Spring 2024

Math Inventory and Success in Algebra 1

Rebecca Schulz
Shawnee State University

SHAWNEE STATE UNIVERSITY

Math Inventory and Success in Algebra 1

A Thesis

By

Rebecca H. Schulz

Department of Mathematical Sciences

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science, Mathematics

June 20, 2024

Accepted by the Graduate Department

6/22/2024

Graduate Director, Date

The thesis entitled '**Math Inventory and Success in Algebra 1**' presented by **Rebecca H. Schulz**, a candidate for the degree of **Master of Science in Mathematics,** has been approved and is worthy of acceptance.

6/22/2024
Date

Graduate Director

6/20/2024
Date

Student

ABSTRACT

Students spend a lot of time being tested on what they have learned, when that time might be better spent on more learning. Because of the stakes attached to state end of year testing, many school districts have started requiring students to take additional benchmark assessments to help them figure out who will need more help in order to pass. One large school system in Virginia has recently begun to require students to take the Math Inventory test at least two times per year. Anecdotal data from teachers who gave it the first year suggested that there may be problems with the test, prompting this research. Data was collected from all Algebra 1 students at an ethnically diverse target high school within the aforementioned school district, including Math Inventory scores for fall and spring, SOL (Virginia Standards of Learning) scores, end of course grades, socioeconomic status, and ethnicity. A linear regression model was found to be statistically significant in predicting spring SOL scores from fall Math Inventory scores. Also, a logistic regression model was found to be statistically significant in predicting a student's success in Algebra 1 (passing both the class and the SOL) from a student's fall Math Inventory score and ethnicity. However, inconsistencies were found between the spring Math Inventory scores and other spring data (SOL scores and end of course grades), suggesting that there may be a problem with the assumptions or perhaps the administration of the test. Recommendations are to use the Math Inventory in the fall to predict SOL scores and success in Algebra 1 at the target high school. It is not recommended that this model be used at other schools, as the target school is different from other schools in the area both ethnically and socioeconomically. However, this process could be used at

other schools to create their own models to predict spring scores. Additionally, the results imply that either further research should be conducted on the spring administration of the Math Inventory (possible inconsistencies in how it is given, student and teacher motivation, etc.), or the spring administration of the Math Inventory should be discontinued.

ACKNOWLEDGMENTS

Thank you to my husband and children for your patience, understanding, and support as I worked towards my degree. I could not have done it without the four of you picking up my slack at home.

Thank you to Dr. Darbro for your patience and guidance during this very long process.

Thank you to my coworkers for your support, especially my principal who pulled some strings in the background so that I could finally get my data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I: Introduction

Chapter 1 will provide an introduction to the Math Inventory assessment. The research problem that was investigated will be presented, as well as the purpose and significance of the study, the research questions, and the hypotheses. Assumptions and key terms will be explained and the organization of the thesis will be outlined at the end of the chapter.

Introduction

Every hour students expend on mandatory testing is one less hour of classroom instructional time. Teachers are expected to cover the same amount of content, but given fewer hours with their students in which to help them gain mastery. This can lead to teacher frustration, especially when the data collected from these assessments does not lend itself to helping teachers improve their practice. Therefore, it is imperative that any testing, be it local, state, or nationally required, is worth the time cost.

In 2022, a large school system in Virginia added the Math Inventory to the list of assessments that students are required to take. Discussions between the researcher and colleagues who administered the test in the spring of 2022 at a high school in that school system indicated concerns with the reliability and predictive ability of the Math Inventory assessment, even though the Math Inventory is a nationally tested and verified instrument (*The Science Behind Quantile Measures*, n.d.). Therefore, this study is intended to compare Math Inventory scores with other accepted measures of student achievement.

Background of the Problem

Every school day, students are assessed in a variety of ways for a variety of purposes. While some assessments evaluate only the individual student, others are used to evaluate teachers, schools, and school systems, and make policy decisions on local, state, and national levels (Ghaicha, 2016).

There is much debate in the field as to how much testing is appropriate and how best to go about it (Di Martino & Baccaglini-Frank, 2017). Considering the far-reaching impact that assessments can have, it is important to verify that the tests are telling us what we think they are.

The Math Inventory is a computer adaptive test that students take multiple times per year to measure their mathematical growth. A wide range of problems are presented to students to pinpoint their current level of mathematical performance. The assessment returns a Quantile score. The Quantile Framework was introduced by MetaMetrics in 2004 as a "measurement system for mathematical understanding, which uses Rasch measurement to conjointly scale both persons and items and anchors the resulting scale in a real-world task continuum." (Williamson, 2016) The Quantile Framework has been validated by multiple sources outside the company in many different locations (*The Science Behind Quantile Measures*, n.d.).

McDonald and Pang compared performance on the Math Inventory and SAT/PSAT and found a moderate positive relationship between the two scores

(McDonald & Pang, 2021). Why, then, did teachers at the target high school have concerns about the results of the Math Inventory assessment? Before that question could be answered it must first be determined if these teacher perceptions were accurate. That is where this study fits in.

Statement of the Problem

In order to comply with Federal Requirements, school systems that receive Elementary and Secondary School Emergency Relief (ESSER) funds must administer periodic assessments to prove that the funding made a positive impact on student education. For this reason, this school system currently requires that every high student enrolled in Algebra 1, Geometry, and Algebra, Functions, and Data Analysis (AFDA) complete the Math Inventory in the fall and spring. While the Math Inventory had been nationally tested and verified (*The Science Behind Quantile Measures*, n.d.), Algebra 1 teachers at the target high school who gave the test in the spring of 2022 communicated concerns about the reliability of the Math Inventory to the investigator. If Math Inventory scores were intended to be used as evidence of student growth, it was extremely important that they accurately reflected the current level of student performance. Before explanations could be investigated, it had to be determined if there was indeed a problem.

Purpose of the Study

The purpose of this study is to quantitatively examine student data from the Math Inventory, state standardized testing, and grades in math class. Math Inventory scores from fall 2022 and spring 2023 were compared to see if students were showing improvement. Spring Math Inventory scores were compared to Virginia Standards of Learning (SOL) test scores and end of course (EOC) grades to determine if Math Inventory scores were consistent with student performance as observed by teachers in the classroom.

The focus of this study is on students who were enrolled in Algebra 1 at the target high school in Virginia for the 2022-23 school year. The variables that were examined include Fall 2022 and Spring 2023 Math Inventory Scores, Virginia Standards of Learning (SOL) test scores, and end of course (EOC) Algebra 1 grades. At the time of this study, the school in question was a small high school with an ethnically diverse population and free and reduced price lunch rates that were high for the area, students' race, socioeconomic status (SES), and language spoken in the students' home were also included as variables in the study.

Fall and Spring Math Inventory scores were included so that it could be determined if students were showing improvement over the course of the year, and if so, how much. SOL scores and EOC grades were included so that it could be determined if students' success in Algebra 1 was consistent with their Spring Math inventory score. For purposes of this study, success in Algebra 1 was defined as passing the Algebra 1 SOL and receiving a passing EOC grade (A through D) in Algebra 1. Ethnicity, SES, and language spoken in the home were included so that if discrepancies were found, potential reasons for them could be explored.

Significance of the Study

The significance of this study depended largely on the results.

If the data showed that there was an inconsistency with Math Inventory scores, it was possible that some of it was explained by race, SES, and/or language spoken in the home, which was examined as part of this study. This would have indicated a need for further research into possible reasons for this discrepancy and could lead to changes in testing policy. Perhaps another way could be found to meet requirements without losing instructional time to additional testing.

If the data showed that there was not an inconsistency in Math Inventory scores, this could help to build teacher confidence in the Math Inventory as a testing instrument. As teacher attitudes can often affect classroom environments and possibly test scores, this is critical. Future research into teacher attitudes about the Math Inventory may be warranted.

Primary Research Questions

1. Are fall Math Inventory scores significant predictors of SOL scores?
2. Are fall Math Inventory scores significant predictors of spring Math Inventory scores?
3. Is the difference in fall and spring Math Inventory scores dependent on student group?
4. Is there a significant correlation between spring Math Inventory scores and SOL scores?
5. Is there a significant correlation between spring Math Inventory scores and end of course (EOC) grades?

6. Are fall Math Inventory scores, ethnicity, and SES significant predictors of success in Algebra 1?

Hypotheses

1. Fall Math Inventory scores are expected to be significant predictors of SOL scores.

2. Based on conversations with teachers who administered the Math Inventory in the spring of 2022, it is hypothesized that fall Math Inventory scores will not be significant predictors of spring Math Inventory scores.

3. Differences in fall and spring Math Inventory scores are not predicted to be dependent on student groups.

4. Based on conversations with teachers who administered the Math Inventory in the spring of 2022, it is hypothesized that there will not be a significant correlation between spring Math Inventory scores and SOL test scores.

5. Based on conversations with teachers who administered the Math Inventory in the spring of 2022, it is hypothesized that there will not be a significant correlation between spring Math Inventory scores and EOC grades.

6. According to the school system's website, 58% of students enrolled at the target high school during the 2021-22 school year were eligible for free or reduced price lunch. The school was ethnically diverse, with 60% of students identifying as Hispanic or Latino, 15.87% identifying as Asian, 16%

identifying as White (not of Hispanic origin), 5% identifying as Black (not of Hispanic origin), and 4% identifying as other. The school had 34% of students classified as English learners, which may indicate a high percentage of families speaking languages other than English in the home. Because the statistics for the target high school differ from other schools in the area, it is expected that ethnicity, SES, and/or language spoken in the home will be significant predictors of a student's success in Algebra 1.

Research Design

The intention of this study was to explore the Math Inventory assessment – whether it was predictive or not and how scores related to other student measures. The population that was chosen for study was students at the target high school who were enrolled in Algebra 1 for the 2022-2023 school year. This school was chosen because at this school teachers have reported that Math Inventory scores did not match their classroom observations. At the target high school, all students enrolled in Algebra 1, Geometry, and AFDA were required to take the Math Inventory. Students who were enrolled in AFDA were eliminated from the study as AFDA was not an SOL course and no SOL scores were available for those students. While Geometry was an SOL course, only one verified math credit was required for graduation, which meant that most Geometry students did not need to take the SOL. For consistency in the data points available for each subject, Geometry students were also eliminated from the study. This left only the Algebra 1 students as participants. While there may have been a few instances of students who failed

Algebra 1 the previous year and repeated it during the 2022-23 school year, it is unlikely that those students would have passed the SOL and not need to retake it. If there were students in this situation (students who passed the SOL but received a failing EOC in Algebra 1 for the 2021-22 school year and repeated the course in 2022-23), they were excluded from the study as well.

The data for this study came from information that was already collected and compiled by the school system: Math Inventory scores (fall 2022 and spring 2023), SOL scores (spring 2023), end of course grades (spring 2023), student free and reduced price lunch status, ethnicity, and language spoken in the home. The data was collected on all students at the target high school that were enrolled in Algebra 1 during the 2022-23 school year.

Once this study was approved by the school system, the data was assembled from existing school system records, scrubbed of identifying data, and turned over for analysis.

Theoretical Framework

The importance of verifying the validity of test scores becomes especially important when viewed through the lens of Labeling Theory. Stemming from Émile Durkheim's book *Suicide*, which was published in 1897, but developed in the 1960s, Labeling Theory suggests that a person's self-identity and behavior is influenced by the terms that are used to describe them, or the labels that are applied to them ("Labeling Theory," 2023). This theory was applied to education by Hargreaves, Rist, and Rosenthal and Jacobson in the late 1960s and early 1970s (K. Thompson, 2017). Rosenthal and Jacobson, particularly, expanded Labeling Theory to the Self Fulfilling Prophecy Theory, showing that classroom outcomes for students were correlated to what teachers were told about students, even when the data that teachers were given had no basis in reality (K. Thompson, 2017). In fact, the impact of labels can be "even more harmful" for those who are labeled incorrectly (Schartung, 2015).

Rist explored the belief that teachers at all levels are influenced in their behavior toward students by the large amount of information that they receive, much of which is second hand (Rist, 1970). Part of the data that informs teacher opinions are test scores. Inaccurately low scores could lead to mis-informed impressions of students, theoretically driving lower student achievement, which is why verification of the validity of assessments, in this case the Math Inventory, is so important.

Assumptions, Limitations, and Scope

In order to compare Math Inventory results from a test given by multiple teachers in different classrooms on different days, some assumptions had to be made. First, the assumption was made that all teachers followed the script and guidelines provided by the school system every time the test was administered. Secondly, it was assumed that there were no significant differences in adherence to directions and student motivation by the teacher. Thirdly, it was assumed that students were given adequate time to complete the Math Inventory assessment.

This study was limited to looking at the consistency of Math Inventory scores to determine if a problem existed with the Math Inventory at the target high school. Scores were compared to results from other (trusted) measuring tools. Consequently, explanations for the results were not the main focus of this study and thus kept to a minimum.

There were many ways that the assumptions of this study could be invalid. Teachers may not have followed instructions, or followed them to varying degrees. Teacher attitudes towards the test may have affected student attitudes/ levels of effort on the test. Similarly, student adherence to directions or motivation to perform could have been affected by their teacher, circumstances, or other intrinsic or extrinsic factors. Any of these reasons could have affected the data, but were not in the scope of this study.

The study was limited to students at the target high school, as teachers at this school had communicated concerns about the Math Inventory. Expanding the scope to additional schools in a future study may be wise, depending on the results of this study.

Definition of Terms

Virginia Standards of Learning Test (SOL) - This test is written by the state and administered in the spring every year. When a student passes a course and its associated SOL exam, they receive a verified credit. High school students in Virginia must have one verified credit in mathematics in order to graduate.

Math Inventory – an online mathematics test written by Houghton-Mifflin and administered by teachers. Currently in a school system in Virginia at the high school level, Algebra 1, AFDA, and Geometry teachers are required to administer the test three times per year (fall, winter, and spring). As only students who are classified as "basic" or "below basic" are required to take it in winter, only the fall and spring scores are included in this study. When taking the test, students are given a variety of high- and low-level math problems to zero in on their current state of mathematical knowledge. The test returns a Quantile score (the mathematical equivalent of a Lexile score for reading).

Quantile score – the score returned by the Math Inventory. This number is intended to tell teachers the current level of their students' mathematical knowledge. Ranges of scores correspond to expected knowledge at the end of each grade level. The Quantile Framework was created by MetaMetrics.

Success in Algebra 1 – for purposes of this study, this is defined as a student having a passing EOC grade (i.e., not an F) and a passing SOL score (≥400)

Socioeconomic status (SES) – for purposes of this study, this is defined by a students' free and reduced price lunch status. Students who receive free lunch are considered the lowest level of SES, while students who pay full price for lunch are considered the highest level of SES.

Rasch model – based on a special case of item response theory and generalized linear model ("Rasch Model," 2022).

Summary

Chapter 1 shared a brief history of the problem and motivation behind this study. It introduced the basics of the goals of the study and how it would be accomplished. In Chapter 2, the literature surrounding this topic will be explored. Chapter 3 will explain the methods used to accomplish the research, while Chapter 4 will share the results of the study. Chapter 5 will provide conclusions and recommendations.

CHAPTER II: Literature Review

Introduction

Chapter 1 introduced the Math Inventory test and the motivation and reasoning behind this research. Chapter 2 starts with a look at the history of standardized testing and the current arguments both for and against it. This leads to an exploration of benchmark testing, with specific attention being paid to the Math Inventory and the Quantile Framework that it is based on. Studies involving the Math Inventory and other tests based on the Quantile Framework will be explored. Finally, the gaps in the current body of research will be explored and the holes that this study fills will be illuminated.

High Stakes Testing Leads to…. More Testing

A history of high stakes testing

To understand why benchmark testing is so widespread today, the history of standardized testing must first be examined. While some authors would have the reader believe that standardized testing goes back thousands of years to the story of the Gilead Guards in the Bible (Cizek, 2001; Mehrens, 1991), the story of standardized testing in America begins a bit more recently.

