Spring 1-1-2017

The Relationship between Testing Frequency and Student Achievement in Eighth-Grade Mathematics: An International Comparative Study Based on TIMSS 2011

Ufuk Guven

THE RELATIONSHIP BETWEEN TESTING FREQUENCY AND STUDENT ACHIEVEMENT IN EIGHTH-GRADE MATHEMATICS: AN INTERNATIONAL COMPARATIVE STUDY BASED ON TIMSS 2011

A Dissertation

Submitted to the School of Education

Duquesne University

In partial fulfillment of the requirements for

the degree of Doctor of Education

By

Ufuk Güven

May 2017

THE RELATIONSHIP BETWEEN TESTING FREQUENCY AND STUDENT ACHIEVEMENT IN EIGHTH-GRADE MATHEMATICS: AN INTERNATIONAL COMPARATIVE STUDY BASED ON TIMSS 2011

By

Ufuk Güven

Approved March 23, 2017

Joseph Kush Professor School of Education, Instruction and Leadership (Committee Chair)	Marie Martin Adjunct Professor School of Education, Instructional Technology and Leadership (Committee Member)
Gibbs Y. Kanyongo, Ph.D. Associate Professor School of Education, Educational Foundations and Leadership Duquesne University (Committee Member)	Rachel Ayieko Assistant Professor School of Education, Instruction and Leadership Duquesne University (Committee Member)

ABSTRACT

THE RELATIONSHIP BETWEEN TESTING FREQUENCY AND STUDENT ACHIEVEMENT IN EIGHTH-GRADE MATHEMATICS: AN INTERNATIONAL COMPARATIVE STUDY BASED ON TIMSS 2011

By

Ufuk Güven

May 2017

Dissertation supervised by Dr. Joseph Kush

The purpose of this study was to examine the relationship between quiz frequency and student achievement in eighth-grade mathematics as measured by TIMSS. The more specific goal of the study was determining the best quiz frequency (daily, weekly, monthly, no quizzes) and student achievement relationship for an eighth-grade mathematics course. The study investigated the above-mentioned relationship in all of the eighth-grade of participant countries combined, as well as in four specific countries: Korea, Singapore, Turkey, and the United States. Another goal of the study was to determine high performing and low performing countries' quizzing practices, and to determine the best relationship of quiz frequency and student achievement in these countries. The study obtained data from the TIMSS 2011 exam and from student, teacher, and school questionnaires. In addition to quiz practices, students' and schools' SES data

were also used in this study as control variables. Quiz frequency data (independent variable) were retrieved from teacher questionnaires, socioeconomic status (SES) data (control variables) were retrieved from student and school questionnaires, and student achievement data were retrieved from the TIMSS 2011 exam. Several multiple linear regressions were performed to determine whether quiz frequency is a significant predictor of student achievement in all countries combined, as well as in individual countries. Regression results indicated that quizzing frequency is not a significant contributor to student achievement in eighth-grade mathematics, either in all countries combined or in individual countries after controlling for SES variables. Furthermore, regression results indicated that weekly quizzes had the best relationship in all countries, monthly quizzes in the top two performing countries (Korea and Singapore), and daily quizzes in Turkey and the United States. Results also indicated that almost all teachers use quizzes. Moreover, the study also found that SES status is a significant contributor to student achievement, and that student achievement significantly and constantly increased as student SES status improve.

DEDICATION

I dedicate this dissertation to my wife, Emine, and to my kids, Ada and Fatih. I hope this research will make up for all of the time that I was unable to spend with you while working on it. I also dedicate this dissertation to my family, patiently waiting for me to complete my studies and get back to them.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my advisor and committee members for their patience, guidance, and encouragement.

Dr. Marie Martin worked very closely with me during my dissertation. She spent most of her time giving critical feedback about my work and making time for Skype meetings to discuss my progress and to suggest articles that helped shape my dissertation. I improved my study tremendously after she stepped in. I cannot thank her enough for everything she did for me over the course of my dissertation.

I would also like to thank Dr. Gibbs Kanyongo for his support and guidance in statistical analyses, as well as for the courses that I took from him. I felt confident in analyzing such complex data as my study required because of his teaching and projects that I undertook in his courses. I am so grateful to have had his continued support as a member of my dissertation committee.

Dr. Rachel Ayieko also greatly supported me in shaping my dissertation by guiding me through the analysis of the complex data using the IDB Analyzer tool. I also benefitted considerably from her sharing of her expertise and experiences in TIMSS.

And finally, I cannot thank Dr. Joseph Kush enough for his support, from day one to date. He was always there whenever I needed his help and he never hesitated to help me as my advisor and committee chair. He always worked closely with me, even when he was on the other side of the planet, to ensure that I got enough support to make it through the dissertation storm. He guided me to study the TIMSS exams which resulted in this product of which I am so proud.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter I: Introduction

Formative assessment is one of the essential parts of every instructional method. Assessment informs students about their own learning, and informs teachers of how students are doing, what works, and where students need help (NCTE, 2013). It also informs interested parties of whether goals and standards of education are being met (Garrison & Ehringhaus, 2016). A range of assessment methods can be used at different frequencies for formative assessment. The literature confirms that quizzes and tests are routinely used in formative assessment in order to assess and promote students' learning (CERI, 2008; Roediger & Karpicke, 2008; Roediger, Putnam, & Smith, 2011). There is, however, a dearth of studies pertaining to the optimal frequency for quizzing as an assessment and learning tool. This absence in scholarship points to a need for further research to help fill this gap.

Frequent Testing

The term 'frequent testing' or 'frequent quizzing' refers to daily or weekly tests and/or quizzes that are conducted for formative evaluation purposes, rather than to summative exams. While some researchers (Gholami & Moghaddam, 2013) have used the term frequent quizzing and some have used the term frequent testing (Shirvani, 2009), both terms are used to describe daily, weekly, or bi-weekly formative assessments. Therefore, the terms tests and quizzes will be used interchangeably in this study, and frequent testing will be used to refer to frequent quizzing through tests and quizzes.

Frequent testing has been defined in very different manners by previous researchers. Some studies (Dineen, Taylor, & Stephens, 1989; Shirvani, 2009) defined frequent testing as being made up of daily quizzes and compared daily quizzes to weekly quizzes; however, other researchers (Kika, Mclaughlin, & Dixon, 1992; Zgraggen, 2009) defined weekly quizzes as including frequent testing and compared weekly quizzes to bi-weekly quizzes. Gholami and Moghaddam (2013) compared weekly quizzes with no quiz group in

investigating the effect of frequent testing. An international assessment organization, Test Enhanced Learning in Classrooms (TELC) (2011), model suggests that teachers implement quizzes frequently, but in particular during the last ten minutes of each class. Another study was conducted to compare the effect of daily quizzes versus weekly quizzes on student mathematics scores and homework grades among high school students, and found that daily quizzes significantly improved students' final exam mathematics scores and homework grades in comparison to weekly quizzes (Shirvani, 2009). Trends in Mathematics and Science Studies (TIMSS) uses the following frequencies to distinguish testing frequencies: "almost everyday", "about half of the lessons", "some lessons", and "never". Frequent testing can therefore refer to a range of different intervals for testing.

However, other studies would suggest that frequent testing, however it is defined, does not always produce better learning or improved scores. A study conducted by Zgraggen (2009) compared weekly quizzes to bi-weekly quizzes in high school math courses and concluded that bi-weekly quizzes are more effective than weekly quizzes in improving students' mathematics scores. Another study conducted by Dineen, Taylor, and Stephens (1989) found no significant difference between using daily quizzes or weekly quizzes. Opponents of frequent testing maintain that frequent testing reduces instruction time, increases student anxiety, stresses test scores rather than learning, and forces teachers to teach for tests and students to study for only what will be on their tests (Gholami & Maghaddam, 2013). Therefore, it is suggested that frequent testing may reduce student learning and achievement. This argument is strongly refuted by Roediger et al. (2011a). They argue that even though frequent testing takes some instructional time, it is not conclusive to say that the activities which will replace quizzes are more effective. Roediger et al. (2011a) also maintain that providing feedback after quizzing avoids the problems that may arise due to frequent testing. The argument that frequent testing increases test anxiety is also refuted by some

researchers. Shirvani (2009) and Agarwal, Bain, and Chamberlain (2012) found that frequent testing as a retrieval practice rather than high stakes testing helps students to reduce their test taking anxiety.

