8-15-2014

Automated Proficiency Indicators to Improve Non-Proficient Scores

Nick C. Maile
Cedarville University

Recommended Citation

Maile, Nick C., "Automated Proficiency Indicators to Improve Non-Proficient Scores" (2014). *Master of Education Research Theses*. 69.
http://digitalcommons.cedarville.edu/education_theses/69

AUTOMATIED PROFICIENCY INDICATORS TO IMPROVE NON-PROFICIENT SCORES

A project submitted in partial fulfillment
of the requirements for the degree of
Masters of Education

By

NICK CHRISTOPHER MAILE
B.A. Middle Childhood Education, Cedarville University, 2002

2014
Cedarville University

ABSTRACT

Maile, Nick C. M. Ed., Education Department, Cedarville University, 2014. *Data-driven intervention to improve non-proficient scores.*

This action-research study compared improvement data for 9th grade Algebra students. The study entailed the use of automated proficiency indicators to guide intervention with students not meeting proficiency standards after an initial assessment. Data from the previous school year was compared and results revealed questions and implications of multiple levels of non and low proficient students. The goal of this study was to show if using a gradebook with automated proficiency indicators would help improve proficiency scores to a greater degree.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank my wife Julie, your patience and support has no rival. Thanks you to my daughters Annabelle, Allison, Samantha, and Charlotte who sacrificed time with daddy during this process.

I thank God for putting such great friends in my life that were my support for this journey. Thank you, Maureen, Joe, Nancy, and Todd for your help.

Thank you, Austin and Lindsey, for co-teaching with me day in and day out. And thank you Lindsey for your spreadsheet, as I would not have had the idea for this study without it.

Chapter 1

Introduction

Currently, the grading systems of some public schools are under scrutiny (Erickson, 2011). Educators, researchers, and writers have investigated grading models for many years, continually trying to arrive at the best model for assessing student achievement. However, large scale grading and reporting reform have been absent as schools often have relied on traditional grading systems. According to Azeem, Afzal, and Majoka (2010) a few commonplace models of traditional grading systems are fixed percent scale, total point method, and the norm-referenced model. However, many educators do not believe traditional grading systems accurately depict students' knowledge and ability. Jung (2011) claimed that the current grading practices are grounded in tradition rather than research in best practice. Often teachers average scores to calculate grades by combining indicators of achievement, behavior, and progress into a single score. Scores may be used directly with a fixed percent scale, or scores may be used to assign each student to a standard deviation curve using a norm referenced model. Educators practice these traditional grading methods despite evidence showing the detrimental consequences.

In some traditional systems, according to Marzano and Heflebower (2011), students earn points for activities, assignments, and behaviors, which accrue throughout a grading period. The teacher adds up the points and assigns a letter grade. A common variation of the point system is to convert points to a percentage score. Percentages are then weighted, combined, and translated to a letter grade. Many teachers use combinations of these traditional grading systems. Although these methods have been utilized for many years, credible educators and researchers assert that these grading methods do not accurately assess what students know and can do- the grades are arbitrary numbers not fixed to specific skills. Spencer (2012) contended that grades tend to be vague and unreliable in traditional grading systems. Some traditional grading systems may be used to reward or punish students for factors unrelated to competency.

Tomlinson (2005a) asserted that grades are representation following a specific amount of time that serves as a summary of evaluations about the students. The goal should be to provide high quality feedback that would “better serve the purpose of providing interested individuals with useful information” (Tomlinson 2005a, p.263). Bagley (2010) stated that grades and assessments should be shifted to communicate more specific information that nurtures reflection and growth. Scriffany (2008) wrote that one grading practice that is gaining popularity is standards-based grading, which involves measuring students’ proficiency on well-defined course objectives.

According to Marzano (2011), standards-based grading conveys how well students have achieved skills and standards. In a standards-based grading system, grades are not about what students earn; grades are about what students learn. Standards-based grading is an innovative grading system that not only breaks down each and every skill students must learn, but provides students with multiple opportunities to show the mastery of skills. Tomlinson (2005a) stated that all students can learn, grow and be challenged at various levels. Standards-based grading disseminates standards which allow reflection to skill-specific attainment. However, growth can only occur if students are frequently assessed, taught, and retaught based on individual needs. To successfully implement standards-based grading, Brookheart (2011) contended that teachers need to develop teaching and learning strategies, formative assessments tied to targeted objectives, and coaching strategies. The development of these strategies and assessments should be as important as the need to develop grading plans. However, one of the most important components of a standards-based grading system is differentiation (Tomlinson, 2000; Tomlinson & Kalbfleisch, 1998; Huang, 2012).

Statement of Issue

During the course of a school day, an average educator teaching at the middle to secondary level may have between 45 and 90 minutes for planning instruction often referred to as “plan time.” In that plan time, teachers have many commitments and responsibilities. Not only must teachers meet with content teams, grade level teams, and other school-related committees, teachers must plan curriculum, write assessments, and evaluate and grade student work. To ensure that continuity exists within the content area, teachers need to collectively reflect on current data in order to design lessons that are differentiated to allow learning at multiple levels. The reflection process is necessary for effective teaching and learning. However, when time becomes a constraint, differentiation and data reflection can easily become neglected (Tomlinson, 2000). In some traditional grading systems, current calculating practices may contribute to skills being ignored or not mastered. Lack of clarity in reporting has a major influence on these unattained skills not being revisited. Standards-based grading, on the other hand, requires the teacher to disseminate between the skills achieved, the proficiency level attained, and the need for future learning (Marzano &

Heflebower, 2011). Standards-based grading magnifies the need for reflection on needed skills which is strengthened when data is at the core of the process (Brimijoin, Marquissee, & Tomlinson, 2003; Tomlinson, 2005b; Frohbieter, Greenwald, Stecher, & Schwartz, 2011). When collection of data is separated by skill, the proficiency level of the skill begs for action from the teacher to differentiate.

The ability to effectively differentiate hinges on formative assessment data (Santangelo & Tomlinson, 2008). The formative assessment data aids focused replanning and reteaching (Acosta-Tello, Shepherd, 2014). Differentiating will assist the teacher in meeting the individual needs of each student wherever that student may be in the learning process (Tomlinson, 2000).

Instilling educators with a data-driven decision making process will allow increased rigor in the classroom while helping avoid the use of gut feelings or opinions (Mandinach, 2012). Carlson, Borman, and Robinson (2011) found significant increases in math achievement after a yearlong study which focused on being data-driven. Use of assessment data can help us reach all levels of students (Kadel, 2010).

As school districts are making changes towards using standards-based grading, the data collected has become more focused. With no more time to plan and reflect built into the school day, a way to efficiently use standards-based data is needed. This study's intent was to show data-driven decision making, which derived from a gradebook that automated proficiency indicators, would result in greater improvement of non-proficient students.

Scope of the Study and Delimitations

The scope of the study included building and using an Excel spreadsheet to analyze the data used for quantifying and sorting scores by proficiency level which aided in differentiation. Automated procedures were built into the gradebooks to allow for the sorting process to occur. All 9th grade Algebra 1 classes at Southview Middle School had access to the gradebooks for the 2013/14 school year. These prolific gradebooks were utilized by all three teachers in the 9th grade math department. The Ankeny school district's power standards for Algebra 1 were used with each power standard being represented by a different cycle.

Significance of the Study

Pairing standards-based grading's consistent emphasis on differentiating for the individual (Tomlinson, 2000), with the gradebook's built-in automated procedures, gaps are quickly revealed. It is apparent that teaching to the middle is becoming not acceptable (Tomlinson & Kalbfleisch, 1998; Tomlinson, 2004; Dixon, Yssel, McConnell, & Hardin, 2014). Advanced students need to be challenged, and struggling students may need scaffolding and/or different approaches to show proficiency in the standards (Tomlinson, 2004, 2005b).

The automated proficiency levels allowed for timely response to assessment data enhancing the ability to challenge each level of student in the differentiation process. Because the timeliness of response is so crucial in the learning spectrum (Tomlinson, 2005a; Qu & Zhang, 2013), teachers would benefit from a quick and efficient tool providing them with the educational data on which to reflect and make the necessary educational adjustments.

Methods of Procedure

The gradebook with built-in automated procedures was created in Microsoft Excel. A series of spreadsheets were used to set up a basic gradebook that would automatically calculate percentages after entering assessment data. Once the basic gradebook was created, I automated the data entered to classify students into the different subgroups: non-proficient, proficient, highly proficient, and not attempted. The comparisons of assessments from the same standard were automated.

After assessments had been evaluated, scores were entered into the gradebook within 48 hours. The spreadsheet program automatically calculated a proficiency rating for each student based on the standard taught and assessed. The proficiency rating for each student was reviewed prior to the reteach of that same standard. The proficiency rating was used to guide interventions with the appropriate students during the bell ringer portion of class. The bell ringer is attempted during the first ten minutes of most Algebra 1 classes. The bell ringer includes various review questions. At least two of these questions reviewing a previous standard. If no new material was scheduled, the automated proficiency data assisted in the creation of groups of various proficiency levels. These data-driven groups allowed differentiated lessons which challenged each level of learner. When the teaching of the new standard was complete, both power standards were assessed. Each standard was assessed during one class period for a testing duration of two days (see Figure 1).

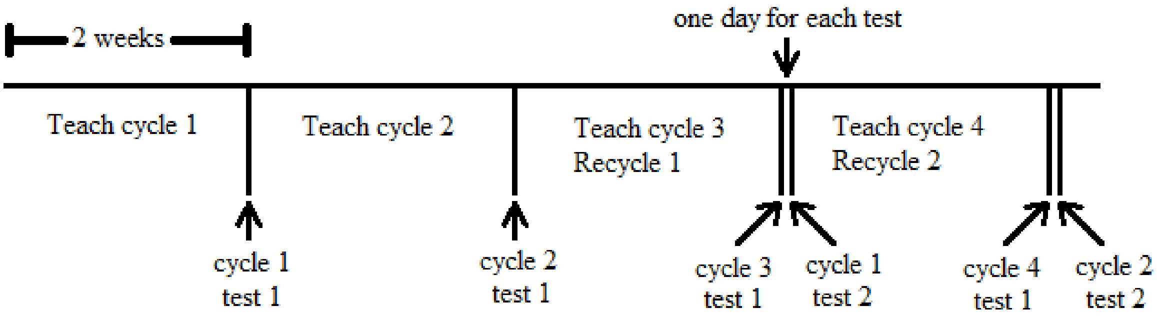


Figure 1. Representation of current assessment process with review of previous standards beginning during cycle 3.

At the completion of the second attempt at a standard, the results were again reviewed to help make decisions on further interventions needed for non-proficient students.

At the conclusion of the study, data was sorted by proficiency level for each of the seven power standards that were assessed during the first semester. Each student had their first test in each power standard compared to their second test in each power standard which resulted in the percent change for that cycle. The statistical analysis used the percent change data from each student for the first seven cycles.

Definition of Terms

Bell Ringer- An activity that students are expected to begin as the bell rings at the beginning of class that ensures immediate engagement (Ankeny Community School District, 2014)

Criterion referenced grading- A grading system that compares each student’s performance to clearly stated performance descriptors that differentiate levels of quality (Guskey, 2001)

Differentiation- To systematically adjust curriculum and instruction to fit each individual’s need (Tomlinson, 2000)

Formative assessment- A type of assessment that is not used for reporting purposes, but to guide decisions on teaching, reteaching, and differentiation (Wormeli, 2006)

Norm referenced grading- A grading practice by which students grades are formed by comparing each students performance to that of others in that student's group or class (Guskey, 2001)

Power standard- An essential skill or standards that was board approved and required learning for a specific content level (Ankeny Community School District, 2014)

Standards-based grading- A grading and reporting practice which measures students' varying achievement levels on specific course objectives (Tomlinson, 2006)

Summative assessment- The assessment whose results represent the proficiency attained by the end of the learning progression and is reported (Wormeli, 2006)

Traditional grading systems- Grading systems that have been used by educational institutions for decades with little or no reform (Jung, 2011).

Chapter 2

Literature Review

Current Grading Conundrum

Schools across America are struggling to not only meet the needs of their diverse students but to assess the students in a grading system which may not be based on the students' knowledge of skills (Dalziel, 1998; Tomlinson, 2005a). Fisher, Frey, and Pumpian (2011) wrote of their teaching conundrum that 55% of the students at Health Sciences High and Middle College in San Diego, California, failed Algebra I. According to Fisher, Frey, and Pumpian (2011), their school was typical of many high schools; grades were derived from a number of categories including tests, quizzes, projects, homework, and classroom behavior. Students were failing for many different reasons such as incomplete homework, attendance issues, and low test scores. Teachers reported that some of the students who failed might have understood the material.

The frustrations that were encountered at Fisher, Frey, and Pumpian's school are not an exception. Erickson (2011) wrote of a high school in Minnetonka, Minnesota, where parents called for grading reform. Teacher surveys showed a large variation in grading factors aside from mastery of course content were used to calculate grades such as attendance, behavior, effort, extra credit, and participation. The district was forced to revise its grading policies after the parents called for more transparency and consistency.

As a result of known discrepancies in current grading practices (Dalziel, 1998; Bagley, 2008), educators and administrators across the country are now scrutinizing their existing grading policies. Efforts are being made to provide consistency in reporting by correlating letter grades with students' knowledge of the content. However, Reeves (2011) stated that "Grading policy is among the most emotional topics in education today" (p. 76). Erickson (2010) created the analogy of the Social Security system to the grading system. He wrote that "although Social Security funds are in decline and no solution is evident, few politicians have the temerity to try to change the system. Because Social Security is the third rail in politics, if you touch it, you'll die" (p. 22). Likewise, in education, the issue that is equally as lethal is grading. Erickson (2010) wrote:

> Grading is one of the most private experiences for students and teachers in the learning process. Usually, a teacher's grading protocols, which can be harmful to students, originate from his or her own experiences as a student. To implement universal and consistent grading practices, strong leaders and the willingness to tackle deeply rooted and harmful grading traditions are required. (p. 22)

The catalyst for change in grading traditions comes at a time when "government, business and industry and the general public are calling for an accountability of student knowledge and abilities" (Campbell, 2012, p. 30). Although teachers do not have control over state-mandated tests, "they do have control over many day-to-day measures, including how classroom assessment is implemented, and their own grade books and what is recorded in them" (Campbell, 2012, p. 30). Reeves (2008) stated "the difference between failure and the honor roll often depends on the grading policies of the teacher" (p. 85).

