizes are greater than 50, it is better to assess normality using a Normal Q-Q Plot because the Shapiro-Wilk's test assesses minor deviations from normality as statistically significant (Laird Statistics, n.d.a). Based on the Normal Q-Q Plots, the data is very close to the line, demonstrating a normally distributed data set for each group of the IV as demonstrated below in Figures 2 through 4, the Normal Q-Q Plots by method.

Figure 2

Normal Q-Q Plot of End-of-Year Score for Method 1

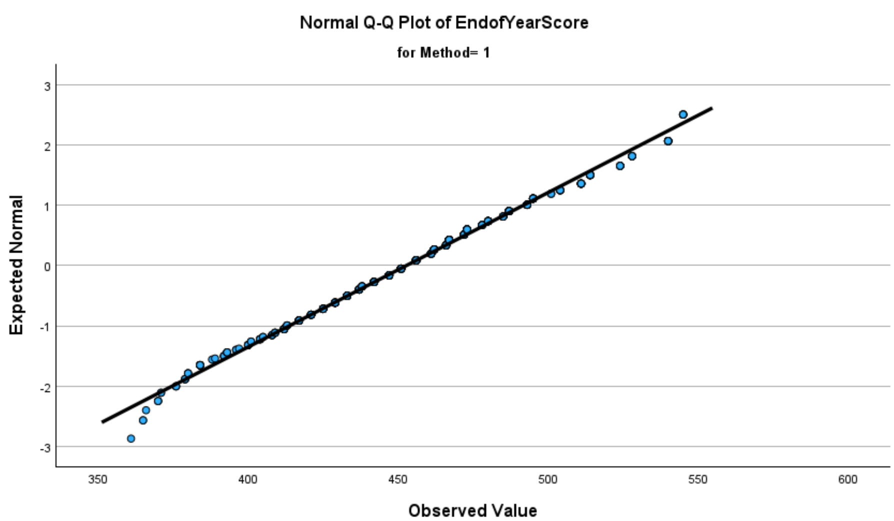


Figure 3

Normal Q-Q Plot of End-of-Year Score for Method 2

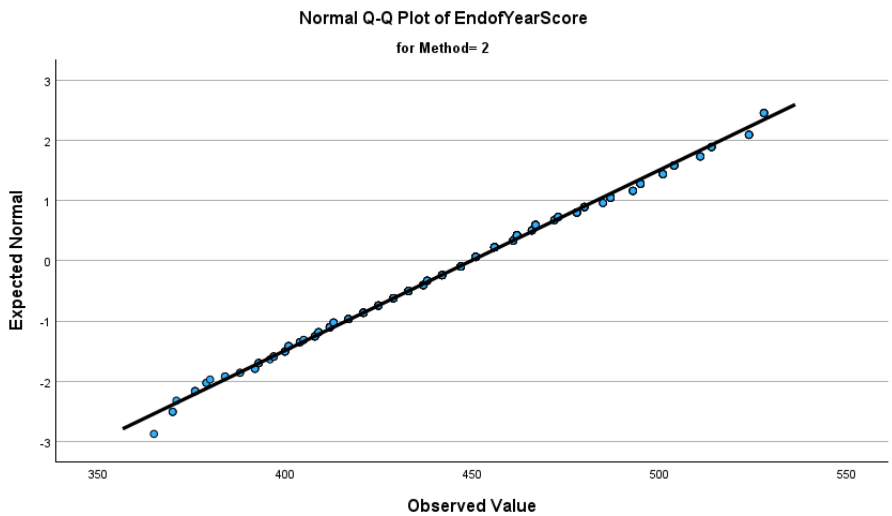
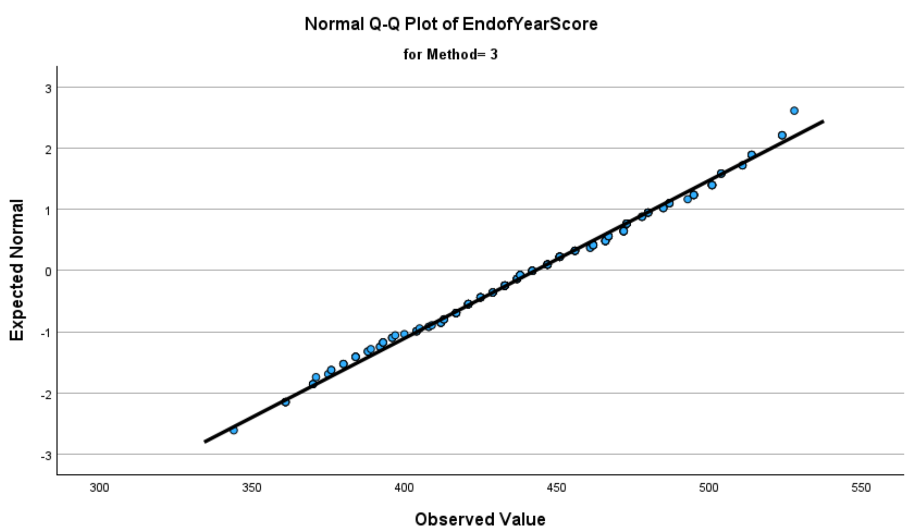


Figure 4

Normal Q-Q Plot of End-of-Year Score for Method 3



The final assumption of homogeneity of variances was tested using Levene's test of equality of variances. The Levene's test for equality of variance determined that the assumption of homogeneity of variances was violated ($p = .001$). Since the assumption of homogeneity of variances was violated, the results were evaluated using the Welch ANOVA and the post hoc results were evaluated using the Games-Howell test.

An overall review of the descriptive statistics demonstrated that the end-of-year state scores decreased as the amount of time given between the original test and its retake was increased, from *2-3 weeks* ($n = 486$, $M = 452.8$, $SD = 39.0$), to *end of quarter* ($n = 491$, $M = 449.8$, $SD = 33.4$), to *anytime* ($n = 220$, $M = 443.0$, $SD = 38.8$). The Welch ANOVA determined that the amount of time used between the original test and its retake was statistically significantly different for different amounts of time, Welch's $F(2, 577.996) = 4.778$, $p = .009$.

The Games-Howell post hoc test determined there was a decrease in the end-of-year state scores from 452.78 ± 39.04 in the *2-3 weeks* to 443.02 ± 38.78 to the *anytime* score, a decrease of 9.77 (95% CI, 2.34 to 17.19) which was statistically significant ($p = .006$). The full set of Games-Howell post hoc test results are provided in Table 6 below.

Table 6

Games-Howell Post Hoc Test Results

	(I) Method	(J) Method	Mean difference (I-J)	Std. error	Sig.	95% confidence interval Lower bound	Upper bound
Games-Howell	1	2	2.99	2.33	.41	-2.47	8.44
		3	9.77*	3.16	.01	2.34	17.19
	2	1	-2.99	2.33	.41	-8.44	2.47
		3	6.78	3.02	.07	-.32	13.88
	3	1	-9.77*	3.16	.01	-17.19	-2.34
		2	-6.78	3.02	.07	-13.88	.32

* The mean difference is significant at the 0.05 level.

Since the group means were statistically significantly different ($p < .05$), we can reject the null hypothesis and accept the alternative hypothesis, there are differences between high school students' geometry end-of-year state exam scores differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*.

Math Growth Scores ANCOVA Results

To answer the second research question, what are the differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*, an ANCOVA analysis was executed on the deidentified data provided by the site. The analysis was performed on the original data set without the outliers generated during the ANOVA analysis, resulting in group sizes of *short-term* (n = 486), *end of the quarter* (n = 291), and *anytime* (n = 220), and representing 1197 students across 19 curriculum teams from 11 high and secondary schools.

Six of ten assumptions, the covariate should be linearly related to the dependent variable at each level of the independent variable, homogeneity of regression slopes, there is normality of within-group residuals, homoscedasticity, homogeneity of variances, and no outliers were assessed after data collection and during the data analysis (Laird Statistics, n.d.b). The results of the assumption that the covariate should be linearly related to the dependent variable at each level of the independent variable determined was a linear relationship between the Fall and Spring Math Inventory scores for each retake method, as determined by visual inspection of a scatterplot. Data analysis determined there was homogeneity of regression slopes as the interaction term was not statistically significant, $F(2, 1177) = .021$, $p = .979$.

The assumption that the dependent variable should be approximately normally distributed for each group of the independent was assessed using the Shapiro Wilk's test. The Shapiro-Wilk's test determined the standardized residuals were normally distributed for all methods, Method 1 ($p = .101$), Method 2 ($p = .215$), and Method 3 ($p = .468$). After a visual inspection of the standardized residuals plotted against the predicted values by method, it was determined that the data demonstrated homoscedasticity. The assumption of homogeneity of variances was tested using Levene's test of equality of error variances. The Levene's test determined there was homogeneity of variances, ($p = .244$).

The results of testing the assumption that there were no significant outliers in the standardized residuals resulted in a finding that there were outliers. The outliers constituted 1% of data set, or 14 out of the original 1197 student samples, (11 data points greater than 3 SDs and 3 data points below -3 SDs). The outliers were removed, and the subsequent analysis confirmed the ANCOVA results were not materially affected by the removal of the outliers. The resulting sample sizes after reducing the original data set without outliers were *short-term* (n = 478), *end of the quarter* (n = 488), and *anytime* (n = 217) for a total of 1183 students across 19 curriculum teams from 11 high and secondary schools. The ANCOVA assumptions analysis provided above was based on the data set without outliers.

An overall review of the descriptive statistics demonstrated that the unadjusted means for the spring math growth scores (post-test) decreased as the amount of time given between the original test and its retake was increased, see Table 7 below. An overall review of the descriptive statistics demonstrated that the adjusted means for the spring math growth scores (post-test) were basically the same for 2-3 weeks and quarterly but decreased when the amount of time given between the original test and its retake was increased to anytime, see Table 7 below.

Table 7

Unadjusted and Adjusted Means and Variability for Spring Math Growth Scores (Post-Test) with Fall Math Growth Scores (Pre-Test) as Covariate

		Unadjusted		Adjusted	
	N	*M*	*SD*	*M*	*SE*
2-3 Weeks	478	1111.86	123.95	1103.08	4.44
End of Quarter	488	1099.31	124.29	1106.08	4.39
Anytime	217	1084.47	132.81	1088.60	6.57

Note: *N* = number of participants, *M* = Mean, *SD* = Standard Deviation, *SE* = Standard Error

An ANCOVA was run to determine the effect of the amount of time given between original school-based test and its retake on the spring math growth score (post-test) after controlling for the fall math growth score (pre-test). After adjustment for the fall math growth score (pre-test), the ANCOVA analysis determined that the amount of time used between the original test and its retake was not a statistically significantly difference for different amounts of time on the spring math growth score (post-test), $F(2, 1179) = 2.537$, $p = .08$, partial $\eta^2 = .004$. Therefore, we cannot reject the null hypothesis and there are no differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*.

Discussion

This study attempted to answer the question if there were differences in learning, as measured by both end-of-year state exam scores or high school math growth scores, depending on the length of time given to students between the original school-based assessment and its retake. The section summarizes the results of the two analyses.

The ANOVA analysis on end-of-year state exam scores determined there was a statistically significant difference in the end-of-year exam scores based on what type of retake was implemented, but the post-hoc analysis revealed that, although there wasn't a statistically significant difference between 2-3 weeks and quarterly, there was a statistically significant difference between the two extremes, 2-3 weeks and anytime. In addition, although the expectation was that scores would increase if students were given additional time to learn the material and prepare for the retake, the analysis demonstrated the scores decreased from the 2-3 week to the anytime scores. It should be noted that, though there was a statistical difference, the actual difference in mean score was only 9.7

points, or approximately 2% of the overall score. This is consistent with the results of studies that demonstrated an improvement in exam results using the two-week timeframe between the original assessment and its retake (Adeniji et al., 2018; Ee et al., 2018; Halperin, 2020; Morphew et al., 2020).

In comparison, the ANCOVA analysis on math growth scores determined there was not a statistically significant difference based on the implemented retake when controlling for the pre-assessment score. Similar to the analysis of the end-of-year state scores, the unadjusted results for the math growth scores showed a decrease in mean from 2-3 weeks to the anytime scores. However, when adjusted for the pre-assessment score, the mean scores demonstrated a slight increase from 2-3 weeks to quarterly but decreased when the amount of time increased to anytime. This is consistent with the results of studies that showed either no measurable difference in grades or mixed results (Carlisle, 2020; Harsy & Hoofnagle, 2020) and the results of previous studies that showed the strongest predictors on end-of-year high stakes testing were prior academic achievement and grades (Early et al., 2016; Lin & Chen, 2020; Pollio & Hochbein, 2015).

Although the theoretical framework does not directly address the length of time between the original assessment and its retake, in Bloom's theory, the retake is supposed to be provided before the start of the next unit, which would imply the use of a shorter timeframe between assessments (Bloom, 1968). As mentioned previously, the site does not implement Bloom's theory with fidelity, as additional academic support and the opportunity to take a retake offered to students are not provided prior to the beginning of new instruction, but in parallel with new instruction.

Conclusion

This study attempted to answer the question if there were differences in learning, as measured by both end-of-year state exam scores or high school math growth scores, depending on the length of time given to students between the original school-based assessment and its retake. Based on the results from the analyses of both end-of-year scores and math growth scores, the scores showed similar or a slight reduction when students are given between 2-3 weeks and the end of the quarter but demonstrate a statistically significant reduction in learning outcomes on the end-of-year state exam when the time allowed between the original assessment and its retake is increased. These results imply that a shorter time frame between the original test and its retake might yield better learning outcomes. However, these results are flawed and inconclusive as the dataset included both the scores for students who took retakes and those who did not.

The district policy, leaving it to the curriculum teams to define the amount of time between the original test and its retake, has resulted in many different practices implemented within the district and a slight decrease in high school students' geometry end-of-year state exam scores. This discrepancy between accepted theory and practical application was the genesis for the identified gap in practice. Implementing the recommended amount of time between the test and retake that maximizes student acquisition and demonstration of content knowledge would lead to improved student learning and mastery of content material while increasing consistency of practice across the district. Based on the results of this study, and the difficulties acquiring a quality data set to analyze the current implementation within the district, a policy paper to inform district leaders of the study results, including recommendations on improved access to a quality data set to evaluate how effectively students are learning from the implementation of test retakes, was the appropriate project for this work.

Section 3 summarizes the project paper with research-based recommendations based on a review of the professional literature and the findings of the study.