No Child Left Behind

In 1983, the National Commission on Excellence in Education published a report entitled *A Nation at Risk*, (National Commission on Excellence in Education, 1983) which suggested that the American education system was lagging behind those of other countries. This led to much political focus, discussion, and research, eventually resulting in the passage of the No Child Left Behind Act (NCLB) in 2002. NCLB required annual testing of students for teacher, school, and district accountability purposes – what is currently referred to as high stakes testing – with the goal of raising academic standards for all students and an emphasis on students with special needs, minorities, and non-native speakers (Minarechová, 2012).

Arguments for No Child Left Behind and High Stakes Testing

Proponents of NCLB support the transparency of test scores standardized by state and published where anyone can access them. They argue that adequate yearly progress requires schools to focus more on lower achieving groups, such as special needs students, who were previously ignored. They view states stepping in to take over and improve failing schools as an appropriate consequence. (Starr & Spellings, 2014)

Supporters of high stakes testing claim that it has many positive outcomes, including improvement of teacher professional development and awareness of special needs and non-native speakers. Because of high stakes testing, teachers are better informed about testing practices and procedures which leads to better assessments at the classroom level. School choice and accountability systems are possible because data (test scores) is accessible. Cizek argues that at some point decisions need to be made regarding students, and when they are they should be "based on sound information" in the form of standardized, generally high-stakes, tests. (Cizek, 2001)

Note that that proponents of testing are generally made up of parents and politicians, while those who argue against testing are usually educators. Testing companies tend to stay out of the debate. (Cizek, 2001)

Negative Consequences of No Child Left Behind and High Stakes Testing

Educators chafe under NCLB and the reality that their job security is based on an assessment of someone else's performance. (Cizek, 2001) While older students feel the pressure to pass their end of year state-mandated exams that determine whether or not they will graduate, younger students feel the stress of testing as well (Erskine, 2014).

At the elementary school levels, many districts have restructured students' days to include more time for reading and math, which means less time for recess, physical education, art, music, science, and social studies. The fun activities have been removed from the school day, and as a result many students no longer enjoy going to school (Berliner, 2011).

In secondary schools, many teachers narrow instruction to focus on the standards that are covered on the test (or in some cases, the standards that have historically been most emphasized on the test) (Bancroft, 2010). Teachers have had to eliminate labs and other fun activities in order to focus on covering the standards (McMillan et al., 1999). Large amounts of classroom time are devoted to test preparation, which leads to boredom on the part of students (Mora, 2011). Lower achieving schools focus on the test to the exclusion of all other topics, leading to graduates who can pass the test but are unprepared for college. This is most often seen in districts with low-incomes and/or high percentages of non-native speakers of English, two of the groups that NCLB is specifically purported to support (Ruecker, 2013).

Every Student Succeeds

Problems with NCLB led to the passage of the Every Student Succeeds Act (ESSA) in 2015, which gave states the responsibility of creating their own accountability systems. While ESSA reduced the amount of testing that was required, high stakes tests were not eliminated completely (Martin, 2021).

Throughout the NCLB era and continuing with the passage of ESSA, school districts needed a way to identify which students would probably not pass the test. Preferably identification would happen early in the school year, allowing time for additional support. Fortunately, testing companies had a solution…. more tests (benchmark assessments).

Enter Benchmark Testing

Benchmark assessments are generally aligned to state standards and are administered at multiple times throughout the school year. They are given to large numbers of students in different classes, schools, and school districts, so that educators have points of comparison to see at what level their students are performing. Multiple data points for the same student across the school year allow educators to see how much progress the student has made. Theoretically, benchmark tests can also help with early identification of students who are likely to not do well on state assessments. (Herman et al., 2010) They are used in many school districts across the country with the goal of raising test scores. (Henderson et al., 2007)

Introducing the Quantile framework

The Quantile Framework for Mathematics was released by MetaMetrics in 2004. (Williamson et al., 2016) It is a scale of mathematical tasks that uses Rasch measurement to conjointly scale persons and items with the goal of linking assessment to instruction. In other words, if a teacher knows where their student falls within the Quantile Framework, then they will know what math that student is ready to learn. (Williamson, 2016)

How it relates to the Math Inventory

The Math Inventory is a computer adaptive test that is based on the Quantile Framework. Students answer between 25 and 45 questions during one administration of the test. The test fluctuates between questions of lower and higher difficulty in order to converge on the student's current level of math knowledge and returns a Quantile score. According to the Math Inventory Technical Guide, the assessment can be completed in 20-40 minutes. (Houghton Mifflin Harcourt, 2014)

Related tests

At the time of this writing, there are several benchmark tests on the market that will return a student's Quantile score, as is shown in Table 1 below.

Table 1: K-12 Assessment Products Reporting Quantile Student Measures (Assessments & Math Programs, n.d.)

Product	Company	Grade Levels Reporting Quantile Measures
Achieve300 Math	Achieve300	K-12
aimsweb Plus	Pearson	1-8
DRC BEACON	Data Recognition Corporation (DRC)	3-8
Cognia Interim Assessments	Cognia	3-8
Exact Path Math Assessment	Edmentum	K-12
Happy Numbers Placement Assessment	Happy Numbers	K-5
Imagine Math	Imagine Learning	K-10
Instructure: CASE Assessments	Instructure/TE21	3-9
i-Ready Diagnostic	Curriculum Associates	K-12
ISIP Math	Istation	K-8
MAP Growth	NWEA	K-12
Math Inventory	Houghton Mifflin Harcourt (HMH)	K-12
MyPath	Imagine Learning	K-12
PAM (Progress Assessment of Math)	Voyager Sopris Learning	1-10
Pathway2Careers	NS4ed	Algebra 1, Geometry
Star Math	Renaissance	1-12

School districts looking for instruments to predict who will not pass the high-stakes assessments have many products to choose from. That is assuming, of course, that these instruments measure what they are purported to when subject to real world usage.

Independent research with assessments based on the Quantile Framework

As no studies could be identified that used the current iteration of the Math Inventory, and few could be found that used the Scholastic Math Inventory, the search for relevant material was expanded to include the benchmark assessments that are listed in Table 1. In order to return a Quantile score, test items must be individually evaluated for Quantile measures. It should be noted that assessments are periodically updated, and questions may have been changed, added, or deleted when the test was aligned with the Quantile Framework. It is possible that the research listed below utilized assessments that had not yet been aligned with the Quantile Framework.

Predictability Study of Istation ISIP (Math and Reading) and Ohio AIR (Math and English Language Arts) Test for 3rd-8th Grade Students in the Youngstown City School District

The focus of LaPlante's study was to determine if the ISIP could predict scores on the Ohio American Institutes for Research (AIR) reading and math assessments. LaPlante used linear regression to fit a model to the April ISIP and spring AIR scores for 3rd-8th grade students in Youngstown City School District. Predictability bands for each of the 5 Ohio achievement levels were then created using a 95% confidence interval, which was done separately for each grade level. LaPlante concluded that ISIP scores were predictive of AIR scores and recommended that, going forward, the prediction bands be used to identify which students will need extra help in order to pass the AIR tests. (LaPlante, 2018)

As this was a report of LaPlante's work, no theoretical framework or literature review was included. Recommending extensive use of prediction bands that were created using only one year's worth of data seems a bit premature. One would hope that the model would be adjusted in future years to reflect additional data collected.

LaPlante's study is relevant as it is looking at the predictive ability of a benchmark test that is similar to the Math Inventory. The location, and thus the state test used in the study, differ from focus of the current study, as does the age group of the students.

Exploring the i-Ready Predictive Capability

Shneyderman's quantitative study compared i-Ready's predictions of student Florida Standards Assessment (FSA) scores to actual FSA scores. The scores from the Fall and Winter administrations of the i-Ready diagnostic test were used by Curriculum Associates (the distributors of i-Ready) to estimate probabilities of students scoring in every achievement level. The probabilities of scoring a 3, 4, or 5 (passing) were added and the result was coded dichotomously (1 for pass if the sum was greater than 0.5, otherwise 0). Also, the student's actual score on the FSA was coded dichotomously (1 for scoring a 3, 4, or 5, or 0 for scoring a 1 or 2). Shneyderman calculated accuracy and Cohen's Kappa for each grade level for English Language Arts (ELA) and math. Sensitivity, specificity, positive predictive value, and negative predictive value were also calculated for each grade level and test. Shneyderman's study showed that, except for 8th grade math, there was a substantial agreement between the predicted and actual scores. In 8th grade math there was a moderate agreement between the predicted and actual scores. Shneyderman concluded that i-Ready diagnostic scores were predictive of FSA scores. (Shneyderman, 2017)

As this was a report of Shneyderman's work, no theoretical framework or literature review was included. Shneyderman's study could be improved by including multiple years of data.

Shneyderman's study is relevant as it is looking at the predictive ability of a benchmark test that is similar to the Math Inventory. The location, and thus the state test used in Shneyderman's study, differ from focus of the current study, as does the age group of the students.

The Relationship Between i-Ready Diagnostic and 10th Grade Students' High-Stakes Mathematics Test Scores

Thompson examined the i-Ready scores of 10th graders at a Washington high school to evaluate their ability to predict scores on the state End of Course Exams (EOCE). Thompson performed multiple linear regression on the September i-Ready scores and spring EOCE scores, with gender, ethnicity, and socioeconomic status included as additional predictor variables. Thompson found that i-Ready scores were statistically significant predictors of EOCE scores, but gender, ethnicity, and SES had no effect. (H. A. Thompson, 2018)

Thompson combined the theories of Piaget; Vygotsky; Ausubel; Anderson, Gagné, and Rumelhart; and Bloom to construct the theoretical framework for his study. High stakes testing, curriculum-based measures, computerized adaptive diagnostics, and progress monitoring were all explored in Thompson's literature review. As part of his dissertation, Thompson wrote a position paper detailing his findings and recommendations to be submitted to the high school that participated in the study, which seemed to be more of a focus than the research itself.

Thompson's paper is relevant because it used scores from a benchmark test to predict scores on a state assessment. Other similarities to the current study include that it was based at a single high school and explored the influences of ethnicity and SES on the prediction of state test scores. The difference in geographic area, however, means that a different state test was used in Thompson's work. Also, there may be differences due to focus on 10th graders, as opposed to Algebra 1 students who are mostly 9th graders.

A Study of the Predictive Validity of the STAR Math Test for the Algebra 1 End of Course Exam

Smith's quantitative study investigated the use of STAR math scores from the end of 8th grade to predict success on the Algebra 1 End of Course (EOC) state test. Smith constructed a simple linear regression model using two years' worth of data for 200 students. Based on the extremely low p-value ($p = 1.07 \times 10^{-31}$), Smith was concluded that the model was a valid tool to predict Algebra 1 EOC scores from STAR math scores. (Smith, 2012)

While no theoretical framework was shared by Smith, several pieces of related literature were explored. A similar study, conducted in Arkansas, was mentioned in the literature review, yet Smith made no effort to explain how her study was different from the Arkansas study or to define the gap in the literature that she was filling. However, using two years' worth of data, as opposed to only one like several other studies did, gives the reader more confidence in the results of Smith's study.

Smith's work is relevant as it looks at the predictive ability of a benchmark test. The same age group (Algebra 1) was examined at only one school, but the difference in location (Missouri) means that the scores in Smith's study were predicted for a different high-stakes test.

The Predictive Validity of Selected Benchmark Assessments Used in the Mid-Atlantic Region

Brown and Coughlin examined data from the publishers of four benchmark tests (STAR, MAP, TerraNova, and Study Island) in order to determine if there was evidence that the benchmark assessments could predict scores on the state tests used in the mid-Atlantic region (Pennsylvania, Delaware, Maryland, New Jersey, and DC). The documentation for each assessment was evaluated for evidence of predictive ability. Brown and Coughlin found that TerraNova showed evidence of predictive ability, with predictive validity coefficients ranging from 0.67 to 0.82, for the Pennsylvania System of School Assessments for 5th, 8th, and 11th grade. Brown and Coughlin found no evidence of predictive ability for any other benchmark test, grade, or state assessment. (Brown & Coughlin, 2007)

Brown and Coughlin's literature review noted that any previous studies were limited in scope of student age or test. Brown and Coughlin's study was limited to state assessments in the Mid-Atlantic region and benchmark tests that were popular in that area. As Brown and Coughlin suggest, it would be advantageous to have this information for all benchmark tests related to all state assessments.

Two of the assessments included in Brown and Coughlin's study are included in Table 1. While neither of them met the requirements of predictive validity for Brown and Coughlin's study (giving evidence of predictive validity for state assessments in the mid-Atlantic region), that does not mean that they did not provide evidence of predictive validity for state assessments in other regions. While Virginia is near the Mid-Atlantic region, it was not included in the scope of Brown and Coughlin's study.

MAP Growth Validation Study

In this study, Gareis examined how well MAP Growth test items align with Virginia Standards of Learning (SOLs). Test items were examined individually to determine if the reviewer agreed with the publisher's stated SOL and depth of knowledge (DOK) level. The reviewers found that while the test was well aligned with them, it did not cover all of the SOLs adequately. It also did not test the highest levels of cognitive demand. While Gareis' study did not specifically examine validity, it concluded that the MAP Growth test should not be used to predict performance on SOL tests because prediction of SOL scores "has not been systemically established." (Gareis et al., 2021)

Gareis' work is relevant because it concerns not only a similar benchmark test, but also the Virginia SOLs. However, predictive validity was specifically excluded from this study.

Comparing the Performance on the MI and SAT/PSAT for the Purpose of Monitoring Student Achievement

McDonald and Pang compared Scholastic Math Inventory and PSAT/SAT scores to see if the Math Inventory scores could be used to predict PSAT/SAT scores. Fall Math Inventory scores were compared to PSAT scores because they are taken around the same time. Similarly, spring Math Inventory scores were compared to SAT scores. Scatterplots of the data were examined, showing a moderate positive correlation. McDonald and Pang also created linear regression models for the MI and PSAT and MI and SAT both for overall scores and including sub-scores. McDonald and Pang concluded that there was a positive correlation between Math Inventory and PSAT/SAT scores based on p-values and regression coefficients. (McDonald & Pang, 2021)

While McDonald and Pang had some good insights into possible problems with their study, namely potential score inaccuracies of the Math Inventory due to student unfamiliarity with the test and the fact that it was administered on the computer instead of on paper, overall the study is weak. Alignment of tests is mentioned but unexplored, more time is spent explaining statistics than applying it to the results, and there are references to cancer patients that make no sense in this setting.

Due to the fact that it explored the validity of using the Math Inventory to predict scores on a national test, McDonald and Pang's research is arguably the most relevant paper included here. An earlier version of the Math Inventory was used, however, and scores for a national test were predicted (as opposed to a state test in the Virginia SOL).

The gap in using benchmark assessments to predict high-stakes assessment scores

This era of high stakes testing has led many school districts to rely on benchmark tests as an early indicator of which students will need additional supports in order to pass the end of year state test. While the Math Inventory is not explicitly marketed as a predictor of end of year assessment success, statements within the published documentation that the Math Inventory can "provide an indication of outcomes on summative assessments (p.12)" and "also be used to identify those students who are "at risk" (p.12)" suggest that it can be used for that purpose (Houghton Mifflin Harcourt, 2014). Assuming that this is true for other benchmark assessments is not unreasonable. If school districts are relying on these instruments to predict student success, it is imperative that the predictive ability of these assessments is confirmed.

No work had been done with the Math Inventory (or Scholastic Math Inventory) predicting scores on end of year state tests. This is the gap in the literature that this study seeks to fill. While the MAP test, which also returns a Quantile score, has been studied in the state of Virginia at the Algebra 1 level, it is not the same test. Also, that study did not look at race or socioeconomic status as indicators of the predictive nature of the test, and it did not consider consistency with end of course grades.