Role of Quizzing in Retrieval Practices

Even though there is a lack of consensus about the value of frequent testing, the literature would suggest that it has an important role to play in formative assessment in the context of retrieval practices for retention and learning (Brame & Biel, 2015; CERI, 2008; Roediger & Karpicke, 2008; Roediger et al., 2011a). Retrieval practice is defined as accessing and retrieving newly acquired information from the memory whenever needed (Nunes & Karpicke, 2015; Roediger in an interview with Francisco, 2014). Quizzing or testing is a widely-used retrieval practice strategy to facilitate learning (Agarwal, Bain, & Chamberlain, 2012; Brame & Biel, 2015; Roediger et al., 2011a). A number of studies have demonstrated that the use of quizzes as a retrieval practice strategy can increase academic achievement (Roediger, Agarwal, McDaniel, & Mcdermott, 2011b; Salas-Morera, Arauzo-Azofra, & García-Hernández, 2012). It can also enhance student learning (Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel, & Metcalfe, 2007; Roediger et al., 2011a), improve long-term retention (Agarwal et al., 2012; Roediger & Butler, 2011), and promote deep learning as opposed to just "rote learning" (Francisco, 2014; Karpicke & Grimaldi, 2012; Roediger et al., 2011a). Moreover, testing as a retrieval practice can enable the transfer of information to new learning (McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Pastötter & Bauml, 2014), and quizzing reduces test-taking anxiety (Agarwal, D'Antonio, Roediger, McDermott, & McDaniel, 2014; Shirvani, 2009) and mitigates the amount of forgotten information (Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel, & Metcalfe, 2007; Roediger & Karpicke, 2008). Brame and Biel (2015) found that frequent quizzing is an effective strategy for retrieval practice, with the caveat that it should be low stakes. Another

positive indirect effect of testing is motivating students to study regularly and keep up with subject (Suda, Bell, & Franks, 2011). Some studies demonstrate that students also perceive quizzing as a very helpful tool (Karpicke & Grimaldi, 2012; Salas-Morera et al., 2012). Moreover, teachers also value testing as a retrieval practice for learning (Agarwal, Bain, & Chamberlain, 2012) and they use testing to encourage students to study periodically (Suda, Bell, & Franks, 2011).

Quizzing and Learning Theories

Quizzing finds a place for itself in various learning theories. Quizzing as a retrieval practice is an active learning strategy (Francisco, 2014; Karpicke & Grimaldi, 2012). Quizzing can also be used to stimulate prior knowledge, as described in the third event of Gagne's nine events of instruction (CITT, 2016; Pelech, 2016). In Bloom's Revised Taxonomy, the category "Knowledge" was renamed "Remember", and described as retrieving relevant information from long-term memory. The cognitive processes involved in this category are recalling and recognizing (Krathwohl, 2002). This matches the function of quizzing in retrieval practice (Armstrong, 2016; Brame & Biel, 2015). Retrieval-based learning is an excellent example of how findings from cognitive science can lead educational activities (Nunes & Karpicke, 2015). Studies that combined cognitive science and education have shown that when students take a quiz or test, cognitive action occurs that enables learning, which is also called the "testing effect" (Pelech, 2015; Roediger & Karpicke, 2006).

Quizzing as a retrieval practice is also related to the Constructivism learning theory, according to which learning occurs as learners are actively involved in a process of meaning and knowledge construction, rather than in passively receiving information (Gray, 1997). Quizzing requires the activation of prior knowledge to construct one's own learning. Students perceive this process as aligning with the constructivist principle of making prior knowledge the starting point of new learning (Pelech, 2015). Pastötter and Bauml (2014) refer to this

phenomenon as the forward – as opposed to the backward – effect of testing. Lastly, Brame and Biel (2015) state that, ideally, quizzing as a retrieval practice should provide feedback. This brings the student to a higher level of understanding through interacting with the teacher (Hein, 1991). Quizzing with feedback is predicated on Vygotsky's Zone of Proximal Development theory, where he speaks of the distance between the actual developmental level of the unaided student and the level of potential development which is attainable with the help of an adult (Vygotsky, 1978). Quizzes can, therefore, be used as a tool to aid students in achieving their potential through an effective feedback mechanism.

Mathematics and Frequent Testing

In the 21st century, all students need to have higher order mathematics and technological thinking skills in order to be productive in their careers and in their personal lives. Mathematics is a pervasive subject in our daily life; from managing money to cooking, determining lengths of time, putting things together, and calculating quantities of purchases, so many daily tasks require some level of mathematical computations. Students should also have a good understanding of mathematics in order to understand daily news and make sense of world events that are often presented through statistics, increases, and decreases, if they are to be effective citizens (Mullis, Martin, Foy, & Aurora, 2012). In the 21st century workplace, companies and institutions are looking for a new type of employee that is tech savvy and experienced in coding, data structures, mathematics, and augmented reality (Amador & Soule, 2015). This has led to a global focus on STEM (science, technology, engineering, and mathematics) courses (ICEF 2014). The U.S. Department of Education has launched projects toward the end of having STEM proficient students and teachers (U.S. Department of Education, 2016). The number of jobs that require students/candidates to use high level mathematics and mathematical thinking has increased with advancements in technology and with modern management methods (Mullis, Martin, Ruddock, O'Sullivan, &

Preuschoff, 2009). Therefore, STEM courses' instructors should design and implement special strategies to increase students' technological mathematical understandings (Amador & Soule, 2015).

Quizzes can be one of the special strategies implemented to increase students' mathematical skills. Roediger et al. (2011a) summarize benefits of frequent quizzing and testing which also apply to mathematics, as testing shows the gap in student knowledge, organizes learning materials, provides feedback to teachers and students, helps transfer knowledge to new concepts, and encourage students to study. The literature would indicate that using quizzes in mathematics courses has been shown to be especially effective because mathematic concepts rely on each other and require a firm foundation for the processing of upcoming concepts (Lowe, 2015). Quizzing as a retrieval practice is an active learning strategy (Francisco, 2014; Karpicke & Grimaldi, 2012) and it activates prior knowledge that helps learners to construct their own learning (Pelech, 2015). Teachers can also use quizzes to identify gaps and weaknesses so that they can plan additional teaching activities, in line with constructivist principles, to make sure students have a firm base for upcoming concepts. Frequent testing also encourages students to attend more classes (Zarei, 2008), which is very important for math courses in order to ensure that students do not have gaps in their knowledge. This practice enables students to get ready for the next topics and transfer their knowledge (McDaniel et al., 2013; Pastötter & Bauml, 2014) to new concepts.

This study will focus on mathematics. Based on data from the Trends in International Mathematics and Science Study (TIMSS), an international math and science assessment program, this study will explore whether there is a relationship between testing frequency and student achievement in mathematics.

TIMSS Overview

There are several international assessment projects, such as the Programme for International Student Assessment (PISA) and Progress in International Reading and Literacy Skills (PIRLS), that assess students worldwide in different subjects and at different grades. These projects are repeated at different time intervals to enable participant education systems to monitor their trends and progress. TIMSS is one of the several international assessment projects that assesses student achievement and collects extensive information from teachers, students, schools, and other educators which potentially effects student achievement in mathematics and science.

TIMSS has been collecting data, usually every four years, since 1995, from 4th and eighth-grade students in approximately 50 countries. The data for this dissertation study will come from TIMSS 2011 assessment scores and questionnaires. TIMSS measures overall student achievement based on four international benchmarks (advanced, high, medium, and low) across major content domains (e.g., number, algebra, and geometry in mathematics, and earth science, biology, and chemistry in science), and by knowing, applying, and reasoning of cognitive domains (IEA, 2016).

Purpose of the Study

The literature reveals a gap in the examination of factors that affect student learning in mathematics – the quiz frequency factor. The first purpose of this study was to help bridge that gap by examining the relationship between testing frequencies and students' achievement in mathematics through TIMSS data. A second purpose of this study was examining the relationship between testing frequency and students' mathematics achievement in high performing and low performing countries, in order to see if the relationship differs from country to country. A further goal of this study was to examine the practice of testing frequencies in different countries in order to see how often high performing and low

performing countries implement formative assessment tests or quizzes. The study also focuses on advantages and disadvantages of question types that can be used in quizzes and tests, as well as advantages and disadvantages of tools that enable teachers to create quizzes. Additionally, the study presents a list of online quiz/test creating tools that teachers can use to develop their own quizzes with various features (see Appendix A).

Statement of the Problem

Educational institutions and teachers have long realized the importance of assessment and routinely include among their strategies numerous types of tests and quizzes in various frequencies. Administering quizzes and tests frequently in mathematics courses is essential because quizzes enable teachers to know what students have learned and where they are having problems so that teachers can modify their instruction to help students eliminate their weaknesses before introducing another topic. This procedure is especially important in mathematics courses because mathematics topics rely on each other and a firm base is required for mastering upcoming topics (Lowe, 2015).

Even though frequent testing is encouraged by researchers (Roediger & Karpicke, 2006; Shirvani, 2009), the ideal frequency of testing for a class has become a matter of controversy. Some teachers utilize quizzes at the end of each course to assess and improve student learning while some teachers barely use quizzes. Therefore, this study investigated the relationship between testing frequency and eighth-grade students' achievement in mathematics on the TIMSS 2011 exam to see if there are significantly different relationships between testing frequencies. Thus, teachers and administrators can make better decisions when investing time to prepare and utilize quizzes.

Research Questions

The following questions were considered in the completion of this study:

RQ1. Is there a significant relationship between testing frequency and the mathematics achievement of eighth-grade students' achievement scores as measured by the TIMSS, when SES variables are controlled for statistically?

RQ2. Is there a significant relationship between testing frequency and the mathematics achievement of eighth-grade students' achievement scores as measured by the TIMSS in different countries (South Korea, Singapore, United States, and Turkey), when SES variables are controlled for statistically?