Section 3: The Project

Introduction

The research study investigated whether there were any differences in learning as demonstrated by either the geometry end-of-year state exam or end-of-year math growth scores to determine if there was a length of time between a site-developed original exam and its retake that improved learning. There were three different lengths of time studied that different schools were using: *short-term (2 – 3 weeks), end of the quarter*, and *anytime*. The results indicate that there may be some benefit to a shorter time period between retakes, but the results are flawed and inconclusive because of problems with the validity of the data. The problem is that the data included students who had and who had not used retake assessments and should have only included those who had retaken assessments to evaluate if their learning had benefitted from the retakes. Even though the results of this study are flawed it is worth noting that findings did indicate that there is a significant difference in end of year test scores between schools that had a policy of 2-3 weeks and anytime, with the longer amount of time decreasing scores. While it appears that longer time frames result in decrease in scores because of the flawed data set it is unclear what timeframe would be best to recommend at this time. Because of this policy recommendations focus on attaining a dataset that only includes students who have retaken tests and improving consistency of policy in syllabi.

A policy paper was developed (see Appendix A) to address these findings and to provide recommendations to the site based on this quantitative research. The policy paper contains three recommendations for the site. The first recommendation is to require all summative grades in the electronic grade book include a code if they have been modified by a retake or grade replacement. The second recommendation is to charge the district's 25-26 Secondary Grading Committee to use the improved summative grade data generated from the first recommendation to validate the results of this study and determine if the window for retakes should be limited to a shorter timeframe. The final recommendation is to require either a minimum set of information be included in every course syllabus or create a common syllabus outline for use in the district. If adopted, these recommendations would provide better insight into who is accessing the retake opportunity, provide better data which the district could use to evaluate its grading policies, and provide increased consistency across the district.

Rationale

The district policy leaving it to the curriculum teams to define the amount of time between the original test and its retake has resulted in many different practices implemented within the district and a slight decrease in high school students' geometry end-of-year state exam scores. This discrepancy between accepted theory and practical application was the genesis for the identified gap in practice. As the results of this research do not support the proposal of a specific initiative, but instead influence the implementation of a policy, a policy paper is the most appropriate project to share the results of this study with district leadership (UMSU Advocacy Service, 2018). The results of this study and the difficulties acquiring a quality data set to analyze the current implementation within the district are the primary foci of the policy paper (Appendix A).

This policy paper followed the recommendations of Stowe (2021) and the American Library Association (2007), providing both the description and urgency of the issue and a proposed course of action supported by the research findings from this study and current literature. The intended audience for this policy paper is the local district leadership, including but not limited to the superintendent, the chief academic officer, and the assistant superintendent, instructional services department. A policy paper recommending specific policy changes was the appropriate project for this study is it provides the background for the identified issue, the research results from this study, relevant current literature, and a set of specific recommendations with an implementation plan to address the issues, with a goal on improving both district processes and increasing the district's ability to evaluate their practices leading to improved student learning.

Review of the Literature

A literature review was conducted to support the development of the recommended policy paper. The focus of the policy paper was the three recommendations: 1) require all summative grades include a code in the electronic gradebook if they have been modified by a retake or grade replacement, 2) charge the district's 25-26 Secondary Grading Committee to use the improved summative grade data generated from the first recommendation to validate the results of this study and determine if the window for retakes should be limited to a shorter timeframe, and 3) require either a minimum set of information be included in every course syllabus or create a common syllabus outline for use by every school in the district. The literature review encapsulates a comprehensive evaluation and assessment of relevant research from scholarly, peer-reviewed journals identified through online searches of databases including ERIC, Education Research Complete, and Google Scholar. Relevant articles between the years 2020 and 2025 were included. The Walden University Library provided access to a wide variety of information from peer-reviewed articles. The following keywords were used: mastery learning, standards-based grading, mastery grading, reassessments, retakes, mastery assessments, policy paper, white paper, policy briefs, educational policy, and syllabus. The review of literature is separated into four sections: policy paper development, mastery learning unintended consequences, lack of research on mastery learning implementation, and syllabus development.

Policy Paper Development

Implementing a common amount of time between the test and retake that maximizes student acquisition and demonstration of content knowledge might lead to improved student learning and mastery of content material while increasing consistency of practice across the district. At this time, it is still unclear what the recommended amount of time should be. There is some evidence that a shorter time period might have some benefit, but the results of this study are inconclusive because the data set is flawed as it includes all students instead of only those that took retakes. Based on the results of this study, and the difficulties acquiring a quality data set to analyze the current implementation within the district, a policy paper to inform district leaders of the study results, including recommendations on improved access to a quality data set to evaluate how effectively students are learning from the implementation of test retakes, was the appropriate project for this work (Antonopoulou et al., 2021). Although there are two main foci for policy papers, designing policy and policy content, this policy paper will be focused on policy content (Siddiki & Curley, 2022). The policy paper needed to objectively focus on the specific recommendations resulting from the research study with the supporting analyses and current literature as the rationale for the recommended policy change (Arnautu & Dagenais, 2021; Mayer, 2021). Additionally, the policy paper needed to be written for the target audience, with a stand-alone executive summary written to both capture the interest of the audience and provide enough information to support the recommendation if there is not enough time to read the entire document (Mayer, 2021; Stowe, 2021). Since education policy matters to many stakeholders, the policy recommendations acknowledged the importance of these stakeholders with their inclusion in the timeline implementation (Carstensen & Emmenegger, 2023; Siddiki & Curley, 2022) Although policy needs to be evidence driven, it also needs to be both equally robust to the intended and unintended consequences of policy alternatives (Desouza & Lin, 2011; Ndihokubwayo, 2023).

Mastery Learning Unintended Consequences

Themes in the current literature continue to provide convergent evidence of the following unintended consequences of mastery learning: 1) student's arriving unprepared for initial exams, 2) student procrastination leading to increased course failures or drop-outs, 3) increased instructional workload, 4) increased burden and time management required of students with the resultant increase in student stress, and 5) increased student focus on grades, not learning (Buraphadeja & Srisarkun, 2024; Chamberlain Jr., 2023; Curley & Downey, 2024; Ford et al., 2024; Herman et al., 2020; Huey et al., 2022;

Knight & Cooper, 2019; Linhart, 2020; McKell & Danowitz, 2020; Venkatesh & Piercey, 2021). In addition, the implementation of mastery learning has continued to result in a change to the distribution of grades, with a higher percentage of students earning an A but also a greater percentage of students who either withdraw from courses or earn a D or F (Herman et al., 2020), and questions on whether mastery learning either reduces student stress (Lewis, 2020; Linhart, 2020) or improves student achievement (Buraphadeja & Srisarkun, 2024; Fergus & Petrick Smith, 2022; Morris & McKenzie, 2025; Song et al., 2022; Supriya et al., 2024). The themes evidenced in the literature were raised as issues by both students and teachers during both the 23-24 and 24-25 Secondary Grading Advisory Committee meetings. Although the mastery learning process has many potential positive impacts on student learning, it has also resulted in unintended consequences that affect its implementation.

Lack of Research on Mastery Learning Implementation

There continues to be a lack of quantitative research on how teachers implement the mastery learning process and if there is a demonstration of improved student learning and retention when using this process (Buraphadeja & Srisarkun, 2024; Huey et al., 2022; Swanson et al., 2025). Current research relies on either discussions of how to implement mastery learning or student perception data (Butler, 2023; Chamberlain Jr., 2023; Curley & Downey, 2024; Harsy et al., 2021; Linhart, 2020; Tripp et al., 2025; Venkatesh & Piercey, 2021). Despite the continued implementation of standards-based grading, the practice itself remains largely unvalidated with limited research on student outcomes (Link & Guskey, 2022; Townsley & Buckmiller, 2020; Zhang, 2024). This study was an attempt to garner data on actual student outcomes but was flawed because it could not capture the data for only those who retook exams. School systems are continuing to modify their instructional design to include mastery grading without the data to validate both the best practice for implementation and to validate if the use of mastery grading is resulting in improved student learning (Fergus & Petrick Smith, 2022; Huey et al., 2022). Research on mastery learning is affected by lack of quality data due to absence of baseline data to compare against, the inclusion of non-representative data samples, and the inability to compare with traditional grading due to variations in course delivery and demographics (Buraphadeja & Srisarkun, 2024). The first recommendation focuses on a method to provide the specific data necessary to allow creation of a quality data set that could be used to evaluate the effectiveness on the retake policy. This data set would be able to differentiate between students who had accessed a retake exam from those who had not. If the district does not collect the correct data to evaluate performance, then the evaluation of benefits to students, staff, and parents, and the ability to validate that there are improved learning outcomes is severely limited.

Syllabus Development

During the evaluation of the syllabi from all the geometry curriculum teams, the specifics of how retakes were implemented could only be determined in 28 of 48 syllabi. This led to the third recommendation to create a common syllabus outline for all curricula to create as much consistency as possible across the school district. A syllabus template provides consistency across curriculum, ensuring students are held to a similar standard (Wagner et al., 2023). A syllabus needs to clearly state the practices and policies of the district, include positive and inclusive language to promote diversity and equity, and provide consistent course requirements and learning outcomes (Chen et al., 2023). Clear and concise communication of course expectations provide clarity to students and parents (Wagner et al., 2023). Also necessary in the syllabi are the inclusion of specific language on mastery requirements. This would include limits on reassessment opportunities, discussions on the minimum standard of work to access reassessment opportunities, and forced delays between reattempt opportunities. The goal would be to ensure student focus on true mastery of content material (Cilli-Turner et al., 2020). A consistent syllabi design will reduce the impression that different schools are implementing different processes, while also providing clear documentation that curriculum teams are following the guidelines provided by the district's grading policies.

Project Description

A policy paper was developed for the superintendent, chief academic officer, and the assistant superintendent, instructional services department at the local site. The content of this paper will be presented to district leadership who are currently working on modifying the district's secondary grading policy. Based on the research conducted, the recommendations from the 23-24 and 24-25 Secondary Grading Committees (District, 2024, 2025), and a review of literature, it is recommended that the district implement the following policy modification and recommendations:

1. Require all teachers to include a code in the electronic grade book on any grade that has been modified by either an improved grade from a retake or by a grade replacement.

2. Charge the district's 25-26 Secondary Grading Committee to use the improved summative grade data generated from the first recommendation to validate the results of this study and determine if the window for retakes should be limited to a shorter timeframe.
3. Recommend either a minimum set of information be included in every course syllabus or create a common syllabus outline that is used by every school in the district.

The project description is broken into three sections: 1) needed resources, existing supports, barriers, and solutions; 2) project implementation and timeline; and 3) roles and responsibilities.

Needed Resources, Existing Supports, Barriers, and Solutions

This discussion of needed resources, existing supports, barriers, and solutions are organized by the numbered recommendations above.

For the first policy modifications, adding a code in the gradebook, there is no additional monetary cost associated with the implementation of these changes. The existing electronic gradebook can implement this recommendation as it is currently configured. There will be a one-time cost associated with writing software to produce a report that will allow access to the data so it can be evaluated at both the local and district levels. There will also be opportunity cost associated with staff time who will have to include codes for all summative grades, and for school-based administrators to learn how to access and interpret the data for use at school sites. Implementation of the codes can be done immediately, but the software to access the data can be implemented within a threemonth window, as it will have to be written and tested. Supports are two-fold, inclusion in the 25-26 district grading policy will ensure all teaching staff are informed of the requirement, and explanation for the change can be disseminated through the already existing middle and high school principals' association to school sites. The primary barrier is teacher buy-in for the need to add this extra step for every summative grade in the gradebook. This should be addressed up front by the need to analyze data to evaluate whether the grading changes being made are successfully creating an increase in student learning.

For the second recommendation, charging the district's 25-26 Secondary Grading Committee to use the improved summative grade data generated from the first recommendation to validate the results of this study and determine if the window for retakes should be limited to a shorter timeframe, there are no additional resources necessary to implement this change. This requirement to analyze and evaluate this data can be included as a task for the 25-26 school year's secondary grading committee, with the results considered for possible inclusion in the 26-27 school year.

The third recommendation, a common course syllabus, would require personnel from across the division to develop the common syllabus, the middle and high school principal's association to approve the document, followed by distribution to all schools for implementation. The new superintendent has been looking for ways to increase consistency across the division, and this recommendation supports this focus. Implementation of this recommendation will require support from administration, but the task is relatively simple, and the impact is minimal to administrators and instructional staff.

Project Implementation and Timeline

The timeline for implementation is based on the projected date for when the policy statement will be provided to the district and the district's willingness to begin initial implementation. The implementation of the first policy recommendation can be made immediately as part of the 25-26 District Grading Policy. The second recommendation can be included as a task for the 25-26 Secondary Grading Committee for implementation in the 26-27 school year. Due to the projected date for when the policy recommendation will be forwarded to the district, the third recommendation should be targeted for development during the 25-26 school year and implementation in the 26-27 school year.

Roles and Responsibilities

Conducting this research project and forwarding the policy statement to the district are my primary role and responsibility. Upon receipt of the policy statement, the superintendent, the chief academic officer, and the assistant superintendent, instructional services department will determine which recommendations are to be adopted. If the district adopts these recommendations, I will be available as an advisor, or member of the advisory committees tasked with implementing these policy recommendations.

Project Evaluation Plan

The evaluation of the first two recommendations will be goals-based. A goals-based evaluation is appropriate as the function of the evaluation is to improve student learning and the evaluation of student learning (Gouvard & Durand, 2023; Zhu et al., 2023). The goal of the first recommendation is to provide a method that allows for the collection of a quality data set that can be used to evaluate the impacts to student learning based on the changes made to the grading system implemented by the district. The goal of the second recommendation is to charge the district's 25-26 Secondary Grading Committee to use the improved summative grade data generated from the first recommendation to validate the results of this study that imply a shorter time frame between the original site-based exam to its retake provides an increase to both the geometry end-of-year state exam and students' math growth scores.

The evaluation of the last recommendation will be outcome-based. The focus of the third recommendation is the creation of a common syllabus outline for the district. An outcome-based evaluation is appropriate as the focus is on measuring performance through the achievement of an outcome, in this case, the creation of and distribution of the common syllabus outline (Dr. Preeti Oza, 2021; V. Lin et al., 2022; Schalock, 2001). A draft common syllabus is included in the policy paper to facilitate the process.

The key stakeholders for all three recommendations are the superintendent, chief academic officer, the assistant superintendent, instructional services department, the middle and high school principal's association, and the instructional and administrative staff.

Project Implications

If adopted, these recommendations would provide better insight into who is accessing the retake opportunity, provide better data which the district could use to evaluate its grading policies, and provide increased consistency across the district. These changes would provide specific and measurable data for the district and the next Secondary Grading Committee to use to evaluate the effectiveness of current policies while increasing the validity of the results of any analysis performed. In addition, this data can be used to assess the validity of concerns raised by instructional staff during both the 23-24 and 24-25 Secondary Grading Committee sessions on the unintended consequences in student behavior that have risen due to the policy, such as arriving unprepared for exams, procrastination leading to increased course failures or drop-outs, focusing on grades not learning, and increased time management required leading to increased student stress.