Conclusion

This chapter explored the history of high-stakes testing the resultant reliance of school systems on benchmark testing, and the companies that produce them, as an early warning system to identify students who will need extra help. Studies involving related benchmark assessments were examined and summarized. The lack of research validating the ability of the Math Inventory to predict success on state tests in general, and the Virginia Algebra 1 SOL in particular, was shown to be a gap that this research can fill.

Chapter three will explain the methods that were used when conducting this study, while chapter four will focus on the results and chapter five will share conclusions.

CHAPTER III: Methodology

Introduction

Chapter 1 introduced the Math Inventory assessment and the potential problems with the results from the assessment. Chapter 2 explored the history of testing to give some context to the current situation and examined the literature for similar studies in order to demonstrate the gap into which this research fits. Chapter 3 will explain the methodologies used while conducting this research.

Setting and Participants

This study took place at a diverse high school in a large school system in an urban part of Virginia. According to the school system, 1813 students were enrolled at the target high school for the 2021-22 school year. English Learner services were provided to 700 (34.30%) of these students. More than half of the students, 1186 (58.11%) received free- or reduced- price meals. The school population was made up of mostly minorities, with 324 (15.87%) students identifying as Asian, 97 (4.75%) of students identifying as Black (not of Hispanic origin), 1223 (59.92%) of students identifying as Hispanic or Latino, 321 (15.73%) students identifying as White (not of Hispanic origin), and 76 (3.72%) of students identifying as Other.

In order to compare Math Inventory scores with Virginia Standards of Learning Test (SOL) scores, the participant pool had to be carefully selected. Students who were enrolled in Algebra I, Geometry, Algebra II, and AFDA (Algebra, Functions, and Data Analysis) were all required to take the Math Inventory in the fall and in the spring. AFDA is a non-SOL course, meaning that there is no end of year SOL test for students to take, so students enrolled in AFDA were eliminated from the potential participant pool. High school students in Virginia are required to have one verified credit in math in order to graduate, meaning that they must pass the course and the associated SOL test for the same class. This means that students in Geometry and Algebra II, having taken Algebra I in middle school, generally already have a verified credit and are not required to take the SOL test at the end of the year. For this reason, only Algebra I students were included in the study, as this group would consistently have Math Inventory scores for fall and spring as well as an SOL score.

All students who were enrolled in Algebra I at the target school during the 2022-23 school year were included in this study. According to Creswell, selecting the largest sample size possible minimizes the chances that the sample will be different from the population (Creswell, 2012)

This study should be generalizable to all schools in the school district. As the student population at the target school contained a higher percentage of free- and reduced- price lunch students and ethnically minority students than the student populations at many other high schools in the district, the results may not be as generalizable if socioeconomic status (SES) or ethnicity are found to be contributing factors.

A power analysis was conducted using G*Power in order to determine the priori statistical power. In his dissertation, Thompson stated that he used a slope of 0.15, $\alpha = 0.05$ and power $= 0.95$ (H. A. Thompson, 2018). As he used a benchmark test to predict scores on a state high-stakes test, the same settings were used to conduct this analysis. As it is more common to use power $= 0.8$, that was used in the calculation instead. Using an effect size of 0.15, $\alpha = 0.05$, a power of 0.8, and one predictor, a minimum sample size was calculated to be $N = 55$. As there were approximately 400 students enrolled in Algebra I at the target school for the 2022-23 school year, adequate power was not anticipated to be an issue for this study.

Instrumentation

Math Inventory

The Math Inventory is a benchmark assessment that is administered via computer multiple times per year (fall, winter, and spring in the target school system) in order to monitor a student's progress. Students are given a variety of high- and low-level math questions in order to zero in on their current math proficiency. The instrument returns a Quantile score (similar to a Lexile score for reading), based on the Quantile Framework.

The Quantile Framework was validated using construct-identification validity, or by comparing Quantile measures to other mathematical achievement measures. According to MetaMetrics, the correlation between a score on the Virginia SOL test and the Quantile measure is between 0.86 and 0.89, depending on which grade level/course is being examined. (Houghton Mifflin Harcourt, 2014) This is a strong correlation, meaning that, according to the creators of the Quantile Framework, students with a high Quantile score should get a high score on their math SOL test.

Table 2: Math Inventory performance level ranges by grade (Houghton Mifflin Harcourt, 2014)

Grade	Below Basic	Basic	Proficient	Advanced
K	EM400–EM185	EM190–5	10–175	180 and Above
1	EM400–60	65–255	260–450	455 and Above
2	EM400–205	210–400	405–600	605 and Above
3	EM400–425	430–620	625–850	855 and Above
4	EM400–540	545–710	715–950	955 and Above
5	EM400–640	645–815	820–1020	1025 and Above
6	EM400–700	705–865	870–1125	1130 and Above
7	EM400–770	775–945	950–1175	1180 and Above
8	EM400–850	855–1025	1030–1255	1260 and Above
9	EM400–940	945–1135	1140–1325	1330 and Above
10	EM400–1020	1025–1215	1220–1375	1380 and Above
11	EM400–1150	1155–1345	1350–1425	1430 and Above
12	EM400–1190	1195–1385	1390–1505	1510 and Above

*Emerging Mathematician

A student's Quantile score should tell a teacher what math a student is ready to learn and can be used to determine performance level by grade (Table 2). According to the Math Inventory Professional Learning Guide, the performance level categories by Quantile score break down as follows:

Advanced: students exhibit superior performance on grade-level-appropriate skills and concepts and are on track for college and career (in terms of their mathematical development)

Proficient: students exhibit competent performance on grade-level-appropriate skills and concepts and are on track for college and career (in terms of their mathematical development)

Basic: students exhibit minimally competent performance on grade-level-appropriate skills and concepts and may be considered marginally on track for college and career (in terms of their mathematical development)

Below Basic: students do not exhibit minimally competent performance on grade-level-appropriate skills and concepts and are not considered on track for college and career (in terms of their mathematical development) (Math Solutions, 2020)

According to MetaMetrics, the company that created the Quantile framework that the Math Inventory is based on, the Quantile Framework has been “externally validated by hundreds of studies in more than 25 states and 24 countries” (*The Science Behind Quantile Measures*, n.d.) Information from reliability studies of the Scholastic Math Inventory (an earlier version of the Math Inventory) was included in the Math Inventory Technical Guide (Houghton Mifflin Harcourt, 2014) and showed marginal reliability ranging from 0.96 to 0.98, meaning that it was consistently able to order students. Data from a test-retest study, where administrations of the test were conducted one week apart showed test-retest correlations of 0.70 to 0.79, which was considered satisfactory (Houghton Mifflin Harcourt, 2014). (Houghton Mifflin Harcourt, 2014). Descriptive statistics from a validation study were also included, indicating that all measures were “as expected” (Houghton Mifflin Harcourt, 2014).

Because the Math Inventory is a computer adaptive test containing thousands of questions, a copy of it is not available to include in the Appendix.

SOL

The Virginia Standards of Learning (SOL) test was developed by the Virginia Department of Education (VDOE) in order to ensure that students in Virginia schools are achieving a minimum level of competency in math (*Virginia SOL Assessment Program | Virginia Department of Education*, n.d.) Standards of learning were assigned to reporting categories, and a test blueprint was created to guide how many questions from each reporting category would be included in the test. Test questions were written to specific Standards of Learning and difficulty levels by educators who are experienced in authoring questions for assessments. They went through multiple reviews for quality, accessibility, and fairness before being field tested as non-scored items on a spring administration of the SOL test. The data from the field test was then reviewed to ensure that no questions were biased. (Virginia Standards of Learning Assessments Technical Report 2021-22 Administration Cycle, n.d.) Questions were then assembled into tests by content specialists and psychometricians with the goal of creating tests that are "equivalent in content representation and psychometric characteristics with year and across years". (Virginia Standards of Learning Assessments Technical Report 2021-22 Administration Cycle, n.d., p.18)

Raw scores, or how many questions a student answered correctly, can be affected by which version of the test was administered, so each test must be converted to a scale score before it can be compared to scores from other students or other administrations of the test. The unidimensional IRT Rasch Partial Credit Model and WINSTEPS software were used to calibrate the difficulty and underlying student proficiency of test questions. A linear transformation was then applied to the proficiency scale to create scale scores. SOL tests return a scaled score between 0 and 600. A score of 400 is the threshold for passing, while 500 or above considered pass/advanced. It is interesting to note that because of the way the scores are scaled "the distance between scale scores does not remain the same for each change in the raw scores" (Virginia Standards of Learning Assessments Technical Report 2021-22 Administration Cycle, n.d., p. 36)

Reliability for the Algebra 1 SOL assessment was calculated using Cronbach's alpha to be 0.90 and 0.91 for the two versions of the test. Decision accuracy and consistency statistics are shown below in Table 3 (Virginia Standards of Learning Assessments Technical Report 2021-22 Administration Cycle, n.d., p.59). These show that decisions based on a student's performance on the assessment would be fairly consistent and accurate.

Table 3: Algebra 1 SOL assessment Decision Consistency and Accuracy Indices

	N	Accuracy	False Positive	False Negative	Consistency
Version 1	53,903	0.93	0.03	0.04	0.90
Version 2	40,030	0.92	0.04	0.04	0.89

A released version of the Algebra 1 SOL is included in Appendix . (*Released Tests & Item Sets (ALL SUBJECTS) | Virginia Department of Education*, n.d.)

Procedure

For each participant, the Fall 2022 Math Inventory score, Spring 2023 Math Inventory score, Spring 2023 Algebra I SOL score, Algebra I end of course grade, race/ethnicity, free- and reduced-price lunch status, and language spoken at home were compiled from archival records by school system employees. All identifying information was stripped from the data before it was released to the researcher.

As the data being utilized in this study came from existing school system records, no extra effort was required on the part of participants. Because identifying information was removed before it was released, participants experienced no harm or discomfort and participant confidentiality was maintained at all times. Once received, the data was not viewed by anyone except the researchers. There was no risk to participants of this research.

Permission to complete this study was obtained from the IRB at Shawnee State University and the school system. As the school system prefers to remain unidentified, documents indicating permission to proceed from both agencies have not been included.

Data Processing and Analysis

Several statistical techniques were used to answer the research questions. The variables used include fall and spring Math Inventory scores, Algebra I SOL scores, EOC grades, SES, race/ethnicity, and language spoken in the home. Quantitative test scores, from both benchmark and high stakes state administered assessments, have often been used as variables in education research (LaPlante, 2018; McDonald & Pang, 2021; Smith, 2012). Iverson used end of course grades when examining the relationship between the Iowa Assessment and different grading practices (Iverson, 2014), as did Summers when attempting to predict success on the Arkansas Benchmark Test (Summers, 2009). SES based on school lunch status has been used as a categorical variable multiple times in research related to benchmark and/or high stakes testing (Lewis, 2013; Martin, 2021; H. A. Thompson, 2018)

Race/ethnicity has also been a common categorical variable in research related to benchmark and/or high stakes testing (Lewis, 2013; Martin, 2021).

Several researchers, including Ruecker, Vela, and Valenzuela, have investigated English Language Learners (ELLs) and their success in school (Ruecker, 2013; Valenzuela, 2006; Vela et al., 2017). According to Vela, "ELLs, or students who are English Language Learners, speak a language other than English in their home." (Vela et al., 2017) Because of delays in ELL level testing due to the coronavirus pandemic (Isbell & Kremmel, 2020), language spoken in the home was used as a categorical variable instead of ESOL level.

The statistical techniques that were used to answer each research question are described below:

a. Are fall Math Inventory scores significant predictors of Virginia Standards of Learning (SOL) scores?

Simple linear regression was used to determine if fall Math Inventory scores are significant predictors of SOL scores. McDonald and Pang, Smith, and LaPlante all used simple linear regression to predict performance on state or nationally administered assessments from benchmark assessments (LaPlante, 2018; McDonald & Pang, 2021; Smith, 2012)

b. Are fall Math Inventory scores significant predictors of spring Math Inventory scores?

Simple linear regression was used to determine if fall Math Inventory scores are significant predictors of spring Math Inventory scores.

c. Is the difference in fall and spring Math Inventory scores dependent on student group?

The difference between fall and spring Math Inventory scores was calculated. Students were counted as having gained knowledge if this difference was a positive number and having not gained knowledge if this difference was negative or 0. A chi square test for independence was used to compare the numbers of students who gained/did not gain knowledge within each category (Hispanic/non-Hispanic, high/low SES, successful/not successful in Algebra 1).

d. Is there a significant correlation between spring Math Inventory scores and Virginia Standards of Learning (SOL) Test scores?

The correlation coefficient between spring math inventory scores and SOL test scores was examined. The correlation between benchmark test scores and high-stakes test scores was used in studies validating the MAP test and predicting performance on the Ohio AIR test using ISIP (Gareis et al., 2021; LaPlante, 2018).

e. Is there a significant correlation between spring Math Inventory scores and end of course (EOC) grades?

The correlation coefficient between spring math inventory scores and EOC grades was examined.

f. Are fall Math Inventory scores, race/ethnicity, and SES significant predictors of success in Algebra 1?

The difference between the spring Math Inventory score and the fall Math Inventory score, or score difference, was calculated for each student. A categorical variable, success in Algebra I, that was true if the student passed the SOL and received a grade of D or higher as their EOC Algebra I grade and false otherwise, was determined for each student.

Direct logistic regression was used to examine the score difference predicted from success in Algebra I, SES, race/ethnicity and language spoken in the home.

Summary

Chapter 3 explained and justified the methods that were used in this research. Chapter 4 will share the results of the study, while Chapter 5 will explain conclusions about them.

CHAPTER IV: Results

Chapter one introduced the problem and research questions, while chapter two gave background and literature review related to the questions. Chapter three explained how each of these research questions would be tested. Chapter four will give details about the data that was received, how it was processed, and share the results of the statistical analysis.

Data

The data was received from the school system in the form of an Excel file containing 561 cases which included a researcher ID, fall Math Inventory score (MIf), spring Math Inventory score (MIs), Algebra 1 end of course grade (EOC), Algebra 1 SOL score (SOL), whether the student identified as Hispanic or not (Hisp), and whether the student received free or reduced price meals or not (SES). Language spoken in the home could not be included, as the school system felt that providing this data for such a small sample size would put student anonymity at risk. Fall and spring Math Inventory scores and SOL scores were continuous quantitative variables. Algebra 1 EOC grade was given as a categorical variable with 11 possible categories (A, A-, B+, B, B-, C+, C, C-, D+, D, F). Whether a student identified as Hispanic and their free or reduced price meal status were given as dichotomous categorical variables.

Upon further analysis, it was discovered that 156 of the cases in the raw data were related to SOL retakes, thus the cases had the same researcher ID but different Algebra 1 SOL score (and no other information). In 40 of these cases the additional

SOL score was lower than the original score and in 116 cases the additional SOL score was higher than the original. For all of these cases the higher of the two SOL scores was retained and the lower score deleted.

Even after removing cases with duplicate SOL scores there was still missing data, which are summarized in Table 4. Once the cases missing more than one piece of data were deleted, $n = 326$ cases remained for analysis. Of these, 29 were missing MIf data and 132 were missing MIs data, leaving a total of 165 cases with complete data. The largest possible data set was used when answering each research question.