Dalziel (1998) contended the main problem with grading is the lack of "direct correspondence between the empirical property being assessed (student performance on a task), and the 'real numbers' of mathematics" (p. 353). The lack of correspondence paired with final grades with vague disseminations can lead to great variance from teacher to teacher (Dalziel, 1998). Campbell (2012) noted that grades may be affected by student motivation, self-esteem, and the social consequences attributed by the teacher.

The accountability of student knowledge comes with the need for grades to reflect the students' understanding of the content. Prior to the development of the Common Core State Standards, most school districts had their own set of skills or standards which every student should master in every discipline every year (K-12); however, those school specific skills were compiled with attendance, behavior, and homework and not assessed individually. Erickson (2011) shares that innovative school districts in California, Minnesota, and Virginia began to "look at one guiding question: What should go into a grade" (p. 66). The answer to the schools' question was very simplistic: "Grades should reflect only what a student knows and is able to do" (p. 66). Now that most states have implemented the Common Core, the succinct skills for each grade level, each discipline, and each state have been written. The creation of the Common Core and increased accountability movement are both impetuses to holding schools responsible for the learning of all students- in essence, the evolution of standards-based learning has been affected by these two movements.

Standards-Based Education

Colby (1999) writes that “as districts and states continue the move toward standards-based education, questions arise around best practices for implementation” (p. 52). First and foremost, teachers need to focus on standards, assessing students’ proficiency with district-developed assessments, and reporting student progress in relation to those standards. Gusky (2001) wrote of the need to move from a norm-referenced to criterion-referenced grading. Norm-referenced grading is considered grading with reference to others and is typically associated with a standard bell-shaped probability curve. Guskey(2001) later mentioned “criterion-referenced grading, in contrast, compare each student’s performance to clearly stated performance descriptions that differentiate levels of quality. Teachers judge students’ performance by what each student does, regardless of how well or poorly their classmates perform” (p, 20).

In the standards-based system, grading and reporting must be criterion-referenced (Tomlinson, 2005a). Teachers, at all levels, must collectively identify the standards students are expected to learn and what evidence will be used to judge that achievement or performance. It is then, and only then, that grades will be based on clearly stated learning criteria. A student’s grade will have direct relevance to a standard, and the reporting will clearly communicate that attainment of the standard (Shippy, Washer, & Perrin, 2013).

Once the course standards are clearly defined, a policy for consistently and objectively reporting student achievement must be established. Qu and Zhang (2013) stated the standard-based grading system, teachers use two assessment categories: formative and summative. Formative assessments are assessments made throughout a learning progression.

Formative assessments inform both the teacher and the student of the student's progress in mastering material that will appear in the upcoming summative exam (Wormeli, 2006).

Teacher use formative assessment data to re-group, reteach, and perhaps extends students' learning (Frohbieter, Greenwald, Stecher, & Schwartz, 2011). To be truly effective, teachers must review the formative assessments in a timely manner because these formative assessments are directly connected to skills which must be mastered (Tomlinson, 2005a; Qu & Zhang, 2013).

Summative assessments, as the word indicates, are the final grade for a skill- The summative score is the grade that is placed in the gradebook as the representation of the student's understanding compared to the standard (Wormeli, 2006; Frohbieter, Greenwald, Stecher, & Schwartz, 2011; Qu & Zhang, 2013).

O'Connor and Wormeli (2011) claimed that students need helpful feedback. "If [teachers] grade the formative steps that students take as they wrestle with new learning, every formative assessment becomes a final judgment, with no chance for revision and improvement. Feedback is diminished, and learning wanes" (p. 44). Formative and summative reports must be distinct from each other to be useful. Formatives should be given during the learning process to help foster growth, and summatives should be given at the end of a learning progression as a declaration of understanding. The data from summatives and formatives must support the intended purpose.

Differentiation

Varying degrees of student readiness, background knowledge, and skill deficiencies must be addressed. Concerns are met when the teacher uses the results from the formative assessments to intentionally group students for additional practice. Along with the additional practice come reteaching, modeling, and feedback (Tomlinson, 2005a; and Frohbieter, Greenwald, Stecher, & Schwartz, 2011). Tomlinson (2003) wrote that "students who are the same age differ in their readiness to learn, their interests, their styles of learning, their experiences, and their life circumstances" (p. 6). Because of these differences, student's learning is impacted; therefore, the pacing and the support must be varied. Tomlinson (2003) contends that curriculum shows educators what to teach, and differentiation tells educators how to teach. Differentiation simply suggests ways in which educators can make that curriculum work best for varied learners.

When planning a lesson, teachers initially choose the skills which must be taught and ultimately mastered. Then the teacher creates a learning progression with multiple formative assessments integrated into the unit. In addition, the teacher will prepare for differentiation, providing materials and tasks on the standard at varied levels of difficulty, with varying degrees of scaffolding, through multiple instructional groups (Tomlinson, 2005a). Teachers can craft lessons in ways that tap into multiple student interests to promote heightened learner interest in the standard (Dixon, Yssel, McConnell, & Hardin, 2014). Tomlinson (2003) suggested that "teachers encourage student success by varying ways in which students work alone or collaboratively, in auditory or visual modes, or through practical or creative means" (p. 9). The end of the unit culminates with a summative assessment, but that does not occur until students are ready for that assessment.

Formative feedback is key to the success of students in a differentiated classroom. Wormeli (2006) claimed that students do not always know what they know or what they do not know. Frequent feedback from formative assessments is key. Continual assessment that allows immediate feedback helps students correct misconceptions. "Tweens learn more when teachers take off the evaluation hat and hold up a mirror to students, helping them compare what they did with what they were supposed to do." (p. 18)

Teachers who employ the instruction strategy of differentiation frequently provide ongoing formative assessments, timely feedback, and intentional groupings. The expected result is to have students whom often have success and master the learning objectives.

Data to Drive Instruction

In the differentiated classroom, assessments are not only frequent, but are diagnostic. Marzano (2003) stated that schools using data are following advice from the world of education and the world of business. "If schools are in the business of helping students learn, then the data used to guide decisions should relate directly to student achievement." (p. 56)

Doubet (2012) stated that teachers should have day-to-day data that can help guide instruction. Teachers should be using formative assessment to gather data on students' grasp of learning objectives. Formative data should then be used to drive instruction, group students, and tailor instruction for small groups.

Data should not just be collected, it should be cherished (Marzano, 2003). "The use of data allows for organized, simplified discussions that merge to create focused priorities and productive action." (p. 56)

In conclusion, data should come directly from formative assessments in the classroom. Because the interpretation of formative assessment data is imperative for students' success, educators must have a system or plan in place for interpreting and using data. In addition to creating a plan, educators need to have a means of acquiring the data in a timely manner so that the continuum of learning will move at a sequential and, hopefully, a rapid pace. This capstone project does just that- it outlines the plan and the methodology for creating, storing, manipulating, and retrieving formative assessment data which the classroom teacher can utilize to advance the knowledge and skills of all his/her students. The proficiency data can successfully mirror the students' acquisition of skills in each and every content area along the learning continuum.

Chapter 3

Methodology

A quantitative analysis was conducted to search for change in achievement from a student's first assessment on a specific power standard (course objective) compared to the second assessment of the standard. The purpose of the research was to see if having automated proficiency indicators on a consistent basis would result in greater improvement on students' assessment scores.

Students took an assessment after each power standard was taught. The scores were entered into the excel spreadsheet (see Figure 2). Once scores were entered into the spreadsheet, the test was toggled on by inputting a 'y' at the top of the column (see Figure 3). Automated percentage scores appeared directly next to inputted assessment values. Once toggled, a missing first assessment would translate to a '?' and a missing second assessment would appear as a '!'. After an assessment was turned on, automating formulas changed the percentage values into proficiency indicators seen at the spreadsheet's far right. This section is referred to as the "print sheet" (see Figure 4). The print sheet had all students listed alphabetically with each student's picture directly beside their specific proficiency scores. The print sheet was arranged so that all students' proficiencies levels for every power standard could be seen and printed on one piece of paper. The following proficiency indicators were used: scores less than 60% were "non-proficient" indicated by a blank space, scores greater than or equal to 60% and less than 80% were “low proficient” indicated by ‘/’, scores greater than or equal to 80% and less than 90% were “proficient” indicated by ‘X’, and scores greater than or equal to 90% were “high proficient” indicated by ‘O’. The print sheet used ‘?’ for a missing first assessment and ‘!’ for a missing second assessment (see Figure 5).

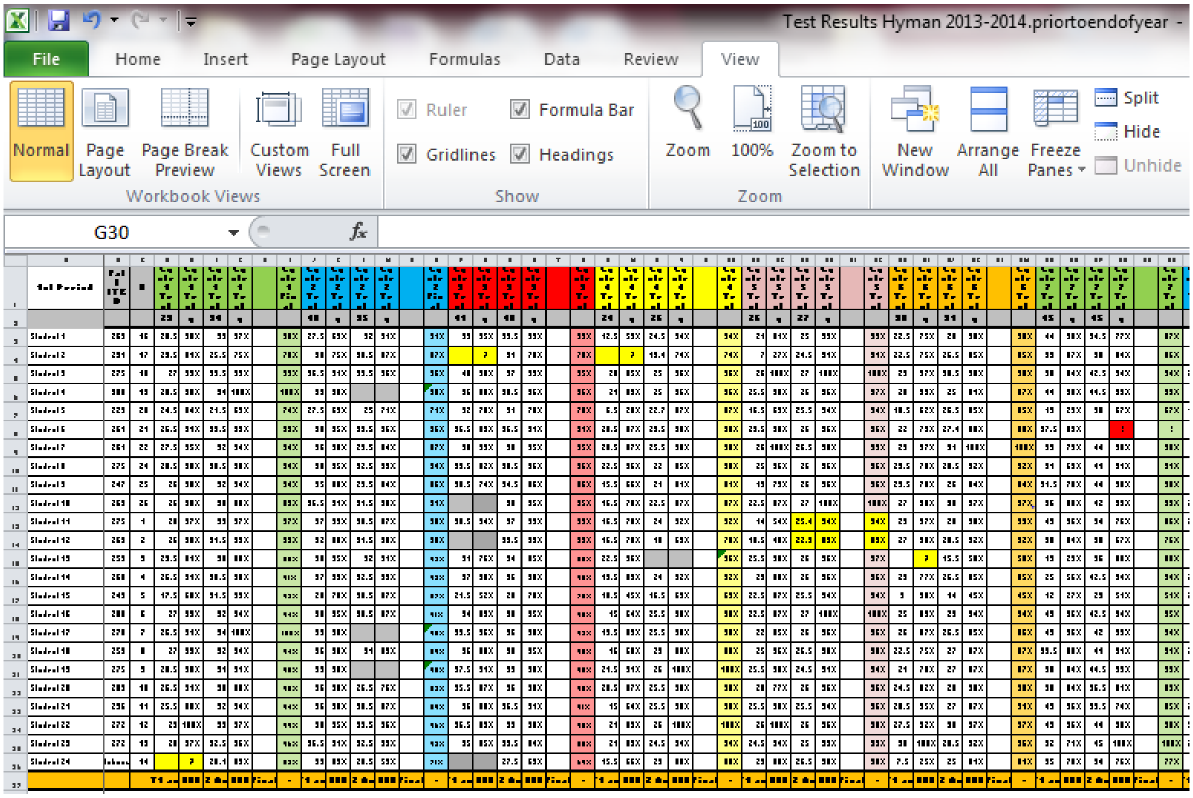
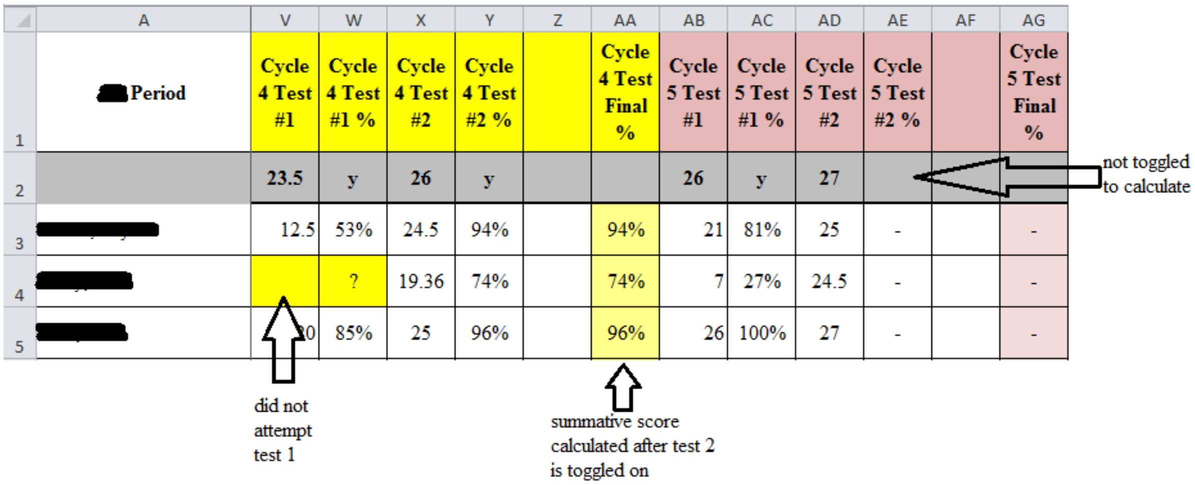


Figure 2. Automated gradebook with first semester assessment scores.



	A	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	[redacted] Period	Cycle 4 Test #1	Cycle 4 Test #1 %	Cycle 4 Test #2	Cycle 4 Test #2 %		Cycle 4 Test Final %	Cycle 5 Test #1	Cycle 5 Test #1 %	Cycle 5 Test #2	Cycle 5 Test #2 %		Cycle 5 Test Final %
2		23.5	y	26	y			26	y	27			
3	[redacted]	12.5	53%	24.5	94%		94%	21	81%	25	-		-
4	[redacted]		?	19.36	74%		74%	7	27%	24.5	-		-
5	[redacted]	20	85%	25	96%		96%	26	100%	27	-		-

did not attempt test 1

summative score calculated after test 2 is toggled on

not toggled to calculate

Figure 3. The assessment data translates to a percentage after being toggled with a ‘y’, and summative scores are tabulated once the second assessment has been entered.