Implementing these recommendations also provides valuable information to parents and guardians on whether their students are accessing their retake opportunities to improve their grades, and if they are, whether their students grades are improving. Once this additional data is created, the district can begin the analysis of the effectiveness of the current grading system, and the groundwork will have been created that allows the evaluation of any subsequent changes the district makes as it moves towards standards-based-grading. The results of these recommendations support positive social change by improving both district processes and increasing the district's ability to evaluate practices leading to improved student learning.

There is support for having a retake policy from parents, students and some staff, but there is little evidence on best method for implementation or the impact on student learning resulting from the practice. The 23-24 and 24-25 Secondary Grading

Committees (district reports, 2024, 2025) attempted to identify best practices for grading to improve consistency across the district but were limited by the data they had available, relying primarily on grade trends and qualitative data. The research conducted attempted to determine the optimum amount of time between the original test and its retake, but the limitations of the data reduced the validity of the results. The district is moving down a path towards implementing standards-based grading without the ability to evaluate if the changes made are improving student learning. If the district does not collect the correct data to evaluate performance, then the evaluation of benefits to students, staff, and parents, and the ability to validate that there are improved learning outcomes is severely limited.

Section 4: Reflections and Conclusions

The following section summarizes the project strengths and limitations and recommendations for alternative approaches. This is followed by my personal reflections on scholarship, project development and evaluation, and leadership and change considerations. Finally, a reflection on the importance of the work, possible implications, applications and directions for future research, and the conclusion are provided.

Project Strengths and Limitations

Project strengths are based on the identification of issues related to the access of quality data that affected not only the research study but affect the site's ability to analyze its own processes, as well as improving the consistency of the documentation used across the district. The issues faced attaining quality data from the site resulted in the subsequent understanding that the site does not access quality data itself to use in evaluations of its own practices. I was able to gather sufficient data to perform an evaluation of the practice of amount of time between the original site exam and its retake, but these results have limited validity. As such, the focus of the policy paper turned to a limited discussion on the results of the data and analysis with a more focused discussion on the need, and potential methods available, to gather quality data. If the district does not collect the correct data to evaluate performance, then the evaluation of benefits to students, staff, parents, and administration, and the ability to validate that there are improved learning outcomes from actions taken by the administration are severely limited.

Recommendations for Alternative Approaches

The purpose of this quasi-experimental quantitative study was to investigate if there were differences between high school students' geometry end-of-year state exam scores or between high school students' geometry math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*. As mentioned, there were issues performing this analysis due to the lack of quality data available from the site. In the professional literature there are numerous discussions of the qualitative value of implementing a retake process, but there is very little quantitative research to further evaluate the process. Several options exist that could have been used instead of the choices made within this research.

An alternate approach considered was to perform the analysis using a reduced set of students at fewer school sites. This design would have individual teachers provide the data on which students specifically in their classes accessed the retake opportunity, allowing for me to include only students who had accessed the retake opportunity in the data set. Although the results from this data set have greater validity, this choice was not taken due to the concern regarding student data protection and generalization issues associated with the much smaller sample size.

Another alternative approach considered was to change to a qualitative approach and ask for teacher and administrator input on the amount of time provided between the original exam and its retake. This approach was not taken as it was the researcher's focus to provide a quantitative basis to the academic discussion of retakes.

The final study implemented is a combination of two alternate approaches for the originally proposed study. First, the analysis was performed on the end-of-year state exam data without differentiating whether a student had or had not accessed their retake opportunity. In addition, the evaluation of math growth data was included as another method for accessing the impact of the amount of time between tests on student learning. Due to the limitations in the original study, better data must be collected and analyzed before alternative solutions can be created for the local problem.

Scholarship, Project Development and Evaluation, and Leadership and Change

The process of working through the EdD degree has been eye opening and personally transformational process from a scholarship, project development and evaluation, leadership, and social change perspective.

Scholarship and Project Development and Evaluation

I have always believed in the importance of education and the opportunities that education can provide. The course work provided an opportunity to interact with students from all different types of educational backgrounds, who worked within education via a wide range of positions and who are from all over the world, created a much deeper insight into the vast array of issues, challenges, and strengths that are encompassed in the field of education. Through the course work's focus on current literature, I was spurred to evaluate my practice from different lens, evaluating areas for improvement and sharing these improvements with my curriculum team and department.

The research study performed was also a source of many lessons, several of which occurred through the problems and challenges that were faced in the execution of the study. I came into the program with a research project in mind, but through the program, I learned how to better create alignment within a research study, and how to look at the literature review as not just something that supports the research process, but how my research might help inform future research studies. Finally, the many challenges I had defining and accessing a data set that could be used to address my research study not only helped me better understand the study site while also providing a much deeper understanding of the statistical methods that could be used to address the research problem.

Leadership and Social Change

Completing this EdD research study and project are not the culmination of my educational experience but the start of a new one. Learning how to ask good questions, defining and if necessary, creating good data sets that can be used to address questions, and providing leadership with quality analyses are key to understanding whether the practices being put into place are producing the intended results.

The original social change significance of this study was that the district had implemented a policy resulting in many different implementations of the amount of time given between the original school-based test and its retake. This ultimately resulted in a slight decrease in high school students' geometry end-of-year state exam scores. Due to the validity issues with the data, it is not completely clear if the effects of amount of time between the original school-based test and its retake influence high school students' geometry end-of-year state exam scores, though it initially appears that a shorter time frame is better. The true social change significance of this research study is not just answering the research questions but identifying that the district is not collecting the correct data it needs to evaluate its processes. Without good data, the district will not be able to improve both district processes and increase the district's ability to evaluate changes in practice that result in improved student learning.

Reflection on Importance of the Work

The research study was important as it started the process of trying to determine what processes should be used to implement test retakes to improve student learning. Most of the literature and research around retakes and standards-based grading is qualitative in nature. This research study took a first step from evaluating what we think is true from a qualitative perspective and towards research that is evaluated based on the results of actual practice. This research study determined that the amount of time between the original test and the retake that had the highest influence on high school students' geometry end-of-year state exam scores was the shorter timeframe of either 2-3 weeks or quarterly. Unfortunately, there were validity concerns with the answer resulting from an inability to access a quality data set. The more important result of this study was the identification of the fact that the district cannot produce a quality data set and is currently analyzing the effects of the changes in grading policy using data that cannot be used to provide specific and accurate analytical results. The district is data rich but is not collecting the data it needs to evaluate its own processes.

Implications, Applications, and Directions for Future Research

The position paper has implications to the study site as it identifies a major issue, that the district is not collecting the correct data to evaluate the impacts of grading processes under use. Additionally, the research study provides a first step towards evaluating best practices in the implementation of retakes, that a shorter timeframe provides better impact on student learning than a longer timeframe, although there were validity concerns with the answer resulting from an inability to access a quality data set.

Future research should replicate this study with a better data set to see if the results are consistent, addressing the validity issues associated with the results of this study. Additionally, another practice used with retakes is whether the retake grade is capped at 80%, 90%, or 100%. Evaluating this retake practice should focus both on determining if there is an impact on student learning and to evaluate if there is a reduction in the unintended consequences displayed by students because of the students' access to a retake. Finally, this study assumed that the remediation provided to students between the original exam and its retake was consistent across the district. Evaluating if there is a remediation practice that improves student learning would be another aspect of the retake practice that could be evaluated.

Conclusion

It's not about collecting data, it's about collecting the right data, which allows the district to make informed decisions on its implemented practices and policies to increase student learning.

There is support for having a retake policy from parents, students and some staff, but there is little evidence on best method for implementation or the impact on student learning resulting from the practice. The 23-24 and 24-25 Secondary Grading Committees attempted to identify best practices in grading to improve consistency across the district but were limited in the data they had available, relying primarily on grade trends and qualitative data. The research conducted attempted to determine the optimum amount of time between the original test and its retake, but the limitations of the data reduced the validity of the results. The district is moving down a path towards implementing standards-based grading without the ability to evaluate if the changes made are improving student learning. If the district does not collect the correct data to evaluate performance, then the evaluation of benefits to students, staff, and parents, and the ability to validate that there are improved learning outcomes is severely limited.

References

Adeniji, S. M., Ameen, S. K., Dambatta, B. U., & Orilonise, R. (2018). Effect of mastery learning approach on senior school students' academic performance and retention in circle geometry. *International Journal of Instruction*, *11*(4), 951–962. https://doi.org/10.12973/iji.2018.11460a

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

American Library Association. (2007). Guidelines for position papers & issue briefs. https://www.ala.org/yalsa/aboutyalsa/yalsahandbook/whitepapers

Antonopoulou, P., Chadwick, P., McGee, O., Sniehotta, F. F., Meyer, C., Lorencatto, F., Meyer, C., O'Donnell, A., Lecouturier, J., Kelly, M., & Michie, S. (2021). Research engagement with policy makers: a practical guide to writing policy briefs. https://discovery.ucl.ac.uk/id/eprint/10141871/1/A practical guide to writing policy briefs.pdf

Arnautu, D., & Dagenais, C. (2021). Use and effectiveness of policy briefs as a knowledge transfer tool: A scoping review. *Humanities & Social Sciences Communications*, *8*(1). https://doi.org/10.1057/s41599-021-00885-9

Babbie, E. (2017). *The basics of social research*. Cengage Learning.

Bälter, O., Zimmaro, D., & Thille, C. (2018). Estimating the minimum number of opportunities needed for all students to achieve predicted mastery. *Smart Learning Environments*, *5*(1), 1–19. https://doi.org/10.1186/s40561-018-0064-z

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*(2), 89–99. https://doi.org/10.1080/00220671.1991.10702818

Barth, E., & Higginbottom, R. S. (2021). The calculus mastery exam: A report on the use of gateway-inspired assessment tools at liberal arts colleges. *PRIMUS*, *31*(8), 883–899. https://doi.org/10.1080/10511970.2020.1776804

Barzagar Nazari, K., & Ebersbach, M. (2019). Distributing mathematical practice of third and seventh graders: Applicability of the spacing effect in the classroom. *Applied Cognitive Psychology*, *33*(2), 288–298. https://doi.org/10.1002/acp.3485

Birhan, A. T. (2018). Effects of mastery learning instruction on engineering students' writing skills development and motivation. *Journal of Language and Education*, *4*(4), 20–30. https://doi.org/10.17323/2411-7390-2018-4-4-20-30

Blackstone, B., & Oldmixon, E. (2019). Specifications grading in political science. *Journal of Political Science Education*, *15*(2), 191–205. https://doi.org/10.1080/15512169.2018.1447948

Bloom, B. S. (1968). Learning for mastery. Instruction and curriculum. Regional Education Laboratory for the Carolinas and Virginia, topical papers and reprints, number 1. *Evaluation Comment*, *1*(2). https://eric.ed.gov/?id=ED053419

Bloom, B. S. (1974). Time and learning. *American Psychologist*, *29*(9), 682–688. https://doi.org/10.1037/h0037632

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research:

Meaning and value in the most common educational measure. *Review of Educational Research*, *86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Buckmiller, T., Peters, R., & Kruse, J. (2017). Questioning points and percentages: Standards-based grading (SBG) in higher education. *College Teaching*, *65*(4), 151–157. https://doi.org/10.1080/87567555.2017.1302919

Buraphadeja, V., & Srisarkun, V. (2024). Mastery learning in CS1: A longitudinal study during and post-pandemic. *Discover Education*, *3*(1), 260. https://doi.org/10.1007/s44217-024-00361-x

Butler, M. (2023). Alternative grading systems and student outcomes: A comparative analysis of motivation, enjoyment, engagement, stress, and perceptions of final grades. *Assessment & Evaluation in Higher Education*, *0*(0), 1–11. https://doi.org/10.1080/02602938.2025.2475068

Campbell, R., Clark, D., & OShaughnessy, J. (2020). Introduction to the special issue on implementing mastery grading in the undergraduate mathematics classroom. *PRIMUS*, *30*(8–10), 837–848. https://doi.org/10.1080/10511970.2020.1778824

Carlisle, S. (2020). Simple specifications grading. *PRIMUS*, *30*(8–10), 926–951. https://doi.org/10.1080/10511970.2019.1695238

Carr, M., Prendergast, M., Breen, C., & Faulkner, F. (2017). How well do engineering students retain core mathematical knowledge after a series of high threshold online mathematics tests? *Teaching Mathematics & Its Applications*, *36*(3), 136–150. https://doi.org/10.1093/teamat/hrw021

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *64*(8), 1–9. https://doi.org/10.1177/016146816306400801

Carstensen, M. B., & Emmenegger, P. (2023). Education as social policy: New tensions in maturing knowledge economies. *Social Policy & Administration*, *57*(2), 109–121. https://doi.org/10.1111/spol.12888

Çevik, M., & Bakioğlu, B. (2022). The effect of STEM education integrated into teaching-learning approaches (SEITLA) on learning outcomes: A meta-analysis study. *International Journal of Progressive Education*, *18*(2), 119–135. https://doi.org/10.29329/ijpe.2022.431.8

Chamberlain Jr., D. (2023). How one instructor can teach a large-scale, mastery-based college algebra course online. *PRIMUS*, *33*(8), 867–888. https://doi.org/10.1080/10511970.2023.2190183

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language*, *102*, 83–96. https://doi.org/10.1016/j.jml.2018.05.007

Chen, J., Hughes, S., & Ranade, N. (2023). Reimagining student-centered learning: Accessible and inclusive syllabus design during and after the COVID-19 pandemic. *Computers and Composition*, *67*, 102751. https://doi.org/10.1016/j.compcom.2023.102751

Childers, A. B., & Lu, L. (2017). Computer based mastery learning in developmental mathematics classrooms. *Journal of Developmental Education*, *41*(1), 2-6,8-9. https://www.jstor.org/stable/44987467

Cho, Y., & Blood, I. A. (2020). An analysis of TOEFL® primary™ repeaters: How much score change occurs? *Language Testing*, *37*(4), 503–522. https://doi.org/10.1177/0265532220927751

Choi, H., & Lee, H. S. (2020). Knowing Is not half the battle: The role of actual test experience in the forward testing effect. *Educational Psychology Review*, *32*(3), 765–789. https://doi.org/10.1007/s10648-020-09518-0

Cilli-Turner, E., Dunmyre, J., Mahoney, T., & Wiley, C. (2020). Mastery grading: Build-a-syllabus workshop. *PRIMUS*, *30*(8–10), 952–978. https://doi.org/10.1080/10511970.2020.1733152

Collins, J. B., Harsy, A., Hart, J., Haymaker, K. A., Hoofnagle, A. M., Janssen, M. K., Kelly, J. S., Mohr, A. T., & OShaughnessy, J. (2019). Mastery-based testing in undergraduate mathematics courses. *PRIMUS*, *29*(5), 441–460. https://doi.org/10.1080/10511970.2018.1488317

Cundiff, P. R., McLaughlin, O., Brown, K., & Grace, K. (2020). In search of greater understanding: The impact of mastery learning on social science education. *Teaching Sociology*, *48*(2), 107–119. https://doi.org/10.1177/0092055X20907979

Curley, B., & Downey, J. (2024). Implementation of alternative grading methods in a mathematical statistics course. *Journal of Statistics and Data Science Education*, *32*(3), 272–282. https://doi.org/10.1080/26939169.2023.2249956

Desouza, K. C., & Yuan, L. (2013). Towards evidence-driven policy design: Complex adaptive systems and computational modeling. *Annual Review of Policy Design, 1*(1), 1-19.