Table 4: Summary of number of cases with missing pieces of data

Situation	Number of cases
No MIf	30
No MIs	129
No MIf or MIs	38
No MIf, MIs, or SOL	3
No MIs or SOL	4
No MIf, MIs, EOC	29
No MIf, MIs, SOL, or EOC	1

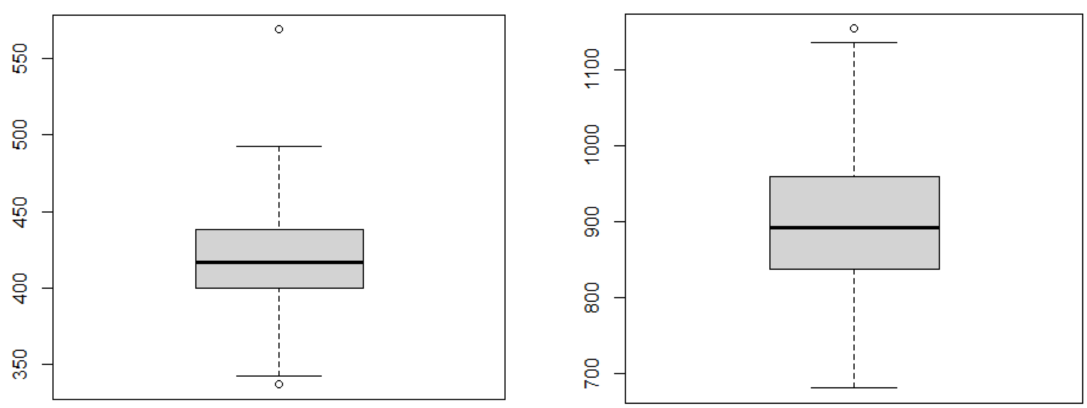
Data analysis

Are fall Math Inventory scores significant predictors of SOL scores?

There were $n = 297$ cases that contained both a fall Math Inventory score and an Algebra 1 SOL score. For this question, the fall Math Inventory score was the independent variable, and it was used to predict the SOL score (M=418.0, SD=32.0), the dependent variable. The descriptive statistics are shown below in Table 5. The mean Math Inventory score falls into the "Below Basic" category for 9th graders, according to the publisher (see Table 2), and the mean SOL score is a passing score.

Boxplots of both the fall Math Inventory score and SOL score are shown below in Figure 1. Note the outlier on the Math Inventory scores and the extreme outlier on the SOL scores.

Figure 1: Box plots of SOL and Fall Math Inventory scores

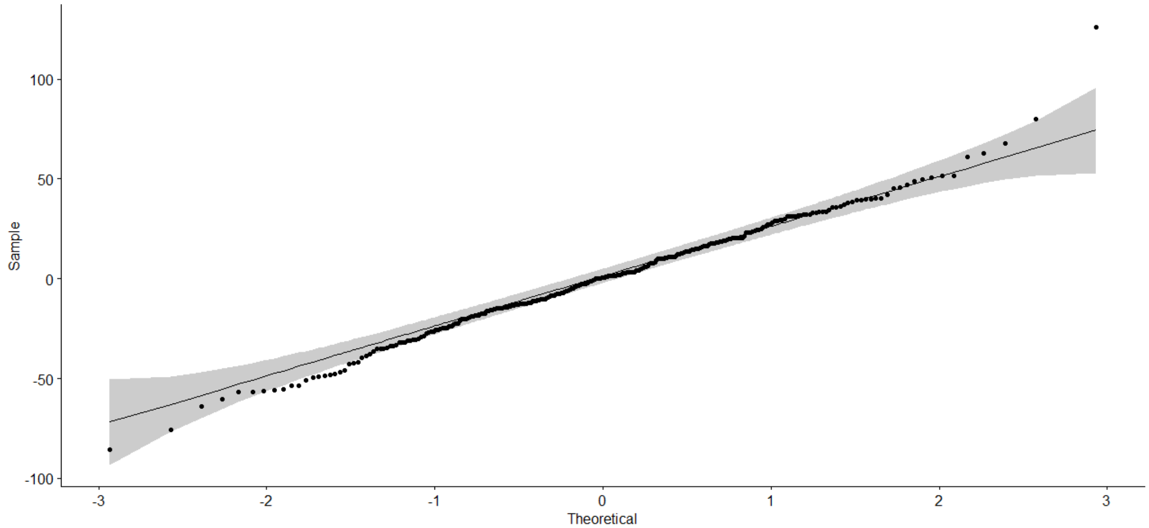


SOL scores Fall Math Inventory scores

Since the number of cases per predictor easily exceeds 15 (Field et al., 2013), there was no concern with adequate sample size. Analysis was performed using R (R Core Team, 2021).

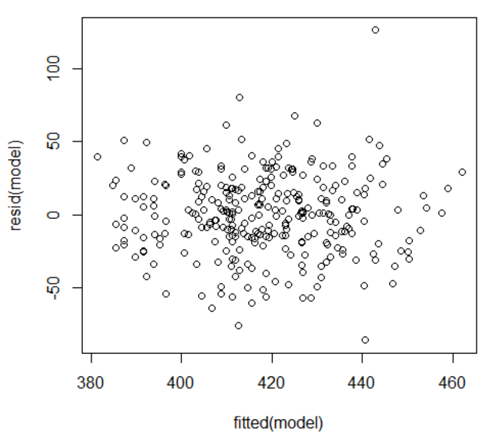
Results of the evaluation of the assumptions indicated some concerns with independence, normality, and equal variances. Independence was tested with the Durbin-Watson test: D-W Statistic $= 0.509$, $p = 0$, and it was determined that there was a problem with independence. Shapiro's test for normality also revealed concerns: $W = .989,\ p = .025$. As seen in Figure 2, the GGQQ Plot shows data points outside of the predicted sample area both between -2 and -1.5 and greater than 2.5 on the theoretical (horizontal) axis.

Figure 2: GGQQ Plot of Residuals



The scatterplot (Figure 3) shows no discernable patterns, indicating that a linear model could be an appropriate fit for the data.

Figure 3: Scatterplot of Fitted vs. Residual values



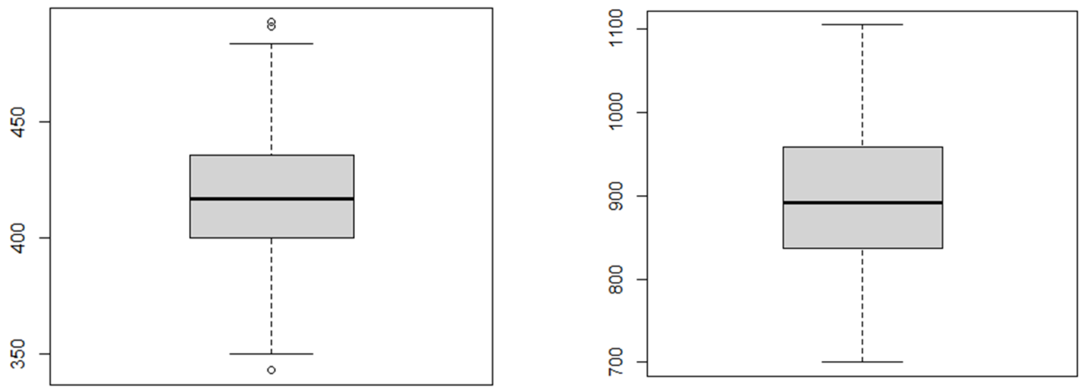
A simple linear regression model was created using Fall Math Inventory Scores to predict SOL scores. The unstandardized regression coefficients and standard errors are shown in Table 5. The multiple R-squared reveals that 25.26% of the variance in SOL scores is explained by the regression on Fall Math Inventory scores. A test of the full model against the intercept-only model was significant; $F(1{,}295) = 99.723, p < .001$. There were 19 cases that were either outliers, had extreme standardized residuals (|standardized residual|>2), or were extremely influential cases (hat values $> 3\left(\frac{k+1}{n}\right)$). These cases were deleted before a second model was created.

Table 5: Descriptive statistics for Initial Model

Variable	Mean	St. Dev.	B	SE
Fall Math Inventory	895.4	94.7	0.170	0.017
Intercept			265.714	15.332

Removal of the 19 problem cases improved the independence, normality, and equal variances of the dataset. The descriptive statistics for SOL (M=417.7, SD=27.7) and Fall Math Inventory score (see Table 6) changed slightly, as did their boxplots (see Figure 4).

Figure 4: Boxplots of SOL and Fall Math Inventory scores after removal of problem cases

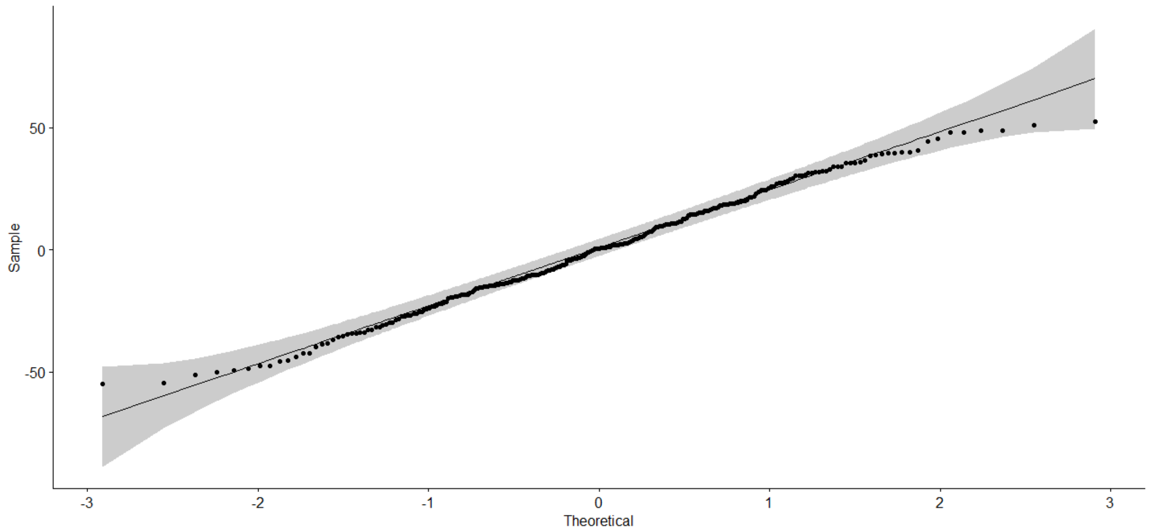


SOL scores

Math Inventory Fall scores

Concerns with the assumptions of independence, normality, and equal variances in the first model were resolved by the second model. Independence was tested with the Durbin-Watson test: D-W Statistic $= 1.964$, $p = 0.778$, and it was determined that there was no longer a problem with independence. Shapiro's test for normality also revealed no more concerns: $W = .992$, $p = .137$. The GGQQ Plot also showed improvement based on these changes (see Figure 5), with no data points outside of the prediction field.

Figure 5: GGQQ Plot for Second Model



The unstandardized regression coefficients and standard errors for the second model are shown in Table 6. The multiple R-squared reveals that 28.17% of the variance in SOL scores is explained by the regression on Fall Math Inventory scores. A test of the full model against the intercept-only model was significant; $F(1{,}276) = 108.23, p < .001$. Examination of the standardized residuals, outliers, and influential cases determined that no additional cases needed to be deleted.

Table 6: Descriptive Statistics for Second Model

Variable	Mean	St. Dev.	B	SE
Fall Math Inventory	892.1	91.4	0.161	0.015
Intercept			273.973	13.885

The model predicting SOL scores from fall Math Inventory scores was statistically significant, so the resulting equation could be used to predict possible SOL scores from fall Math Inventory scores:

Equation 1 $$SOL = 265.714 + 0.170MIf$$

Utilizing this formula, Quantile scores for predicting SOL cutoff scores were calculated, as shown in Table 7. The Quantile labels were chosen using the 9th grade cutoff ranges in Table 2, as the majority of Algebra 1 students were 9th graders. Based on this, one could conclude that students scoring Below Basic on the fall Math Inventory could be considered for extra intervention. Or to be more specific, students with a Quantile score below 800 should be considered for extra help. This is similar to LaPlante's prediction bands to identify students who may not pass the Ohio AIR test based on their scores on the ISIP (LaPlante, 2018).

Table 7: Estimated Quantile scores for SOL cutoff scores

SOL score band	Minimum SOL score	Estimated Quantile score	Quantile label
Automatic retake	350	606	Below Basic
Automatic retake	375	642	Below Basic
Pass	400	790	Below Basic
Pass Advanced	500	1378	Advanced

Another way to interpret the equation is to convert the Quantile score ranges to predicted SOL scores, as shown in Table 8. Again, these are based on the 9th grade levels in Table 2. It is interesting to note that the Pass Advanced score on the SOL is predicted to be above the minimum Quantile score for Advanced. This may be because the one student in the data set who scored in the Pass Advanced range on the SOL was an extreme outlier and deleted before the model was re-run.

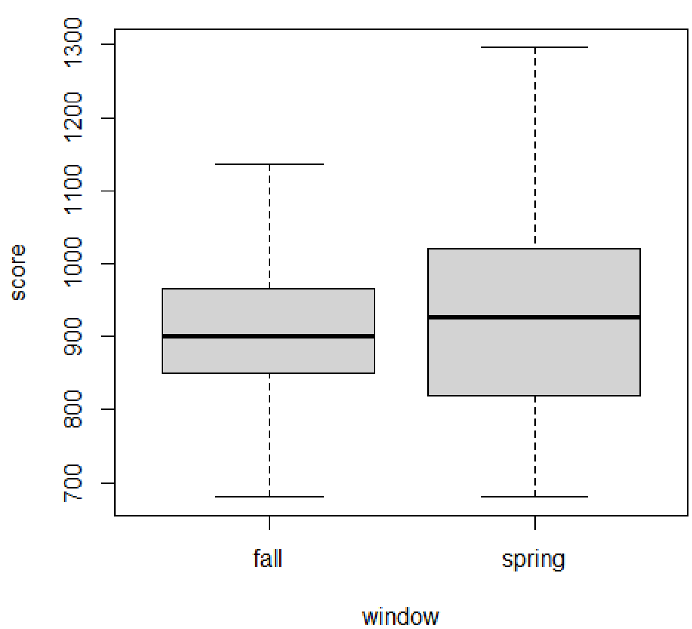
Table 8: Quantile labels converted to predicted SOL score ranges

Quantile label	Quantile Range	Predicted SOL range	Corresponding SOL label(s)
Below Basic	EM400-940	Below 425	Did not Pass/Pass
Basic	945-1135	426-458	Pass
Proficient	1140-1325	459-490	Pass
Advanced	1330 and above	Above 491	Pass/Pass Advanced

Are Fall Math Inventory Scores Significant Predictors of Spring Math Inventory Scores?

A repeated measures ANOVA was conducted on the n=165 students who had both fall (mean=903.818, sd=89.332) and spring (mean=926.218, sd=134.340) Math Inventory scores. Boxplots (Figure 6) of the data show no outliers.

Figure 6: Boxplots of fall and spring Math Inventory scores



The Shapiro-Wilk test (W=0.994, p=.775 for fall, W=0.985, p=.078 for spring) indicated no issues with normality. Levene's Test indicated that there was an issue with homogeneity of variance (F=25.695, p<.001). Because there are only two levels of testing, sphericity was not a concern.

The ANOVA showed a small effect size, $\eta^2 = 0.0096$ with $F(1{,}164) = 5.074, p < .001$. This means that there was not a significant difference between the fall and spring measurements. The Bonferroni test (p=.026) showed that the difference between the fall and spring scores was statistically significant at the .05 level. Clearly, these two results are inconsistent with each other.

Because of the issues with homogeneity of variance, Friedman's non-parametric rank sum test was also conducted, $\chi^2(1{,}165) = 3.286, p = .070$, showing that the differences between the fall and spring Math Inventory scores were not statistically significant at the .05 level. This is confirmed by the Friedman multiple comparison test, which showed an observed difference of 23 and a critical difference of 25.176, p=.220. Thus fall Math Inventory scores could not reliably predict spring Math Inventory scores.

A priori sample size was calculated in G*Power (Faul et al., 2009) using an effect size of 0.0989, 2 groups with 165 measures each, and correlation of 0.405. The software calculated that a sample size of 546 was needed in order to achieve adequate power, which this study fell far short of.