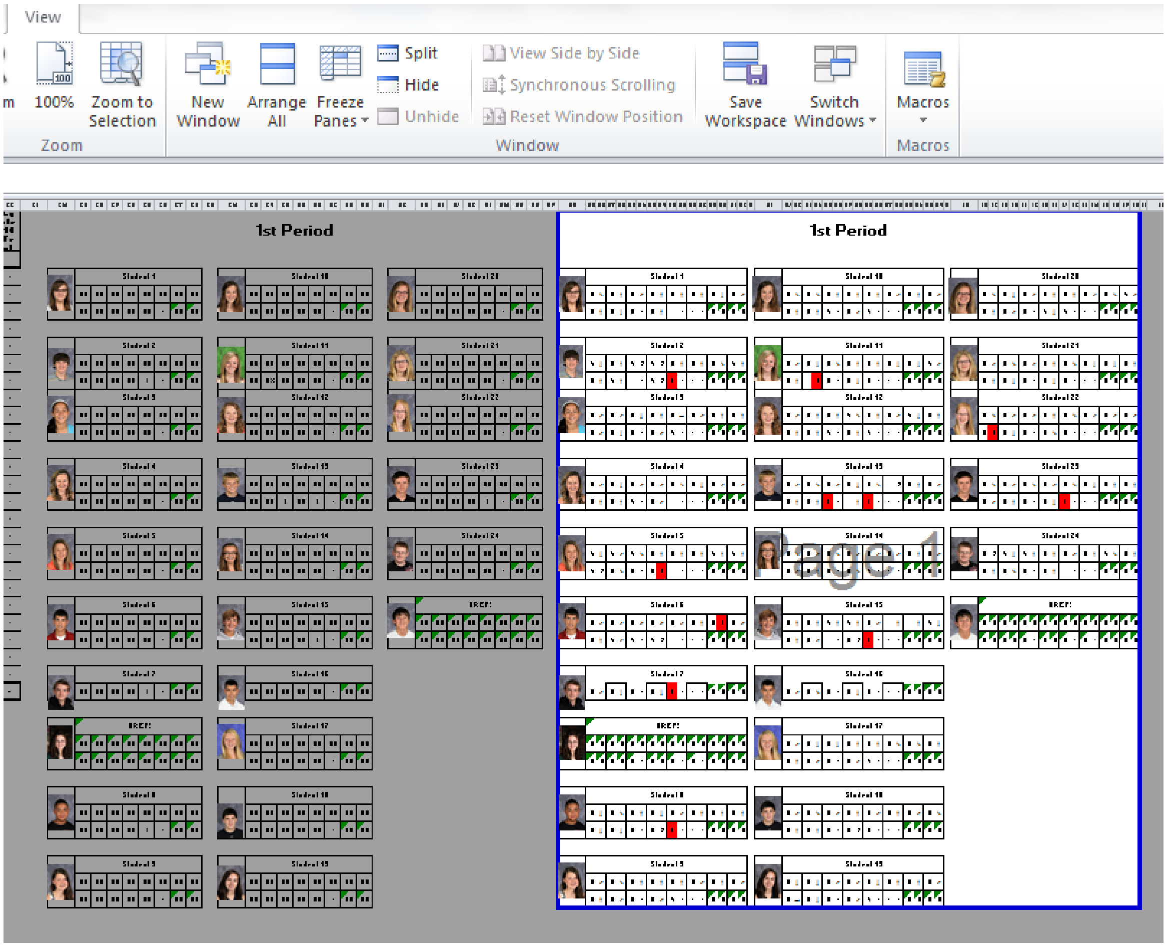
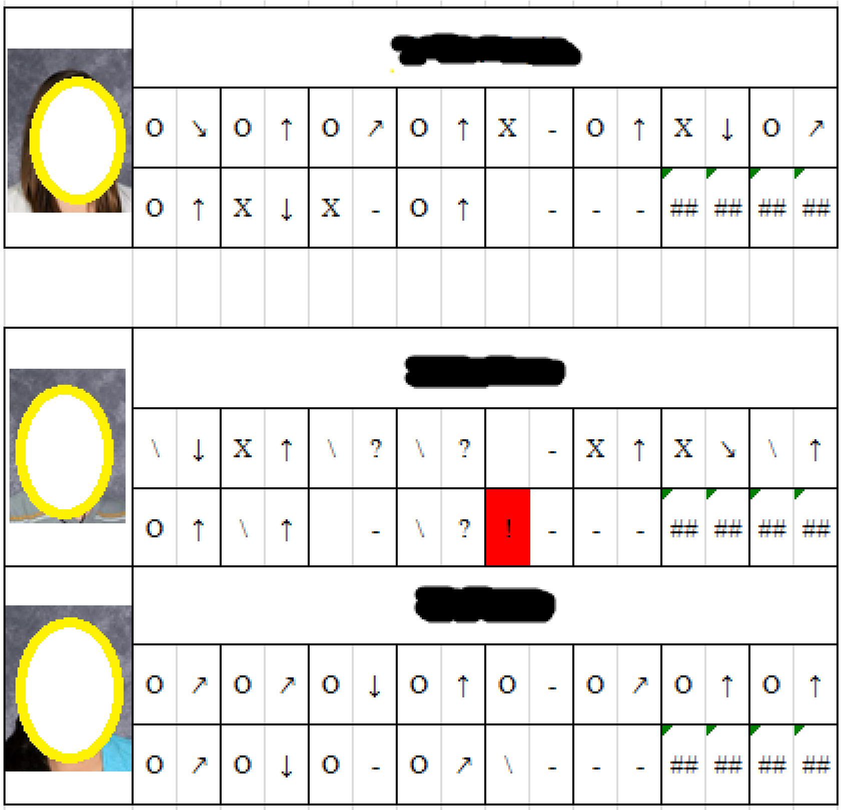


Figure 4. The print sheet showed an entire class' automated proficiency indicators for each cycle.



[redacted]															
O	↘	O	↑	O	↗	O	↑	X	-	O	↑	X	↓	O	↗
O	↑	X	↓	X	-	O	↑		-	-	-	##	##	##	##
[redacted]															
\	↓	X	↑	\	?	\	?		-	X	↑	X	↘	\	↑
O	↑	\	↑		-	\	?	!	-	-	-	##	##	##	##
[redacted]															
O	↗	O	↗	O	↓	O	↑	O	-	O	↗	O	↑	O	↑
O	↗	O	↓	O	-	O	↗	\	-	-	-	##	##	##	##

Figure 5. An example of proficiency indicators automated onto the print sheet.

Each power standard was revisited during at least one subsequent cycle with an average length of about two weeks. Students were reintroduced to the power standard approximately two to four weeks later through warm-up problems referred to as bell ringers. During approximately the first ten minutes of class, students attempted the review questions from the previous standard along with a few questions about the material taught in the previous few days from a new power standard and one question on material that would be new learning for that day. At the completion of the new cycle, two tests were given. The new material was tested on one day, and the recycled material from a previous power standard was tested on the next day. Both tests were scored. The new material was entered under its specific power standard as test 1, and the recycled material was entered in the previous power standard as test 2. Once a student had both tests completed from a specific standard, the summative score was then tabulated. If the second test in a power standard had a higher percentage than the first, then the second test's percentage was entered as the final score. If the second test was worse, the average of the two tests was used which is standard practice of the 9th grade math team (see Figure 2).

Comparison data between each set of tests was compiled from the first semester of two consecutive years. The difference was calculated between the test 2 score and the test 1 score for each of the first seven power standards which represented all power standards assessed and reported during the first semester. For each year, every score was filtered by classification: all classes, co-taught every day, most comparable, and high needs; and tier: below 80%, below 70%, and below 60%. The average of each tier was tabulated and a one-tailed *t* test assuming unequal variance with an alpha of .05 was performed comparing the means of 2013/14 to 2012/13.

Rationale for the Method

A quantitative study was performed to evaluate whether automating scores into proficiency indicators could enhance improvement scores on standards that were reassessed. The *t* test was chosen because of the variable population sizes, and a normal distribution can be assumed as the population is regionally consistent. A one-tailed *t* test was chosen because a positive correlation was expected. An alpha level of .05 was used due to its general acceptability as a standard measure of significance.

Population of the Study

The population of the study was Algebra 1 students that were in a co-taught by Mrs. Hyman and I during the school years of 2012/13 and 2013/14. All of the classes are taught in a suburban school district in Ankeny, Iowa, just north of Des Moines, Iowa. The classroom population is inclusive of special education, at-risk identified, and English language learners. The 2012/13 population consisted of 54 students with 85.1% white, 5.6% Hispanic, 3.7% Asian, 3.7% multi-racial, and 1.9% black. The population had 24.1% at risk, 13.0% special education including behavior disorders, and 7.4% English language learners. The population had 50.0% males and 50.0% females. The 2013/14 population consisted of 112 students with 89.2% white, 5.4% Hispanic, 2.7% black, 1.8% Asian, and .9% multi-racial. The population had 8.0% at risk, 11.6% special education including behavior disorders, and 2.7% English language learners. The population had 51.8% males and 48.2% females. All students during both years were 9th graders aged 14 and 15. Ankeny Community School District is a large district in Iowa with enrollment near 10,000 students. Ankeny is an upper middle class community. In 2011, the median household income was greater than $70,000 compared to the state median of just less than $50,000. Post high school education was sought by over 86% of graduating seniors for the past four years. The graduation rate of the Ankeny school district has been above 92.5% for at least the past six years.

Sample

Sample criteria. The students that were present the entire first semester of either school year were included in the study. If a student moved in more than two weeks after the first day or transferred out before the end of the semester, that student's results were removed.

The sample used consists of three different tiers: students who scored “less than 80%” on their first assessment, students who scored “less than 70%” on their first assessment, and students who scored “less than 60%” on their first assessment.

Assessment data was only used for individuals that had both test scores represented for a given cycle.

Rationale for sample. The sample was restricted to students that were present the majority of the semester to ensure the process of teaching and reteaching could occur. The data was categorized into different tiers so that trends could be identified, understanding that lower tiers are represented in all data. Proficient at grade level is marked at 80%, so the "less than 80%" designation was used to show all students that did not meet the expected proficiency level on their first attempt. 60% is Algebra 1's passing standard and marks the lowest point of basic proficiency. The "less than 60%" represents students who are non-proficient after the first assessment. "Less than 70%" was used to see if there were any trends in the students who showed only basic proficiency after test 1.

Method of sampling. Opportunity sampling was used for the population of the study. In 2012/13, any student in the Ankeny school district in 9th grade Algebra could have been placed in one of my classes. In 2013/14, any 9th grade Algebra student in the south half of the Ankeny school district could have been placed in one of my classes.

Procedure

Instruments. The comparative data was derived from assessments created in the 2011/12 school year. First semester's power standards were used represented by Cycles 1-7. Each assessment had approximately 10 to 12 questions. A variety of question styles were used such as multiple-choice, writing frame, and error analysis; although, the majority of the questions were open-ended.

Validity Measures. As the tests were written and used in 2011/12, the math team had a chance to reword, rearrange, and improve them. Some major changes may had been made between 2011/12 and 2012/13, but only minor changes were made between 2012/13 and 2013/14 which allowed for good comparison data. Only 1st semester tests were used because in 2013/14 the math department used a different method of intervention during most of second semester. Scores of students who did not complete both tests or students who did not attend the entire semester (with the exception of the first two weeks) were removed from calculations.

Only students' scores that had Mrs. Hyman and me as co-teachers were used. I co-taught with one other Algebra 1 teacher in 2012/13 and a different Algebra 1 teacher in 2013/14.

Data Collection Method. The data collected was the percentages earned on cycle tests. Percentages were calculated by dividing the number of points earned divided by the total points possible. Mrs. Hyman graded the tests given and entered the scores in a spreadsheet in 2012/13. In 2013/14 Mrs. Hyman graded the tests, and either of us entered the scores into the gradebook with automated proficiency indicators. Students took the cycle test 1 only after all the material was taught and reviewed. Cycle test 2 was taken after recycling the material in the bell ringers during a subsequent unit. If students missed a test, they had twice as many days as were missed to make up the test but were encouraged assess as soon as possible.

All data used in the study came directly from the spreadsheet gradebooks used.

Each student's percent change was tabulated between test 1 and test 2 of the first seven cycles. The data was arranged for comparing means of classes or groups of classes dependent upon class-specific criteria. A *t* test was used with the resultant means assuming different levels of variance. The *t* test was chosen due to the use of two variables: test 1 and test 2 from a given cycle.

Relevant Ethical Considerations. In 2013/14, my belief was that concentrated focus on differentiating for students based on proficiency data was the best method to reach students with non-proficient scores. In that light, the study was performed using past data instead of a control group. Different years' data would sacrifice some validity to the study, but I could not see using automated data for some and not for others.

Treatment Variables. The independent variable was the use of an Excel spreadsheet as a gradebook with built-in automated formulas. The gradebook translated data from the first attempted assessment of a standard to the level of proficiency attained. The proficiency interpretation showed itself at the far right of the spreadsheet in the area referred to as the "print sheet". The print sheet showed the level of proficiency attained. Symbols were used to represent the different levels of proficiency. '?' was used to represent a student had not taken the assessment, a blank was used for scoring less than 60%, '/' was for scoring at least 60% but less than 80%, 'X' was for scoring at least 80% but less than 90%, and 'O' is used for scoring at or above 90%. The symbols were used so the sheet could be carried around class without inadvertently sharing scores of other students. No key describing the meaning of the symbols was present on the print sheet. The print sheet helped guide my interventions to be with students that needed the most remediation during the bell ringer.

The dependent variable was a presumed effect that students with higher need being reached in a timely focused manner would increase the improvement from a first assessment in a power standard to second assessment in the same power standard.

Safeguards to Internal and External Validity. To ensure the validity, all class types were used whether high needs, low needs, beginning of the day, end of the day, with or without higher percentages of identified students. Consistency was held by using the same teaching team and the same methods of scoring the assessments.

Chapter 4

Results and Analysis

Introduction. The purpose of the study was to compare the effectiveness of reaching non-proficient/low proficient students using automated data. To limit uncontrolled variables, only classes fitting the following criteria were used: 9th grade Algebra 1 classes co-taught by both Mrs. Hyman and me. Classes that met the criteria were filtered into four classifications each containing three tiers.