Dudley, U. (2010). What is mathematics for. *Notices of the AMS*, *57*(5), 608–613.

Dweck, C. S. (2000). *Self-theories: Their Role in Motivation, Personality, and Development*. Psychology Press.

Early, D. M., Berg, J. K., Alicea, S., Si, Y., Aber, J. L., Ryan, R. M., & Deci, E. L. (2016). The impact of Every Classroom, Every Day on high school student achievement: Results from a school-randomized trial. *Journal of Research on Educational Effectiveness*, *9*(1), 3–29. https://doi.org/10.1080/19345747.2015.1055638

Ebersbach, M., & Barzagar Nazari, K. (2020). No robust effect of distributed practice on the short- and long-term retention of mathematical procedures. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00811

Ee, M. S., Yeoh, W., Boo, Y. L., & Boulter, T. (2018). Examining the effect of time constraint on the online mastery learning approach towards improving postgraduate students' achievement. *Studies in Higher Education*, *43*(2), 217–233. https://doi.org/10.1080/03075079.2016.1161611

Elsinger, J., & Lewis, D. (2020). Applying a standards-based grading framework across lower level mathematics courses. *PRIMUS*, *30*(8–10), 885–907. https://doi.org/10.1080/10511970.2019.1674430

Emery, A., Sanders, M., Anderman, L. H., & Yu, S. L. (2018). When mastery goals meet mastery learning: Administrator, teacher, and student perceptions. *Journal of*

Experimental Education, *86*(3), 419–441. https://doi.org/10.1080/00220973.2017.1341863

Erbes, S., Wizner, M., & Powlis, J. (2021). Understanding the role of traditional & proficiency-based grading systems upon student learning and college admissions. *Journal of Higher Education Theory & Practice*, *21*(10), 54–68. https://doi.org/10.33423/jhetp.v21i10.4625

Fergus, S., & Petrick Smith, C. (2022). Characteristics of proficiency-based learning and their impacts on math anxiety in the middle grades. *Research in Middle Level Education Online*, *45*(4), 1–19. https://doi.org/10.1080/19404476.2022.2045810

Fernandez, O. E. (2021). 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), 855–868. https://doi.org/10.1080/10511970.2020.1772915

Ford, J., Erickson, R., Le, H., Vick, K., & and Downey, J. (2024). Relating consistent improvement to overall performance in a calculus I course that utilizes standards-based grading. *PRIMUS*, *34*(8), 792–804. https://doi.org/10.1080/10511970.2024.2361374

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*, *109*(8), 1067. https://doi.org/10.1037/edu0000197

Gouvard, P., & Durand, R. (2023). To be or not to be (typical): Evaluation-mode heterogeneity and its consequences for organizations. *Academy of Management Review*, *48*(4), 659–680. https://doi.org/10.5465/amr.2020.0314

Guo, L. (2021). Effects of the initial test interval and feedback timing on L2 vocabulary retention. *Language Learning Journal*, *49*(3), 382–398. https://doi.org/10.1080/09571736.2018.1551416

Guskey, T. R. (2005). Formative classroom assessment and Benjamin S. Bloom: Theory, research, and implications. *Online Submission*. https://eric.ed.gov/?id=ED490412

Guskey, T. R. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "learning for mastery." *Journal of Advanced Academics*, *19*(1), 8–31. https://doi.org/10.4219/jaa-2007-704

Guskey, T. R. (2010). Lessons of mastery learning. *Educational Leadership*, *68*(2), 52.

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Corwin Press.

Guskey, T. R., Townsley, M., & Buckmiller, T. M. (2020). The impact of standards-based learning: Tracking high school students' transition to the university. *NASSP Bulletin*, *104*(4), 257–269. https://doi.org/10.1177/0192636520975862

Halperin, A. (2020). Mastery-based testing in calculus with a final exam component. *PRIMUS*, *30*(8–10), 1017–1039. https://doi.org/10.1080/10511970.2019.1700575

Harsy, A., Carlson, C., & Klamerus, L. (2021). An analysis of the impact of mastery-based testing in mathematics courses. *PRIMUS*, *31*(10), 1071–1088. https://doi.org/10.1080/10511970.2020.1809041

Harsy, A., & Hoofnagle, A. (2020). Comparing mastery-based testing with traditional testing in calculus II. *International Journal for the Scholarship of Teaching and Learning*, *14*(2). https://doi.org/10.20429/ijsotl.2020.140210

Herman, G. L., Cai, Z., Bretl, T., Zilles, C., & West, M. (2020). Comparison of grade replacement and weighted averages for second-chance exams. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, 56–66. https://doi.org/10.1145/3372782.3406260

Heubach, S., & Krinsky, S. (2020). Implementing mastery-based grading at scale in introductory statistics. *PRIMUS*, *30*(8–10), 1054–1070. https://doi.org/10.1080/10511970.2019.1700576

Houghton Mifflin Harcourt. (2020). *Math Inventory Professional Learning Guide*. Houghton Mifflin Harcourt Publishing.

Huey, M. E., Silvey, P. R., Vaughan, A. G., & Fisher, A. L. (2022). Assessing the impact of standards-based grading policy changes on student performance and practice work completion in secondary mathematics. *Studies in Educational Evaluation*, *75*, 101211. https://doi.org/10.1016/j.stueduc.2022.101211

Japee, G., & Oza, P. (2021). Curriculum and evaluation in outcome-based education. *Psychology and Education Journal*, *58*(2), 5620–5625. https://doi.org/10.17762/pae.v58i2.2982

Jones, P. A. (2020). Implementing specifications grading in MPA courses: A potential strategy for better work-life balance. *Journal of Public Affairs Education*, *26*(4), 531–547. https://doi.org/10.1080/15236803.2020.1773713

Katzman, S. D., Hurst-Kennedy, J., Barrera, A., Talley, J., Javazon, E., Diaz, M., & Anzovino, M. E. (2021). The effect of specifications grading on students' learning and attitudes in an undergraduate-level cell biology course. *Journal of Microbiology & Biology Education*, *22*(3), e00200-21. https://doi.org/10.1128/jmbe.00200-21

Kliegl, O., Carls, T., & Bäuml, K.-H. T. (2019). How delay influences search processes at test. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000701

Knight, M., & Cooper, R. (2019). Taking on a new grading system: The interconnected effects of standards-based grading on teaching, learning, assessment, and student behavior. *NASSP Bulletin*, *103*(1), 65–92. https://doi.org/10.1177/0192636519826709

Kulik, C.-L. C., Kulik, J. A., Bangert-Drowns, R. L., & Slavin, R. E. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*(2), 265. https://doi.org/10.3102/00346543060002265

Laird Statistics. (n.d.a) One-way ANOVA in SPSS statistics. https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

Laird Statistics. (n.d.b) One-way ANCOVA in SPSS statistics. https://statistics.laerd.com/spss-tutorials/ancova-using-spss-statistics.php

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review*, *33*(3), 959–987. https://doi.org/10.1007/s10648-020-09572-8

Lehman, E., De Jong, D., & Baron, M. (2018). Investigating the relationship of standards-based grades vs. traditional-based grades to results of the scholastic math inventory at the middle school level. *Leadership Review of Doctoral Research*, *6*, 1–16.

Lenarz, J., & Pelatt, K. E. (2020). A transition to mastery-based testing with the hope of increasing student persistence. *PRIMUS*, 1–10. https://doi.org/10.1080/10511970.2020.1843209

Lewis, D. (2020). Student anxiety in standards-based grading in mathematics courses. *Innovative Higher Education*, *45*(2), 153–164. https://doi.org/10.1007/s10755-019-09489-3

Lin, V., Liu, G.-Z., Hwang, G.-J., Chen, N.-S., & Yin, C. (2022). Outcomes-based appropriation of context-aware ubiquitous technology across educational levels. *Interactive Learning Environments*, *30*(8), 1515–1538. https://doi.org/10.1080/10494820.2019.1703012

Lin, Y.-M., & Chen, M. Y. (2020). Understanding writing quality change: A longitudinal study of repeaters of a high-stakes standardized English proficiency test. *Language Testing*, *37*(4), 523–549. https://doi.org/10.1177/0265532220925448

Linhart, J. M. (2020). Mastery-based testing to promote learning: Experiences with discrete mathematics. *PRIMUS*, *30*(8–10), 1087–1109. https://doi.org/10.1080/10511970.2019.1695236

Link, L. J., & Guskey, T. R. (2022). Is standards-based grading effective? *Theory Into Practice*, *61*(4), 406–417. https://doi.org/10.1080/00405841.2022.2107338

Livingston, J. A., & Gentile, J. R. (1996). Mastery learning and the decreasing variability hypothesis. *The Journal of Educational Research*, *90*(2), 67–74. https://doi.org/10.1080/00220671.1996.9944447

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2020). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. Educational Psychology Review, 32(1), 277–295. https://doi.org/10.1007/s10648-019-09489-x

Mangum, A. (2020). Implementation of mastery-based testing in calculus I. *PRIMUS*, *30*(8–10), 869–884. https://doi.org/10.1080/10511970.2019.1709587

Mayer, D. (2021). The connections and disconnections between teacher education policy and research: Reframing evidence. *Oxford Review of Education*, *47*(1), 120–134. https://doi.org/10.1080/03054985.2020.1842179

McKell, K. C., & Danowitz, A. (2020). Exploring the effect of standards-based grading on student learning. *2020 IEEE Frontiers in Education Conference (FIE)*, 1–7. https://doi.org/10.1109/FIE44824.2020.9273889

Morphew, J. W., Silva, M., Herman, G., & West, M. (2020). Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering. *Applied Cognitive Psychology*, *34*(1), 168–181. https://doi.org/10.1002/acp.3605

Morris, S. R., & McKenzie, S. C. (2025). Stuck in the same old grades: Unchanging disparities in economically disadvantaged students' course failures. *The*

Educational Forum, *89*(1), 58–74. https://doi.org/10.1080/00131725.2024.2312299

Ndihokubwayo, K. (2023). Know how to write an influential policy brief: A systematic guide to writers and readers. *Journal of Classroom Practices*, *2*(2), 1–10. https://doi.org/10.58197/prbl/UDHD5751

Nelson, P. M., Parker, D. C., & Van Norman, E. R. (2018). Subskill mastery among elementary and middle school students at risk in mathematics. *Psychology in the Schools*, *55*(6), 722–736. https://doi.org/10.1002/pits.22143

Nevid, J. S., & Gordon, A. J. (2018). Integrated learning systems: Is there a learning benefit? *Teaching of Psychology*, *45*(4), 340–345. https://doi.org/10.1177/0098628318796920

Olsen, B., & Buchanan, R. (2019). An investigation of teachers encouraged to reform grading practices in secondary schools. *American Educational Research Journal*, *56*(5), 2004–2039. https://doi.org/10.3102/0002831219841349

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pelánek, R., & Řihák, J. (2018). Analysis and design of mastery learning criteria. *New Review of Hypermedia and Multimedia*, *24*(3), 133–159. https://doi.org/10.1080/13614568.2018.1476596

Peterson, D. J., & Wissman, K. T. (2018). The testing effect and analogical problem-solving. *Memory*, *26*(10), 1460–1466. https://doi.org/10.1080/09658211.2018.1491603

Pollio, M., & Hochbein, C. (2015). The association between standards-based grading and standardized test scores as an element of a high school reform model. *Teachers College Record*, *117*(11), 1–28. https://doi.org/10.1177/016146811511701106

Racsmány, M., Szőllősi, Á., & Marián, M. (2020). Reversing the testing effect by feedback is a matter of performance criterion at practice. *Memory & Cognition*, *48*(7), 1161–1170. https://doi.org/10.3758/s13421-020-01041-5

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology. General*, *140*(3), 283–302. https://doi.org/10.1037/a0023956

Rawson, K. A., Dunlosky, J., & Janes, J. L. (2020). All good things must come to an end: A potential boundary condition on the potency of successive relearning. *Educational Psychology Review*, *32*(3), 851–871. https://doi.org/10.1007/s10648-020-09528-y

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology. Applied*, *24*(1), 57–71. https://doi.org/10.1037/xap0000146

Scarlett, M. (2018). “Why did I get a C?”: Communicating student performance using standards-based grading. *InSight: A Journal of Scholarly Teaching*, *13*, 59–75. https://doi.org/10.46504/14201804sc

Schalock, R. L. (2001). *Outcome-Based Evaluation*. Springer Science & Business Media.

Scheopner Torres, A., Brett, J., Cox, J., & Greller, S. (2018). Competency education implementation: Examining the influence of contextual forces in three New Hampshire secondary schools. *AERA Open*, *4*(2). https://doi.org/10.1177/2332858418782883

Selbach-Allen, M. E., Greenwald, S. J., Ksir, A. E., & Thomley, J. E. (2020). Raising the bar with standards-based grading. *PRIMUS*, *30*(8–10), 1110–1126. https://doi.org/10.1080/10511970.2019.1695237

Siddiki, S., & Curley, C. (2022). Conceptualising policy design in the policy process. *Policy and Politics*, *50*(1), 117–135. https://doi.org/10.1332/030557321X16346727541396

Song, M., Garet, M. S., Yang, R., & Atchison, D. (2022). Did states’ adoption of more rigorous standards lead to improved student achievement? Evidence from a comparative interrupted time series study of standards-based reform. *American Educational Research Journal*, *59*(3), 610–647. https://doi.org/10.3102/00028312211058460

Stange, K. E. (2018). Standards-based grading in an introduction to abstract mathematics course. *PRIMUS*, *28*(9), 797–820. https://doi.org/10.1080/10511970.2017.1408044

Stowe. (2021, October 18). How to write a policy brief (step by step). FiscalNote. https://fiscalnote.com/blog/guide-writing-policy-brief

Supriya, K., Bang, C., Ebie, J., Pagliarulo, C., Tucker, D., Villegas, K., Wright, C., & Brownell, S. (2024). Optional exam retakes reduce anxiety but may exacerbate score disparities between students with different social identities. *CBE—Life Sciences Education*, *23*(3), ar30. https://doi.org/10.1187/cbe.21-11-0320

Swanson, R., Bingham, A., Sanders, M., & Moulton, C. (2025). Mastery-based testing in linear algebra: An entry point to alternative grading. *Teaching & Learning Inquiry*, *13*. https://eric.ed.gov/?id=EJ1464435

Townsley, M. (2019). Considering standards-based grading: Challenges for secondary school leaders. *Journal of School Administration Research and Development*, *4*(1), 35–38. https://doi.org/10.32674/jsard.v4i1.1941

Townsley, M., & Buckmiller, T. (2020). Losing As and Fs: What works for schools implementing standards-based grading? *Educational Considerations*, *46*(1). https://doi.org/10.4148/0146-9282.2204

Townsley, M., & Varga, M. (2018). Getting high school students ready for college: A quantitative study of standards- based grading practices. *Journal of Research in Education*, *28*(1), 92–112.