Is the difference in fall and spring Math Inventory scores dependent on student group?

In order to look at which groups are making gains in Math Inventory scores over the course of the year, first the difference between spring and fall Math Inventory scores was calculated for all $n = 165$ students who had both scores. That was then expressed as a categorical variable (gain) which was true if the spring Math Inventory score was higher than the fall Math Inventory score and false if it was not. Next, a series of chi square tests of independence were conducted in order to determine if any of these relationships were related. The counts of each comparison group are shown in Table 9.

Table 9: Counts for chi square independence comparison

	Not Hispanic	Hispanic	No free or reduced price lunch	Free or reduced price lunch	Not successful in Algebra 1	Successful in Algebra 1
No gain	23	50	15	58	14	59
Gain	29	63	22	70	11	81

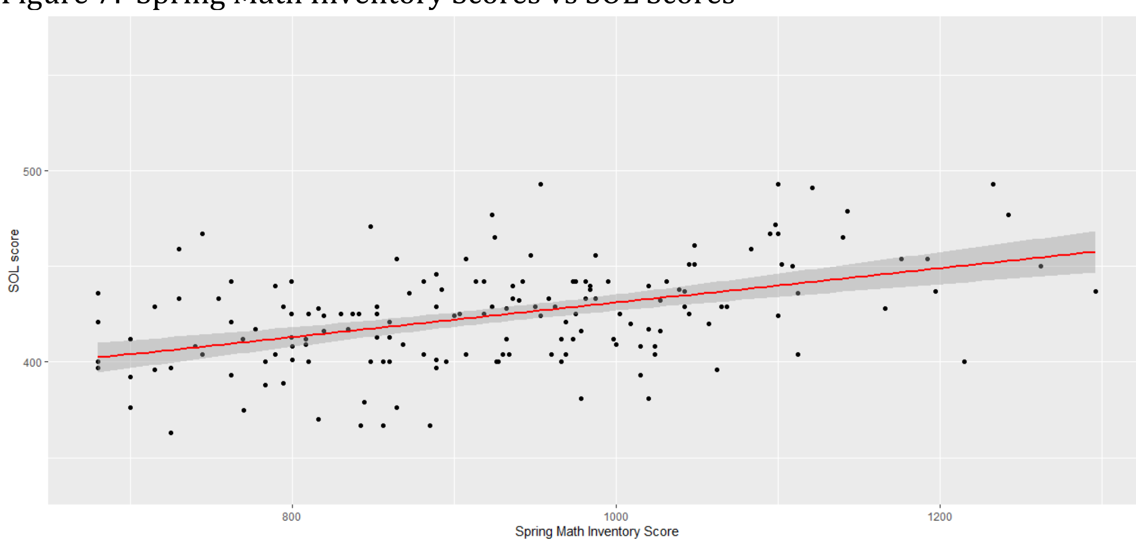
Pearson's chi squared test returned $\chi^2(1, N = 165) = 0, p = 1$ for the comparison of Hispanic/not Hispanic groups, indicating that the groups were independent. The same test yielded $\chi^2(1, N = 165) = 0.107, p = .744$ for the free/reduced price lunch comparison, indicating that these groups were also independent of each other. The same test was run to look at groups of students who were/were not successful in Algebra 1 and did/did not make gains over the course of the school year, yielding $\chi^2(1, N = 165) = 1.137, p = .286$. This indicates that these groups are also independent.

Yates' continuity correction was used for all comparisons because there were only two choices in each group. Standard errors were examined for all groups, and there were no concerns as none were found to be greater than 2.

Is there a significant correlation between spring Math Inventory scores and Virginia Standards of Learning (SOL) Test scores?

There were $n = 194$ cases that contained both a spring Math Inventory score and an SOL score. Spring Math Inventory scores and SOL scores were found to be moderately positively correlated, $r(192) = .458$, with 95% CI [.339,.562] using Pearson's product-moment test. A scatterplot of the data, as compared to a line of best fit, is shown in Figure 7.

Figure 7: Spring Math Inventory Scores vs SOL Scores



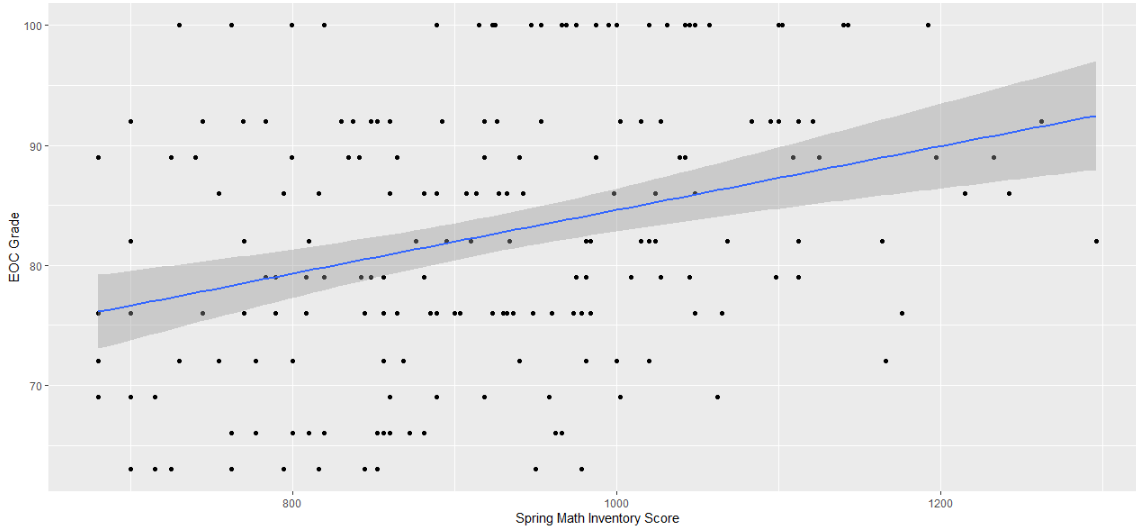
Is there a significant correlation between spring Math Inventory scores and end of course (EOC) grades?

There were $n = 194$ cases that contained both a spring Math Inventory score and an end of course grade. End of course grades were received as letter grades but were converted to numeric values by assigning each letter a value equivalent to the top of that grading interval, as shown in **Error! Not a valid bookmark self-reference.**. Spring Math Inventory Scores and EOC grades were found to be moderately positively correlated, $r(192) = .318$, with 95% CI [.185, .439] using Pearson's product-moment test. A scatterplot of the data, as compared to a line of best fit, is shown in Figure 8.

Table 10: Converting EOC letter grades to numeric values

Letter grade	Range (by school system)	Conversion value
A	93-100	100
A-	90-92	92
B+	87-89	89
B	83-86	86
B-	80-82	82
C+	77-79	79
C	73-76	76
C-	70-72	72
D+	67-69	69
D	64-66	66
F	≤ 63	63

Figure 8: Spring Math Inventory Scores vs. EOC Grades



Are Fall Math Inventory scores, race/ethnicity, and socioeconomic status significant predictors of success in Algebra 1?

After deleting records with no fall Math Inventory scores, n=303 cases remained. If a student both passed the SOL (achieved a score greater than or equal to 400) and passed the class (achieved an EOC grade that was not an F), they were considered to have been successful in Algebra 1. Cases that were missing either an SOL or EOC score (or both) were then deleted, leaving a total of n=297 cases for this investigation.

A direct logistic regression analysis was performed on success in Algebra 1 as the outcome and three predictors: fall Math Inventory score (MIf), whether or not a student identified as Hispanic (Hisp), and a student's free and reduced-price lunch status (SES).

The first model was created using n=297 students who had MIf scores ranging from 680-1155, mean=895.448, sd=94.737. Of these 297 students, 223 (75.084%) were successful in Algebra 1, 211 (71.044%) identified as Hispanic, and 226 (76.094%) received free or reduced-price lunch.

The information from the first model is shown in Table. Although neither Hisp or SES were shown as statistically significant predictors, only SES was removed for the second model because Hisp was much closer to being statistically significant. The information from the second model is shown in Table 11.

Table 11: Logistic regression analysis of predicting success in Algebra 1, first model ($n = 297$)

Variable	B	Wald (z-ratio)	p-value	Odds Ratio (OR)	95% CI Lower, OR	95% CI Upper, OR
MIf	0.010	5.559	<.001	1.010	1.007	1.014
Hisp	-0.578	-1.455	.146	0.561	0.249	1.194
SES	-0.137	-0.350	.726	0.872	0.395	1.856
(Constant)	-7.194					

While both models were statistically significant when compared to a constant-only model, $\chi^2(2, N = 297) = 45.661, p < .001$ for model 2 and $\chi^2(3, N = 297) = 45.785, p < .001$ for model 1, there was not a statistically significant difference between the two models, $\chi^2(1, N = 297) = 0.123, p = .725$.

Table 12: Logistic regression analysis of predicting success in Algebra 1, second model ($n = 297$)

Variable	B	Wald (z-ratio)	p-value	Odds Ratio (OR)	95% CI Lower, OR	95% CI Upper, OR
MIf	0.010	5.556	<.001	1.010	1.007	1.014
Hisp	-0.631	-1.713	.0867	0.532	0.250	1.067
(Constant)	-7.228					

Casewise diagnostics revealed several cases that were problematic in both models. Those cases were removed and the models were re-run.

A logistic regression model was created using $n = 286$ students who had MIf scores ranging from 705-1155, mean= 899.951, sd=92.641. Of these 286 students, 220 (76.923%) were successful in Algebra 1, 210 (73.427%) identified as Hispanic, and 221 (77.273%) received free or reduced-price lunch.

The full model ($AIC = 266.61$), predicting success in Algebra 1 from MIf, Hisp, and SES, was statistically significant when compared to a constant-only model ($AIC = 311$), $\chi^2(3, N = 286) = 50.388, p < .001$. The second model ($AIC = 264.81$), predicting success in Algebra 1 from only MIf and Hisp, was also statistically significant when compared to a constant-only model, $\chi^2(2, N = 286) = 50.182, p < .001$. The two models were not statistically significant from each other, $\chi^2(1, N = 286) = .206, p = .650$. Note that when the full model was optimized using the backwards method, the resulting model was identical to the second (no SES) model. This is consistent with the AICs of the three models, where a lower AIC indicates a better fit to the data. The variance in success status accounted for is small, with McFadden's rho=0.163, df=3 for the full model and McFadden's rho=0.162, df=2 for the second (no SES) model.

Using the second (no SES) model, prediction success (using 0.5 as the threshold) was acceptable with 230 of 286 cases (80.420%) predicted accurately. Sensitivity and specificity for that model were 0.955 and 0.303, respectively.

Table 13 below shows the regression coefficients, Wald statistics, odds ratios, and 95% confidence intervals for the odds ratios for the predictors in the second model.

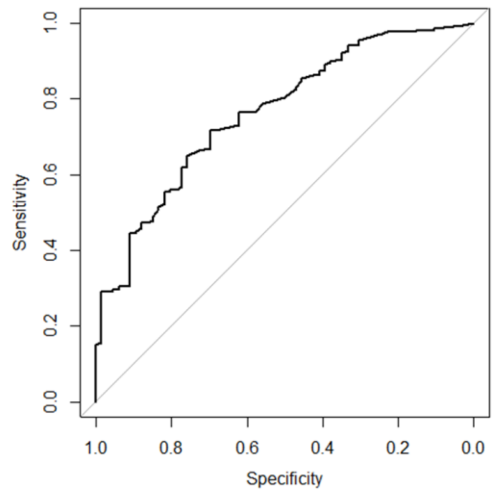
Table 13: Logistic regression analysis of predicting success in Algebra 1, second model ($n = 286$)

Variables	B	Wald (z-ratio)	p-value	Odds Ratio (OR)	95% CI Lower, OR	95% CI Upper, OR
MIf	0.00949	4.816	<.001	1.010	1.006	1.014
Hisp	-1.537	-2.793	.00523	0.215	.0622	0.569
(Constant)	-5.835					

Variance inflation factors (VIF) values were 1.016 (MIf) and 1.016 (Hisp), indicating that multicollinearity is not a problem. Examination of the significance levels of the interaction between MIf and the log of itself (Hosmer & Lemeshow, 1989) indicated that linearity between the predictor and the logit of itself could be assumed.

Using the two-predictor model, a receiver operating characteristic curve (ROC) is shown below in Figure 9. The AUC was found to be 0.763, which indicates average model accuracy (Tape, 2015).

Figure 9: ROC Curve for predicting success in Algebra 1 (second model)



The equation from the logistic regression model is shown below:

Equation 2 $$P(Success\ in\ Algebra\ 1) = \frac{e^{-5.835+.00949MIf-1.537Hisp}}{1+e^{-5.835+.00949MIf-1.537Hisp}}$$

In other words, after controlling for all other variables, students who identify as Hispanic are 21.5% more likely to be successful in Algebra 1 than students who do not identify as Hispanic.

Conclusion

The results of many tests of the Math Inventory were shared in chapter four. It was shown that the fall Math Inventory score could be used to reliably predict a student's SOL score, but not their spring Math Inventory score. A moderately positive correlation was demonstrated for both spring Math Inventory/SOL scores and spring Math Inventory scores/EOC grades. Gains made over the course of the year (as measured by a difference in Math Inventory scores) were shown to be independent of student groupings. A model to predict success in Algebra 1 from fall Math Inventory scores and a student's identification as Hispanic/not Hispanic showed average results.

Chapter five will contain analysis and possible explanation of these results, along with further discussion and next steps.

CHAPTER V: Summary

Introduction

Every year, countless hours of instructional time are lost to benchmark and/or standardized testing. Over the course of one school year, Algebra 1 students are required by the district to take two district-created benchmark tests and sit for the Math Inventory three times (fall, winter, and spring). Freshmen and students who have not yet earned a verified credit in math are also required to sit for the SOL at least once. In a school year where a teacher is scheduled to meet with their students approximately 90 times, that is 6% of total instructional time lost to testing for all students. This calculation does not include the SOL or final exams, not to mention the various unit tests that students take across the course of the year. Needless to say, every day of instructional time that is lost makes it that much harder to get through the required content.

Because so much instructional time is lost to testing, it is imperative that the testing tools that are used are carefully chosen, providing meaningful, applicable, actionable data to our teachers, schools, and school districts. The Math Inventory provides a Quantile score for each student who takes it. This Quantile score can be converted to a label based on the students' grade level, as shown previously in Table 2.

But what does it all mean? In order for this data to be useful, teachers need context. Knowing that a student scored "below basic" on the fall Math Inventory doesn't mean much to a teacher who does not have extensive experience with

Quantile scores. Does it mean that the student is likely to pass the SOL? The class? This study aims to provide the context that teachers need to make their students' Quantile scores meaningful.

Students often have many labels applied to them: multilingual learner, honors student, still needs a verified credit, special needs, gifted, etc. According to labeling theory, these labels can affect how people, especially teachers, treat students, eventually becoming self-fulfilling prophecies (Rist, 1970). Before we add yet another label to our math students (basic, below basic, proficient, or advanced), we should make the effort to understand exactly what that label means.

Implications

The ability to look at a fall Math Inventory score and predict an SOL score would be an incredible tool for teachers trying to identify where to focus extra help early in the school year. As detailed in Chapter 4, Equation 1 has been shown to predict SOL scores from fall Math Inventory with statistical significance at the target school. Using the information shown in Table 7 and Table 8, teachers can identify students who may need assistance in passing the SOL and begin targeted remediation early in the school year.

Unfortunately, every school and the students attending it are different, and it is highly likely that, while the process for developing the equation, and thus the cutoff numbers shown in Table 7 and Table 8, is sound, the equations appropriate for predicting SOL scores in different schools would vary greatly. So while the equation is appropriate for use at the target school, it is likely not generalizable to other schools or school systems. The process of creating the equation, however, and the application of using it to predict performance on SOL scores, should be generalizable to any individual school in the state of Virginia.