Significance of the Study

This study examined the relationship of quiz frequency and eighth-grade students' achievement in mathematics. This study has chosen eighth-grade mathematics achievement scores because of their influence on education (Rodriguez, 2004). Eight-grade is of particular significance in the student learning journey. It is a gateway to high school in most countries, and nationwide standardized exams at the end of eighth-grade are a common practice for placing students in prestigious high schools while some private schools in the U.S. administer their own high school placement exams in order to select students (Turner, 2014). Eighth-graders are in transition to high schools, as well as between the 'concrete operational period' and the 'formal operation period' according to Piaget's cognitive development stages (Anthony, 2016). Eighth-graders complete the 'concrete operational stage' in which they can: exhibit logical thinking; arrange figures in serial order without trial and error: digest number, volume, and mass; and approach problems with a strategic method. Then, they move to the formal operation stage where they develop logical thought, deductive reasoning skills, and advanced memory and decision-making skills (Anthony, 2016). Students may need help and/or guidance to completely acquire such skills. As Vygotysky's Zone of Proximal Development suggests, students achieve more with a person who has already mastered the skills that are being learned (Coffey, 2009). Therefore, through providing feedback via

frequent tests or quizzes, the teacher helps students to improve their cognitive skills and develop their learning.

The study showed the relationship between students' mathematics achievement and using quizzes at different frequencies so that teachers can make an informed decision regarding when they plan to create and administer quizzes. The study also compared quiz practices in different countries so that all stakeholders in education can see how other countries are using quizzes, and the relationship of testing frequency and student scores on the TIMSS exam. The significance of the study goes beyond determining the relationship of testing frequency and mathematics achievement. It will also serve as a guide for teachers to select best quiz types by providing lists of question types, and the associated advantages and disadvantages of such question types (see Appendix B). This study also provides lists of online quiz-generator tools for various platforms such as websites, applications, and software. The list also includes functions of those tools that can serve as guidance for teachers when they decide to integrate quiz tools into their classroom activities.

The findings of this study will serve as a guide for teachers regarding the value of tests or quizzing at different frequencies as a teaching strategy to improve student achievement in mathematics. Results of this study will inform teachers about the relationship between testing frequency and student achievement in mathematics. The study will also inform teachers about optimal testing frequency so that they can adjust their testing frequency practices.

Selection of the Countries

The study used data from all of the TIMSS 2011 participating countries for the first research question. For the second research question, the researcher chose to investigate the relationship of testing frequency and student achievement in Korea, Singapore, the United States, and Turkey. Korea and Singapore are the two top performing countries and so the

researcher selected them to compare their quiz frequency practices with those of the United States and Turkey. The United States performed well above the TIMSS Scale centerpoint (500), and Turkey performed well below the TIMSS Scale centerpoint average. These two countries are also selected as representative of above and below average performing countries and compare their testing frequency practices with top two performing countries. Table 1 shows countries that are of interest to this study and their rankings, as well as their average math scores. The TIMSS 2011 website provides rankings and average scores for each of the TIMSS 2011 participating countries.

Table 1

Selected countries

Scale Centerpoints	Countries	Rankings	Average Scores
Above Average	Korea	1st	613
	Singapore	2nd	611
	United States	9th	509
Below Average	Turkey	24th	452

Definition of Terms

Eighth-grade students: TIMSS defines eighth-grade students as all students enrolled in an eighth year of schooling that started from the first year of The International Standard Classification of Education (ISCED) Level 1, providing the mean age of at least 13.5 years at the time of testing.

Frequent Testing: This term refers to daily, weekly, bi-weekly, and monthly formative assessment tools (e.g. tests and quizzes).

Test and/or quiz: The terms are used interchangeably in this study and they refer to daily, weekly, bi-weekly, and monthly formative assessment tools.

High performing countries: For the purpose of this study, countries that achieve average scores of above the TIMSS scale centerpoint (500) are considered as high performing countries.

Low performing countries: For the purpose of this study, countries that achieve average scores of below the TIMSS scale centerpoint (500) are considered as low performing countries.

TIMSS Scale Centerpoint: TIMSS defines the scale centerpoint as, at each grade level, the scale centerpoint of 500 that is set to correspond to the mean of the overall achievement distribution. Achievement data from subsequent TIMSS assessment cycles were linked to these scales so that increases or decreases in average achievement might be monitored across assessments. TIMSS uses the scale centerpoint as a point of reference that remains constant from assessment to assessment. (p.36)

Chapter II: Literature Review

Introduction

This study was conducted to examine the relationship between testing frequency and student achievement in eighth-grade mathematics. The study used the TIMSS international assessment project as its main data source. This study sought to answer the following two research questions:

RQ1. Is there a significant relationship between testing frequency and the mathematics achievement of eighth-grade students' achievement scores as measured by the TIMSS, when SES variables are controlled for statistically?

RQ2. Is there a significant relationship between testing frequency and the mathematics achievement of eighth-grade students' achievement scores as measured by the TIMSS in different countries (South Korea, Singapore, United States, and Turkey), when SES variables are controlled for statistically?

The first part of the literature review offers a brief overview of assessment, focusing on the role of formative assessment. This section is followed by a review of international assessment projects: PISA, PIRLS, and TIMSS. A comparison of these three programs then provides the rationale for this study's selection of TIMSS as the main data source. The second part of the literature review focuses on the findings of previously conducted TIMSS studies to analyze factors that affect mathematics achievement scores. Socioeconomic factors and their influence on student achievement are also discussed in detail. A comparison of SES variables used in international assessment programs and in the TIMSS program since it was first administered in 1995 is represented in tables. The third part of the literature review will discuss the effects of frequent testing in general, as

well as in mathematics in particular, as they are revealed in the literature. The main focus of the fourth part of this chapter is effective assessment. This includes quizzing, feedback, active learning, and teacher-generated online and traditional testing methods. A summary of the chapter, including the need for this study, is also provided at the end of this literature review.

Formative Assessment

This study focuses on quizzing in the context of formative assessment. The terms "formative assessment" and "summative assessment" were first introduced by Scriven (1967) to distinguish the two main types of assessment. According to this distinction, summative evaluation is defined as assessment of academic progress at the end of a specific time period in order to establish a student's academic standing relative to some pre-determined criterion. Formative evaluation is used to gather information for assessing the effectiveness of a curriculum and inform school systems in decisions to adopt a curriculum and then how to improve it (Scriven, 1967). In their critical review of formative evaluation, Dunn and Mulvenon (2009) define the purpose of formative evaluation as assessment to provide feedback, to inform policy, and to inform all educational stakeholders about the teaching and learning process. This distinction between formative and summative evaluation subsequently became known as Assessment For Learning and Assessment Of Learning, respectively (Stiggins, 2005; William, 2011).

Discussion, questioning, observation, entry or exit slips, a probe, and quizzes are some examples of formative assessment tools which are often used by teachers. These forms of assessment are considered to be low stakes because they usually do not have

point values related to students' grades, which runs parallel to the purpose of formative assessment (Eberly Center, 2015) and retrieval practices (Pelech, 2015). This study investigated the relationship of quizzes, a formative assessment tool, with student achievement, as measured through an international summative assessment project.

Since this study's interest is eighth-grade mathematics, PIRLS data is not appropriate, as the program measures only reading comprehension. PISA assesses 15-year-old students in mathematics, science, reading, and financial literacy (OECD, 2016) but they do not collect data from teachers about their testing frequency practices. This study has selected TIMSS assessment data because one of the main focuses of TIMSS is mathematics, it assesses several grades, including eighth-grade, and most importantly it collects data about teachers' instructional practices. With regard to contextual factors, TIMSS employs teacher and NRC surveys in addition to student and principal surveys, while PISA does not collect data from teachers. This militates against research into the relationship between teacher practices and student achievement, a relationship that is at the core of this study.

The TIMSS not only measures student achievement, but also collects contextual data about the country, school, teacher, student, and home environments. This data creates a natural laboratory where countries can learn from one another by analyzing such data (IEA, 2016). TIMSS assessments also collect and provide detailed information about curriculum and curriculum implementation, instructional methods, and school resources, as well as policy-relevant information about curriculum emphasis, technology usage levels, and teacher preparation and training programs (IEA, 2016). Thus, policy makers

can see what is going on in their educational systems and in other systems so that they can provide solutions to their problems through TIMSS data.

Review of TIMSS Studies

TIMSS assessments and questionnaire data have been a primary resource for many international research studies. TIMSS research studies have focused on factors (student, school, home, teacher, and instruction-related factors) that influence student achievement scores in mathematics and science (Bofah & Hannila, 2015; Lay, Ng, & Chong, 2015; Patnam, 2013).

Student-Related Factors

Learning occurs in a context, and student-related factors are one of the important pieces of the learning context. A TIMSS study that investigated the effect of self-efficacy in 4th and eighth-grade mathematics students, and it found that students' self-efficacy remarkably affects students' achievement in mathematics (Evans, 2015). Another TIMSS study that used the 1999 and 2007 administration data from Turkish students' answers to the student questionnaire showed that students' attitudes, value, and self-efficacy towards mathematics increased positively from 1999 to 2007 (Bilican, Demirtasli, & Kilmen, 2011). The percentage of students that think mathematics is essential to getting a good job or getting into a desirable university also increased from 1999 to 2007, which indicates an increase of the importance of mathematics in today's life (Bilican et al., 2011).

Another student-related factor to which studies have been devoted is whether gender influences student achievement. The biological argument that boys do better than girls in mathematics is diminished by Evans' study (2015). Her study used U.S. students'

TIMSS 2011 results to investigate the effect of gender and self-efficacy on mathematics achievement, and she found that gender itself does not have a significant effect on students' mathematics scores, but that boys who have high self-efficacy towards mathematics do better than girls who have high self-efficacy for mathematics. Another study that used TIMSS 2003 data also found no significant difference between boys and girls at 4th and eighth-grade levels in overall mathematics achievement (Dindyal, 2008). However, other studies have found significant differences in some content areas between boys and girls. Dindyal (2008) found that, at eighth-grade, girls did significantly better than boys in Algebra while boys did significantly better than girls in Measurement. No significant differences were found in Number, Geometry, and Data between boys and girls. The study also found that, at 4th grade, girls did better than boys in Geometry and Data while boys did better than girls in Measurement, and no difference was found between genders in Number and Patterns and Relationships.