The four classifications were 1) “all classes” co-taught by Mrs. Hyman and me, 2) “most comparable” classes being similar in ability level and amount of time that I was in the classroom, 3) classes “co-taught every day”, and 4) Classes considered “high needs”. The three tiers each category was separated into were: 1) scores of “less than 80%”, 2) scores of “less than 70%”, and 3) scores of “less than 60%”.

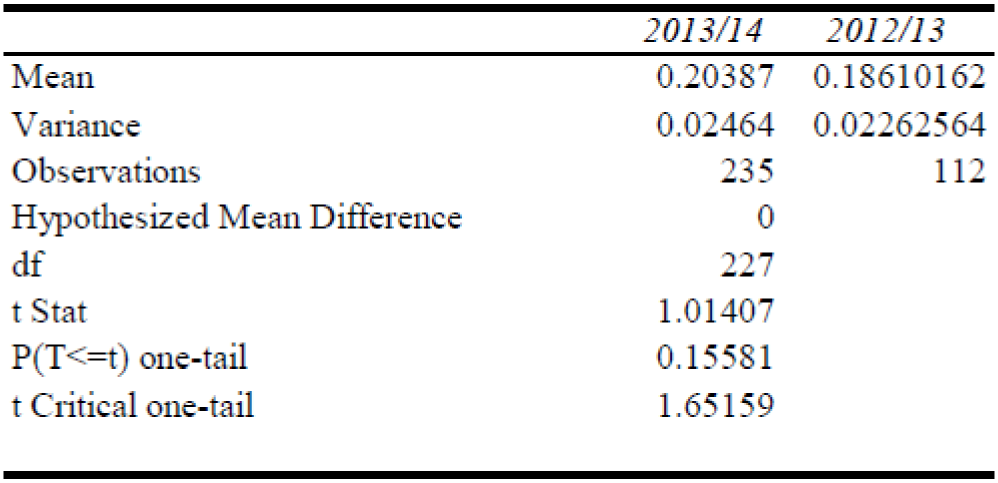
Description and Data Analysis for “all classes” co-taught by Mrs. Hyman and me. In the 2012/13 school year, I co-taught with Mrs. Hyman two classes, and in 2013/14 I co-taught with her four classes. For each of the following tests, the null hypothesis being $\mu 1 = \mu 2$ with $\alpha = .05$, a one-tailed *t* test was performed using the means of the change in percentage from test 1 to test 2 for each cycle in each given class.

Comparing students scoring less than 80%, there was a relatively proportional number of observations at 235 in 2013/14 compared to 112 in 2012/13, that is to say 235 test

1 scores for all of cycle 1 through cycle 7 were lower than 80% in 2013/14. These are relatively proportional considering in 2012/13 there were two classes and in 2013/14 there were four classes with slightly larger class sizes. The test showed a result of 1.014 standard deviations from the null hypothesis with a *p* value 15.58% (see Table 1). The *t* test result showed significant value while still not reaching the critical *t* score of 1.65 which would have resulted in the rejection of the null hypothesis (see Figure 6).

Table 1

All Classes in the Study: Scores < 80%



	2013/14	*2012/13*
Mean	0.20387	0.18610162
Variance	0.02464	0.02262564
Observations	235	112
Hypothesized Mean Difference	0	
df	227	
t Stat	1.01407	
P(T<=t) one-tail	0.15581	
t Critical one-tail	1.65159	

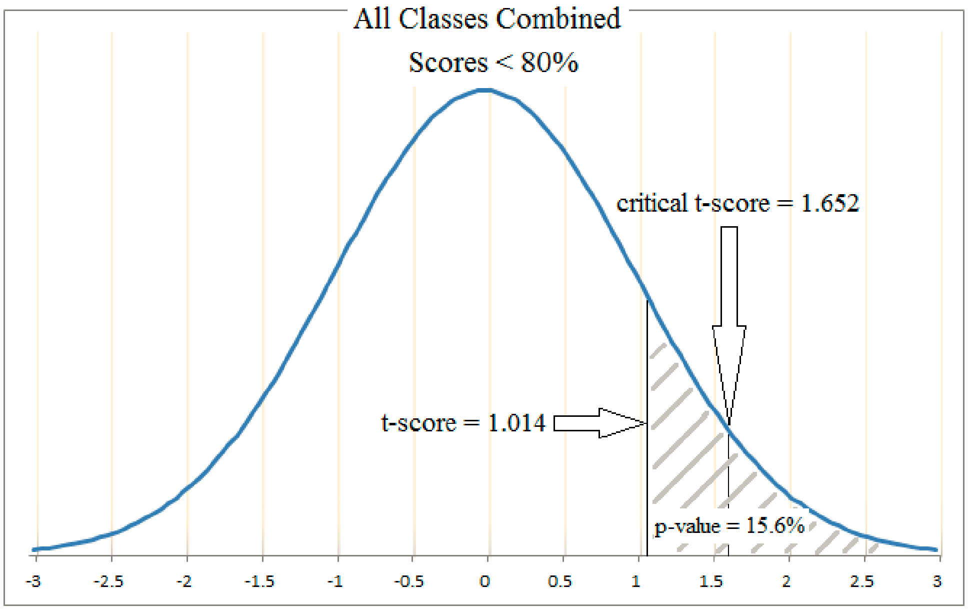
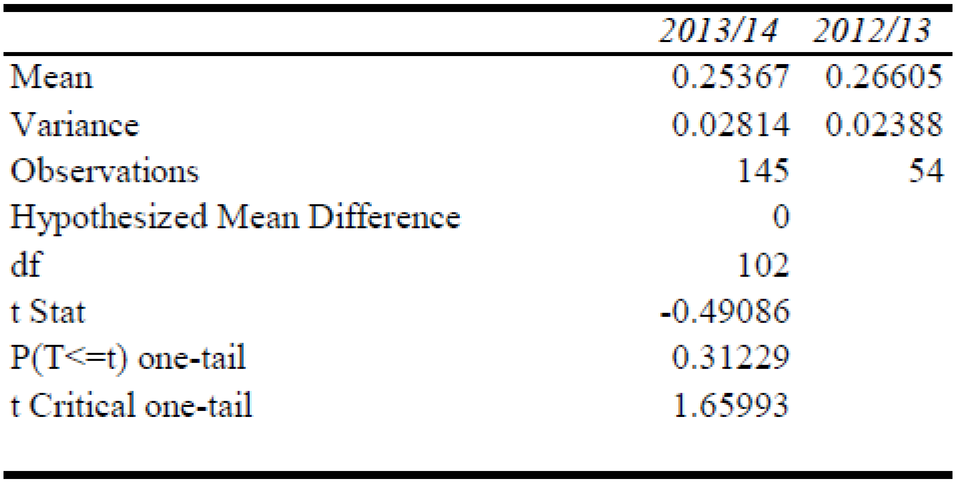


Figure 6. Standard deviation curve representation of Table 1.

Comparing students scoring less than 70% had a slightly disproportionate number of observations at 145 in 2013/14 compared to 54 in 2012/13. The test showed a result of -.491 standard deviations from the null hypothesis with a *p* value of 68.8% (see Table 2). The *t* test result shows insignificant placement compared to the null hypothesis at less than one standard deviation (see Figure 7). The *t* score comparing students less than 70% for “all classes” co-taught in the study is the first occurrence where a negative correlation was noted in the comparison data.

Table 2

All Classes in the Study: Scores < 70%



	2013/14	*2012/13*
Mean	0.25367	0.26605
Variance	0.02814	0.02388
Observations	145	54
Hypothesized Mean Difference	0	
df	102	
t Stat	-0.49086	
P(T<=t) one-tail	0.31229	
t Critical one-tail	1.65993	

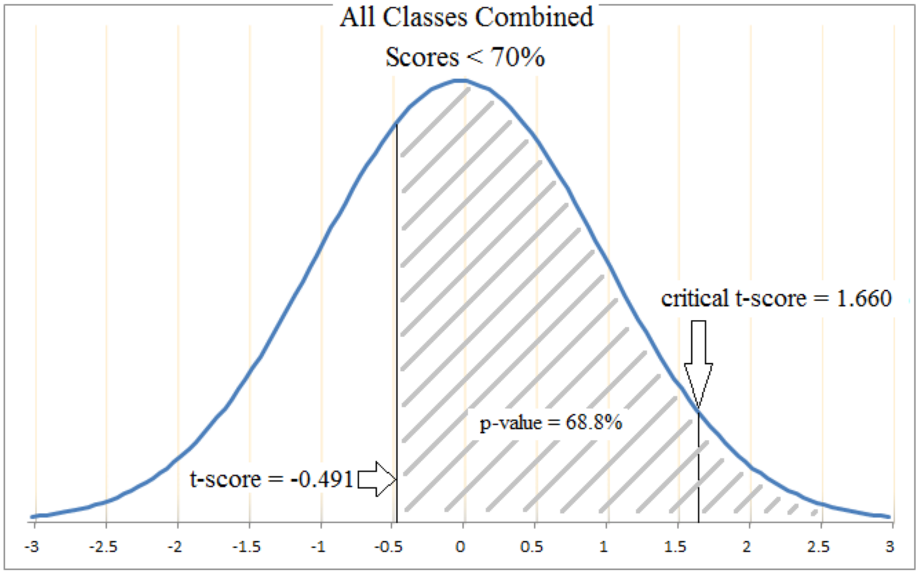
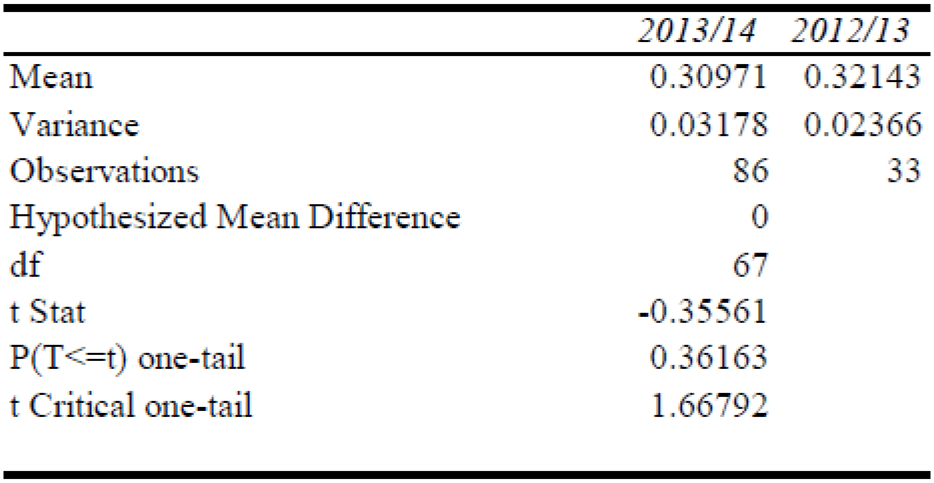


Figure 7. Standard deviation curve representation of Table 2.

Looking at students scoring less than 60%, there was still a disproportionate amount of observations. In 2013/14, there were 86 while the 2012/13 school year had only 33. The *t* test resulted in a *t* score of -0.356 with a *p* value of 63.8% (see Table 3). At less than one standard deviation from the null hypothesis, the *t* score result is insignificant (see Figure 8).

Table 3

All Classes in the Study: Scores < 60%



	2013/14	2012/13
Mean	0.30971	0.32143
Variance	0.03178	0.02366
Observations	86	33
Hypothesized Mean Difference	0	
df	67	
t Stat	-0.35561	
P(T<=t) one-tail	0.36163	
t Critical one-tail	1.66792	

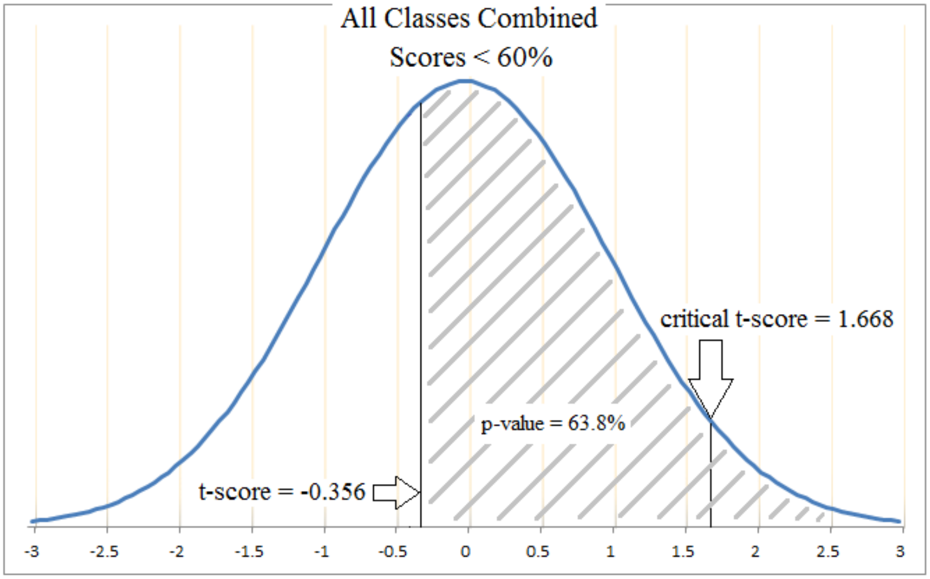


Figure 8. Standard deviation curve representation of Table 3.

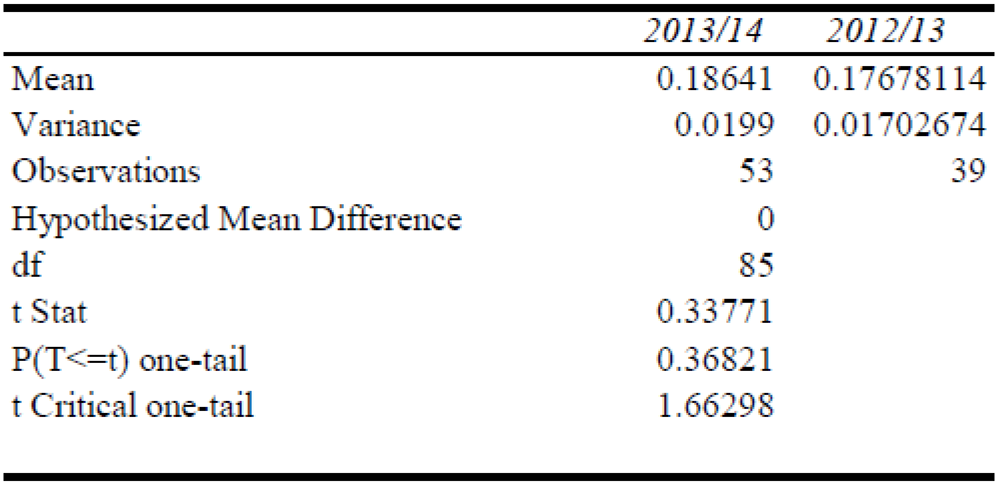
Description and Data Analysis for my "most comparable" classes between the two years. These classes are "most comparable" because in each I co-taught every day, the class period was the same, and the class size was comparable at 25 in 2013/14 compared to 21 in 2012/13. For each of the following tests, the null hypothesis being with

, a one-tailed *t* test was performed using the means of the change in percentage from test 1 to test 2 for each cycle in each given class.