Tripp, B., Ravi, A., Pang, E., & Furrow, R. E. (2025). Undergraduate STEM students' perceptions of grading practices reveal that quiz retakes positively impact drivers of self-determination. *CBE—Life Sciences Education*, *24*(1), ar4. https://doi.org/10.1187/cbe.24-06-0167

Trumbo, M. C. S., McDaniel, M. A., Hodge, G. K., Jones, A. P., Matzen, L. E., Kittinger, L. I., Kittinger, R. S., & Clark, V. P. (2021). Is the testing effect ready to be put to work? Evidence from the laboratory to the classroom. *Translational Issues in Psychological Science*, *7*(3), 332–355. https://doi.org/10.1037/tps0000292

UCLA Advanced Research Computing. (2021) One-way ANOVA power analysis/g*power data analysis examples. https://stats.oarc.ucla.edu/other/gpower/one-way-anova-power-analysis/

UMSU Advocacy Service. (2018) How to be credible, convincing, and persuasive: A handy guide to writing compelling proposals and position papers. https://umsu.unimelb.edu.au/pageassets/support/advocacy/about-the-advocacy-service/reports-and-publications/ADV_HANDY-GUIDE-How-to-be-Credible-Convincing-Persuasive-16PP_DIGITAL.pdf

Venkatesh, A., & Piercey, V. (2021). Developing professional mindset with mastery grading. *PRIMUS*, *31*(7), 826–836. https://doi.org/10.1080/10511970.2020.1750078

Wagner, J. L., Smith, K. J., Johnson, C., Hilaire, M. L., & Medina, M. S. (2023). Best practices in syllabus design. *American Journal of Pharmaceutical Education*, *87*(3), ajpe8995. https://doi.org/10.5688/ajpe8995

Weir, R. J. (2020). Rethinking precalculus and calculus: A learner-centered approach. *PRIMUS*, *30*(8–10), 995–1016. https://doi.org/10.1080/10511970.2019.1686669

Williams, K. (2018). Specifications-based grading in an introduction to proofs course. *PRIMUS*, *28*(2), 128–142. https://doi.org/10.1080/10511970.2017.1344337

Winget, M., & Persky, A. M. (2022). A practical review of mastery learning. *American Journal of Pharmaceutical Education*, 8906. https://doi.org/10.5688/ajpe8906

Wisch, J. K., Ousterhout, B. H., Carter, V., & Orr, B. (2018). The grading gradient: Teacher motivations for varied redo and retake policies. *Studies in Educational Evaluation*, *58*, 145–155. https://doi.org/10.1016/j.stueduc.2018.06.005

Wyse, A. E. (2021). How days between tests impacts alternate forms reliability in computerized adaptive tests. *Educational and Psychological Measurement*, *81*(4), 644–667. https://doi.org/10.1177/0013164420979656

Wyse, A. E., & Anderson, D. (2020). What do we mean when we say "pass rate"? *Educational Measurement: Issues & Practice*, *39*(1), 61–67. https://doi.org/10.1111/emip.12297

Zhang, B. (2024). A didactic explanation of standards-based grading. *New Waves - Educational Research & Development*, *27*(2), 42–56.

Zhou, J., & Cao, Y. (2020). Does retest effect impact test performance of repeaters in different subgroups? *ETS Research Report Series*, *2020*(1), 1–15. https://doi.org/10.1002/ets2.12300

Zhu, X., Yang, Y., & Cao, M. (2023). Effect of a mutual goal-based continuous care program on self-management behaviour and health outcomes in patients after percutaneous coronary intervention: A randomized controlled trial. *Nursing Open*, *10*(6), 3707–3718. https://doi.org/10.1002/nop2.1625

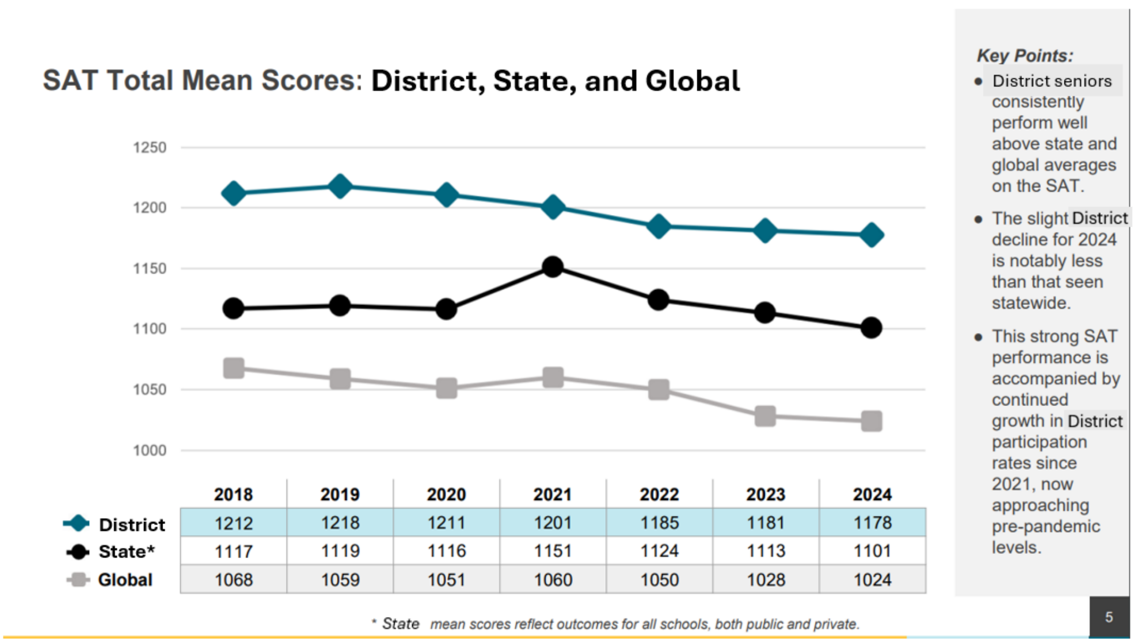
Zimmerman, J. K. (2020). Implementing standards-based grading in large courses across multiple sections. *PRIMUS*, *30*(8–10), 1040–1053. https://doi.org/10.1080/10511970.2020.1733149

Appendix: The Project

Recommendations to Improve Learning through Retake Implementation

or

Good Decisions Require Good Data



Source: Superintendent's 'Weekly Reflections', October 13, 2024

Dr. Christina N. Walton
August 2025

Executive Summary

A retake policy was implemented in 2016 leaving to curriculum teams the discretion to determine the amount of time between the original test and retake. In the nine years since:

- Many different practices are being used across the district
- Minimal to no improvement in high school students' geometry end-of-year state exam scores from 2015 – 2018
- A reduction in SAT mean scores from 2018-2024
- Resulted in unintended consequences for students including: arriving unprepared for exams, procrastination leading to increased course failures or drop-outs, focusing on grades not learning, and increased time management required leading to increased stress.

A quantitative study investigated if there were differences between high school students' geometry end-of-year state exam scores or between high school students' geometry math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*.

- The timeframes were determined by curriculum teams in the 2023-2024 school year

Although the expectation was scores would increase if students were given additional time to learn material and prepare for the retake, the analysis demonstrated scores decreased when the amount of time between the original exam was increased from the 2-3 week to anytime. However, these results are flawed and inconclusive as the dataset included both the scores for students who took retakes and those who did not.

Recommendation: The district implement the following policy modifications.

1. Require all summative grades include a code in SIS if they have been modified by a retake or grade replacement.
2. Charge the 25-26 Secondary Grading Committee to use the improved summative grade data generated from the first recommendation to validate the results of this study and determine if the time frame between the original exam and its retake should be limited to a shorter time frame.

During the evaluation of the syllabi from all the geometry curriculum teams, the specifics of how retakes were implemented could only be determined in 28 of 48 syllabi.

Additional recommendation:

3. Require either a minimum set of information be included in every course syllabus or create a syllabus outline for use by every school in the district.

Implementation: As part of the 25-26 District Grading Policy, require all summative grades include a code in SIS if they have been modified by a retake or grade replacement.

- This change would provide specific and measurable data for the district

- Informs parents that students are accessing retake opportunities to improve their grades

- Provides ability to collect quality data so the district can analyze the effectiveness of the current grading system and any future changes the district implements

- Charge the district's 25-26 Secondary Grading Committee to use the improved summative grade data generated from the first recommendation to validate the results of this study and determine if the window for retakes should be limited to a shorter timeframe

- Common syllabus: A draft common syllabus is attached at Appendix B for consideration by the Middle and High School Principals Association

    - A consistent syllabus design reduces the impression that different schools are implementing different processes, while providing clear documentation that curriculum teams are following the district's grading policies.

Summary: The 23-24 and 24-25 Secondary Grading Committees attempted to identify best practices in grading to improve consistency across the district but were limited in the data they had available, relying primarily on grade trends and qualitative data. If the district does not collect the correct data to evaluate performance, then the evaluation of benefits to students, staff, and parents, and the ability to validate that there are improved learning outcomes is severely limited.

It's not about collecting data, it's about collecting the right data, allowing the district to make informed decisions on the practices and policies implemented to increase student learning.

Introduction

A retake policy in a Northeastern suburban public school district was implemented in 2016 leaving to curriculum teams the discretion to determine the amount of time between the original test and retake. In the nine years since initial implementation, this policy resulted in many different practices being used across the district and ultimately resulted in minimal to no improvement in high school students' geometry end-of-year state exam scores. Schools are currently executing a variety of practices without the data to determine which practice is most effective. As a result, it is unclear if it is acceptable that different practices are being used or if a singular one should be adopted that provides better access to mastery learning while minimizing the unintended consequences of the policy. The research study attempted to fill this gap in practice by investigating what the differences in high school students' geometry end-of-year state exam scores and math growth scores were given the implementation of different retake time frames, and which of these practices currently in use by schools in the district have the highest influence on student learning.

Description of the Problem

Evidence of the problem at the local level included evidence of the different practices that have resulted from the policy, evidence of the minimal improvement of high school students' geometry end-of-year exam scores, and reduction in district SAT mean scores.

A data discovery was performed by requesting syllabi from secondary and high school's Geometry and Geometry Honors curriculum teams for the 23-24 school year

(N = 48). The syllabi were used to determine how much time individual curriculum teams across the district used between the original school-based test(s) and their retakes at their respective schools, see table A1 below. Syllabi were not requested from either the alternative high schools or the governor's school as they were outside the scope of the study. The time frames identified in the research problem were chosen from those identified as being in use by curriculum teams within these syllabi.

Table A 1

Summary of time frames used in a Northeastern Suburban School District

School year	Amount of time provided	Number of curriculum teams	Details
2023-2024	End of Quarter	8	Retake of multiple units taken at end of quarter
	2-3 Weeks	14	Retake 2-3 weeks after original exam
	Anytime	6	Retake available anytime during the school year
	Not Enough Information	20	Syllabi provided did not identify amount of time provided or used a different method

The retake policy was intended to provide students with additional time to work towards mastery of material and an additional opportunity to demonstrate this knowledge (District policy, 2021). Using the state end-of-year geometry exam as a measure of course material mastered during the school year, and therefore the effectiveness of a student's use of the retake to master material, Table A2 below shows there was a slight decrease in the percentage of high school students' who passed the geometry end-of-year state exam between school years 2014-2015 (the year prior to implementation of the retake) and

2017-2018 (three years after implementation). The increase in percentage in the 2018-2019 year was the result of a policy change that required only one math exam per student, resulting in a higher proportion of honors students being included in the pool of students who took the exam. Future years are not included in the table as the baseline of students was different due to the impacts of COVID on student learning.

Table A 2

Percentage of Students Meeting or Exceeding the High School Geometry End-of-Year State Exam

School year	Number of students	Percentage of students who passed
2014-2015	13,752	85
2015-2016	14,019	85
2016-2017	14,506	83
2017-2018	14,607	82
2018-2019*	13,704	85

Note. From Northeastern State Department of Education (n.d.).
* Increased percentage resulted from new policy requiring only one math exam per student increasing the proportion of honors level students in comparison to all students who took the exam.

Recent evidence of the reduction in learning was provided in the Superintendent's "Weekly Reflections" email dated October 13, 2024, (see table on front cover). Although the district's SAT mean scores continue to out-perform both the state and global SAT mean scores, this chart shows a reduction in SAT scores starting in 2018, two years after implementation of the district policy that required retakes, through 2024. In addition, as the district represents approximately 14.5% of the state's student population, even though the state numbers are now approximately the same as 2018, (before COVID), if the district's numbers were removed from the state's numbers so the data was independent, the state would show an increase in 2024 while the district was demonstrating a decrease.

Theoretical Framework and Current Literature

Providing students multiple opportunities to demonstrate their knowledge using a retake is assumed to be a best practice, but there is no consensus on the best implementation strategy for a retake process (Brookhart et al., 2016; Olsen & Buchanan, 2019; Wisch et al., 2018). Wisch et al. (2018) identified that there is a discrepancy between theory and practice, highlighting that there are differences of opinion on what constitutes best practice. Guskey (2010) acknowledged students did not consistently improve on subsequent retakes, possibly resulting from trying to simultaneously keep up with new content while trying to remediate previous content. Brookhart et al.'s (2016) literature review on assessment identified that standards-based grading did not improve student knowledge as assessed by high-stakes end-of-year testing and that there were few empirical studies on the effectiveness of the standards-based grading pedagogy. There continues to be a lack of quantitative research on how teachers implement the mastery learning process and if there is a demonstration of improved student learning and retention when using this process (Buraphadeja & Srisarkun, 2024; Huey et al., 2022; Swanson et al., 2025). Despite the continued implementation of the standards-based grading, the practice itself still remains largely unvalidated with limited research on student outcomes (Link & Guskey, 2022; Townsley & Buckmiller, 2020; Zhang, 2024). School systems are continuing to modify their instructional design to include mastery grading without the data to validate both the best practice for implementation and to validate if the use of mastery grading is resulting in improved student learning (Fergus & Petrick Smith, 2022; Huey et al., 2022).