Frequent Testing Literature

Frequent Testing and Student Achievement

Studies on the effects of frequent testing on student achievement go back to the 1930s (Hertzberg, Heilman, & Leuenberger, 1932; Keys, 1934; Kulp, 1933; Turney, 1932). In Keys'(1934) experimental study, he gave weekly tests and homework to an experimental group, and monthly tests and homework to a control group, in an educational psychology course. At the end of the intervention, students received an announced final exam. Results showed that there was no significant difference between the experimental group and control groups. On the other hand, Keys (1939) also gave them an unannounced exam. Results indicated that the experimental group significantly

outperformed the control group. Keys (1939) attributes the results of the announced exam to the fact that all students were prepared for the announced final exam. However, the results of the unannounced exam indicated, as Roediger's and Karpicke's findings also attested, that frequent testing improves retention of knowledge for a long time.

In 1951, a study of the effect of weekly quizzes as compared with monthly quizzes in a college level Government course was conducted at Purdue University (Fitch, Drucker, & Norton, 1951). The researchers gave weekly quizzes to an experimental group and monthly quizzes to a control group for a semester. Their results indicated that students who took weekly quizzes achieved significantly higher knowledge retention than the students who took monthly quizzes. Their results also indicated that frequent quizzes increased student motivation to attend in-class discussions.

Dustin (1971) also conducted a study to investigate the effects of weekly quizzes versus monthly quizzes on students' achievement scores. He also investigated the issue of exam anxiety through a student questionnaire in his study. The experimental group was given weekly quizzes and the control group received monthly quizzes. After the experiment, Dustin gave retention exams to both groups following 7 and 10 weeks of intervention to see if there was any difference in students' retaining the information. Results indicated that the experimental group scored higher on both exams. The difference was significant on the 7^{th} week test, but not on the 10^{th} week test. The results of the questionnaire indicated that frequent testing decreased student test anxiety (Dustin, 1971). Other studies also found that frequent testing increases student achievement (Salas-Morera et al., 2012). They also found that frequent testing helps students to keep

up with a subject and strengthens their involvement. Students also showed a high level of interest in participating in frequent tests (Salas-Morera et al., 2012).

Another study was conducted by a group of researchers to find the effect of frequent testing on students' study habits rather than on test scores (Mawhinney, Bostow, Laws, Blumfield, & Hopkins, 1972). They conducted their study in an undergraduate level Educational Psychology course. They divided their students into three groups; one group received daily quizzes, a second group took weekly quizzes, and the last group took one quiz in three weeks. Their within-subject design analysis showed that students who took daily quizzes developed more consistent learning skills and studied more daily (Mawhinney, et al. 1972). Martin and Srikameswaran (1974) conducted a study in a college Chemistry course to compare the effect of weekly quizzes versus no quizzes. Results indicated that the weekly group scored significantly higher, indicating that frequent testing leads to an improvement in students' study habits.

In 1991, Bangert-Drowns, Kulik, & Kulik (1991) conducted a meta-analysis study on the effect of frequent testing on student learning. They used only studies that were carried out in real classroom settings. They excluded studies that were carried out in lab settings and studies with paid volunteers in their meta-analysis study. Based on these criteria, they selected 35 previously conducted research studies on frequent tests. They found that 29 of the 35 studies demonstrated the positive effect of frequent testing on student achievement, but only 13 of these 29 studies found a significant difference. The remaining six studies found a negative effect, but only one of them indicated that frequent testing had a significantly negative effect on student achievement. Bangerts-Drown et al. (1991) concluded that students who took tests scored higher than students

who did not take tests, but the amount of difference decreased as the number of given tests to students increased. The reason for this finding might be because of another finding; too much testing reduces time for instruction (Gholami & Moghaddam, 2013). Bangerts-Drown et al. (1991) found, however, that student attitudes are more favorable to instruction when they are tested frequently.

In the 2000s, there was an increase in the interest in investigating the relationship between testing frequency and student achievement. Connor-Greene (2002) conducted a study to compare daily quizzes versus announced tests. Connor-Greene used student surveys in her study. She found that having relatively few announced tests throughout the semester led to procrastination and last-minute preparation. On the other hand, daily quizzes encouraged students to complete reading assignments. She concluded that frequent testing is a major factor in motivating students to learn material.

Quizzing as a Retrieval Practice

Roediger's TELC model focused on the effect of quizzes to enhance student learning, in addition to assessing student learning. (Roediger & Karpicke, 2006). They found that implementing quizzes or tests after study improves the recall of learned information for a longer time than restudying (Roediger & Karpicke, 2006). This effect is also known as the *testing effect* (Roediger & Karpicke, 2008). The theory behind the testing effect is known as retrieval practice, a learning strategy that focuses on getting information from memory (Agarwal, 2016). By practicing retrieval, learners strengthen memory for the retrieved information and forgetting is less likely to occur (Agarwal, 2016), and frequent quizzing or testing is a widely-used retrieval practice (Agarwal, Bain, & Chamberlain, 2012; Brame & Biel, 2015; Roediger, Putnam, & Smith, 2011a).

Quizzing as a retrieval practice has been found to have direct and indirect effects on student learning (Roediger et al., 2011a). Direct effect here refers to the testing effect on retention while indirect effect refers to other effects (e.g. motivating students to study regularly) testing might have. Roediger et al. (2011a) summarized the ten benefits of direct and indirect effects of testing as:

1. retrieval aids later retention
2. testing identifies gaps in knowledge
3. testing causes students to learn more from the next study episode
4. testing produces better organization of knowledge
5. testing improves transfer of knowledge to new contexts
6. testing can facilitate retrieval of material that was not tested
7. testing improves metacognitive monitoring
8. testing prevents interference from prior material when learning new material
9. testing provides feedback to instructors
10. frequent testing encourages students to study (pp. 31)

Roediger and Pyc (2012), however, say that frequent testing is an under-used tool in education despite its benefits.

Frequent testing as a retrieval practice enhances student learning both directly and indirectly (Karpicke & Grimaldi, 2012). The researchers also maintain that retrieval practice produces direct effects on the learning process because engaging in the process of retrieval itself produces learning. “Every time we retrieve knowledge, that knowledge is altered, and the ability to reconstruct that knowledge again in the future is enhanced” (p. 404). Each practice of retrieval changes learners’ knowledge and enhances the ability

to retrieve information again in the future (Nunes & Karpicke, 2015). Testing frequency also affects student learning indirectly. For example, when learners attempt to retrieve knowledge from memory, the result of the retrieval attempt gives the learners feedback that they need in order to better manage study time or change encoding methods (Karpicke & Grimaldi, 2012). Quizzing has been found to be effective in engaging students in retrieval practice and in motivating them to study more effectively (Karpicke & Grimaldi, 2012; Connor-Greene, 2002). Furthermore, Shirvani's (2009) EMMAR model (EMMAR standing for Engagement, Monitoring, Motivation, Anxiety, and Retention) suggests that frequent testing increases student engagement, motivation, retention, and monitoring of study times, and reduces test-taking anxiety. In all cases, retrieval practices improve student learning indirectly by better organizing the process of information encoding (Karpicke & Grimaldi, 2012).

Improvements in long-term retention are a direct effect of frequent testing (Agarwal et al., 2012; Roediger & Butler, 2011). Students in laboratory settings received different treatments in Roediger and Karpicke's study (2006). One group of students studied a passage four times (SSSS), a second group studied it three times and received a test (SSST), and a last group studied once and received tests three times (STTT) (note that S here refers to study and T refers to test). Students received two criterion tests: one after five minutes of treatment and another one a week later. Students in the SSSS group retrieved more information than any treatment groups in five minutes after the test while the STTT group retrieved more information than any other groups and the SSS group retrieved the least information as shown in an exam one week later. Roediger and Karpicke (2006) concluded that frequent testing improves long-term retention and this

finding was later confirmed by many other studies (Agarwal et al., 2012; Roediger & Butler, 2011). Roediger and Pyc (2012) also say that reading, highlighting, and flashcard study methods may help students to prepare for an exam, but they will not help students to retain knowledge for a long period of time. He strongly advocates frequent testing for knowledge retention, suggesting that students study with self-testing and that teachers integrate more quizzes and tests in their teaching activities.

Besides improving long-term retention of information, testing as a retrieval practice mitigates the amount of forgotten information (Pashler et al., 2007; Roedgier & Karpicke, 2006). A study that was conducted with high school students in history lessons found that reviewing historical facts with testing as a retrieval practice reduced the amount of forgotten information in comparison to reviewing without any retrieval activity (Roediger & Karpicke, 2006). This study also found that the amount of forgotten information increased as intervals between retrieval practices increased. It can be concluded that more frequent retrieval practice is better than less frequent retrieval in order to reduce the amount of forgotten information.