Looking at scores of less than 80%, the number of observations is disproportionate with 2013/14 having 53 while 2012/13 having only 39. The *t* score of 0.338 (see Table 4) shows insignificant value compared to the null hypothesis with a resulting *p* value of 36.8% (see Figure 9).

Table 4

Most Comparable Classes: Scores < 80%



	2013/14	*2012/13*
Mean	0.18641	0.17678114
Variance	0.0199	0.01702674
Observations	53	39
Hypothesized Mean Difference	0	
df	85	
t Stat	0.33771	
P(T<=t) one-tail	0.36821	
t Critical one-tail	1.66298	

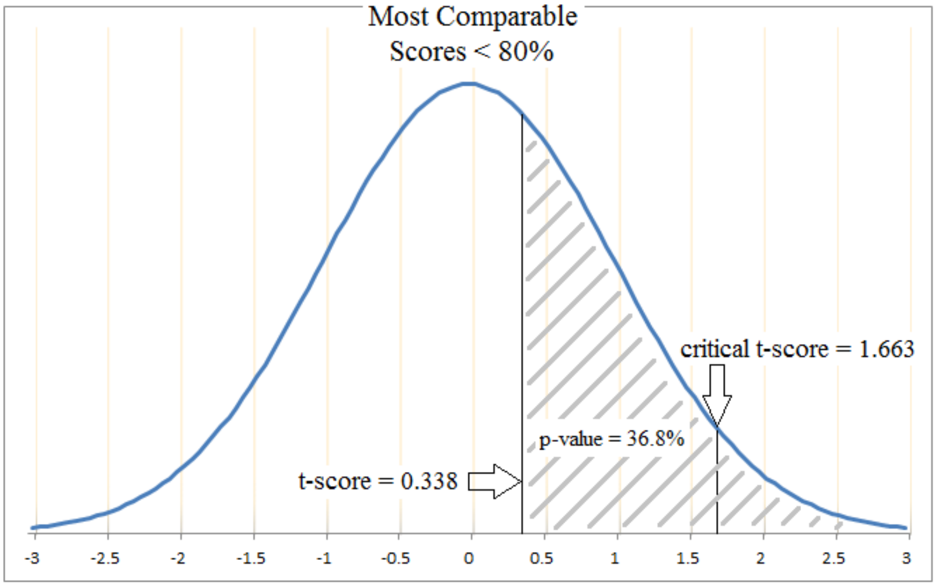
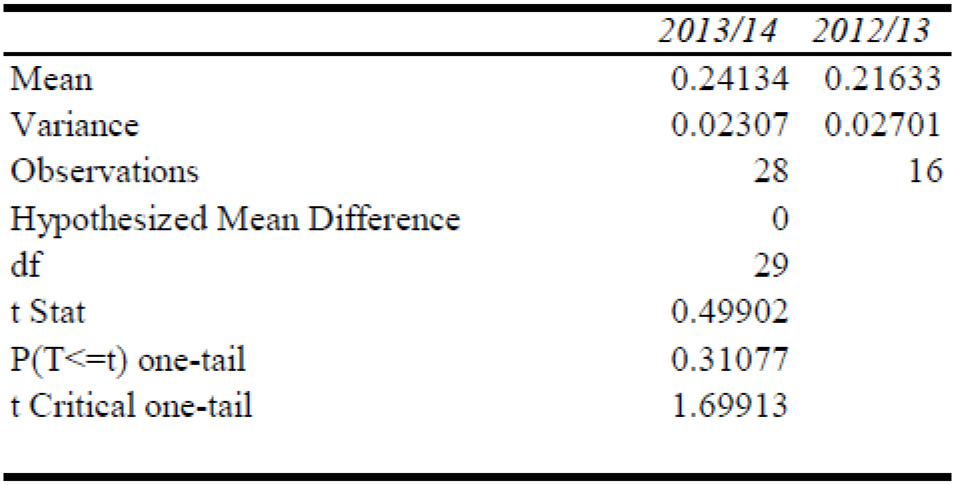


Figure 9. Standard deviation curve representation of Table 4.

Scores of less than 70% were more disproportionate in the numbers of observations. The 2013/14 school year had 28 occurrences while 2012/13 had only 16. The *t* test resulted in a *t* score of 0.499 with a *p* value of 31.1% (see Table 5). At roughly half a standard deviation from the null hypothesis, the *t* score shows to be insignificant (see Figure 10).

Table 5

Most Comparable Classes: Scores < 70%



	2013/14	*2012/13*
Mean	0.24134	0.21633
Variance	0.02307	0.02701
Observations	28	16
Hypothesized Mean Difference	0	
df	29	
t Stat	0.49902	
P(T<=t) one-tail	0.31077	
t Critical one-tail	1.69913	

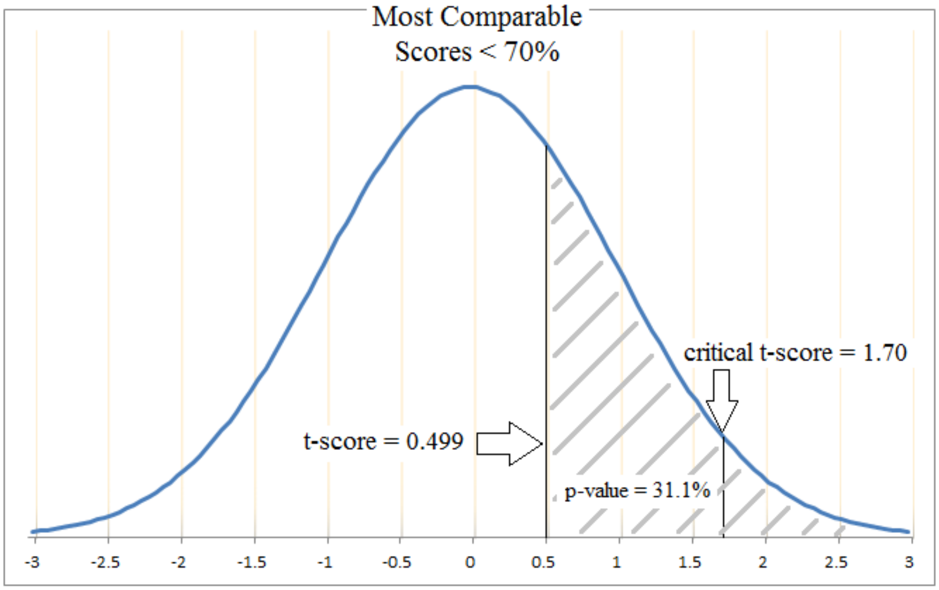


Figure 10. Standard deviation curve representation of Table 5.

Restricting the guidelines to less than 60% for "most comparable" classes revealed an even greater disproportion in observations with the 2012/13 school year having only 6 while the 2013/14 school year had three times that amount at 18. The *t* test showed a *t* score of -0.207 with a *p* value of 57.9% (see Table 6). The *t* score has insignificant value at less than one standard deviation (see Figure 11).

Table 6

Most Comparable Classes: Scores < 60%

	2013/14	*2012/13*
Mean	0.28099	0.30046
Variance	0.02494	0.04476
Observations	18	6
Hypothesized Mean Difference	0	
df	7	
t Stat	-0.20702	
P(T<=t) one-tail	0.42095	
t Critical one-tail	1.89458	

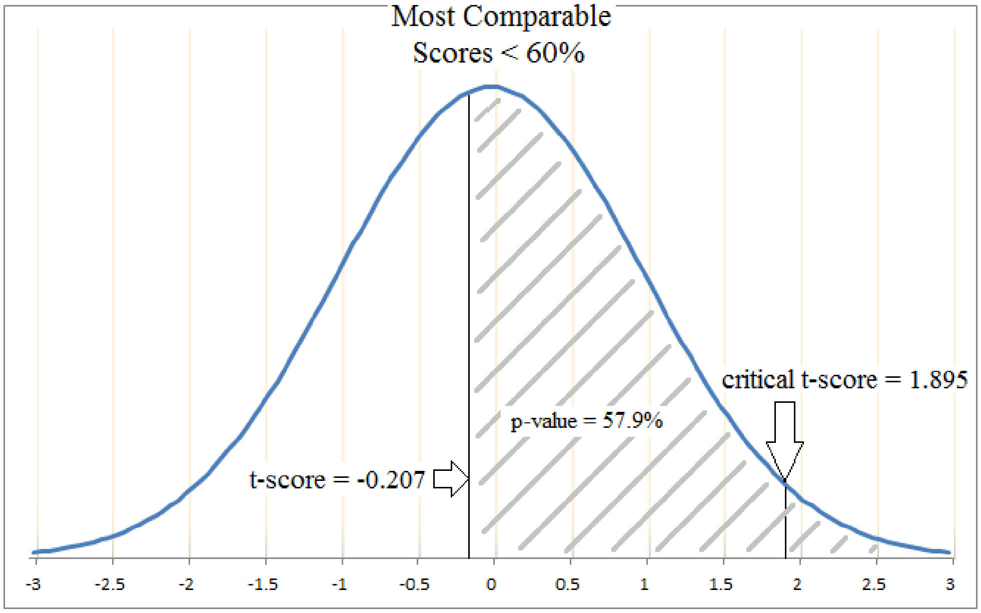


Figure 11. Standard deviation curve representation of Table 6.

Description and Data Analysis for classes I was "present every day". In the 2012/13 school year, I co-taught with Mrs. Hyman two classes per day but was present every day in both classes. In the 2013/14 school year, I co-taught with her four classes, but was present every day in only one class. The rest of the classes I rotated between her and another Algebra 1 teacher approximately every other day. For each of the following tests, the null hypothesis being $\mu 1 = \mu 2$ with $\alpha = .05$, a one-tailed *t* test was performed using the means of the change in percentage from test 1 to test 2 for each cycle in each given class.

Starting with scores below 80%, the number of observations seems fairly proportional with 53 in 2013/14 and 112 in 2012/13. The *t* test resulted in a *t* score of 0.013 with a *p* value of 49.5% (see Table 7). The *t* score shows no significant findings as it shows the smallest *t* score represented in the study (see Figure 12).

Table 7

Present Every Day: Scores < 80%

	2013/14	*2012/13*
Mean	0.18641	0.18610162
Variance	0.0199	0.02262564
Observations	53	112
Hypothesized Mean Difference	0	
df	108	
t Stat	0.01263	
P(T<=t) one-tail	0.49498	
t Critical one-tail	1.65909	

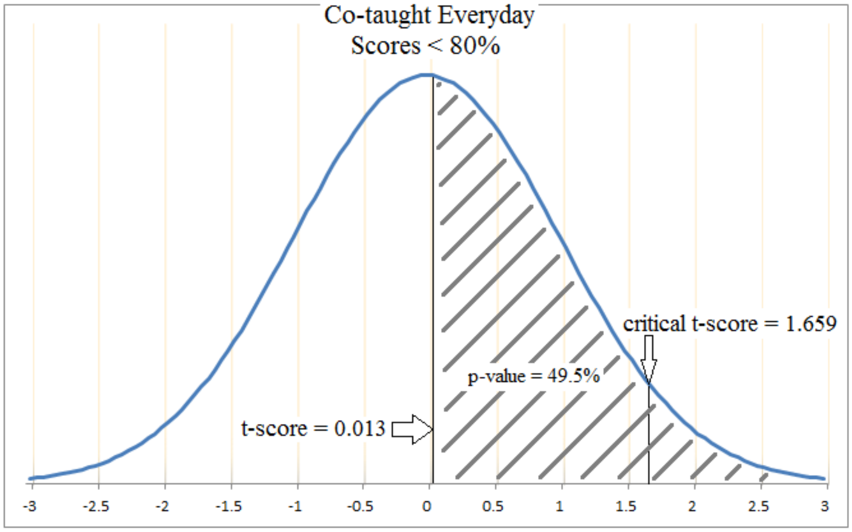


Figure 12. Standard deviation curve representation of Table 7.

Observations of scores below 70% remain proportional with 28 in 2013/14 and 54 in 2012/13. The *t* test showed a *t* score of -0.0694 with a *p* value of 75.5% (see Table 8). The negative *t* score is insignificant at less than a standard deviation from the null hypothesis (see Figure 13).