The theoretical framework used for this research study was Bloom's theory of mastery learning (Bloom, 1968). Although the theoretical framework does not directly address the length of time between the original assessment and its retake, in Bloom's theory, the retake is supposed to be provided before the start of the next unit, which would imply the use of a shorter timeframe between assessments (Bloom, 1968). The site does not implement Bloom's theory with fidelity, as additional academic support and the opportunity to take a retake offered to students are not provided prior to the beginning of new instruction, but in parallel with new instruction. Guskey (2007, 2010) cautioned this implementation of Bloom's theory could result in students not consistently improving on subsequent retakes, as they are trying to simultaneously keep up with new content while remediating previous content, and the students who could benefit most from the corrective process are the least likely to take part in it.

Although the mastery learning process has many positive impacts on student learning, it has also resulted in unintended consequences that affect its implementation. The themes from research on the factors and considerations for mastery learning are mostly convergent demonstrating increased student ownership of learning, making learning more purposeful, and making assessment practices clearer (Harsy & Hoofnagle, 2020; Knight & Cooper, 2019; Weir, 2020), but questions on whether the implementation of mastery learning improves student achievement are beginning to be raised in current literature (Buraphadeja & Srisarkun, 2024; Fergus & Petrick Smith, 2022; Morris &

McKenzie, 2025; Song et al., 2022; Supriya et al., 2024). There are unintended consequences that have resulted with the implementation of mastery learning. Table A3 below provides convergent evidence of the unintended consequences identified in current literature. The themes evidenced in current literature were raised as issues by both students and teachers during the 23-24 and 24-25 Secondary Grading Advisory Committee meetings.

Table A 3

Unintended Consequences of Mastery Learning in Current Literature

Unintended consequence	Current literature
Student's arriving unprepared for initial exam	Emery et al., 2018; Fernandez, 2021; Huey et al., 2022; Knight & Cooper, 2019; Olsen & Buchanan, 2019; Scheopner Torres et al., 2018
Student procrastination leading to increased course failures or drop-outs	Buraphadeja & Srisarkun, 2024; Carlisle, 2020; Collins et al., 2019; Ee et al., 2018; Emery et al., 2018; Fernandez, 2021; Ford et al., 2024; Harsy & Hoofnagle, 2020; Herman et al., 2020; Huey et al., 2022; Knight & Cooper, 2019; Linhart, 2020; Mangum, 2020; McKell & Danowitz, 2020; Scheopner Torres et al., 2018; Venkatesh & Piercey, 2021; Weir, 2020
Increased burden and time management required of students with a resultant increase in student stress	Buraphadeja & Srisarkun, 2024; Curley & Downey, 2024; Emery et al., 2018; Lewis, 2020; Linhart, 2020; McKell & Danowitz, 2020
Increased student focus on grades, not learning	Herman et al., 2020; Huey et al., 2022; Olsen & Buchanan, 2019
Increased instructional workload	Adeniji et al., 2018; Barth & Higginbottom, 2021; Blackstone & Oldmixon, 2019; Buraphadeja & Srisarkun, 2024; Carlisle, 2020; Chamberlain Jr., 2023; Collins et al., 2019; Cundiff et al., 2020; Curley & Downey, 2024; Elsinger & Lewis, 2020; Emery et al., 2018; Fernandez, 2021; Halperin, 2020; Heubach & Krinsky, 2020; Linhart, 2020; Mangum, 2020;

	McKell & Danowitz, 2020; Olsen & Buchanan, 2019; Scheopner Torres et al., 2018; Stange, 2018; Williams, 2018; Wisch et al., 2018
Change in the distribution of course grades, with a higher percentage of students earning an A but also a greater percentage of students who either withdraw from courses or earn a D or F	Carlisle, 2020; Harsy & Hoofnagle, 2020; Herman et al., 2020; Jones, 2020; Katzman et al., 2021; Lenarz & Pelatt, 2020; Stange, 2018; Weir, 2020
No change or a reduction in content retention both within and in follow-on courses	Barth & Higginbottom, 2021; Buraphadeja & Srisarkun, 2024; Carr et al., 2017; Childers & Lu, 2017; Harsy & Hoofnagle, 2020; Katzman et al., 2021; Lyle et al., 2020; Stange, 2018; Weir, 2020; Wisch et al., 2018; Zimmerman, 2020;

Current literature on mastery-based assessments do not provide a clear picture on the practice, evidencing multiple practices associated with the amount of time between the original test and retake, and demonstrating mixed results from the practice (Cundiff et al., 2020; Harsy & Hoofnagle, 2020; Heubach & Krinsky, 2020). A complicating factor is most research is performed on read passages, word lists, or word pairs (Adesope et al., 2017). The working memory demand is stronger for a mathematical procedure than simpler tasks like recalling word lists, as mathematical procedures are more complex requiring both recalling underlying concepts such as formulas and the ability to apply these concepts within a solving task (Ebersbach & Barzagar Nazari, 2020). Although there are many recent qualitative reviews or discussions on the implementation of mastery learning in the classroom there are fewer that provide statistical analyses and fewer still that compare across implementation choices to determine the factors that have an impact on student retention of content or focus on mathematics as the content knowledge.

Overview of the Research

The purpose of this quantitative study was to investigate if there were differences between high school students' geometry end-of-year state exam scores or between high school students' geometry math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*. A quantitative causal-comparative approach was used to analyze the data. The timeframes were determined by curriculum teams at individual schools within the target school district. In the 2023-2024 school year, fourteen curriculum teams implemented retakes within two to three weeks of the original test (*short-term*), six curriculum teams implemented retakes anytime during the course (*anytime*), and eight curriculum teams implemented at the end of the quarter (*end of the quarter*). It was assumed that the remediation provided to students between the original exam and its retake was consistent across the district, although this is an area for further research.

The original plan was to include in the data set only students who had accessed a retake on a locally developed, school-based exam. Upon receipt of the original request for data it was determined that though the research met the technical merits and needs of the district, the request was denied because the data was not readily available, and the district could not justify spending public funds to access and provide the requested data. In conjunction with my committee, Walden University school administration, and the district, it was determined that two data sources, student's end-of-year geometry exam scores and student's math growth scores were available and could be used to address the research problem, though the results would have reduced validity.

Two statistical analyses were used on the data. An analysis of variance (ANOVA) was used on the state end-of-year geometry exam scores as the goal was to determine if there was a statistically significant difference between the three timeframes. An analysis of covariance (ANCOVA) was used on the math growth scores as the goal was to determine if there was a statistically significant difference in the math growth scores between the three timeframes when controlling for the effects of the fall test (the pre-test). Details on the variables used in the ANOVA and ANCOVA analyses are provided in Table A4 below.

Table A 4

Variables

	Research question	Variables	Data analysis
RQ1	What are the differences between high school students' geometry end-of-year state exam scores differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*?	IV: Retake Timing *Short-term* 1 *End of Quarter* 2 *Anytime* 3 DV: End-of-Year Exam Scores Continuous from 0 - 600	One-way ANOVA
RQ2	What are the differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing: *short-term* (2-3 weeks), *anytime*, or *end of the quarter*?	IV: Retake Timing *Short-term* 1 *End of Quarter* 2 *Anytime* 3 DV: Post-test (spring) math growth scores Continuous from 0 – 1600 Covariate: Pre-test (fall) math growth scores Continuous from 0 – 1600	One-way ANCOVA

Research Findings

Although the expectation was scores would increase if students were given additional time to learn material and prepare for the retake, the analysis demonstrated scores decreased when the amount of time between the original exam was increased from the 2-3 week to anytime. However, these results are flawed and inconclusive as the dataset included both the scores for students who took retakes and those who did not.

This study attempted to answer the question if there were differences in learning, as measured by both end-of-year state exam scores or high school math growth scores, depending on the length of time given to students between the original school-based assessment and its retake. Table A5 below summarizes the results of the two analyses.

Table A 5

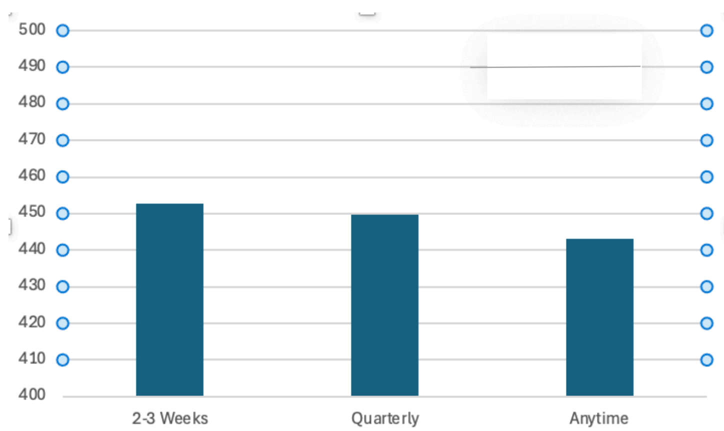
Summary of Results

Data source	Results	Statistically significant
Geometry End-of-Year State Scores	• There was a statistical difference in the end-of-year exam scores • There was a statistically significant difference between the two extremes, 2-3 weeks and anytime o Difference represents 9.7 points or 2% of overall score • There wasn't a statistically significant difference between 2-3 weeks and quarterly or between quarterly and anytime	Yes
Math Growth Scores	• There was not a statistical difference on the math growth scores when controlling for the pre-assessment score • Unadjusted results for the math growth scores showed a decrease in mean from 2-3 weeks to the anytime scores • When controlling with the pre-assessment score, o Mean scores similar for 2-3 weeks and quarterly o Scores decreased when time increased to anytime	No

The ANOVA analysis on end-of-year state exam scores determined there was a statistically significant difference in the end-of-year exam scores based on what type of retake was implemented, but the post-hoc analysis revealed that, although there wasn't a statistically significant difference between 2-3 weeks and quarterly or between quarterly and anytime. There was a statistically significant difference between the two extremes, 2-3 weeks and anytime. Although there was a statistical difference, the actual difference in mean score was only 9.7 points, or approximately 2% of the overall score, see Figure A1 below. This is consistent with the results of studies that demonstrated an improvement in exam results using the two-week timeframe between the original assessment and its retake (Adeniji et al., 2018; Ee et al., 2018; Halperin, 2020; Morphew et al., 2020).

Figure A 1

Mean Geometry End-of-Year State Scores by Retake Timing



In comparison, the ANCOVA analysis of math growth scores determined there was not a statistically significant difference based on the implemented retake timing when controlling for the pre-assessment score. Similar to the analysis of the end-of-year state scores, the unadjusted results for the math growth scores showed a decrease in mean from 2-3 weeks to the anytime scores. However, when adjusted for the pre-assessment score, the mean scores demonstrated similar results for 2-3 weeks and quarterly but a decreased score when the amount of time given increased to anytime, see Table A6 below. This is consistent with the results of studies that showed either no measurable difference in grades or mixed results (Carlisle, 2020; Harsy & Hoofnagle, 2020) and the results of previous studies that showed the strongest predictors on end-of-year high stakes testing were prior academic achievement and grades (Early et al., 2016; Lin & Chen, 2020; Pollio & Hochbein, 2015).

Table A 6

Unadjusted and Adjusted Means and Variability for Spring Math Growth Scores (Post-Test) with Fall Math Growth Scores (Pre-Test) as Covariate

		Unadjusted		Adjusted	
	N	*M*	*SD*	*M*	*SE*
2-3 Weeks	478	1111.86	123.95	1103.08	4.44
End of Quarter	488	1099.31	124.29	1106.08	4.39
Anytime	217	1084.47	132.81	1088.60	6.57

Note: *N* = number of participants, *M* = Mean, *SD* = Standard Deviation, *SE* = Standard Error

These results imply that a shorter time frame between the original test and its retake might yield better learning outcomes. However, these results are flawed and inconclusive as the dataset included both the scores for students who took retakes and those who did not. Additional details on the statistical analyses and results of the both the ANOVA and ANCOVA analyses are provided in Appendix A.

Policy Recommendations

The district has a significant issue that the data it uses to base decisions on, which includes the data used in the above research, is limited as it includes both the students who accessed the retake process and those who did not. The quantitative data provided as justification in both the 23-24 and 24-25 Secondary Grading Advisory Committee reports were based on grade trends only, using data that did not directly apply to the questions that the district was trying to answer. This lack of useful and measurable data needs to be addressed if the district is going to be able to address the issues teachers have raised regarding the increases in workload, increased student stress resulting from recent grading policy changes, whether there is a positive correlation with graduation rates, or whether grading policy changes are making a measurable impact on student learning that counteracts the demonstrated unintended consequences on student behavior these policies are evidencing.

Based on the research conducted, the recommendations from the 23-24 and 24-25 Secondary Grading Committees, and a review of literature, it is recommended that the district implement the following policy modification and recommendations:

1. Require all teachers to include a code in SIS on any grade that has been modified by either an improved grade from a retake or by a grade replacement.
    - This would provide the specific data to be able to evaluate the effectiveness on the retake policy, implementation improvements if applicable, and impacts or improvements on students learning based on demographic or prior academic achievement.
2. Charge the 25-26 Secondary Grading Committee to use the improved summative grade data generated from the first recommendation to validate the results of this study and determine if the window for retakes should be limited to a shorter timeframe.

- Although the analysis of the data provided by the district implies that a shorter time frame between the original test and its retake yields better outcomes, there are validity issues with the results. Rerunning this analysis with an improved data set is recommended before making a final recommendation.

During the evaluation of the syllabi from all the geometry curriculum teams, the specifics of how retakes were implemented could only be determined in 28 of 48 syllabi. One of the superintendent's challenges to the district has been to create as much consistency as possible across the school system. An additional recommendation would be to:

3. Require either a specific minimum set of information be included in every course syllabus or create a syllabus outline that is used by every school in the district. (A draft common syllabus is provided in Appendix B.)
    - A consistent syllabi design will reduce the impression that different schools are implementing different processes, while also providing clear documentation that curriculum teams are following the guidelines provided by the district's grading policies.

Implementation

As part of the 25-26 District Grading Policy, require all teachers include a code in the student information system (SIS) on any grade that is modified by either a retake or a grade replacement.

- RTK (Retake) – This grade improved as demonstrated by the student's retake exam.
- GRR (Grade Replacement) – This grade improved as demonstrated by a student's work on another assignment
- RTL (Retake but Lower) – Student's demonstration of knowledge did not improve on the retake. Student keeps original higher score.