Another positive indirect effect of frequent testing is increasing student attendance (Wilder, Flood, & Stromsnes, 2001; Zarei, 2008). Wilder et al. (2001) found that the use of random quizzes increased attendance by 10% in an undergraduate course. Another study found that the use of frequent quizzes motivates students toward better attendance (Zarei, 2008). Student attendance is especially important in mathematics courses, as math concepts are strongly linked with each other and students' existing knowledge should be taken into consideration in teaching math (Campbell, 2008). This finding is an important indirect effect of frequent testing because students will not have

gaps in their existing knowledge to militate against new learning in mathematics if they attend courses regularly.

Some studies have distinguished between the backward effect and the forward effect. Backward effect refers to the claim that testing previously learned information can enhance its long-term retention. Forward effect suggests that testing previously learned information can also increase the long-term retention of subsequently presented new information, whether it is related to previous information or not (Pastötter & Bauml, 2014). The forward effect of testing occurs by keeping student attention high when transitioning from previous to new lecture content, encouraging task-relevant strategies and discouraging irrelevant activities, reducing test anxiety, and reducing experienced mental effort (Pastötter and Bauml, 2014). The forward effect is also called test-potentiated learning and was first investigated by Izawa in 1966 (Roediger et al., 2011a). Izawa found that testing can increase the amount of information learned from future study sessions (Roediger et al., 2011a). Brame and Biel (2015) also said that frequent testing facilitates the learning of upcoming materials, whether they are related or not. Wissman, Rawson, and Pyc (2011) conducted a study with undergraduate students to examine the forward effect of testing. Students in a control group only read three passages in a text and then took an interim test. Students in an experiment group also read three passages in a text, but they also took free recall tests after the first two reading passages. Students in the experiment group also took an interim test after reading the third passage. Results indicated that students who took the recall test between reading passages recalled as much as two times more than the read-only group of students. This

result was observed even when reading passages were not related to each other (Wissman et al., 2011).

It must be noted, however, that many studies (FairTest, 2007; Fulton, 2016; Spector, 2015) claim that frequent testing impacts negatively on students' senses of self-efficacy, and also causes stress and anxiety. Much of this research has typically been focused on appropriate methods of measuring the impact of test anxiety on student performance in summatively assessed high stakes examinations such as academic and standardized tests where stakes are high and teachers teach to tests (Cassady & Johnson, 2002; Fulton, 2016; Marshal, 2007; Trifoni & Shahini, 2011). This has resulted in an increase in the late 20th and early 21st century phenomenon of test anxiety. "Examination stress and test anxiety have become pervasive problems in modern society." (p. xiii) (Spielberger & Vagg, 1995). This view is supported by a recent report on the impact of accountability measures in schools in England, with the challenging title of "Exam Factories?" This report found that students are suffering from increasingly high levels of school-related anxiety because of increased pressure from exams, and greater awareness at younger ages of their own 'failure' (Hutchings & Kazmi, 2015). A similar Australian-based report provides evidence of the "stress, anxiety, pressure and fear experienced by students" because of the current emphasis on frequent high stakes tests (Polesel, Dulfer, & Turnbull, 2012).

In contrast, many studies have found that frequent low stakes testing actually reduces students' test taking anxiety (Agarwal, et al., 2012; Dempster, 1992; Dustin, 1971; Shirvani, 2009), and this conclusion is a common finding among testing frequency studies. Testing anxiety is negatively correlated with student achievement (Shirvani,

2009) and the above cited studies found that frequent tests reduce students' test taking anxiety. Students who participated in Agarwal et al.'s (2012) study reported that taking frequent low stakes quizzes decreased students' test anxiety. Researchers found that students' overall perspectives are positive regarding frequent testing as a retrieval practice (Agarwal et al., 2012; Salas-Morera et al., 2012; Suda et al., 2011). Salas-Morera, et al. (2012) found that students highly appreciated the use of quizzes, giving this practice an average rating of between 4.06 and 4.20 out of 5, because it helped them keep up with subject, strengthened their involvement in other activities, and had a very positive impact on academic scores. Karpicke and Grimaldi (2012) also found that students viewed frequent tests as a valuable learning tool and liked them. Students in Suda et al.'s (2011) study reported that unannounced quizzes forced students to study regularly, which is a mediating effect of frequent testing (Shirvani, 2009).

Teachers' perspectives on the use of quizzing or frequent testing are similar to those of students. They value quizzing as a tool to encourage students to study regularly (Agarwal et al., 2012; Suda et al., 2011). A teacher who participated in Agarwal et al.'s (2012) study reported that he used quizzes every day after realizing the potential of quizzing. A principal also reported that quizzes increased student grades significantly (Agarwal et al., 2012).

Despite the positive effects of frequent testing, some researchers have found it may have unwanted effects on student learning, as well. Testing without feedback may result in students perceiving incorrect answers as being correct answers, especially if closed-ended questions are used in the tests (Roediger et al., 2011a). Gholami and Moghaddam (2013) summarized the disadvantages of frequent testing as follows:

implementing and grading tests can be time consuming for teachers, and it may take too much time from instruction. Time shortages may cause the prevention of integration of larger units (Bangert-Drowns et al., 1991; Gholami & Moghaddam, 2013). Roediger et al. (2011a) also say that even though frequent testing may take some instructional time away, there is no evidence to suggest that activities which can replace testing are more effective than testing. Using tests frequently may also become tedious for students and may decrease their interest in topics and learning in general (Bangert-Drowns et al., 1991; Gholami & Moghaddam, 2013). Marshall (2007) also criticized the use of excessive tests in British schools. She says that too much testing does not provide effective and lifelong learning because the main focus is high test scores rather than learning. Therefore, teachers teach to the test and teach only required information that students need to do well on the tests, and students study only for the material that will be on the tests (Marshall, 2007).

Frequent Testing in Mathematics

Since the late 20th century, a number of studies have been devoted to the effect of frequent testing in mathematics. Most studies have found such testing to be beneficial. Townsend and Wheatley (1975) conducted a study in analytic geometry and calculus classes with 442 students for a quarter of an academic year. They compared four groups which took tests at deferent frequencies. The results showed that daily quiz groups achieved significantly higher results than other groups on a midterm exam.

Ma (1995) investigated the frequent testing effect in high school algebra and geometry courses through informal oral testing, while other studies have used written tests. Ma (1995) composed two academically equal groups based on students' prior

grades. The experimental group was treated with frequent oral testing throughout a semester. The results indicated a significant difference only in the geometry course, in favor of the experimental group. Ma (1995) concluded that frequent oral testing significantly improves 11th grade students' problem solving skills in geometry.

Another study was conducted to investigate students' preferences for frequent testing and the associated effect on their learning by Kika, Mclaughlin, and Dixon (1992). The study took place in an 11th grade algebra course. Kika et al. (1992) gave a weekly test to one group and bi-weekly tests to a second group of students for two months. After two months, the second group took weekly quizzes and the first group took bi-weekly tests for two months. At the end of the intervention, students filled out a questionnaire to indicate their preference for test frequency. Results indicated that students prefer to take weekly tests, and students who took weekly quizzes scored higher than the bi-weekly tested group in each session of the study. The study also found that low and medial-level students showed more improvement through frequent testing than high achieving students (Kika et al. 1992).

Johnson (2006) investigated the effect of frequent testing in an introductory calculus course in two academically similar classes. He gave weekly short tests to an experimental class, which consisted of 28 students, and no tests were given to a control class, which consisted of 23 students. The results of the study indicated that the experimental group which took weekly short tests scored significantly higher than the control group in the final examination. His study also found that weekly tests work as a good predictor of students' final examination grades (Johnson, 2006).

One of the most recent studies that investigated the effect of frequent testing on students' achievement in mathematics is Shirvani's (2009) study, in which he found that daily quizzes significantly improve student achievement more than weekly quizzes. Four high school geometry classes participated in Shirvani's study. The control group received weekly quizzes every Friday while the experimental group received quizzes every day in the last minutes of each lesson for a term of six weeks. Students in both groups were also assigned the same number of homework during the six-week term. At the end of the treatment period, students in both groups received the same final exam. Results indicated that students in the experimental group outperformed the control group students in both achievement scores and homework grades. Shirvani (2009) theorized the benefit of frequent testing as being EMMAR, which stands for engagement, motivation, monitoring study times, anxiety (reduce anxiety for test taking), and retention of the learned material for a longer time.

Even though most studies found that frequent testing is effective to increase student achievement in mathematics, some studies have found none or very little benefit of frequent testing in student achievement in mathematics. Zgraggen (2009) used two control groups and two experimental groups in his study. Experimental groups received weekly quizzes while control groups received quizzes every two weeks for the period of a term and, at the end of the term, students had a final exam, and one month after the term, students had an unannounced retention exam. Results indicated that students in control groups scored better than in experimental groups on both the final exam and the retention exam, which suggests that bi-weekly quizzes are more effective than weekly quizzes in improving students' mathematics scores.

In contrast to Roediger's findings that testing is more effective than restudying, Burk (1987) found that frequent testing is not more effective than reviewing in geometry. Another study that found no significant difference between daily quizzes versus weekly quizzes was conducted by Dineen, Taylor, and Stephens (1989). They gave daily quizzes to an experimental group and weekly quizzes to a control group over the course of a semester in a high school mathematics course. Even though the experimental group scored higher than the control group, the improvement was not statistically significant. This result is similar to that of Bangert-Drowns, Kulik, and Kulik's (1991) study. Their meta-analysis study found that most previously conducted studies found that frequent testing improves student scores in comparison to infrequent testing, but that the effect size of improvements is small. Another meta-analysis of prior studies was conducted in 2009 by Basol and Johanson. They analyzed 79 previous frequent testing studies by putting them in high, medium, and low frequency testing categories. Basol and Johanson (2009) found more improvements with high frequency testing, but no significant improvement between three testing frequencies was detected.