Table 8

Present Every Day: Scores < 70%

	2013/14	*2012/13*
Mean	0.24134	0.26605
Variance	0.02307	0.02388
Observations	28	54
Hypothesized Mean Difference	0	
df	56	
t Stat	-0.69444	
P(T<=t) one-tail	0.24514	
t Critical one-tail	1.67252	

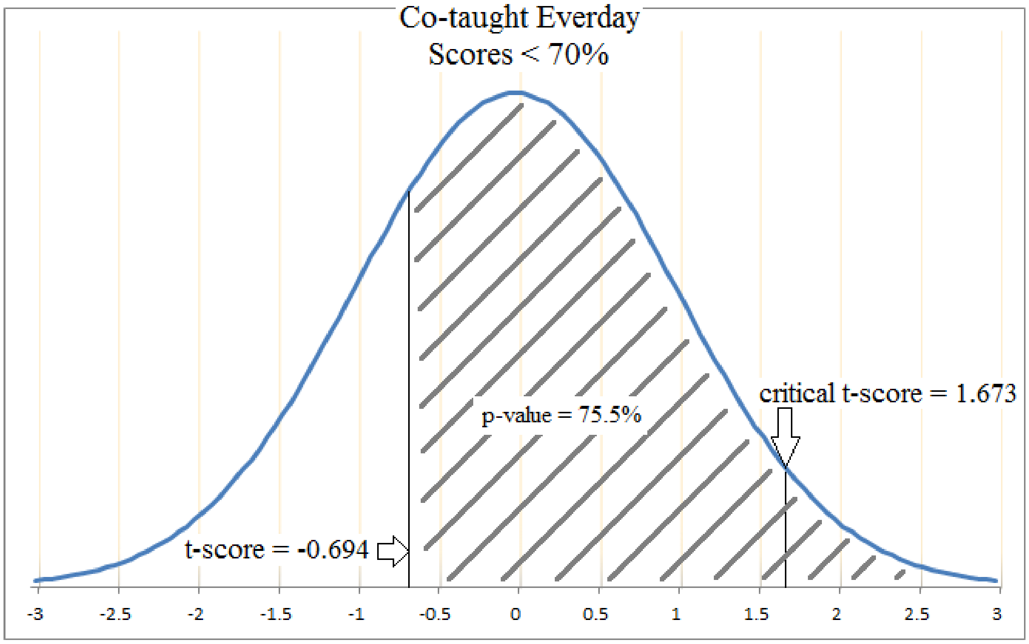
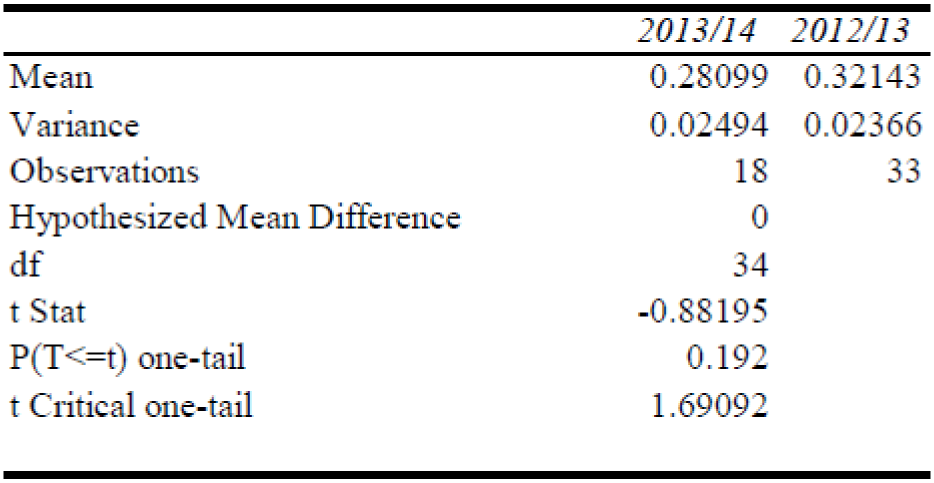


Figure 13. Standard deviation curve representation of Table 8.

Scores of below 60% again remain relatively proportional with regards to the number of observances. The 2013/14 year had 18 while 2012/13 had 33. The *t* test performed on the scores below 60% data revealed a *t* score of -0.882 with a *p* value of 80.8% (see Table 9). The *t* score represents a negative correlation by almost a full standard deviation but would still be considered not significant in a one-tailed *t* test that assumes a positive correlation (see Figure 14).

Table 9

Present Every Day: Scores < 60%



	2013/14	*2012/13*
Mean	0.28099	0.32143
Variance	0.02494	0.02366
Observations	18	33
Hypothesized Mean Difference	0	
df	34	
t Stat	-0.88195	
P(T<=t) one-tail	0.192	
t Critical one-tail	1.69092	

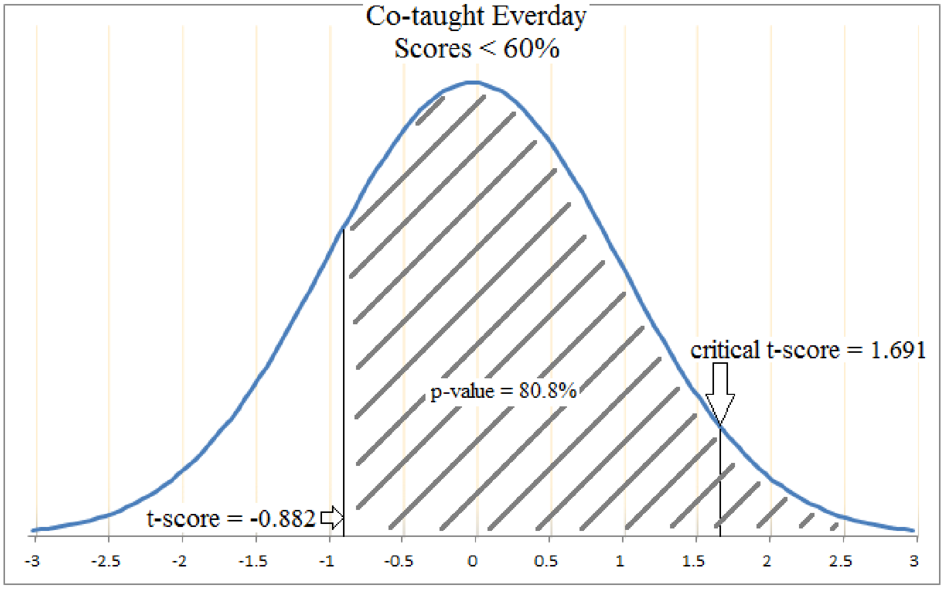


Figure 14. Standard deviation curve representation of Table 9.

Note that when looking at the 2012/13 school year data, there were only two classes and the data is pulled progressively more from the “high need” classroom. At less than 80% nearly twice as many occurrences appeared in the “high needs” class at less than 80%. At

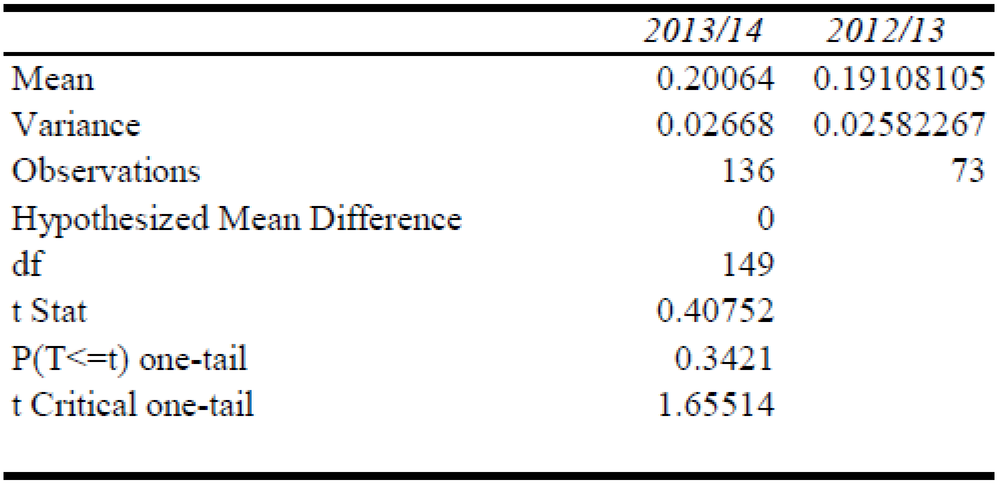
70% there were nearly two and half times as many occurrences, and at less than 60% there were four and a half more occurrences.

Description and Data Analysis for "High Need" Classrooms. The classification of "high needs" is a general notation being used to refer to classes that have three or more students that need extensive interventions and more individual attention than typically afforded during a standard class period. During the 2012/13 school year, only one of the classes met these criteria, but in 2013/14, there were two. For each of the following tests, the null hypothesis being with , a one-tailed *t* test was performed using the means of the change in percentage from test 1 to test 2 for each cycle in each given class.

When looking at scores that were restricted to less than 80%, the number of observations was slightly disproportionate with 136 instances in 2013/14 while having 73 in 2012/13. The *t* test showed a positive *t* score of 0.408 with a *p* value of 34.2% (see Table 10). Although a positive correlation was shown, at less than half a standard deviation from the null hypothesis, the data shows statistically insignificant (see Figure 15).

Table 10

High Need Classes: Scores < 80%



	2013/14	*2012/13*
Mean	0.20064	0.19108105
Variance	0.02668	0.02582267
Observations	136	73
Hypothesized Mean Difference	0	
df	149	
t Stat	0.40752	
P(T<=t) one-tail	0.3421	
t Critical one-tail	1.65514	

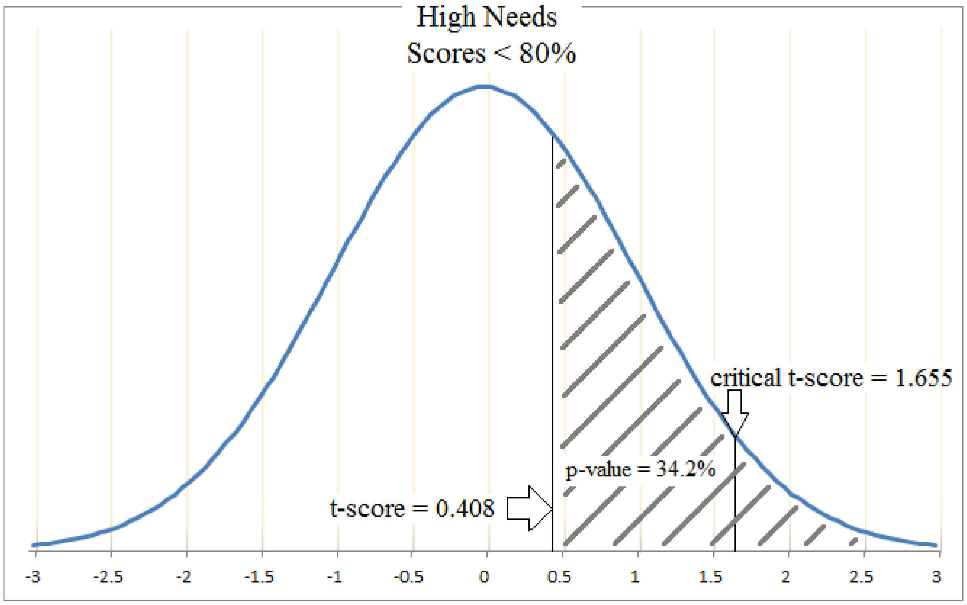
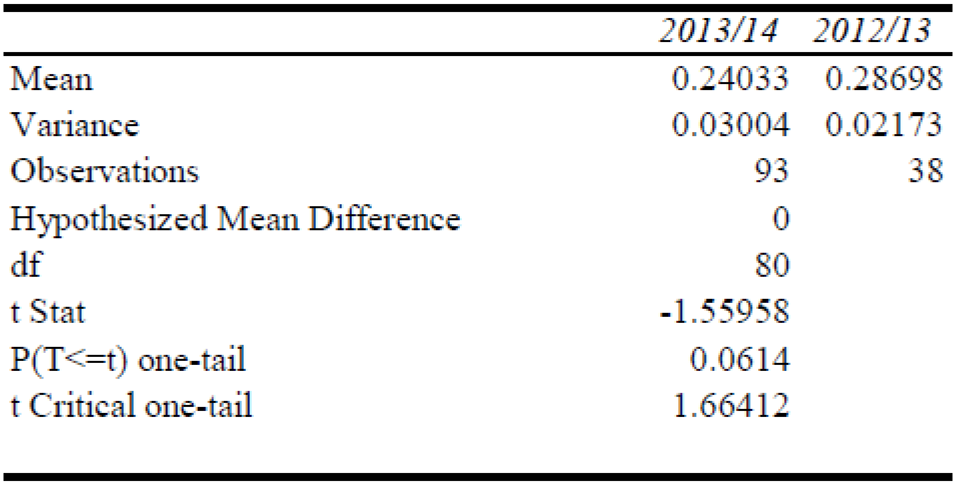


Figure 15. Standard deviation curve representation of Table 10.

Reducing the values to less than 70% switched the disproportionate amount of occurrences to having more in 2013/14 with 93 compared to 38 in the 2012/13 school year. After performing the one-tailed *t* test, a *t* score of -1.560 resulted with a *p* value of 93.9% (see Table 11). The resulting data shows a negative correlation by over one and a half standard deviations. Still, this *p* value concerning the findings is statistically insignificant because of the test being one-tailed assuming a positive correlation (see Figure 16).

Table 11

High Need Classes: Scores < 70%



	2013/14	2012/13
Mean	0.24033	0.28698
Variance	0.03004	0.02173
Observations	93	38
Hypothesized Mean Difference	0	
df	80	
t Stat	-1.55958	
P(T<=t) one-tail	0.0614	
t Critical one-tail	1.66412	

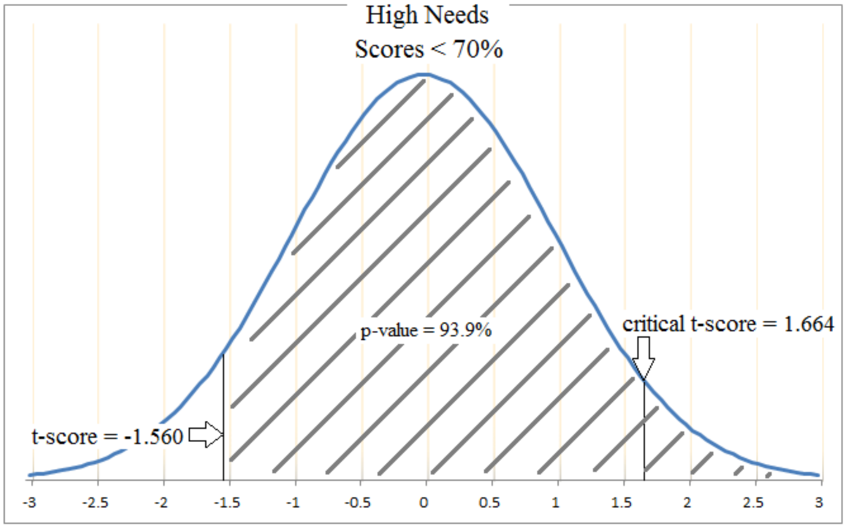
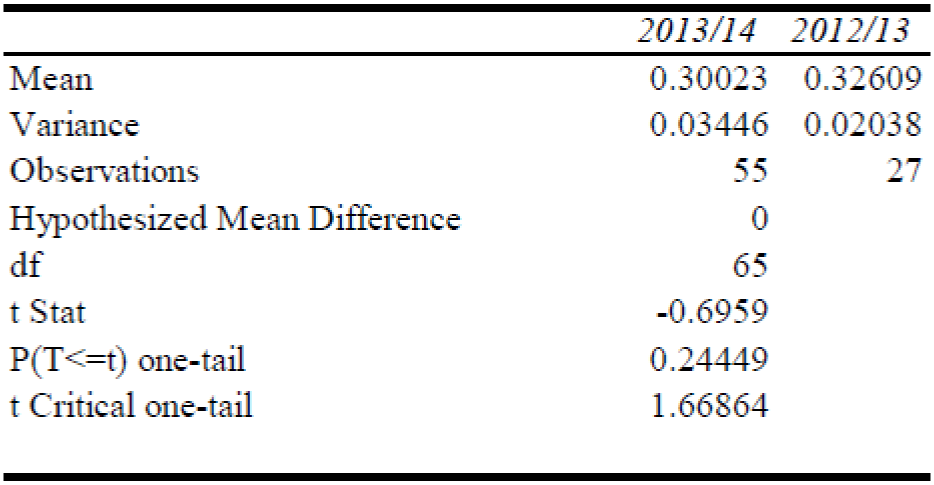


Figure 16. Standard deviation curve representation of Table 11.