These findings, that a Quantile-based benchmark assessment can be used to predict performance on a state assessment, are consistent with the work of Smith and LaPlante (LaPlante, 2018; Smith, 2012).

A logistic regression model was also created to investigate if fall Math Inventory scores, race, and SES could predict success in Algebra 1. While SES was not a significant predictor of success in Algebra 1, whether or not a student identified as Hispanic was, as is shown in Equation 2. This is an interesting result, as 94.7% of the non-Hispanic students were successful in Algebra 1, compared to 70.5% of the Hispanic students. This may be due to the fact that the Hispanic students had a much lower mean fall Math Inventory score (881.995 versus 949.566 for non-Hispanic students), although the standard deviations for both groups were similar (87.775 for Hispanic and 88.011 for non-Hispanic). The fact that ethnicity was a statistically significant predictor here, while neither SES nor ethnicity were in Thompson's work with the i-Ready exam (H. A. Thompson, 2018), suggests that Equation 2 may not be generalizable to other schools and/or school systems.

Regardless, this model could be used as an early indicator to help identify which students may need extra help in order to be successful in Algebra 1. So as with the equation that was developed to predict SOL scores from fall Math Inventory scores, the unique demographics of the target school make it unlikely that the equation is generalizable to other schools, but the process used to develop it should be applicable in any school in the state of Virginia. This is consistent with the work done by Thompson and Shneyderman (Shneyderman, 2017; H. A. Thompson, 2018).

Based on the data, however, it is not possible to predict spring Math Inventory scores from fall Math Inventory scores. It is interesting to note the small change in mean scores (23 points) from fall to spring administrations of the same test, versus the very large change in standard deviation (44 points). Additionally, there were approximately twice as many data points for the fall administration of the Math Inventory ($n = 297$) as there were for the spring ($n = 165$). This suggests a possible problem with the data, specifically the assumptions that were made. At the beginning of the school year, students and staff are often full of energy and seem to be serious about learning, but as the year goes on they seem to run out of energy and look forward to the end of the year. Considering the change in atmosphere and attitude in a high school in May versus September, it is highly possible that either teachers were not consistent in administering the test or that there were significant differences in student adherence to directions and student motivation.

Problems with the spring Math Inventory data can also be found when looking at correlations with SOL scores and EOC grades. While there is a slight positive correlation between spring Math Inventory scores and SOL scores and an even smaller positive correlation between spring Math Inventory scores and EOC grades, it is much lower than one would expect. McDonald and Pang found a much higher correlation when comparing a spring administration of the Math Inventory and SAT scores (McDonald & Pang, 2021). While differences in teacher grading practices may contribute to the lower correlation to EOC grades, the low correlation to SOL scores, which are taken during the same time of year, may be explained by teacher and student motivation and attitude toward the Math Inventory test itself.

Approximately half of the students who took both the Math Inventory in both fall and spring did not improve their score from one administration to the next. This is independent of race, SES, or whether or not the student was successful in Algebra 1. While it is possible that students are truly not learning any math over the course of the year, it seems more likely that student and/or teacher motivation is affecting spring Math Inventory scores and thus the perceived gain in mathematical knowledge.

Generalizability and Limitations

As stated previously, the unique demographics of the target school mean that Equation 1 and Equation 2 are not likely to be generalizable to other schools. The process of using linear (or logistic) regression to predict performance on the SOL from the fall Math Inventory (and other factors) should be generalizable in schools and school systems throughout the state of Virginia. This is consistent with and extends the previous work of LaPlante, Shneyderman, Thompson, and Smith, who used benchmark tests that returned a Quantile score to predict performance on end of year high stakes assessments given by their respective states (LaPlante, 2018; Shneyderman, 2017; Smith, 2012; H. A. Thompson, 2018).

The information that can be gleaned from the collected data is limited by the inherent problems with the spring Math Inventory scores. While more research needs to be done into the causes of the problem, it is clear that a much smaller number of students took the Math Inventory in the spring than did in the fall. Also, in general the spring Math Inventory scores were not consistent with students' performance on the SOL and/or EOC grade. It is likely that some of the initial assumptions of this study, that students would do their best and teachers would be consistent in their administration of the Math Inventory test, may have been incorrect.

Recommendations

Predicting SOL scores from fall Math Inventory scores appears to be a useful application of the Quantile score, as it can help identify at risk students early in the year, thus maximizing the amount of extra help the student can receive. The predictions are only as good as the model however. Ideally, models could be created for each high school in the school system (as each school has its own demographics and average level of achievement), which could then be refined on an annual basis using cumulative data from all school years during which the Math Inventory was administered. Depending on the goals of the school, using a logistic regression to predict if a student will pass or not, as opposed to predicting an actual score, may be more desirable. Due to the wide range of demographics across the school system and the state, a system-wide model would probably be less useful than a school-by-school approach.

Motivation of both students and teachers have the potential to drastically affect student scores from one administration of the Math Inventory exam to the next. Further exploration of teacher and student motivation, possibly via surveys or interviews, may be an appropriate next step. If motivation and ability to follow directions is found to be low in the spring, it may be that administering the Math Inventory in the spring is not a productive use of time.

At the time that this study was being developed, it was not known that an additional administration of the Math Inventory would be required in the winter for students who had scored in the basic or below basic ranges during the fall administration of the Math Inventory. An analysis of this data could be helpful in developing a stronger model for identifying students who may be at risk of not passing the SOL. Additionally, if is found that there is a change in student and teacher motivation from one administration of the Math Inventory to the next, a comparison of the change in scores from fall to winter versus winter to spring could yield some insight as to when that might be occurring.

Conclusions

The Math Inventory is a tool that can be used to take a picture of a student's mathematical performance on any given day. Unfortunately, there are many factors that can affect a student's performance, only one of which is their actual mathematical ability (Popham, 1999). For as-yet-undetermined-reasons, these factors seem to come into play in the spring, as demonstrated by increased variability and decreased correlation of ability with scores, making a spring administration of the Math

Inventory an ineffective use of time. If there are reasons that the test must be given at that time, some work should go into investigating other factors such as the motivation of students and teachers. Without students being motivated and able to actually try and fully apply the mathematical ability that they possess, the data from the test is worthless.

Now that fall Math Inventory scores have been put into context, they are a piece of data that can be used to predict success. Educators need to be wary of the labels accompanying these scores, however, lest they become self-fulfilling prophecies.

REFERENCES

Bancroft, K. (2010). Implementing the mandate: The limitations of benchmark tests. *Educational Assessment, Evaluation & Accountability*, *22*(1), 53–72. https://doi.org/10.1007/s11092-010-9091-1

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, *41*(3), 287–302. https://doi.org/10.1080/0305764X.2011.607151

Brown, R. S., & Coughlin, E. (2007). *The Predictive Validity of Selected Benchmark Assessments Used in the Mid-Atlantic Region* (017; Issues & Answers). National Center for Education Evaluation and Regional Assistance.

Cizek, G. J. (2001). More Unintended Consequences of High-Stakes Testing. *Educational Measurement: Issues and Practice*, *20*(4), 19–27. https://doi.org/10.1111/j.1745-3992.2001.tb00072.x

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed). Pearson.

Di Martino, P., & Baccaglini-Frank, A. (2017). BEYOND PERFORMANCE RESULTS: ANALYZING THE INFORMATIONAL AND DEVELOPMENTAL POTENTIALS OF STANDARDIZED MATHEMATICS TESTS. *For the Learning of Mathematics*, *37*(3).

Erskine, J. L. (2014). It Changes How Teachers Teach: How Testing Is Corrupting Our Classrooms and Student Learning. *Multicultural Education*, *21*(2), 38–40.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Field, A., Miles, J., & Field, Z. (2013). *Discovering statistics using R*. Sage.

Gareis, C. R., McMillan, J. H., Smucker, A., & Huang, K. (2021). *MAP Growth Validation Study: An Evaluation of Selected MAP Growth Assessments to the Virginia Standards of Learning and an Exploration of the Utility of MAP Growth Reports for Determining Student Performance Relative to Grade Level*.

Ghaicha, A. (2016). Theoretical Framework for Educational Assessment: A Synoptic Review. *Journal of Education and Practice*, 20.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (039; Issues & Answers). National Center for Education Evaluation and Regional Assistance.

Herman, J. L., Osmundson, E., & Dietel, R. (2010). *Benchmark Assessment for Improved Learning*.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. Wiley.

Houghton Mifflin Harcourt. (2014). *Math Inventory Technical Guide*.

Isbell, D. R., & Kremmel, B. (2020). Test Review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Iverson, M. (2014). *Relationships of the Iowa assessment results to traditional and standards-based English and math end-of-course grades—ProQuest*.

https://www.proquest.com/openview/c4387a2efa05a4d412a2cb92e9684ee3/1?pq-origsite=gscholar&cbl=18750

Labeling theory. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Labeling_theory&oldid=1136325109

LaPlante, J. (2018). *Predictability Study of Istation ISIP (Math and Reading) and Ohio AIR (Math and English Language Arts) Tests for 3rd – 8th Grade Students in the Youngstown City School District*. Youngstown City Schools.

Lewis, B. L. (2013). *An analysis of mathematics achievement disparities between Black and white students and socioeconomically disadvantaged and advantaged students across content strands by elementary and middle school level in a diverse Virginia school district* [ProQuest Dissertations Publishing]. https://www.proquest.com/docview/1362253345/abstract/B5399DAADFF84DDEPQ/1

Martin, R. (2021). *Examining the Effects of Implementation of i-Ready to Fidelity on Reading and its Effects on Free or Reduced Lunch Students in Seven Middle Schools*. University of Central Florida.

Math Solutions. (2020). *Math Inventory Professional Learning Guide Grades K-Algebra II*. Houghton Mifflin Harcourt. http://samresources.scholastic.com/ResourceManager/previewresource.spr?_page=0&objectId=16642

McDonald, M., & Pang, Y. (2021). Comparing the Performance on the MI and SAT/PSAT for the Purpose of Monitoring Student Achievement. *International*

Journal of Information and Education Technology, *11*(12), 590–596. https://doi.org/10.18178/ijiet.2021.11.12.1569

McMillan, J. H., Myran, S., & Workman, D. (1999). *The Impact of Mandated Statewide Testing on Teachers' Classroom Assessment and Instructional Practices*. 19.

Mehrens, W. A. (1991, April 11). Using Performance Assessment for Accountability Purposes: Some Problems. *Annual Meeting of the American Educational Research Association*. https://files.eric.ed.gov/fulltext/ED333008.pdf

Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy / Pedagogický Casopis*, *3*(1), 82–100.

Mora, R. (2011). "School is So Boring": High-Stakes Testing and Boredom at an Urban Middle School. *Perspectives on Urban Education*, *9*(1), 1–9.

National Commission on Excellence in Education. (1983). *A Nation at Risk: The Imperative for Educational Reform*. http://edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf

Popham, W. J. (1999). Why Standardized Tests Don't Measure Educational Quality. *Educational Leadership*, *56*(6). https://www.ascd.org/el/articles/why-standardized-tests-dont-measure-educational-quality

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch model. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Rasch_model&oldid=1122640520

Released Tests & Item Sets (ALL SUBJECTS) | Virginia Department of Education. (n.d.). Retrieved April 29, 2023, from https://www.doe.virginia.gov/teaching-learning-assessment/student-assessment/sol-practice-items-all-subjects/released-tests-item-sets-all-subjects

Rist, Ray. C. (1970). Student Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education. *Harvard Educational Review*, *40*(3), 411–451.

Ruecker, T. (2013). High-stakes testing and Latina/o students: Creating a hierarchy of college readiness. *Journal of Hispanic Higher Education*, *12*(4), 303–320.

Schartung, J. (2015). *Examining the Influence of Negative Labeling on Educational Aspirations*.

Shneyderman, D. A. (2017). *Exploring the i-Ready Predictive Capability* (Research Brief 1608). Miami-Dade County Public Schools.

Smith, M. (2012). *A Study of the Predictive Validity of the STAR Math Test for the Algebra 1 End of Course Exam*. Northwest Missouri State University.

Starr, J. P., & Spellings, M. (2014). Examining High-Stakes Testing. *Education Next*, *14*(1), 70–77.

Summers, P. (2009, May). *Middle school grading practices and the ability to predict achievement on the Arkansas Benchmark Test—ProQuest*. https://www.proquest.com/openview/97ce5960886e5298254dba7ea5d67fa2/1?pq-origsite=gscholar&cbl=18750

Tape, T. G. (2015). *Interpreting diagnostic tests: The Area Under an ROC Curve*. https://darwin.unmc.edu/dxtests/roc3.htm

The Science Behind Quantile Measures. (n.d.). Quantile. Retrieved April 29, 2023, from https://www.quantiles.com/parents-students/understanding-quantile-measures/the-science-behind-quantile-measures/

Thompson, H. A. (2018). *The Relationship Between i-Ready Diagnostic and 10th Grade Students' High-Stakes Mathematics Test Scores*. Walden University.

Thompson, K. (2017, November 1). *Teacher Labelling and the self-fulfilling prophecy #class notes*. ReviseSociology. https://revisesociology.com/2017/11/01/labelling-self-fulfilling-prophecy-education/

Valenzuela, A. (2006). The Disappearance of High School English Language Learners from Texas High Schools. *Williams Review, 1*(1), 166–200.

Vela, A., Jones, D., Mundy, M.-A., & Isaacson, C. (2017). *Determining the effectiveness of bilingual programs on third grade state exam scores*. *33*.

Virginia SOL Assessment Program | Virginia Department of Education. (n.d.). Retrieved April 29, 2023, from https://www.doe.virginia.gov/teaching-learning-assessment/student-assessment/virginia-sol-assessment-program

Virginia Standards of Learning Assessments Technical Report 2021-22 Administration Cycle. (n.d.). Virginia Department of Education.

Williamson, G. L. (2016). Novel Interpretations of Academic Growth. *JAEPR*, *2*(2), 15–35.

Williamson, G. L., Sanford-Moore, E. E., & Bickel, L. (2016). *The Quantile® Framework for Mathematics Quantifies the Mathematics Ability Needed for College and Career Readiness* [MetaMetrics Research Brief].

Appendix A

A released version of the Algebra 1 SOL. This test is based on the 2009 version of the Virginia Standards of Learning and was released in 2015.

VIRGINIA STANDARDS OF LEARNING

TEST ITEM SET

Algebra I

2009 Mathematics Standards of Learning

Released Spring 2015

Property of the Virginia Department of Education

SAMPLE A

What is the solution to $3(2x-1)=3$?

- A $x=\frac{1}{3}$
- B $x=\frac{2}{3}$
- C $x=1$
- D $x=5$

Directions: Type your answer in the box. Your answer must be in the form of a fraction in simplest form. Use "/" for the fraction bar.

SAMPLE B

What is the value of $\frac{3}{x+2}$ when $x=4$?

Your answer must be in the form of a fraction in simplest form.

Which expression represents four less than half a number, n?

- **A** $4-\frac{1}{2}n$
- **B** $\frac{1}{2}n-4$
- **C** $\frac{1}{2}(4-n)$
- **D** $\frac{1}{2}(n-4)$

1 of 44

Which of the following binomials is a factor of x^2-x-6 ?

- **A** $x-1$
- **B** $x-2$
- **C** $x-3$
- **D** $x-6$

2 of 44

Directions: Click on all the correct answers.

Identify each expression that is in simplest radical form.

$x\sqrt{50y}$	$64\sqrt{x}$	$7x^2y\sqrt{2xy}$	$\sqrt{12x^3y^4}$

Which expression is equivalent to $\frac{1}{6}(30x - 24y) - \frac{1}{8}(32x - 16y)$?