Rodriguez (2010) examined the relationship between classroom assessments and student achievement by using American students' and teachers' data from the TIMSS 2003. He used homework, teacher-made objective tests and open-ended tests, projects, observations, student responses in class, and externally created exams as assessment practices. While the findings of his study found a significant positive relationship between homework and student achievement, he found that teacher-made tests had a slight negative relationship in relation to student achievement. The challenge of creating high-quality tests or quizzes may have influenced this result because low-quality tests

may negatively affect students' academic self-efficacy, motivation, and effort (Rodriguez, 2010).

Quizzing and Learning Theories

Quizzing as a retrieval practice can be underpinned by various learning theories. The section below discusses the relationship of quizzes as retrieval practices and learning theories.

Quizzing and Bloom's Revised Taxonomy

A retrieval practice is defined as accessing and retrieving newly learned information from the memory when needed (Francisco, 2014; Nunes & Karpicke, 2015), and the "Remember" category of Bloom's revised category is defined as retrieving relevant information from long-term memory (Krathwohl, 2002). The definitions of the two terms are so similar that comparison leads to a natural conclusion that quizzing as a retrieval practice is a very useful strategy for teachers to implement for "Remember" category objectives. The "Remember" category includes recalling and recognizing subcategories. Quizzing after study has been proven to improve later recall more than restudying (Roediger & Karpicke, 2006), and quizzes are good opportunities for students to practice recall, but only low-stake quizzes enhance student learning (Brame & Biel, 2015). Moreover, testing of recall can improve learning of subsequently presented new information (Pastötter et al., 2014).

Quizzing is also a useful strategy for improving students' second category objectives in the revised taxonomy: "Understand". It is found that testing improves student performance on questions that require students to make inferences (Karpicke & Blunt, 2011; Smith & Karpicke, 2014). Inferring is a subcategory of the second category

(understand; determining the meaning of instructional messages, including oral, written, and graphic communication) in the revised taxonomy (pp. 215, Krathwohl, 2002). Other studies have also found that frequent testing benefits go beyond memorizing or *rote* learning, also providing deep learning (Brame & Biel, 2015; Francisco, 2014; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Smith & Karpicke, 2014). Smith and Karpicke (2014) investigated the effect of quizzing as a retrieval practice to see if testing only improves recall of information or if it also improves meaningful learning in comparison to a study only condition. After one week of intervention, students took a final exam. Their results found that the testing group performed better on questions that only required the recall of information, as well as on questions that required them to make inferences. McDaniel et al.'s (2013) study also found that quizzing with feedback promotes learning that is deeper than just retaining the information, but which can also be transferred into new learning.

Transferring knowledge into new concepts through testing is described as performing well on summative exam questions that were related but not the same as quiz questions (McDaniel et al., 2013). Roediger et al. (2011a) defined such a transfer as applying information learned in one situation to new or other situations. Transferring knowledge is a crucial goal of education because educators would like students to apply their knowledge in furthering their education and in daily life (Roediger et al., 2011a). McDaniel et al. (2013) found that quizzing improved student learning on summative exam questions that required transfer knowledge from quizzes in comparison to the study only condition. Butler (2010) also found that frequent testing improved student learning,

in comparison to repeated study, in retrieving information and transferring knowledge as measured by a summative exam.

Quizzing and Cognitive Science

Retrieval-based learning is an excellent example of how findings from cognitive science can lead educational activities (Nunes & Karpicke, 2015). Studies that combined cognitive science and education have shown that, when students take a quiz or a test, cognitive action occurs that enables learning, and this is also called the "testing effect" (Pelech, 2015; Roediger & Karpicke, 2006). Cognitive methods are used in education to make sure students achieve a particular goal, such as understanding a reading passage, but metacognition strategies are used to check whether a cognitive goal has been met, such as by quizzing oneself to evaluate one's own understanding of the aforementioned reading passage (Livingston, 1997). Students may think they know a lot if they only study the material, but testing informs students about how much they know so that they can make a better prediction of their knowledge (Roediger et al., 2011a). Moreover, Roediger et al. mentioned that

> Students' ability to accurately predict what they know and do not know is an important skill in education, but unfortunately students often make inaccurate predictions. When students reread material repeatedly, they are often overconfident in how well they know the material. Taking a test, however, can lead to students becoming less confident, a finding known as the underconfidence-with-practice effect (Koriat, Scheffer, & Ma'ayan, 2002; see also Finn & Metcalfe, 2007, 2008). Testing can help compensate for the tendency

to be overly confident, which results in a more accurate assessment of learning. (pp. 20)

It is clear that students' metacognitive awareness of what they know is an important skill in education, and frequent testing is an effective method to inform students about their knowledge. Therefore, teachers should implement frequent quizzes into classroom activities and students should use the self-testing method to increase their knowledge of their own understanding of a topic (Roediger et al., 2011a).

Quizzing and Constructivism

Constructivism is a learning theory that explains the way people learn and was developed from the work of Piaget, Vygotsky, Dewey, and Bruner (Pelech, 2016). Constructivists believe that knowledge is actually constructed by learners, rather than transferred from one to others, through an active and complex cognitive process of development, and learners are the creators of meaning and knowledge in the constructivist approach (Gray, 1995). Gray (1995) described constructivism by referring to its four principles: knowledge consists of past constructions where our perceptions are interpreted through a logical framework formed by prior experiences; constructions occur as we adapt and alter our old concepts; learning is an organic process of invention rather than a mechanical accumulation of facts and associations; meaningful learning occurs through reflection and resolution of cognitive conflict (p. 3).

Active participation in the learning process (also known as active learning), social interaction between students and with teachers, and building on students' prior knowledge and experiences are important principles of constructivism. Quizzing as a retrieval practice is an effective tool to stimulate prior knowledge as a starting point for

learning (Pelech, 2015; Hein, 1991). Learning as a social activity refers to interaction of learners with peers, teachers, or someone that has already mastered the subject matter (Hein, 1991). Vygotsky's Zone of Proximal Development (ZPD) also speaks to the need for an adult or more expert learner who can interact with students to provide educational assistance in order to help them achieve their potential (Vygotsky, 1978). Quizzing with feedback can construct a social interaction between learners and teachers to help students towards achieving their potential. Additionally, the constructivist approach to assessment is formative rather than summative; it encourages assessment to improve student learning, and not to provide grades (Socrates Programme, 2009), which is also the goal of quizzing as a retrieval practice (Roediger & Karpicke, 2006). Learning continuing during assessment is another approach of constructivism (Brooks & Brooks, 1993) and is similar to the approach of the retrieval practice in relation to assessment.

Active Learning

Active learning theory was first introduced in the 1990s by Bonwell and Eison. They described active learning as asking students to do things and think about what they do. Active learning incorporates instructional activities that put students at the center of their activities and engages them in the learning process (Prince, 2004). Students must do more than just sit and listen; they must read, write, discuss, or be engaged in solving problems, and instructional techniques should guide students to engage in analysis, synthesis, and evaluation activities (Bonwell & Eison, 1991). They suggest that students should be actively involved in learning processes rather than passively listening to instructors. Teachers should create such learning environments that encourage students to engage in hands-on learning and know the reasons for their activity while learning.

Discussion should be preferred instead of lecturing to activate students' thinking processes and engage students with learning material (Bonwell & Eison, 1991). They also pointed out that students need to be actively engaged with higher-order skills such as analysis, synthesis, and evaluation in order to be defined as active learners. To enable students to gain these skills, instructors need to design instructional activities such that students actively do things and also think about the things that they do. Many studies (Halpern & Hakel, 2003; Michael, 2006; and Prince, 2004) have measured the effectiveness of active learning strategy as opposed to the traditional lecturing method. A study that analyzed 225 studies to compare the effectiveness of active learning versus traditional lecturing found that active learning was more effective in improving student performance in Science, Technology, Engineering, and Mathematics (STEM) courses, in comparison to the traditional teacher-centered teaching method (Freeman, Eddy, McDonough, Smith, Okoroafore, Jordt, & Wenderoth, 2014).

When active learning strategies are implemented in a classroom, teachers will probably spend most of the class time helping students to increase their understanding and promoting deep learning, and spend less time on transferring information through lecturing (Eison, 2015). Students will have opportunities to apply and demonstrate their newly acquired knowledge and skills, and they will receive immediate feedback from classmates and teachers when active learning strategies are in place (Eison, 2015). Quizzing as a retrieval practice is also considered to be an active learning strategy by researchers (Francisco, 2014; Karpicke & Grimaldi, 2012; Socrates Programme, 2009). Shirvani (2009) notes that frequent testing has an indirect effect of increasing student engagement. He further says that, when the frequency of testing is increased in a

classroom, there will also be an increase in student involvement with discussions and an increase in students responding to questions. Another study found that engagement time is the second most important factor that affects student learning, after student ability, and when students are engaged with the learning material, they academically perform better than passive students (Shirvani, 2009). Therefore, tests and quizzes can also be used as active learning strategies in classrooms in order to increase student learning.