Occurrences in the scores less than 60% show proportionality with the 2013/14 school year having 55 while 2012/13 had 27. The less than 60% *t* test revealed a *t* score of -.696 with a *p* value of 75.6% (see Table 12). Although the result of the *t* test shows a negative correlation, it is not near what the less than 70% data showed. The *p* value still shows insignificant at less than one standard deviation from the null hypothesis (see Figure 17).

Table 12

High Need Classes: Scores < 60%



	2013/14	*2012/13*
Mean	0.30023	0.32609
Variance	0.03446	0.02038
Observations	55	27
Hypothesized Mean Difference	0	
df	65	
t Stat	-0.6959	
P(T<=t) one-tail	0.24449	
t Critical one-tail	1.66864	

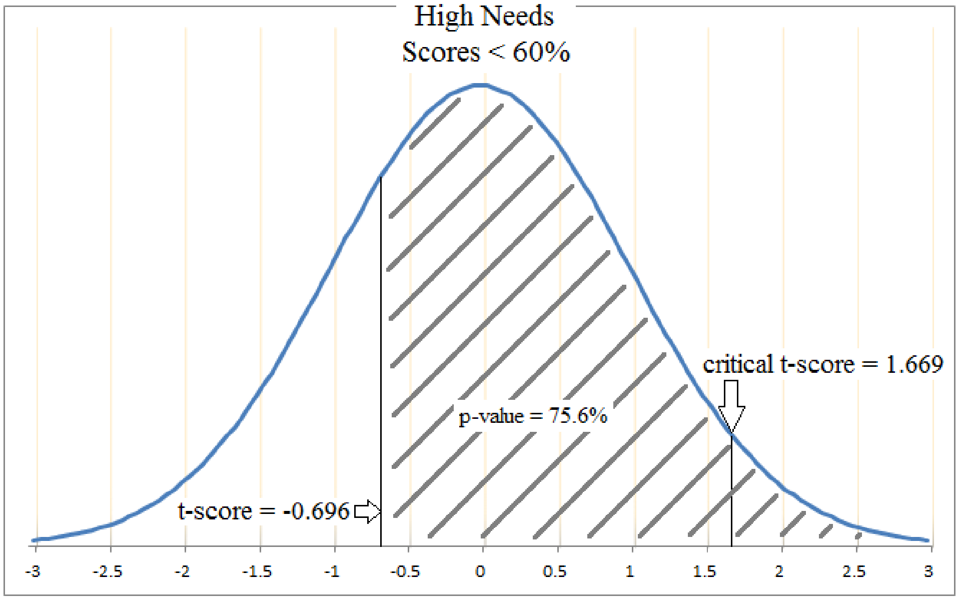


Figure 17. Standard deviation curve representation of Table 12.

Conclusions. Through the process of performing twelve one-sided *t* tests there were positive and negative correlations statistically insignificant in eleven of the twelve tests. The only positive correlation that showed more than one positive standard deviation was the broadest test which had the highest amount of observations and thus the lowest critical *t* score, that is, the scores of less than 80% for "all classes" co-taught by Mrs. Hyman and me in the study. While the *t* test result did not reach the critical *t* score to trigger the rejection of the null hypothesis, at more than a one standard deviation from the null hypothesis, the *t* score does represent a significant finding.

Possibly even more significant is the negative correlation discovered in the "high needs" classroom at more than one and a half standard deviations in the unexpected direction, and that the negative correlation did not affect the significance of the overall data for "all classes" less than 80%.

Chapter 5

Discussions and Implications

Automated proficiency levels to allow data-driven decisions proved a powerful resource during the return to previous standards. That being said, results did not always show what was expected. Throughout the process, a number of elements that influenced the day to day success of students came to light.

Upon collection of the percentage change data, expectations of the results seemed obvious. With data-driven decisions leading the way for differentiation, those most targeted should see greater improvement. The most targeted would represent the lowest scoring students on any given assessment.

Nearly the opposite showed. In each category, without fail, the lowest tier of “less than 60%” had a negative correlation. These unintended results begged the question, “How could this have happened?”

Unfortunately, the majority of the issues lie in the inability to scientifically restrict the variables. The inability to restrict the variables makes it difficult to be certain the results can be trusted. Upon reflecting on the study's results, I am still appreciative of the proficiency levels that were at my fingertips which helped my ability to reach the non-proficient students.

Interpretation of the Results. When reflecting on the study's results, the only significant findings were evident from the broadest scope. Comparing "all classes" in the study, the *t* test revealed a *t* score more than one standard deviation from the null hypothesis. The 2013/14 school year showed a positive correlation in scores "less than 80%". However, a negative correlation, all be it insignificant in its comparison to the null hypothesis, was revealed in the tiers "less than 70%" and "less than 60%". The significant growth at "less than 80%" contrasted by the negative correlation in the other two tiers represents huge improvement for students scoring between 70% and 80%. It could be hypothesized that the improvement may have resulted from the specified interventions directed at students that were not in the 70% to 80% range. This aversion from direct intervention may have forced the students in the 70% to 80% range to wrestle through topics for which enough background knowledge was present to attempt the skills. The students in the 70% to 80% range may have gained success through repetitions and working with peers during the ten minute bell ringer. These students were forced to struggle and were not given the support of scaffolding and intervention unless initiated by the student.

The significant results of the broad group of students scoring "less than 80%" in "all classes" echoed through each of the subsequent categories. Although none of the other categories showed significant levels, at "less than 80%" all still showed positive correlation between growths from the first assessment to the second.

Two classes had many similarities which led to the category of "most comparable". The criteria for these classes were as follows: both were held during the same time of the day, had similar class size, and were considered a relatively low-need class. Disproportionate observation size in the "less than 60%" tier may have shown that the low-need designation given to the 2013/14 class may have been a misclassification. The negative correlation could easily be due to so few extremely low scores in 2012/13, and the students who scored extremely low were easy to pinpoint for remediation. The comparison between "most comparable" is the only finding in the study that showed positive correlation in the "less than 70%" tier. But as specifications were narrowed, limited observations made drawing specific conclusions from the "less than 70%" comparison difficult.

One of the most intriguing finding was in every classification the "less than 60%" tier showed a negative correlation. The negative correlation from "less than 60%" stands completely opposite the pre-study hypothesis. The automated data highlighted skill deficiency for students not reaching proficiency. Therefore, non-proficient students were the most likely to receive direct intervention during the bell ringer. The most insignificant negative correlation showed in the "most comparable" classification. All other categories in 2012/13 had one specific class in common and a higher negative correlation which, at times, approached nearly a full standard deviation in the negative direction. This class appeared in all categories except "most comparable" was classified "high-needs" because a core of four students consistently needed heavy intervention. The studies premise was automated data would pinpoint students who needed more intervention in a timely manner. In the 2012/13 class, four students consistently proved non-proficient after the first attempt through the material. With 27 observances in the "less than 60%" tier, a high percentage came from those four students over the course of the seven cycles.

The recognition of this anomaly was revealed in the most significant findings of the study. Comparing classes designated "high needs" at "less than 60%" showed a negative correlation of -0.670 with 55 observances in 2013/14 and 27 observances in 2012/13. The number of observations is almost perfectly proportional considering two classes from 2013/14 were classified as "high needs" while only one in 2012/13. Where the huge discrepancy compared to expectation was at the "less than 70%" tier. The *t* score showed more than one and a half standard deviations in the negative direction. The *t* score is clearly significant in a one-tailed test with complete expectations of positive correlation. The number of observations in the "less than 70%" tier is clearly disproportionate with only 38 in 2012/13 which represents a difference of 11 scores between 60 and 70% for all seven cycles. "High needs" classes in 2013/14, on the other hand, had 93 observances with 38 scoring between 60 and 70%. That represents an average of more than 72% more test scores represented between 60 and 70% for each of the two classes in 2013/14. The classes in 2013/14 had much higher need in the "less than 70%" tier and more variability in the students represented by "less than 60%". Although having automated data can help pinpoint specific need, it is apparent that the "high need" designated class in 2012/13 had low variability in students who needed direct intervention and less students that needed significant reteaching. The 2012/13 class's data affects three of the four classifications in the study.

The final noticeable result may have tied both significant findings together. Significant positive correlation in the broadest category of scores “less than 80%” in “all classes” proved even more outstanding despite negative correlations in other tiers.

Potential Application of the Findings. After reflecting on the findings, making the bell ringer time as efficient as possible for all students should be the top priority. Currently the process is a one-size-fits-all mentality where each student works through the review problems. Each student is able to work with others around them if so chosen. Students then may ask questions and correct their answers as the teacher reveals the solutions.

Automating the data pinpointed the specific students that reached various levels of proficiency. The majority of intervention focused on the lowest scorers but was mainly limited to one-on-one attention working through specific problems that would help them advance in the understanding of the power standard being reviewed. As the intervention may have been very useful for the few individuals that were pinpointed during that time, it is quite clear that other students that were in need of reteaching were not adequately assisted.

Restructuring the bell ringer time into differentiated groups could prove to reach more students. Students with proficiency levels of 70% and above could be grouped heterogeneously. Within that subgroup, different levels of proficiency could be represented in each team of three or four students. The heterogeneous grouping would allow students most likely to remember the majority of the process and types of problems to succeed within learning groups. Students could ask for help as needed without intervention of a teacher.

Changing the intervention strategy would entail two specific pieces. First, the intervention should include all students who scored less than 70% on their first assessment.

Although the subgroup between 60-70% was not specifically avoided during the process of reteaching during the bell ringer, there was more chance those students would not be involved in the intervention as the majority of time was spent helping the lowest tier. As the findings of the study have shown, the 60-70% range did not show the tendency to improve as much without intervention and could benefit from restructuring into a teacher-led small group. Second, the intervention would change from working through review problems to a teacher-led model and practice method. This method would avoid the counter productivity of having students practice a standard that was not learned on the first attempt.

The next potential application requires background knowledge of the students through communication from the preceding teacher or institution. Purposeful scheduling could prove most successful by putting a few students who have shown repeated tendency to score at extremely low proficiency levels into a class that otherwise has very few students that show strong tendencies for struggle. Purposeful scheduling would allow the teacher to spend more time with high-need students while not worrying that others were not being reached. It was shown, when students have higher ranges of proficiency, success may indeed be improved by allowing struggle.

When returning to the original goal of this study to show if using a gradebook with automated proficiency indicators would help improve proficiency scores to a greater degree, the results indicate that correlations do exist. However, with so many other variables present, the degree to which the proficiency indicators helped cannot be quantified.

Biblical Integrative Component and Implications.

Key Scriptures

Ephesians 5:15-16 Look carefully then how you walk, not as unwise but as wise, making the best use of the time, because the days are evil.

Proverbs 11:1 A false balance is an abomination to the LORD, but a just weight is his delight.

Proverbs 20:23 Unequal weights are an abomination to the LORD, and false scales are not good.

As a Christian educator, I see two major biblical integrative connections with my current research. The first is made under the assumption that data is used to guide instruction. Figuring the proficiency levels by hand can be time consuming. In Ephesians, the Bible tells us to make the best use of the time given. If time is being used to calculate proficiency levels and other helpful statistics, then that time is not being used to design differentiated lessons, work with students, communicate with parents, or collaborate with colleagues. Automating data can improve efficiency and allow time to be used more wisely.

The second connection is found twice in Proverbs where it mentions God's distaste for unjust balances. In the field of education, as in other facets of life, preconceived notions can play a large part in the decisions that are made. As Christian educators, as well as educators in general, it is our duty to treat every student fairly. In lacking the data, or time to find the data needed to drive decisions; an educator's likelihood is to use gut instinct, which is often affected by preconceived notions. When allowing the use of data in decision making, the data can help us avoid snap-judgments and reduces the likelihood of being blindsided by unexpected results. Data-driven decisions help remove unintended partiality and reduce the window of error in choices. The data-driven reflection process can help an educator be as just and fair as possible.

Relation of Results to Theory and Other Literature. The structure of the Algebra 1 assessments, compared to the power standards approved by the board of education, allow the ability to look at differentiation needs for each specific standard as asserted by Tomlinson (2003).

The true focus of this study is the effectiveness of timely differentiated instruction. O'Connor and Wormeli (2011) contend that formative assessments allow for reteaching opportunities that can be tailored for the need of the student. Aside from other formatives used, the first assessment of a power standard is a built-in formative. This assessment is a point of guaranteed feedback that ensures students know exactly where they fell short on the given power standard. Significant improvement scores throughout all levels indicate the power in purposeful reflection and timely intervention.

Finally, data was needed to help drive my decisions (Doubet, 2012). Automating assessment scores was a bi-product of not having enough time to reflect. The automated proficiency indicators proved efficient in ease of use.

Strengths of the Study. In educational action research, there are bound to be a plethora of moving parts and thus a number of variables difficult to keep constant. A definite strength was the day-to-day consistency students experienced in this study. Consistency was driven by having the same teaching core. Mrs. Hyman was considered the lead teacher as I was the interventionist. Although interchanging positions was an option at any time, Mrs. Hyman and I kept our roles fairly consistent through the length of this study. The 9th grade Algebra team was held consistent from 2012/13 to 2013/14 with core input from Mrs. Hyman, Mr. Brasche and me.