This simple change would provide specific and measurable data for the district and the next Secondary Grading Committee to use to evaluate the effectiveness of current policies and validate the results of the study analysis performed to determine if the window for retakes should be limited to a shorter timeframe. In addition, this data can be used to assess the validity of concerns raised by instructional staff on the unintended consequences in student behavior that have risen due to the policy. These codes would also provide useful information to parents and guardians, providing information on whether their students are accessing their retake opportunities, and if they are accessing this opportunity, whether or not they are improving their grades. Once this additional data is created, the district can begin the analysis of the effectiveness of the current grading system, and the groundwork will have been created that allows the evaluation of any subsequent changes the district makes as it moves towards standards-based-grading.

For the final recommendation, a draft common syllabus is attached at Appendix B for consideration by the Middle and High School Principals Association for implementation in the 26-27 school year.

Summary

There is support for having a retake policy from parents, students and some staff, but there is little evidence on best method for implementation or the impact on student learning resulting from the practice. The 23-24 and 24-25 Secondary Grading Committees attempted to identify best practices in grading to improve consistency across the district but were limited in the data they had available, relying primarily on grade trends and qualitative data. The research conducted attempted to determine the optimum amount of time between the original test and its retake, but the limitations of the data reduced the validity of the results. The district is moving down a path towards implementing standards-based grading without the ability to evaluate if the changes made are improving student learning. If the district does not collect the correct data to evaluate performance, then the evaluation of benefits to students, staff, and parents, and the ability to validate that there are improved learning outcomes is severely limited.

It's not about collecting data, it's about collecting the right data, allowing the district to make informed decisions on the practices and policies implemented to increase student learning.

References

Adeniji, S. M., Ameen, S. K., Dambatta, B. U., & Orilonise, R. (2018). Effect of mastery learning approach on senior school students' academic performance and retention in circle geometry. *International Journal of Instruction*, *11*(4), 951–962. https://doi.org/10.12973/iji.2018.11460a

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Barth, E., & Higginbottom, R. S. (2021). The calculus mastery exam: A report on the use of gateway-inspired assessment tools at liberal arts colleges. *PRIMUS*, *31*(8), 883–899. https://doi.org/10.1080/10511970.2020.1776804

Blackstone, B., & Oldmixon, E. (2019). Specifications grading in political science. *Journal of Political Science Education*, *15*(2), 191–205. https://doi.org/10.1080/15512169.2018.1447948

Bloom, B. S. (1968). Learning for mastery. Instruction and curriculum. Regional Education Laboratory for the Carolinas and Virginia, topical papers and reprints, number 1. *Evaluation Comment*, *1*(2). https://eric.ed.gov/?id=ED053419

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, *86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Buraphadeja, V., & Srisarkun, V. (2024). Mastery learning in CS1: A longitudinal study during and post-pandemic. *Discover Education*, *3*(1), 260. https://doi.org/10.1007/s44217-024-00361-x

Carlisle, S. (2020). Simple specifications grading. *PRIMUS*, *30*(8–10), 926–951. https://doi.org/10.1080/10511970.2019.1695238

Carr, M., Prendergast, M., Breen, C., & Faulkner, F. (2017). How well do engineering students retain core mathematical knowledge after a series of high threshold online mathematics tests? *Teaching Mathematics & Its Applications*, *36*(3), 136–150. https://doi.org/10.1093/teamat/hrw021

Chamberlain Jr., D. (2023). How one instructor can teach a large-scale, mastery-based college algebra course online. *PRIMUS*, *33*(8), 867–888. https://doi.org/10.1080/10511970.2023.2190183

Childers, A. B., & Lu, L. (2017). Computer based mastery learning in developmental mathematics classrooms. *Journal of Developmental Education*, *41*(1), 2-6,8-9. https://www.jstor.org/stable/44987467

Collins, J. B., Harsy, A., Hart, J., Haymaker, K. A., Hoofnagle, A. M., Janssen, M. K., Kelly, J. S., Mohr, A. T., & OShaughnessy, J. (2019). Mastery-based testing in undergraduate mathematics courses. *PRIMUS*, *29*(5), 441–460. https://doi.org/10.1080/10511970.2018.1488317

Cundiff, P. R., McLaughlin, O., Brown, K., & Grace, K. (2020). In search of greater understanding: The impact of mastery learning on social science education. *Teaching Sociology*, *48*(2), 107–119. https://doi.org/10.1177/0092055X20907979

Curley, B., & Downey, J. (2024). Implementation of alternative grading methods in a mathematical statistics course. *Journal of Statistics and Data Science Education*, *32*(3), 272–282. https://doi.org/10.1080/26939169.2023.2249956

District Instructional Services Department. (2024). Secondary grading advisory committee final report, school year 2023-2024.

District Instructional Services Department. (2025). Secondary grading advisory committee final report, school year 2024-2025.

Early, D. M., Berg, J. K., Alicea, S., Si, Y., Aber, J. L., Ryan, R. M., & Deci, E. L. (2016). The impact of Every Classroom, Every Day on high school student achievement: Results from a school-randomized trial. *Journal of Research on Educational Effectiveness*, *9*(1), 3–29. https://doi.org/10.1080/19345747.2015.1055638

Ebersbach, M., & Barzagar Nazari, K. (2020). No robust effect of distributed practice on the short- and long-term retention of mathematical procedures. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00811

Ee, M. S., Yeoh, W., Boo, Y. L., & Boulter, T. (2018). Examining the effect of time constraint on the online mastery learning approach towards improving postgraduate students' achievement. *Studies in Higher Education*, *43*(2), 217–233. https://doi.org/10.1080/03075079.2016.1161611

Elsinger, J., & Lewis, D. (2020). Applying a standards-based grading framework across lower level mathematics courses. *PRIMUS*, *30*(8–10), 885–907. https://doi.org/10.1080/10511970.2019.1674430

Emery, A., Sanders, M., Anderman, L. H., & Yu, S. L. (2018). When mastery goals meet mastery learning: Administrator, teacher, and student perceptions. *Journal of Experimental Education*, *86*(3), 419–441. https://doi.org/10.1080/00220973.2017.1341863

Fernandez, O. E. (2021). 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), 855–868. https://doi.org/10.1080/10511970.2020.1772915

Ford, J., Erickson, R., Le, H., Vick, K., & and Downey, J. (2024). Relating consistent improvement to overall performance in a calculus I course that utilizes standards-based grading. *PRIMUS*, *34*(8), 792–804. https://doi.org/10.1080/10511970.2024.2361374

Guskey, T. R. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "learning for mastery." *Journal of Advanced Academics*, *19*(1), 8–31. https://doi.org/10.4219/jaa-2007-704

Guskey, T. R. (2010). Lessons of mastery learning. *Educational Leadership*, *68*(2), 52.

Halperin, A. (2020). Mastery-based testing in calculus with a final exam component. *PRIMUS*, *30*(8–10), 1017–1039. https://doi.org/10.1080/10511970.2019.1700575

Harsy, A., & Hoofnagle, A. (2020). Comparing mastery-based testing with traditional testing in calculus II. *International Journal for the Scholarship of Teaching and Learning*, *14*(2). https://doi.org/10.20429/ijsotl.2020.140210

Herman, G. L., Cai, Z., Bretl, T., Zilles, C., & West, M. (2020). Comparison of grade replacement and weighted averages for second-chance exams. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, 56–66. https://doi.org/10.1145/3372782.3406260

Heubach, S., & Krinsky, S. (2020). Implementing mastery-based grading at scale in introductory statistics. *PRIMUS*, *30*(8–10), 1054–1070. https://doi.org/10.1080/10511970.2019.1700576

Huey, M. E., Silvey, P. R., Vaughan, A. G., & Fisher, A. L. (2022). Assessing the impact of standards-based grading policy changes on student performance and practice work completion in secondary mathematics. *Studies in Educational Evaluation*, *75*, 101211. https://doi.org/10.1016/j.stueduc.2022.101211

Jones, P. A. (2020). Implementing specifications grading in MPA courses: A potential strategy for better work-life balance. *Journal of Public Affairs Education*, *26*(4), 531–547. https://doi.org/10.1080/15236803.2020.1773713

Katzman, S. D., Hurst-Kennedy, J., Barrera, A., Talley, J., Javazon, E., Diaz, M., & Anzovino, M. E. (2021). The effect of specifications grading on students' learning and attitudes in an undergraduate-level cell biology course. *Journal of Microbiology & Biology Education*, *22*(3), e00200-21. https://doi.org/10.1128/jmbe.00200-21

Knight, M., & Cooper, R. (2019). Taking on a new grading system: The interconnected effects of standards-based grading on teaching, learning, assessment, and student behavior. *NASSP Bulletin*, *103*(1), 65–92. https://doi.org/10.1177/0192636519826709

Lenarz, J., & Pelatt, K. E. (2020). A transition to mastery-based testing with the hope of increasing student persistence. *PRIMUS*, 1–10. https://doi.org/10.1080/10511970.2020.1843209

Lewis, D. (2020). Student anxiety in standards-based grading in mathematics courses. *Innovative Higher Education*, *45*(2), 153–164. https://doi.org/10.1007/s10755-019-09489-3

Lin, Y.-M., & Chen, M. Y. (2020). Understanding writing quality change: A longitudinal study of repeaters of a high-stakes standardized English proficiency test. *Language Testing*, *37*(4), 523–549. https://doi.org/10.1177/0265532220925448

Linhart, J. M. (2020). Mastery-based testing to promote learning: Experiences with discrete mathematics. *PRIMUS*, *30*(8–10), 1087–1109. https://doi.org/10.1080/10511970.2019.1695236

Link, L. J., & Guskey, T. R. (2022). Is standards-based grading effective? *Theory Into Practice*, *61*(4), 406–417. https://doi.org/10.1080/00405841.2022.2107338

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2020). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. Educational Psychology Review, 32(1), 277–295. https://doi.org/10.1007/s10648-019-09489-x

Mangum, A. (2020). Implementation of mastery-based testing in calculus I. *PRIMUS*, *30*(8–10), 869–884. https://doi.org/10.1080/10511970.2019.1709587

McKell, K. C., & Danowitz, A. (2020). Exploring the effect of standards-based grading on student learning. *2020 IEEE Frontiers in Education Conference (FIE)*, 1–7. https://doi.org/10.1109/FIE44824.2020.9273889

Morphew, J. W., Silva, M., Herman, G., & West, M. (2020). Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering. *Applied Cognitive Psychology*, *34*(1), 168–181. https://doi.org/10.1002/acp.3605

Olsen, B., & Buchanan, R. (2019). An investigation of teachers encouraged to reform grading practices in secondary schools. *American Educational Research Journal*, *56*(5), 2004–2039. https://doi.org/10.3102/0002831219841349

Pollio, M., & Hochbein, C. (2015). The association between standards-based grading and standardized test scores as an element of a high school reform model. *Teachers College Record*, *117*(11), 1–28. https://doi.org/10.1177/016146811511701106

Scheopner Torres, A., Brett, J., Cox, J., & Greller, S. (2018). Competency education implementation: Examining the influence of contextual forces in three New Hampshire secondary schools. *AERA Open*, *4*(2). https://doi.org/10.1177/2332858418782883

Stange, K. E. (2018). Standards-based grading in an introduction to abstract mathematics course. *PRIMUS*, *28*(9), 797–820. https://doi.org/10.1080/10511970.2017.1408044

Townsley, M., & Buckmiller, T. (2020). Losing As and Fs: What works for schools implementing standards-based grading? *Educational Considerations*, *46*(1). https://doi.org/10.4148/0146-9282.2204

Venkatesh, A., & Piercey, V. (2021). Developing professional mindset with mastery grading. *PRIMUS*, *31*(7), 826–836. https://doi.org/10.1080/10511970.2020.1750078

Weir, R. J. (2020). Rethinking precalculus and calculus: A learner-centered approach. *PRIMUS*, *30*(8–10), 995–1016. https://doi.org/10.1080/10511970.2019.1686669

Williams, K. (2018). Specifications-based grading in an introduction to proofs course. *PRIMUS*, *28*(2), 128–142. https://doi.org/10.1080/10511970.2017.1344337

Wisch, J. K., Ousterhout, B. H., Carter, V., & Orr, B. (2018). The grading gradient: Teacher motivations for varied redo and retake policies. *Studies in Educational Evaluation*, *58*, 145–155. https://doi.org/10.1016/j.stueduc.2018.06.005

Zhang, B. (2024). A didactic explanation of standards-based grading. *New Waves - Educational Research & Development*, *27*(2), 42–56.

Zimmerman, J. K. (2020). Implementing standards-based grading in large courses across multiple sections. *PRIMUS*, *30*(8–10), 1040–1053. https://doi.org/10.1080/10511970.2020.1733149

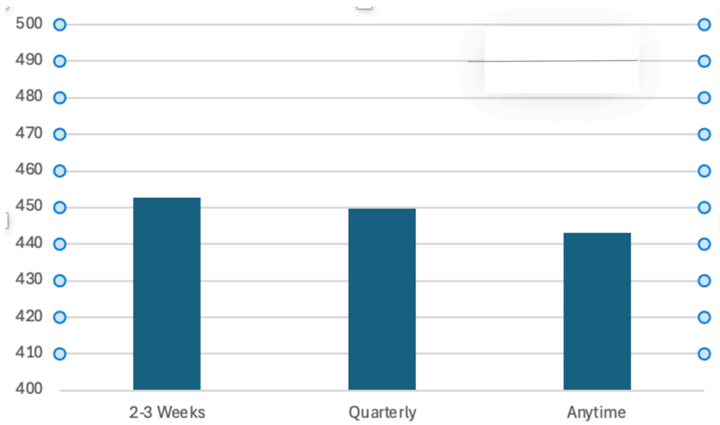
Appendix A: Research Findings

Geometry End-of-Year State Exam Scores

After reducing the original data set provided by the site to only those students who had an end-of-year state score and a pre-test (fall) and post-test (spring) math score, the resulting sample sizes were *short-term* ($n = 486$), *end of the quarter* ($n = 491$), and *anytime* ($n = 220$) for a total of 1197 students across 19 curriculum teams from 11 high and secondary schools. An overall review of the descriptive statistics demonstrated that mean scores on the end-of-year state Geometry exam decreased as the amount of time given between the original test and its retake was increased, from *2-3 weeks* ($M = 452.8$, $SD = 39.0$), to *end of quarter* ($M = 449.8$, $SD = 33.4$), to *anytime* ($M = 443.0$, $SD = 38.8$), see Figure A2 below. The Welch ANOVA determined that the amount of time used between the original test and its retake was statistically significantly different for different amounts of time, Welch's $F(2, 577.996) = 4.778$, $p = .009$.

Figure A 2

Mean Geometry End-of-Year State Scores by Retake Timing



The Games-Howell post hoc test determined there was a statistically significant decrease in the end-of-year state scores of 9.77 (95% CI, 2.34 to 17.19) from the *2-3 weeks* to the anytime score. There was not a significantly significant difference between 2-3 weeks and quarterly or between quarterly and anytime.