Feedback

As mentioned previously, one of the goals of formative evaluation methods is to provide feedback to students and teachers. Chickering and Gamson (1987) point out the importance of providing feedback in their seven principles of good practice work: feedback reduces the discrepancy between desired knowledge and end product, student knowledge at the end of an instruction. Providing feedback, the seventh event, is also pointed out in Gagne's nine events of instruction model. Immediate feedback is especially important in mathematics because mathematic concepts rely on each other in a way that requires a firm foundation for each following concept (Lowe, 2015). If feedback is not given, or given late, students may learn false information while taking quizzes in mathematics (Lowe, 2015; Roediger et al., 2011a). Roediger et al. (2011a) further says that closed-ended questions (e.g. true/false, multiple choice) in quizzes may lead to erroneous learning if feedback is not given. Roediger and Butler (2011) also maintain that feedback in frequent testing increases learning because it enables students to correct errors and maintain correct responses. If students are to learn from the tests, they need to successfully retrieve the correct information, but if students do not retrieve correct information or do not learn it, then the benefits of frequent testing can be very limited or

absent (Roediger & Butler, 2011). Therefore, students need to receive feedback after each retrieval attempt, whether students' attempts are successful or not, to be able to successfully retrieve such information in the future attempts (Roediger & Butler, 2011).

Feedback can come in many forms and kinds; immediate feedback, short-answer feedback, explanation feedback, individualized feedback, and automated feedback are some feedback methods (Butler, Godbole, & Marsh, 2013). Researchers have found that various factors affect the effectiveness of feedback, but the content of the feedback message is probably the most influential aspect of any feedback mechanism because feedback messages enable learners to correct errors and reinforce correct information (Butler, Godbole, & Marsh, 2013).

Feedback is a significant part of any assessment method and it provides chances for learners to decrease their gap between actual knowledge and desired knowledge (Butler, Godbole, & Marsh, 2012). Moreover, feedback functions as a motivator for students (Barker, 2011). Though students are less satisfied with current feedback mechanisms than any other course aspects, institutions have been working to provide well-constructed feedback (Nicol, 2011). However, there is not an easy solution for providing constructive feedback because it is not possible to provide feedback for each student every time, and especially if a student/teacher ratio is high and tests are delivered in the traditional paper-based format because feedback loads extra work for teachers who wish to provide feedback for each student and for every question (Marden, Ulman, Vilson, & Velan, 2013). Another problem with providing feedback is timing; feedback is usually not timely or meaningful in crowded classrooms (Barker, 2011).

It is well documented in the literature that providing feedback is significant, and that it needs to be well constructed to increase student learning. Feedback that is not well planned or not given in a timely manner has very little effect on student learning (Chickering & Gamson, 1987). The content of the feedback message is the most important aspect of the feedback providing process (Butler, et al., 2013), so delivery is vital to the process. It is obvious that there are several aspects of feedback delivery methods that contribute to the effectiveness of feedback, which directly affects teaching and learning. In order to increase student learning, feedback needs to be constructive, timely, appropriate, useful, accurate, individual, detailed, delivered in context, and should lead students' learning forward (Barker, 2011).

Another key aspect of providing effective feedback is doing so in a nonthreatening environment; teachers should not assign credits to assessment instruments if the goal is giving feedback (Marden, Ulman, Wilson, & Velan, 2013). Pelech (2016) says that in order for quizzing to increase student learning, it should be low stakes. However, it is very likely for students to skip the quizzes or not take them seriously if quizzes have no effect on students' grades. Assigning a small percentage of grades to the quizzes might increase student engagement even though quizzes are intended to serve as a formative assessment method, so that students will take these quizzes seriously (Marden, et al., 2013).

Quiz Types and Tools

Several question types (e.g. multiple choice, fill in the blank, matching, ordering, true/false, and short answer) can be used in quizzes as retrieval practices. Studies have found that question types do not differ in improving student learning in quizzing

(McDermott et al., 2014; Smith & Karpicke, 2014) and that they can be administered online or face-to-face (Brame & Biel, 2015). Smith and Karpicke (2014) found that multiple-choice quizzes, short-answer quizzes, and hybrid quizzes are equally effective for increasing student learning. However, it is important to note that closed-ended questions (multiple choice, fill in the blank, matching, ordering, and true/false) without feedback may result in students learning false information and retaining it (Roediger et al. 2011a). To prevent this problem, Roediger et al. (2011a) suggest providing feedback to students after quizzes. Each quiz type has its own distinct features that have advantages (e.g. objective grading) and disadvantages (e.g. failing to show students' thinking processes) as an instructional tool. Appendix B discusses each quiz type with its benefits and also its downsides for teaching and learning. Teachers also have a range of quiz-creating and implementing tools to create quizzes besides traditional paper and pen quizzes. Advancements in technology empower teachers to use mobile devices, applications, websites, software, and learning management tools to generate and deliver their own quizzes. The review of these tools is also included in Appendix B.

Chapter Summary

There are several international organizations that conduct or organize international assessment projects and provide contextual data so that countries can compare their practices and results with one another to improve their educational systems. PISA, PIRLS, and TIMSS are three of the best-known international assessment projects. This study used TIMSS data because of its main focus on mathematics achievement in comparison to the other two organizations. Furthermore, in their investigation of contextual factors, TIMSS surveys teachers in addition to students,

principals, and NCRs. This reach enables researchers to investigate the relationship between teacher practices (e.g. quiz frequency) and student achievement. The review of previous TIMSS studies indicates that none of the contextual factors alone has a major effect on students' math achievement scores, though SES is identified as a significant, and maybe the most important, contextual factor impacting student academic performance.

The literature review of frequent testing is traceable to the beginnings of the 20th century. Frequent testing opponents claim that this practice could take a significant amount of time away from instruction, resulting in teachers teaching to tests, and increase student anxiety. Proponents of frequent testing suggest, however, that even though the exact frequency of testing is controversial, it is a good practice to test students frequently. Low stakes frequent testing has been found to reduce test anxiety, and improve recall of learned material for a longer period of time than restudying. Other benefits include motivating students to study, increasing student engagement and attendance, better organization when studying, potentiating further study, and being perceived as beneficial by both students and teachers.

Need for the Study

Tests or quizzes are routinely used as a formative assessment strategy across disciplines, including mathematics. Yet, the literature reveals that there is no consensus on the relationship between testing frequency and student achievement in mathematics and that even proponents differ on the optimal quiz frequency. These issues need to be explored with further research. That is the aim of this study.

The literature also reveals that most testing frequency studies in any discipline are conducted in a single institution, in a small number of institutions, in a single country, or in a limited range of countries. There is a need for a wider and deeper study across a broader range of both high and low performing countries. The literature also indicates that most of these studies were conducted at the college or university level, but not to any significant extent with 8th graders. Additionally, an examination of the contextual factors included in all TIMSS assessments since it was first administered in 1995 reveals that no studies have been conducted on quiz frequency as a factor influencing student achievement in mathematics. Therefore, there is a gap in the research with eighth-grade students of mathematics across countries. This gap needs to be filled, given the status of math as an essential 21st century skill and the importance of this grade as a foundational and transitional stage in the academic progress of students.

This current study addresses these needs by examining the relationship between testing frequency and student achievement in eighth-grade mathematics, and comparing frequent testing practices in all participants in TIMSS 2011 with a focus on an international comparison of these practices in disparate economies, cultures, and education systems.

Chapter III: Methodology

Introduction

This chapter describes the research methodology of the study, including research questions and demographic information about the participants. An explanation of the research design is followed by a detailed explanation of the procedures utilized to complete the study. A data analysis section is also included at the end of the chapter.

Research Questions

This study sought answers for following research questions:

RQ1. Is there a significant relationship between testing frequency and the mathematics achievement of eighth-grade students' achievement scores as measured by the TIMSS, when SES variables are controlled for statistically?

RQ2. Is there a significant relationship between testing frequency and the mathematics achievement of eighth-grade students' achievement scores as measured by the TIMSS in different countries (South Korea, Singapore, the United States, and Turkey), when SES variables are controlled for statistically?

Research Hypothesis

1. Null Hypothesis: There is no significant relationship between testing frequency and student achievement in eighth-grade mathematics scores as measured by the TIMSS, when controlled for SES variables in all TIMSS 2011 participant countries combined.

2. Null Hypothesis: There is no significant relationship between testing frequency and student achievement in eighth-grade mathematics scores as measured by

the TIMSS, when controlled for SES variables in Korea, Singapore, Turkey, and the United States.

TIMSS International Assessment Project

TIMSS provides data on the mathematics and science achievement of students in participating countries and it enables participating countries to compare their results with students in other countries. TIMSS assessments have been assessing students, usually every four years since 1995, from 4th and eighth-grade students in approximately 50 countries. The most recent TIMSS assessment was in 2015 and it included students from 4th, 8th, and 12th grades. The results of TIMSS 2015 were released at the beginning of December 2016 (NCES, 2016) and could not therefore be included in the research for this study.

Experts from different countries and fields input their knowledge and experiences to develop TIMSS assessments. These assessment materials are endorsed by all participating countries (Martin & Kelly, 1996). Test questions are created so that most topics in mathematics and science are represented in the tests, enabling students to show their skills and knowledge over a wide range of topics. The TIMSS achievement tests do not just include opportunities for selecting the best choice among others, but also include test items that let students provide a short answer to a question or elaborate upon a response (Martin & Kelly, 1996).