- **A** $x - 6y$
- **B** $x - 2y$
- **C** $2x - 4y$
- **D** $9x - 6y$

Which is equivalent to $\sqrt[3]{48}$ in simplest form?

- **A** $2\sqrt[3]{6}$
- **B** $6\sqrt[3]{2}$
- **C** 16
- **D** 24

What is the value of $\sqrt{128}$ in simplest radical form?

- **A** $8\sqrt{2}$
- **B** $64\sqrt{2}$
- **C** $4\sqrt{8}$
- **D** $16\sqrt{8}$

Which polynomial is equivalent to this expression if $n \neq {}^{-}1$?

$$\frac{3+n-2n^2}{1+n}$$

- **A** $2n-3$
- **B** $3-2n$
- **C** $3-2n^2$
- **D** $4-2n^2$

Which is a factor of $2n^2-5n-42$?

- **A** $2n-7$
- **B** $2n-6$
- **C** $n-7$
- **D** $n-6$

Which of the following is equivalent to $\dfrac{a^{12}b^2}{a^3b^6}$?

- **A** $\dfrac{a^9}{b^4}$
- **B** $\dfrac{b^4}{a^9}$
- **C** $\dfrac{a^4}{b^3}$
- **D** a^9b^4

9 of 44

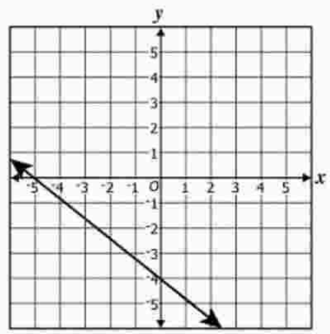
What is the value of this expression when $n = {}^-15$?

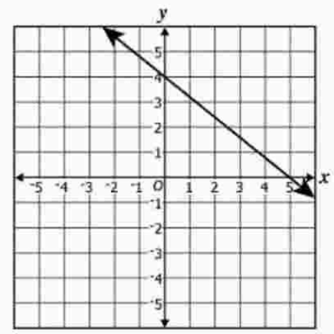
$$^-2|n+6|$$

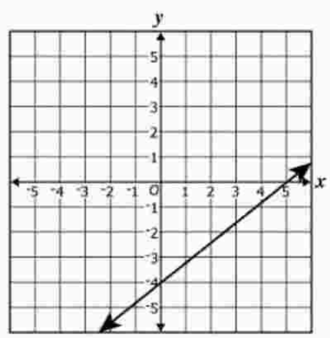
- **A** $^-42$
- **B** $^-18$
- **C** 18
- **D** 42

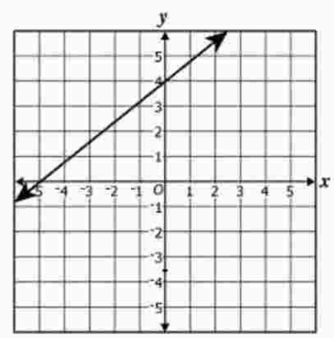
10 of 44

Which graph best represents the equation $4x + 5y = {}^{-}20$?

A 

C 

B 

D 

A formula to find the angle measures of an isosceles triangle is shown.

$$180 = 2x + y$$

Which equation can be used to find x ?

- **A** $x = \dfrac{180 - y}{2}$
- **B** $x = \dfrac{180 + y}{2}$
- **C** $x = 90 - y$
- **D** $x = 90 + y$

Which equation represents the line that passes through the points $(^-4, 4)$ and $(8, ^-2)$?

- **A** $y = {}^-2x + 14$
- **B** $y = {}^-2x - 4$
- **C** $y = \frac{^-1}{2}x + 2$
- **D** $y = \frac{^-1}{2}x - 2$

13 of 44

For which system of inequalities is $(^-3, 1)$ a solution?

- **A** $\begin{cases} x + y < {}^-2 \\ 2x - 3y < {}^-9 \end{cases}$
- **B** $\begin{cases} x + y < {}^-2 \\ 2x - 3y \le {}^-9 \end{cases}$
- **C** $\begin{cases} x + y \le {}^-2 \\ 2x - 3y < {}^-9 \end{cases}$
- **D** $\begin{cases} x + y \le {}^-2 \\ 2x - 3y \le {}^-9 \end{cases}$

14 of 44

What is the solution to this system of equations?

$$\begin{cases} 2x + 4y = 22 \\ 7x + y = 12 \end{cases}$$

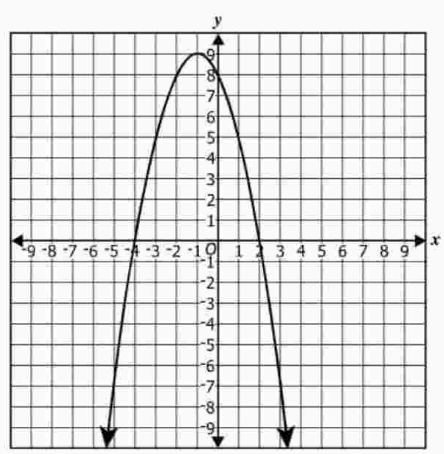
- **A** (3, 4)
- **B** (2, ⁻2)
- **C** (1, 5)
- **D** (⁻1, 6)

15 of 44

Directions: Click on the grid to plot each of the solutions. You must plot all solutions.

The graph of $y = {}^{-}x^2 - 2x + 8$ is shown.

On the grid, identify each of the solutions to ${}^{-}x^2 - 2x + 8 = 0$.



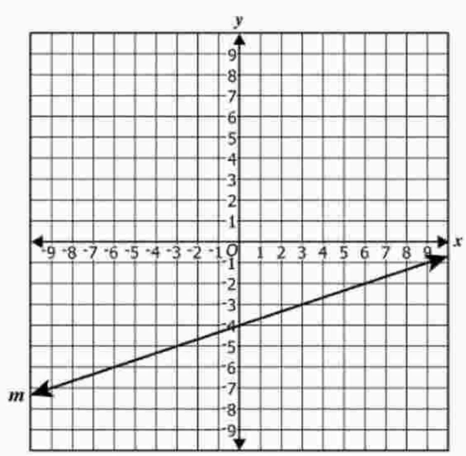
16 of 44

What value of x makes this equation true?

$$3x - 20 = {}^{-}2x$$

- **A** $^{-}20$
- **B** $^{-}4$
- **C** 4
- **D** 20

Which equation best represents line m ?



- **A** $y = {}^{-}3x - 4$
- **B** $y = \frac{^{-}1}{3}x - 4$
- **C** $y = \frac{1}{3}x - 4$
- **D** $y = 3x - 4$

Directions: Click and drag the answers to the correct boxes.

Christopher incorrectly solved an inequality as shown.

Step 1: $^-4(x-7)+1 \le {}^-3$
Step 2: $^-4(x-7) \le {}^-4$
Step 3: $^-4x+28 \le {}^-4$
Step 4: $^-4x \le {}^-32$
Step 5: $x \le 8$

Between which two consecutive steps did Christopher make a mistake?

[ ] and [ ]

Step 1
Step 2
Step 3
Step 4
Step 5

Directions: Type your answer in the box.

Solve for n:

$$\frac{3n-7}{6} = \frac{2n+5}{3}$$

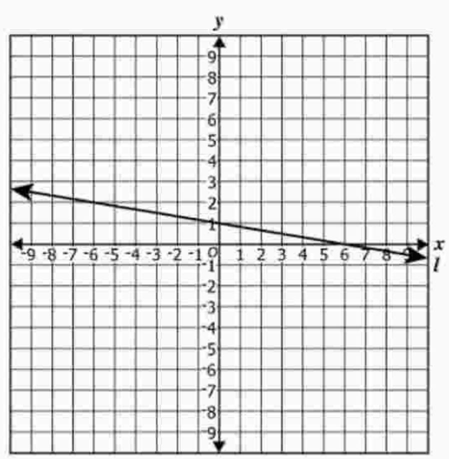
$n =$ [ ]

What values of x are solutions of $3x^2 + 11x = 20$?

- A $-\frac{4}{3}$ and 5
- B $-\frac{5}{3}$ and 4
- C -4 and $\frac{5}{3}$
- D -5 and $\frac{4}{3}$

21 of 44

The graph of line l is shown.



Which number is closest in value to the slope of line l ?

- A -6
- B $-\frac{1}{6}$
- C $\frac{1}{6}$
- D 6

22 of 44

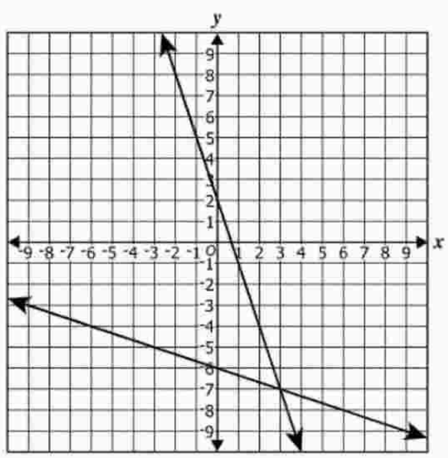
Directions: Type your answer in the box.

Based on the transitive property, complete this statement.

If $2(y-3) \geq 3x-4$ and $3x-4 \geq 6-y$, then $2(y-3) \geq$ __?__

This system of linear equations is graphed as shown.

$$\begin{cases} 3x + y = 2 \\ x + 3y = {}^{-}18 \end{cases}$$



What is the solution to this system of equations?

- **A** (2, ⁻6)
- **B** (3, ⁻7)
- **C** (⁻6, 2)
- **D** (⁻7, 3)

Renee is going bowling.

- **The cost per game is $2.50.**
- **Renee will need to rent a pair of bowling shoes for $1.50.**
- **She can spend up to $16.00 to bowl and rent a pair of shoes.**

What is the maximum number of games that Renee can bowl?

○ **A** 4

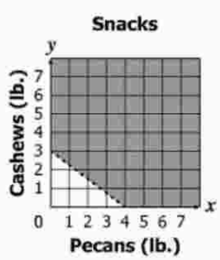
○ **B** 5

○ **C** 6

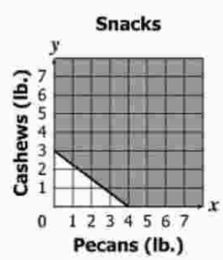
○ **D** 9

Malik can spend no more than $24 to buy pecans and cashews. He will pay $6 per pound for pecans and $8 per pound for cashews. Which graph best represents the number of pounds of pecans and the number of pounds of cashews Malik can buy?

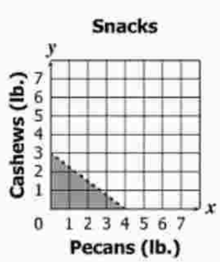
○ **A**



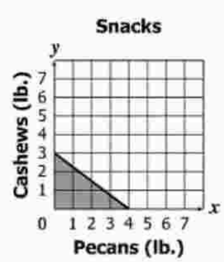
○ **C**



○ **B**

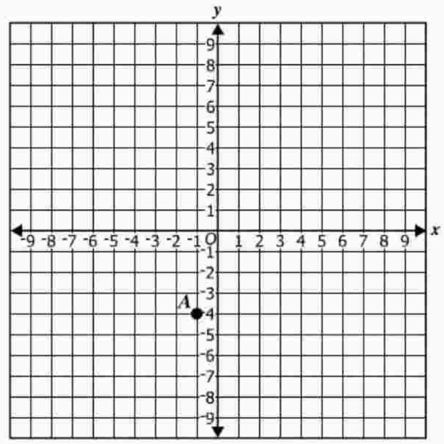


○ **D**



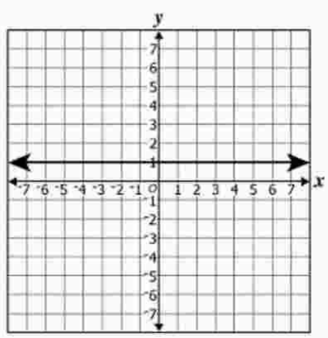
Directions: Click on the grid to plot two points. The coordinates of the points must be integers.

Point A is an element of a direct variation. Plot two points, other than A, that are elements of this direct variation. The coordinates of the points must be integers.

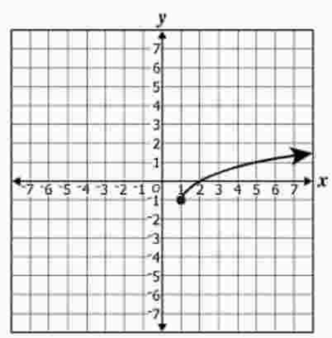


Which graph has exactly one x-intercept and one y-intercept?

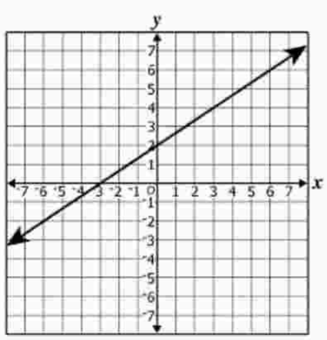
A



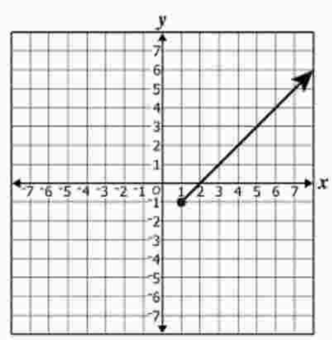
C



B



D



Which equation best represents this data set?

$$\{(^-4, ^-4.8), (^-3, ^-8.2), (^-2, ^-9.1), (^-1, ^-8.1), (0, ^-4.7), (1, 0.3)\}$$

- **A** $y = 1.1x^2 + 4.2x + 4.9$
- **B** $y = 1.1x^2 + 4.2x - 4.9$
- **C** $y = 1.1x - 4.2$
- **D** $y = 1.1x + 4.2$

A relationship between x and y is shown in this table.

x	y
0	1
1	2
2	5
3	10

Which equation represents this relationship?

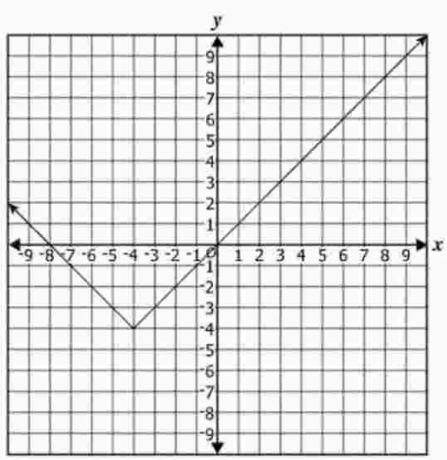
- **A** $y = 2x + 1$
- **B** $y = 5x - 5$
- **C** $y = x^2 + 1$
- **D** $y = (x + 1)^2$

Ms. Scott will pay $2,000 to have her house painted. The amount each painter earns, A, varies inversely for the number of painters, n, that will paint the house. Which equation best represents this situation?

- **A** $A = 2{,}000 + n$
- **B** $2{,}000 = A + n$
- **C** $A = 2{,}000n$
- **D** $2{,}000 = An$

31 of 44

The following graph shows a relation.



Which of the following best describes the range of this relation?

- **A** All real numbers
- **B** All real numbers between -10 and 10
- **C** All real numbers less than or equal to -4
- **D** All real numbers greater than or equal to -4

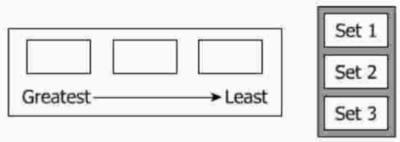
32 of 44

Directions: Click and drag the answers to the correct boxes.

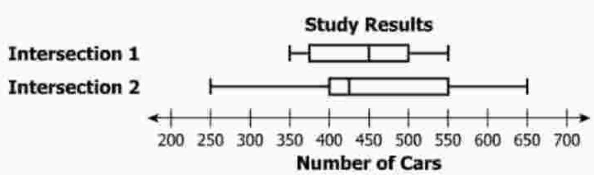
Each of these data sets has a mean of 20.