Math Growth Scores

After reducing the above data set to remove 14 data points that were determined to be outliers, (and when removed, did not materially affect the results of the analysis), the descriptive statistics were calculated and provided in Table A7 below. Without including the effects of the pre-test, the unadjusted means for the spring math growth scores (post-test) decreased as the amount of time given between the original test and its retake was increased. When the effects from the pre-test were included, the adjusted means for the spring math growth scores (post-test) were basically the same for 2-3 weeks and quarterly but decreased when the amount of time given between the original test and its retake was increased to anytime.

Table A 7

Unadjusted and Adjusted Means and Variability for Spring Math Growth Scores (Post-Test) with Fall Math Growth Scores (Pre-Test) as Covariate

		Unadjusted		Adjusted	
	N	*M*	*SD*	*M*	*SE*
2-3 Weeks	478	1111.86	123.95	1103.08	4.44
End of Quarter	488	1099.31	124.29	1106.08	4.39
Anytime	217	1084.47	132.81	1088.60	6.57

Note: *N* = number of participants, *M* = Mean, *SD* = Standard Deviation, *SE* = Standard Error

An ANCOVA was run to determine the effect of the amount of time given between original school-based test and its retake on the spring math growth score (post-test) after controlling for the fall math growth score (pre-test). After adjustment for the fall math growth score (pre-test), the ANCOVA analysis determined that the amount of time used between the original test and its retake wasn't a statistically significantly difference for different amounts of time on the spring math growth score (post-test), $F(2, 1179) = 2.537$, $p = .08$, partial $\eta^2 = .004$. Therefore, there are no differences between high school students' math growth scores when controlling for the pre-assessment score differentiated by three schemes of retake timing.

Appendix B: Example Common Syllabus Outline

During the evaluation of all the geometry curriculum teams' syllabi for the 24-25 school year, the specifics of how retakes were implemented could only be determined in 28 of 48 syllabi. One of the superintendent's challenges to the district has been to create as much consistency as possible across the school system. Attached is a recommended draft of a common syllabus outline for consideration as a standardized document for use in all high and secondary schools across the district. The following information pertains to this draft:

1. All hyperlinks were disabled within the document to maintain the site's request for anonymity. Titles for the hyperlinks were left to inform the reader on the recommended content.
2. The recommended common syllabus is based on the 24-25 Secondary Grading Policy (district, 2025). The document has been updated with the planned modifications released by the district on June 9, 2025, for the 2025-2026 school year. This document must be updated to reflect changes implemented in future policy documents.
3. The general grading information section summarizes the 24-25 Secondary Grading Policy (district, 2025). The course-specific grading information section is where curriculum teams will include specifics based on curriculum and course.
4. The syllabus outline includes the recommended gradebook codes.

5. All items in red are items that are school or course specific and must be updated by school-based curriculum teams.
6. Schools within the district can still use a traditional quarterly system for the 2025-2026 school year. Appropriate language for schools using a rolling gradebook were included in the attached recommended common syllabus. If schools are using a traditional quarterly gradebook, the following statement for the top left of the General Grading Information section should be used:

Quarterly Gradebook
All 'school name' teachers use a quarterly gradebook. Students will receive a quarter grade based on performance in the individual quarter and start over at the beginning of each quarter. Except for AP and DE courses, final grades are determined by an average of quarter grades and a culminating activity/final exam. Quarter 1 - 22.5% Quarter 2 - 22.5% Quarter 3 - 22.5% Quarter 4 - 22.5% Final Assessment - 10% Total - 100%

School Logo	**2025-2026 *Course Name* Syllabus**

Below is contact information for the teachers of this course. If you have questions or concerns about this course, please contact your classroom teacher directly via email first.		
Teacher Name	**Email/Phone**	**Room Number**

Course Description: [Enter text here.]

Course Outline: This is a general course outline and is subject to change.
1st Quarter: [Enter text here.]
2nd Quarter: [Enter text here.]
3rd Quarter: [Enter text here.]
4th Quarter: [Enter text here.]

Materials/texts Required: The following supplies are required:
[Enter text here.]

- **Textbook:** [Enter text here. Include current replacement costs for any planned district textbooks or resources.]

General Grading Information

Rolling Gradebook	District Grading Scale
All *School Name* teachers will use a rolling (cumulative) gradebook, which allows for progress towards a final mark that captures the total picture of student grades. **Cumulative** – Grades from each quarter carry over and build upon the previous quarter **Quarter Grades** are a snapshot of a student's current progress in the course and demonstrate performance to that date, not just the quarter. **Final Grades** are determined by a combination of all assignments and assessments throughout the year, regardless of the quarter the work was completed and will be calculated using the following weights: Gradebook – 90% Final Assessment – 10% Total - 100%	A (100-93) A- (92-90) B+ (89-87) B (86-83) B- (82-80) C+ (79-77) C (76-73) C- (72-70) D+ (69-67) D. (66-60) F (59-50)

Grading Categories

All *School Name* classes will use assignments in three weighted categories. All assignments within a category will be weighted equally.

Summative Assessments - 70% of overall grade

○ Includes culminating assessment tasks that measure mastery of standards. Assignments in this category **are** eligible for one attempt at reassessment.

Formative Assessments - 20% of overall grade

○ Includes assessments that provide students with feedback toward mastery of standards. Assignments in this category **are not** eligible for reassessment.

Practice - 10% of overall grade

○ Includes smaller assignments and homework that allow students to practice content and skills. Assignments in this category **are not** eligible for reassessment.

Courses must include a minimum of 7 graded assignments per quarter with at least two items in each category. Within a quarter, no single assignment can count for more than 35% of the overall grade.

Teachers are expected to grade assignments and post grades within seven school days after the due date with the understanding that major projects/papers may require additional time to ensure quality feedback.

Retakes of Summative Assessments. For summative assessments, one new opportunity to demonstrate mastery will be provided. Teaching teams and departments will communicate reasonable guidelines and timeframes that must be met for a student to be eligible to retake a summative assessment. A retake may be accomplished in one of two ways, depending on the course and the nature of the assessment:

- **Retake**: Students who score below an 90% can complete remediation as designated by the teacher and retake a version of the assessment for up to a 90%
- **Grade Replacement**: If the same skills/concepts are reassessed on an upcoming assessment, the teacher may replace the original assessment grade with the higher score

If a summative assignment provides students opportunity for feedback from a teacher prior to the final deadline, there does not need to be any additional opportunity to redo the assignment. The feedback should include a scored rubric and clear enough guidelines such that students can adjust their original attempt to earn the full 100%.

Note: Final exams, midterms, and formative assessments (quizzes, classwork, and homework) are not eligible for retakes.

Use of letter grades as a way of entering assignments. Letter grades may be used in certain courses for assignments that are designed using rubrics and assess students based on levels of understanding. This might include assignments like performance-based assessments, projects, or essays. Talk to your teacher directly if you have questions about this method of entering assignment grades.

Minimum Grade. Students who make a reasonable attempt at completing an assignment will receive a minimum of a 50%. A “reasonable attempt” occurs when a student submits an assignment with some evidence that an attempt was made to demonstrate knowledge. At the end of each quarter, any quarter grade lower than a 50% will be converted to a 50%.

- **Use of the Zero.** If a student does not turn in an assignment or does not make a “reasonable attempt” to show evidence of their learning, a zero may be entered into the gradebook. If a zero is entered into the gradebook for a summative assessment, a personalized two-way communication must occur with the parent/guardian.

Late Assignments. Teachers expect students to complete and turn in assignments even when late. Late work will be accepted within two weeks of the due date with a 10% penalty. After two weeks, late assignments may be accepted until the end of the quarter, depending on the department's policy. If a student does not turn in an assignment on time, the placeholder NTI (Not Turned In) should be entered into the grade book until the late assignment is turned in.

- **Make Up Work.** Students have one school day for each excused absence to make up missed work for full credit. Students should check Schoology and reach out to their teachers individually to keep up with schoolwork when they are absent.

Final Exams/Culminating Activity. The final examination or culminating activity will account for 10% of the total course grade. Per district policy, midterms and final exams are not eligible for retakes, including in Honor Code cases. The lowest score for a final exam will be a 50%.

SIS Comment Codes. Teachers may use these optional codes in SIS to clarify the status of certain assignments in the gradebook.

- **NTI (Not Turned In)** - Student has not turned in the assignment. This comment carries a score of 0.
- **LA (Late) -** The student handed in the assignment late. There is a 10% penalty for late work.
- **AB (Absent) -** The student was not present in class, and the missing work must be completed. This comment code may be changed to a grade of NTI if the assignment does not get turned in.
- **EXC (Excused) -** The student has been excused from doing this assignment.
- **NYG (Not Yet Graded) -** The student has turned in the assignment, but the assignment has not yet been graded.
- **RTK (Retake)** – This grade improved as demonstrated by the student's retake exam.
- **GRR (Grade Replacement)** – This grade improved as demonstrated by a student's work on another assignment
- **RTL (Retake but Lower)** – Student's demonstration of knowledge did not improve on the retake. Student keeps original higher score.

Extra Credit. Students cannot earn extra credit for tasks that are not related to the curriculum (e.g. bringing in classroom materials, providing parent signatures, participating in fundraising/charitable events or participating in non-curricular activities).

Course-specific Grading Information

Course-specific Grading Information: [Enter text here. Include additional information on grading and assessment practices specific to the course beyond what is included in the general guidelines above. Details should include but are not limited to:

1) course content and structure, 2) information on how late assignments and/or assessments are handled, 3) how and when retakes will be scheduled once an assessment is returned, 4) how much time is provided to schedule and take a retake, and 5) what remediation requirements are necessary to access the retake opportunity.]

Additional Help: We want every student to be successful. There are several opportunities to get additional help if a student is struggling in class: [Modify as appropriate.]

- Teachers are available during Advisory and after school hours (by appointment).
- Math Lab: Our math honors students provide free tutoring during advisory. There is a sign-up sheet in the classroom for students to schedule time with these tutors.
- The district is partnering with Varsity Tutors to provide students and families with extra learning resources. Every student can take advantage of Varsity Tutors' services at no cost, including weekly live online classes; on-demand 24/7 chat tutoring; essay editing; and study resources such as recorded content, practice problems, and diagnostic tests. [district provided access code]

Testing Information: The state end-of-year assessment in Geometry will be administered in May. All freshmen students will be required to take the exam. Additionally, any student who has not yet received a verified math credit (though the passing of another math state assessment) will be required to take the exam. [Modify as appropriate.]

School Name Honor Code

The *School Name* Honor Code reflects the fundamental beliefs that

- Every student has the right to learn in an academic environment free from any form of intellectual dishonesty.
- Honesty and integrity of all members of the school community are crucial in sustaining academic excellence.

The functions of the *School Name* Honor Code are to

- Communicate the meaning and importance of intellectual honesty to all members of the school community
- Explain and support the highest standards of conduct in academic and co-curricular affairs; and
- Educate, identify, and sanction those who do not follow these standards.

Advisory & *School Name* Time

Advisory. Every student is enrolled in an Advisory class for 4th period. This course meets from 9:46-10:16 on Blue days and is designed based on the principles of Responsive Classroom and social-emotional learning.

Advisory time. Advisory time is our academic enrichment and intervention period that occurs just after Advisory on most Blue days. During this time, students may be assigned to an intervention group based on academic, attendance, or wellness needs. Students who are not assigned to a group may be requested by their classroom teachers or may remain in their Advisory classroom to work quietly.

Technology

Digital Learning Consent. At *School Name*, our teachers will use a variety of digital resources to support student learning and promote communication and collaboration with students and families. Some of the digital resources your child may use this year will require your consent before they can be used. Please complete the Digital Consent Form for your student.

SIS & Schoology. SIS is the official gradebook for all district courses. Teachers update SIS weekly or bi-weekly. Parents and students can access SIS for grade, attendance, and other information via ParentVUE and StudentVUE.

All *School Name* courses use Schoology as the primary learning platform where students can find information about daily lessons, assignments, and upcoming assessments. Students should familiarize themselves with each of their Schoology classes as well as the Schoology calendar feature to stay organized throughout the school year. Parents with a ParentVUE account also have access to their children's Schoology courses.

Cell Phones & Personal Devices. Per the District & *School Name* Cell Phone Policy, use of cell phones, headphones, and smart watches during class time and in restrooms are prohibited.

District-issued Devices. Every student is issued a Chromebook and charger at the beginning of the year. We expect students to use these devices responsibly for academic purposes and abide by the District Acceptable Use Policy. We also expect students to take care of their devices by keeping them clean and free from graffiti, stickers, and other damage beyond normal wear and tear. Lost or damaged devices will incur a repair or replacement fee.

Artificial Intelligence (AI). Artificial Intelligence (AI) is a type of technology that involves computer programming and data. There are many different types of AI tools. Some can use data to make predictions, decisions, and/or generate content. Your teacher will specify when it is acceptable for you to use AI on any part of an assignment. Your teacher will also review expectations for citing AI and/or how students should explain their use of AI. Students should follow the Honor Code to maintain academic integrity. You should be able to demonstrate that you understand the information you turn in. AI tools can produce false information called "AI hallucinations." You should always check what AI creates. Make sure it fits the needs of the project and audience. You remain responsible for the content you submit. Students should not use AI on an assignment without the teacher's explicit permission.

- Before using AI tools on assignments, students should ask me (your teacher):
  - if AI can be used on any part of an assignment,
  - how it can be used,
  - how to show your work, and
  - how to cite or explain their use of AI.

- In general, students should not enter any of the following information into AI Tools:
  - personal information (name, date of birth, etc.);
  - education records (grades, transcripts, etc.); and
  - work that belongs to others.

You can search the District Digital Ecosystem Library (DEL) using "AI" (in quotes) to see which AI tool features are approved for use on district devices or networks. The DEL provides information on approved audience(s) for each tool. We may use these tools as appropriate in this course. Your use of AI must comply with the Student Rights and Responsibilities (SR&R). This includes the Acceptable Use Policy in Appendix A of the SR&R.

Personal devices may enable access to other AI tools (not approved in the DEL).
- Use of other AI tools will not be required in this class.
- Parents / Caregivers decide what access to AI tools is appropriate on personal devices.
- Parents and students are encouraged to read the AI tool's terms of service and privacy policies. Find out what kind of information the AI tool can collect, use, and share and what age students must be to use the tool. Check whether you retain rights to your data, the ability to delete your data, and the ability to opt out of data sharing. Be aware that some AI tools may keep the information you enter and use it without permission. You might not keep all rights to your work.
- Be careful with any information you give an AI tool.

CommonSense Media Resources on AI and Parent Tips and FAQs provides supportive resources that parents can explore.

Regardless of whether you use a personal device or a district device, students must follow teachers' expectations, the Honor Code, and SR&R.

Syllabus Acknowledgement Form