Students were from the same school district in the same class at the same age. That is to say all students were 9th grade Algebra 1 students at Ankeny Community Schools. While part of the leadership team did change from 2012/13 to 2013/14, the heart of positive leadership with a focus on differentiation was held constant from both Northview's and Southview's administration.

Unintended bias was kept to a minimum by limited knowledge of the study. Students had absolutely no knowledge of the study, so in no way did the study enhance the students' motivation or effort. While automated proficiency indicators were being used first semester, Mrs. Hyman was not informed of the intent to use the test scores in the study until after the end of the first semester.

The time of the year proved consistent as well. All material was taught during first semester. First semester of both years I coached golf and missed approximately the same amount of class days.

Comparable assessment from year to year is one of the most powerful strengths of this study. These assessments were created the 2011/12 school year. Throughout 2011/12 and before giving the assessments in 2012/13, most corrections and changes were made to ensure clear quality questions that fairly represented the power standards being assessed. From 2012/13 to 2013/14 there were very few changes needed on the assessments.

Consistency was present due to my control of the gradebook. I created the gradebook with built-in automated formulas. I wrote the equations, chose the symbols, and controlled the output. There was no confusion with how the gradebook worked, and if there was any issue, I had the knowledge to fix it.

Finally, the bell ringer intervention was not dependent on others. I was the only one who knew about the study, and it was completed in the course of my daily work. Regardless of research purposes, the automated data would have been used to help guide interventions.

Limitations of the Study. With regards to internal and external validity, there are a number of concerns. Though attempts were made to keep as many variables as constant as possible, some instances proved challenging.

Threats to internal validity began with the split of the Ankeny district into two separate school systems. In 2012/13 Mrs. Hyman, Mr. Brasche, and I taught at Northview Middle School, which served a population of every 9th grade student in the Ankeny Community School District. In 2013/14, Mrs. Hyman, Mr. Brasche, and I were all transferred to Southview Middle School which served 9th grade students from the south side of the district and any open-enrolled students. Although both parts of the district are similar, there are some discrepancies between different subgroups within the district. There are higher percentages of special education, English language learners, and at-risk students in the south portion of the district.

The school year was not held constant resulting in completely different students participating in the study from 2012/13 to 2013/14. Although a control group could have been utilized in 2013/14, the decision to use automated proficiency indicators for all students was made to ensure maximum ability of directed intervention.

A major drawback of this two-year study was the level of experience for Mrs. Hyman and me. The 2012/13 year was Mrs. Hyman's first year as a core math teacher, although she had previously taught for five years in the business department. I was new to my role of math interventionist having just transferred from a lead role the previous year. Those changes, coupled with the growth of a co-teaching partnership, could allow significant variation in the instructional methods and procedures.

The number of times classes co-taught with Mrs. Hyman varied between years. In 2012/13, co-teaching with Mrs. Hyman consistently occurred every day in the same two classes. In 2013/14, co-teaching with Mrs. Hyman occurred in as many as four classes on any given day. During 2013/14 co-teaching occurred every day in one class while the other three classes were co-taught approximately every other day.

Not always being present led to one of the largest threats to internal validity. During 2012/13, co-teaching occurred in the "high needs" class every day; however, in 2013/14 co-teaching occurred approximately every other day in both "high needs" classes. The data may have shown more significant results had that variable been held constant.

Southview is in a different location in the district. The implications of a different building do not necessarily imply extreme changes, but in Southview's case there are quite a few. Northview, where Mrs. Hyman and I co-taught during the 2012/13 school year, was the largest middle school in the state of Iowa with approximately 1400 students. Southview in 2013/14 was a much smaller school of approximately 600. The Northview building was built in 1972 and can be remembered for being cramped, outdated, with poor heating and cooling systems, with no doors on the classrooms, and with very thin walls. In stark contrast, Southview was a two year-old building with spacious rooms, new desks and chairs, mounted projectors with sound systems, and Mimio boards in every classroom. The learning environment experienced great improvements from 2012/13 to 2013/14.

Class size increased from 2012/13 to 2013/14 by approximately four students per class. That would represent an increase in class size by over 15%. The change in class size was mostly due to the 2012/13 9th grade math team having four and a half teams serving the population, while after the split Southview only had two teams. While the initial split may have been slightly more proportional to the distribution of teacher positions, all open enrolled students were funneled into the south feeder system.

The variable that I had the most control over was me. That being stated, I did not have a specified routine that I rigidly followed. I reviewed the data when I felt it was necessary and intervened with students as I saw need. The consistency in the use of the automated proficiency indicators was left up to my discretion. As I value the data considerably, I did my best to implement clear timely interventions, but there was my personal variability left in the study.

Finally, as this study measures change from the first assessment to the second assessment, it does not focus on the overall achievement. Students could achieve significantly better or worse overall on the first assessment, but the merit in this study is given to improvement. As an example: better results would show for a student who improved from 5% on the first assessment to 60% on the second showing 55% growth. In comparison, if the student improved from a 55% to 80% there would show only 25% growth. It is fair to say that most would value higher end-achievement as compared to overall growth in this comparison.

Threats to external validity were limited in scope due to a number of factors. The school district is located in the Midwest. As the demographics may be similar to other Midwestern schools, transferability to schools in other regions may be challenging. The Ankeny school district's standardized test scores are high with a range of 89% to 92% proficient for the 9th grade class in both school years across both sides of the district. Schools with different levels of standardized scores may be affected differently by the same type of interventions.

As stated previously, the school is in an upper-middle class community with a median household income 40% higher than the state of Iowa. Reliability may be affected even within the common region.

This study was completed in only Algebra 1 classes. Although timely interventions are necessary in all subject areas, the results from the study may vary throughout different contents.

Finally, the Ankeny school district used funding from the state of Iowa to help pay for co-teaching positions that focus on specific at-risk populations within the student body. Both co-teachers were highly qualified in the subject area. These positions offer a chance to reduce the student to teacher ratio significantly allowing easier differentiation opportunities. With more students relying on a single teacher, results may diminish as the student to teacher ratio rises.

Suggestions for Future Research. In concluding this study, there have been more questions unearthed than answers found. Further research could help clarification in some areas.

The gradebook with built-in automated proficiency indicators, although created for Algebra 1 mathematics classrooms, could easily be transferable into a different subject area. As long as a class teaches a specific topic, gives an assessment, reviews and reteaches needed material, and gives a comparable second assessment, this gradebook would be usable.

Holding constant a number of unexpected variables derived from the changing district may give clearer results. Keeping as much constant and comparing data from one year to the next while not changing population pools, buildings, and surroundings

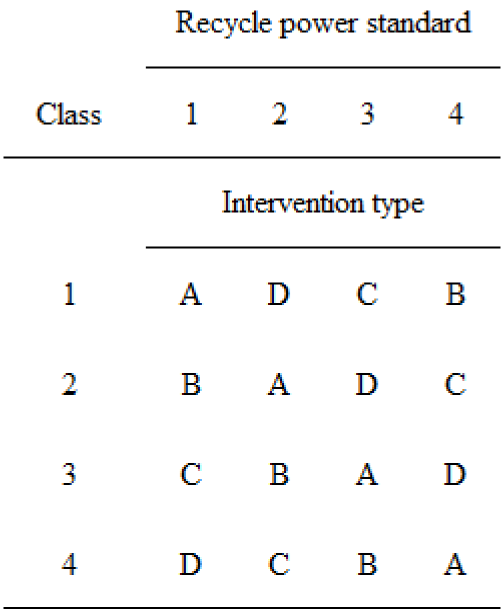
The use of a control group could eliminate the variables created from researching across multiple years. Using a control group may give a better glimpse of the difference between teacher intuition and data-driven decision making.

Assuming that data-driven intervention has a positive impact, comparing classes that have co-teaching every day versus every other day could shed light on the effectiveness for different proficiency levels. This study might reveal more findings about students scoring between 60-70%.

Shifting research focus onto interventions could reveal important findings. During each cycle use a different intervention method with each class. Alternating the methods of intervention used as the cycles progress (see Table 13). Alternating methods may uncover which types of interventions work better in a global sense, or which individual intervention(s) work better for specific classes.

Table 13

Intervention Rotation for Recycled Material



	Recycle power standard			
Class	1	2	3	4
	Intervention type			
1	A	D	C	B
2	B	A	D	C
3	C	B	A	D
4	D	C	B	A

Note. Each letter A through D represents a different intervention style including: small group in class, small group out of class, individual in class, review recycled material two days after new material is assessed.

Finally, a piece that would be interesting to research is how overall achievement is affected by. Raw scores could be used from the second assessment as well as final scores from the semester tests. The change from using only improvement data to using both improvement and achievement may give a better picture of total attainment compared to the standards.

REFERENCES

Acosta-Tello, E., & Shepherd, C. (2014). Equal access for all learners: differentiation simplified. *Journal Of Research In Innovative Teaching*, *7*(1), 51-57.

Azeem, M., Afzal, M., & Majoka, M. (2010). How and why grading. *International Journal Of Learning*, *17*(3), 579-596.

Bagley, S. S. (2008). Growth, personalization, and dialogical exchange in high school. i*nteractions: UCLA Journal Of Education & Information Studies*, *4*(1), 1-26.

Brimijoin, K., Marquissee, E., & Tomlinson, C. (2003). Using data to differentiate instruction. *Educational Leadership*, *60*(5), 70-73.

Brookhart, S. M. (2011). Starting the conversation about grading. *Educational Leadership*, *69*(3), 10-14.

Campbell, C. (2012). Learning-centered grading practices. *Leadership*, *41*(5), 30-33.

Carlson, D., Borman, G., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33(3)*, 378-398.

Colby, S. A. (1999). Grading in a standards-based system. *Educational Leadership*, *56*(6), 52-55.

Dalziel, J. (1998). Using marks to assess student performance: some problems and alternatives. *Assessment & Evaluation In Higher Education*, *23*(4), 351.

Dixon, F., Yssel, N., McConnell, J., & Hardin, T. (2014). Differentiated instruction, professional development, and teacher efficacy. *Journal For The Education Of The Gifted*, *37*(2), 111-127.

Doubet, K. J. (2012). Formative assessment jump-starts a middle grades differentiation Initiative. *Middle School Journal*, *43*(3), 32-38.

Dunn, K. E., Airola, D. T., Lo, W., & Garrison, M. (2013). What teachers think about what they can do with data: development and validation of the data-driven decision making efficacy and anxiety inventory. *Contemporary Educational Psychology*, *38*(1), 87-98.

Erickson, J. A. (2011). How grading reform changed our school. *Educational Leadership*, *69*(3), 66-70.

Erickson, J. A. (2010). Grading practices: the third rail. *Principal Leadership*, *10*(7), 22-24.

Fisher, D., Frey, N., & Pumpian, I. (2011). No penalties for practice. *Educational Leadership*, *69*(3), 46-51.

Frohbieter, G., Greenwald, E., Stecher, B., Schwartz, H. (2011). Knowing and doing: what teachers learn from formative assessment and how they use the information. CRESST Report 802. *National Center For Research On Evaluation, Standards, And Student Testing (CRESST)*,

Guskey, T. R. (2001). Helping standards make the grade. *Educational Leadership*, *59*(1), 20-27.

Huang, S. (2012). Like a bell responding to a striker: instruction contingent on assessment. *English Teaching: Practice And Critique*, *11*(4), 99-119.

Jung, L., & Guskey, T. R. (2011). Fair & accurate grading for exceptional learners. *Principal Leadership*, *12*(3), 32-37.

Kadel, R. (2010). Data-driven decision making--not just a buzz word. *Learning & Leading With Technology*, *37*(7), 18-21.

Mandinach, E. B. (2012). A perfect time for data use: using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71-85. doi:10.1080/00461520.2012.667064

Marzano, R. J., & Heflebower, T. (2011). Grades that show what students know. *Educational Leadership*, *69*(3), 34-39.

Marzano, R. J. (2003). Using data: two wrongs and a right. *Educational Leadership*, *60*(5), 56-60.

O'Connor, K., & Wormeli, R. (2011). Reporting student learning. *Educational Leadership*, *69*(3), 40-44.

Qu, W., & Zhang, C. (2013). The analysis of summative assessment and formative assessment and their roles in college English assessment system. *Journal Of Language Teaching & Research*, *4*(2), 335-339.

Reeves, D. B. (2011). Taking the grading conversation public. *Educational Leadership*, *69*(3), 76-79.

Reeves, D. B. (2008). Effective grading practices. *Educational Leadership, 65(5), 85-87.*

Santangelo, T., & Tomlinson, C. (2008). The application of differentiated instruction in postsecondary environments: benefits, challenges, and future directions. *International Journal Of Teaching & Learning In Higher Education*, *20*(3), 307-323.

Scriffiny, P. L. (2008). Seven reasons for standards-based grading. *Educational Leadership*, *66*(2), 70-74.

Shippy, N., Washer, B., & Perrin, B. (2013). Teaching with the end in mind: The Role of Standards-Based Grading. *Journal Of Family & Consumer Sciences*, *105*(2), 14-16.

Spencer, K. (2012). Standards-based grading: new report cards aim to make mastery clear. *Education Digest: Essential Readings Condensed For Quick Review*, *78*(3), 4-10.

Tomlinson, C. (2000). Reconcilable differences: standards-based teaching and differentiation. *Educational Leadership*, *58*(1), 6-11.

Tomlinson, C. (2004). The Möbius Effect: Addressing learner variance in schools. *Journal Of Learning Disabilities*, *37*(6), 516-524.

Tomlinson, C. (2005a). Grading and differentiation: paradox or good practice? *Theory Into Practice*, *44*(3), 262-269.

Tomlinson, C. (2005b). Quality curriculum and instruction for highly able students. *Theory Into Practice*, *44*(2), 160-166.

Tomlinson, C., & Kalbfleisch, M. (1998). Teach me, teach my brain: a call for differentiated classrooms. *Educational Leadership*, *56*(3), 52-55.

Tomlinson, C., &McTighe, J. (2006). Integrating differentiated instruction and understanding by design. Alexandria, VA: ASCD

Wormeli, R. (2006). Differentiating for tweens. *Educational Leadership*, *63*(7), 14-19.