Set 1: $\{18, 19, 20, 21, 22\}$
Set 2: $\{20, 20, 20, 20, 20\}$
Set 3: $\{16, 18, 20, 21, 25\}$

Order the sets from greatest standard deviation to least standard deviation.



A study was conducted to determine the number of cars that passed through two intersections each day for 20 days. The results are summarized in these box-and-whisker plots.



Which statement is best supported by these data?

- **A** The range of the data for Intersection 2 is twice the range of the data for Intersection 1.
- **B** The lower quartile for Intersection 1 is greater than the lower quartile for Intersection 2.
- **C** The interquartile range for Intersection 1 is the same as the interquartile range for Intersection 2.
- **D** The total number of vehicles that passed through Intersection 2 is greater than the total number of vehicles that passed through Intersection 1.

Which of these functions has exactly two different zeros?

- A $f(x) = \frac{1}{10}x + 4$
- B $g(x) = \frac{3x - 10}{3}$
- C $h(x) = x^2 - 4x + 4$
- D $k(x) = x^2 + 11x + 24$

35 of 44

In which table does y vary directly with x ?

A

x	y
1	3
2	3
3	3

C

x	y
1	5
2	7
3	9

B

x	y
1	4
2	8
3	12

D

x	y
1	9
2	7
3	5

36 of 44

Which equation could represent a graph with x-intercepts of (4, 0) and ($^-7$, 0) ?

- **A** $y = x^2 + 3x - 28$
- **B** $y = x^2 - 3x - 28$
- **C** $y = x^2 + 3x + 28$
- **D** $y = x^2 - 3x + 28$

37 of 44

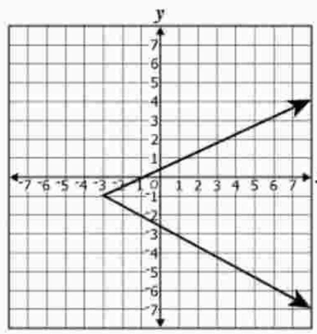
Which number is a zero of the function h ?

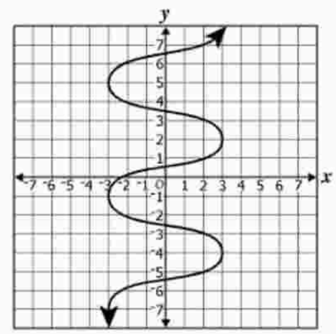
$$h(x) = x^2 + 3x - 18$$

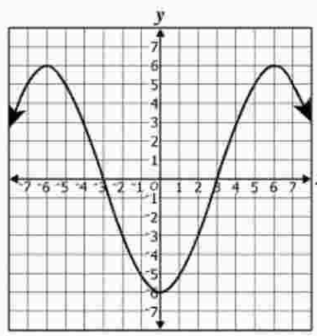
- **A** $^-6$
- **B** $^-3$
- **C** 0
- **D** 6

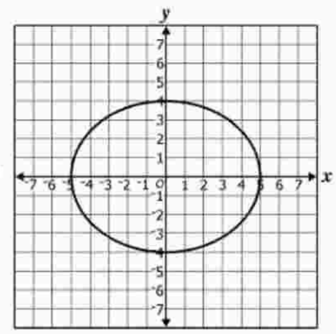
38 of 44

Which of the following graphs appears to be a function?

○ **A** 

○ **C** 

○ **B** 

○ **D** 

If $f(x) = (x-3)^2 + 1$, what is $f(6)$?

○ **A** $^-2$

○ **B** 7

○ **C** 10

○ **D** 16

Which number is NOT an element in the domain of this relation?

$$\{(^-2, 3), (0, 4), (1, 1), (6, 0)\}$$

- **A** 4
- **B** 1
- **C** 0
- **D** $^-2$

41 of 44

$$\{(^-5, 9), (2, 31), (9, 143), (11, 151), (0, 42), (5, 97)\}$$

Using the equation of the line of best fit, which number is the best prediction of the output when the input is 13 ?

- **A** 127
- **B** 159
- **C** 170
- **D** 178

42 of 44

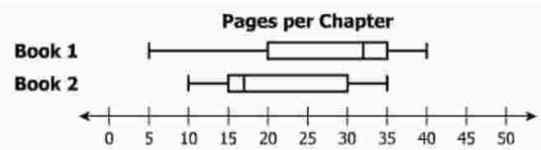
A data set has a mean of 720 and a standard deviation of 6. Which is closest to the z-score for an element of this data set with a value of 709 ?

- **A** 11.00
- **B** 1.83
- **C** $^{-}11.00$
- **D** $^{-}1.83$

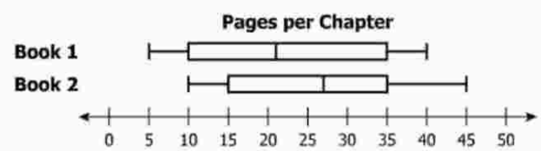
43 of 44

Ramon drew box-and-whisker plots to summarize the number of pages in each chapter of two books. The values of the interquartile ranges for these box-and-whisker plots are the same. Which box-and-whisker plots could represent these data?

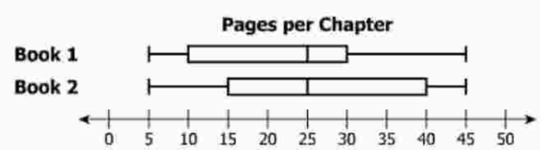
A



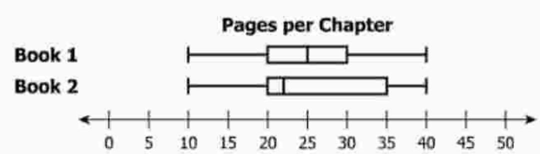
B



C

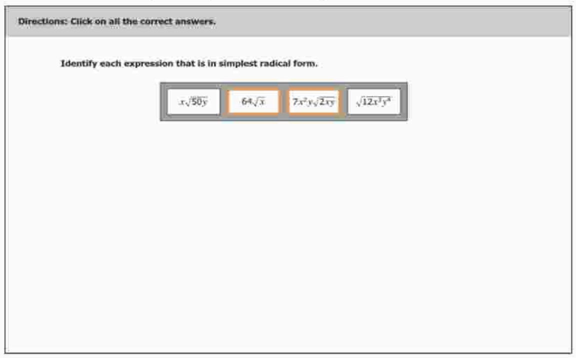


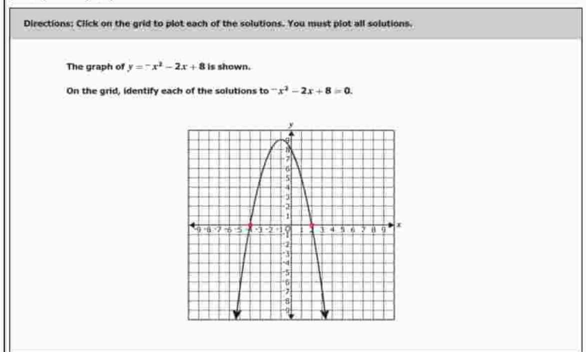
D

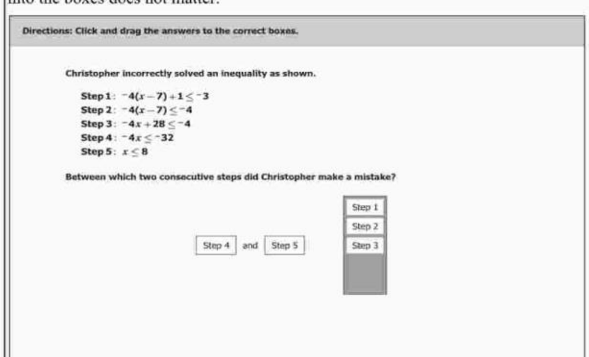


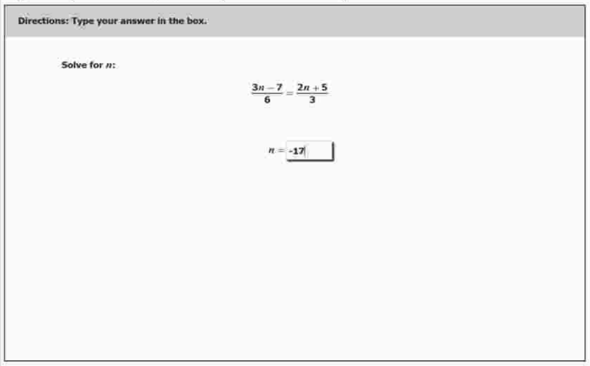
44 of 44

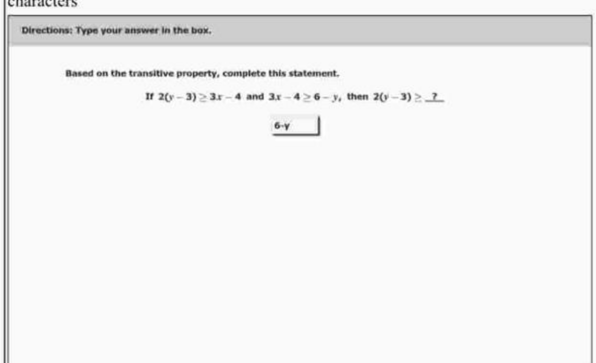
Algebra I
Released Test Item Set Spring 2015
Answer Key

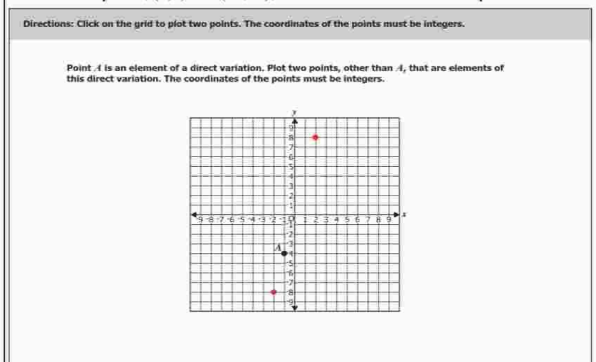
Sequence Number	Item Type: Multiple Choice (MC) or Technology-Enhanced Item (TEI)	Correct Answer	Reporting Category	Reporting Category Description
1	MC	B	001	Expressions and Operations
2	MC	C	001	Expressions and Operations
3	TEI	$64\sqrt{x}$ (second from left) & $7x^2y\sqrt{2xy}$ (third from left) Both of these answers, and only these answers, must be selected. 	001	Expressions and Operations
4	MC	B	001	Expressions and Operations
5	MC	A	001	Expressions and Operations
6	MC	A	001	Expressions and Operations
7	MC	B	001	Expressions and Operations
8	MC	D	001	Expressions and Operations
9	MC	A	001	Expressions and Operations
10	MC	B	001	Expressions and Operations
11	MC	A	002	Equations and Inequalities
12	MC	A	002	Equations and Inequalities
13	MC	C	002	Equations and Inequalities
14	MC	D	002	Equations and Inequalities
15	MC	C	002	Equations and Inequalities

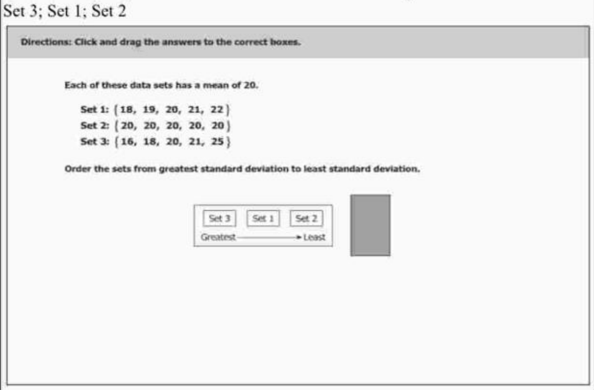
Sequence Number	Item Type: Multiple Choice (MC) or Technology-Enhanced Item (TEI)	Correct Answer	Reporting Category	Reporting Category Description
16	TEI	Both of these points, and only these points, must be plotted on the coordinate plane: (–4,0) and (2,0).  	002	Equations and Inequalities
17	MC	C	002	Equations and Inequalities
18	MC	C	002	Equations and Inequalities

Sequence Number	Item Type: Multiple Choice (MC) or Technology-Enhanced Item (TEI)	Correct Answer	Reporting Category	Reporting Category Description
19	TEI	Step 4 and Step 5 must be placed into the boxes. The order in which they are placed into the boxes does not matter.  	002	Equations and Inequalities

Sequence Number	Item Type: Multiple Choice (MC) or Technology-Enhanced Item (TEI)	Correct Answer	Reporting Category	Reporting Category Description
20	TEI	Typed response: -17 (and all equivalent answers)  Directions: Type your answer in the box. Solve for n: $\frac{3n-7}{6}=\frac{2n+5}{3}$ $n=$ -17	002	Equations and Inequalities
21	MC	D	002	Equations and Inequalities
22	MC	B	002	Equations and Inequalities

Sequence Number	Item Type: Multiple Choice (MC) or Technology-Enhanced Item (TEI)	Correct Answer	Reporting Category	Reporting Category Description
23	TEI	Typed response: 6-y OR any equivalent expression that does not exceed six characters  Directions: Type your answer in the box. Based on the transitive property, complete this statement. If $2(y-3)\geq 3x-4$ and $3x-4\geq 6-y$, then $2(y-3)\geq$? 6-y	002	Equations and Inequalities
24	MC	B	002	Equations and Inequalities
25	MC	B	002	Equations and Inequalities
26	MC	D	002	Equations and Inequalities

Sequence Number	Item Type: Multiple Choice (MC) or Technology-Enhanced Item (TEI)	Correct Answer	Reporting Category	Reporting Category Description
27	TEI	Any <u>TWO</u> of these points must be plotted on the coordinate plane: (−2,−8), (0,0), (1,4), or (2,8) Two of these points, (2,8) and (−2,−8), are shown on the coordinate plane below. 	003	Functions and Statistics
28	MC	B	003	Functions and Statistics
29	MC	B	003	Functions and Statistics
30	MC	C	003	Functions and Statistics
31	MC	D	003	Functions and Statistics
32	MC	D	003	Functions and Statistics

Sequence Number	Item Type: Multiple Choice (MC) or Technology-Enhanced Item (TEI)	Correct Answer	Reporting Category	Reporting Category Description
33	TEI	Answers must be placed in the correct order from left to right: Set 3; Set 1; Set 2 	003	Functions and Statistics
34	MC	A	003	Functions and Statistics
35	MC	D	003	Functions and Statistics
36	MC	B	003	Functions and Statistics
37	MC	A	003	Functions and Statistics
38	MC	A	003	Functions and Statistics
39	MC	B	003	Functions and Statistics
40	MC	C	003	Functions and Statistics

Sequence Number	Item Type: Multiple Choice (MC) or Technology-Enhanced Item (TEI)	Correct Answer	Reporting Category	Reporting Category Description
41	MC	A	003	Functions and Statistics
42	MC	C	003	Functions and Statistics
43	MC	D	003	Functions and Statistics
44	MC	A	003	Functions and Statistics

868948 1 2 3 4 5 A B C D E Printed in the USA ISD6978

BIBLIOGRAPHY

Rebecca H. Schulz

Candidate for the Degree of

Master of Science Mathematics

Thesis: MATH INVENTORY AND SUCCESS IN ALGEBRA 1

Major Field: Mathematics

Biographical: Rebecca worked as an electrical engineer for an architecture company in a former life. She currently teaches high school math in northern Virginia.

Personal Data: Rebecca is a mother of three children.

Education: Bachelor of Science Electrical Engineering from Georgia Institute of Technology

Completed the requirements for the Master of Science in Mathematics, Portsmouth, Ohio on June 20, 2024.

6/22/2024

ADVISER'S APPROVAL: Douglas Darbro