The TIMSS developers use a matrix design system to create tests. In this design, test items are distributed as blocks of items among multiple test booklets and these booklets are distributed across students in a country so that a country's students are measured in most of the topics, but individual students do not have to answer all of the

questions. For example, 8 test booklets were created for 8th graders in TIMSS 1995, all of which together would require about 5 hours to answer; however, nobody takes all of the booklets. Each student received only one booklet, which required 90 minutes for answering the questions (Martin & Kelly, 1996). Therefore, TIMSS tests cover most of the content in mathematics and science, but reduce the burden on students (Martin & Kelly, 1996).

Participants

Data for this study was obtained from publicly available records of the Trends in International Mathematics and Science Study (TIMSS) 2011 assessments. The sample included the eighth-grade students who took the TIMSS 2011 mathematics exam. More than 200,000 students (50.1% female, 49.9% male) from 42 countries participated in the eighth-grade mathematics TIMSS 2011 exam. Table 2 shows participant information for each country. No information was available describing the number of students who had been retained or who were receiving special education.

Table 2

Sample size summary of TIMSS 2011 participating countries

Countries	Number of schools	Number of teachers	Number of students
Armenia	153	918	5,846
Australia	277	1,013	7,556
Bahrain	95	818	4,640
Chile	193	834	5,835
Chinese Taipei	150	1,218	5,042
England	118	1,012	3,842
Finland	145	1,028	4,266
Georgia	172	874	4,563
Ghana	161	655	7,323
Hong Kong SAR	117	1,178	4,015
Hungary	146	990	5,178
Indonesia	153	771	5,795
Iran	238	829	6,029
Israel	151	1,026	4,699
Italy	197	993	3,979
Japan	138	1,135	4,414
Jordan	230	811	7,694
Kazakhstan	147	970	4,390
Korea	150	1,220	5,166
Lebanon	147	899	3,974
Lithuania	141	980	4,747
Macedonia	150	843	4,062
Malaysia	180	863	5,733
Morocco	279	743	8,986
New Zealand	158	978	5,336
Norway	134	952	3,862
Oman	323	730	9,542
Palestine	201	804	7,812
Qatar	109	819	4,422
Romania	147	913	5,523
Russia	210	1,077	4,893
Saudi Arabia	153	790	4,344
Singapore	165	1,215	5,927
Slovenia	186	1,004	4,415
Sweden	153	970	5,573
Syrian	148	718	4,413
Thailand	172	852	6,124
Tunisia	207	850	5,128
Turkey	239	905	6,928
Ukraine	148	934	3,378
United Arab Emirates	458	905	14,089
United States	501	1,024	10,477

As can be seen in Table 2, TIMSS participants greatly vary from one country to another. The differences in number of schools, teachers, and students are related to

countries' populations in general. TIMSS excludes some schools and students from participation if school sizes are extremely small, accessibility is difficult to schools because of their remote location, schools offer an education that is radically different from the mainstream educational system, or if schools only provide education to special needs students. Students with functional and intellectual disabilities as well as students whose language is different than the test language were also excluded from the TIMSS. Botswana, Honduras, and South Africa were excluded from this study because they tested 9th grade students instead of eighth-grade students in the TIMSS 2011 exam.

This study used eighth-grade students' mathematics achievement scores, student questionnaires, school questionnaires, and teacher questionnaires of the TIMSS 2011. The TIMSS 2011 employed a two-stage stratified cluster design to select students from each participating country that represent a total population of the students. In the first sampling phase, schools were selected based on their probabilities proportional size (PPS) from a list of nationally representative schools and two additional schools were randomly selected for each school in case initially selected schools refused to participate. About 150 schools were sampled in most countries, including one or two classrooms from each school so that representation of a sample size of at least 4,500 students was attained for each country. In the second sampling phase, one or two intact classrooms from the target classrooms were selected. To ensure an efficient sampling design and implementation, the National Research Coordinators conduct class sampling in each country via WinW3S software that was developed by IEA and Statistics Canada.

Setting

This study used pre-existing data from the TIMSS 2011. TIMSS does not apply any treatment conditions, but rather collects data on the regular practices of teaching and learning activities. Teachers implement their tests or quizzes without any intervention during the school year and TIMSS collects data from teachers through teacher questionnaires regarding instructional practices that include quiz practices, academic and professional background, classroom resources available for teaching, attitudes toward teaching mathematics, and many other types of information. TIMSS also implements school questionnaires that are usually filled out by school principals to collect data about school settings and resources available for teaching and learning.

In the TIMSS 2011, 37% of students were attending schools in large sized cities (population of above 100,000), 28% were attending schools in medium sized cities or towns (population between 15,001 and 100,000), and 35% in small towns or rural areas (less than 15,000 people). TIMSS 2011 eighth-grade students were distributed fairly equally across the three types of schools. 32 % of the eighth-grade students attended relatively affluent schools while 36 % attended more disadvantaged schools. (Mullis, Martin, Foy, & Arora, 2012).

Sixty-nine percent of eighth-grade students were in schools where almost all students (more than 90%) spoke the language of the TIMSS assessment as their native language, 13% were in schools where the majority of students (51-90%) were native speakers of the TIMSS assessment language, and 17% were in schools where half the students (or less) spoke the language of the assessment as their native language.

Internationally, only 25% of schools were able to provide all resources that were necessary for instructional activities, 69% of schools had some resources, and 6% of schools had almost none of the resources that were required for instruction in eighth-grade.

The TIMSS 2011 eighth-grade test had four content domains (Number, Algebra, Geometry, Data and Chance) and three cognitive domains (Knowing, Applying, and Reasoning) that describe expected behaviors as students engage with the mathematics content (Mullis, Martin, Foy, & Arora, 2012). Table 3 shows the eighth-grade mathematics tests' percentage of content and cognitive domains in the TIMSS 2011 test.

Table 3

TIMSS 2011 eighth-grade mathematics framework

Content Domains	Percentage	Cognitive Domains	Percentage
Numbers	30%	Knowing	35%
Algebra	30%	Applying	40%
Geometry	20%	Reasoning	25%
Data & Chance	20%		

The following section summarizes selected countries' educational settings.

Country Settings

Republic of Korea

The Ministry of Education, Science, and Technology (MEST) is the responsible government institution in Korea. Even though new modes of educational operation have decentralized educational administration and promoted local autonomy, the country implements a national curriculum and regional guidelines. However, there is some flexibility in national curriculum and guidelines to allow individual school characteristics and school objectives to be implemented (Mullis, Martin, Minnich, Stanco, Arora, Centurino, & Boston College, 2012). The country uses a single-track 6-3-3-4 system with

6 years of primary, 3 years of middle school, 3 years of high school, and 4 years of university education. The education language is Korean and the first 9 years of education (elementary and middle school) are free and mandatory in Korea (Mullis et al., 2012).

The national common curriculum and the high school elective-centered curriculum are two different national curricula of Korea. The national common basic curriculum includes subject matters, optional activities, and extracurricular activities. The subject matter consists of ten courses: Korean language, moral education, social studies, mathematics, science, practical arts, physical education, music, fine arts, and foreign language, which is English. The Korean national curriculum is revised periodically to meet the demands of an always changing world (Mullis et al., 2012).

Students take mathematics courses from 1st grade through 10th grade. In each grade, mathematics courses are organized as two levels, and each level has six content domains. Students take one level of mathematics for a period of a semester and, if students fail at one level, they have to take additional courses to move on to the next level. Instructional time is 40 minutes in elementary schools, 45 minutes in middle schools, and 50 minutes in high schools. An average Korean middle school student takes 136 hours of mathematics courses in a year. In middle schools, student evaluations are conducted at the end of each semester and students need to reach a certain level in mathematics and English to be able to move to the next level. Students who do not meet the certain level must take supplementary courses before moving to the next level (Mullis et al., 2012).

Singapore

The Ministry of Education is the responsible agency for all educational activities in Singapore, and defines the goal of education as helping every child to realize their full potential, develop passion to learn, and be good citizens for their community and country. Singapore implements a national curriculum and offers education freely in public schools. The language of instruction is English, but Malay, Chinese (Mandarin), and Tamil are also offered as content domains based on the mother tongue of classrooms' populations. Primary school is 6 years and mandatory. Students take a standardized exam, Primary School Leaving Examination (PSLE), in mathematics, English, their mother tongue, and science after primary school. Students use their scores from the PSLE exam to guide their secondary education (Mullis et al., 2012).

Secondary education is not mandatory in Singapore, but it is highly common for students to go on to secondary schools. There are three different course of studies in secondary schools: Express, Academic, and Technic courses of studies. Secondary school takes about 4-5 years based on the course of study students select (Mullis et al., 2012).

Mathematics is a compulsory course from 1st grade through 10th grade in Singapore and students may take additional math courses by selecting electives. A single curriculum framework is used in all grade levels, but there is minor differentiation on emphases at every level. The common curriculum framework consists of five components, these being concepts, skills, processes, metacognition, and attitudes. The curriculum frameworks give directions to teachers about teaching, learning, and assessment strategies. Teachers are encouraged to use mathematical tools such as

calculators, graphing software, dynamic geometry software, and spreadsheets from grade 5 on (Mullis et al., 2012).

Student assessment, both formal and informal, starts from primary school in Singapore. Starting from grade 3, students are assessed by at least two summative assessments each year, one at the end of each semester. Oral presentations, written tests, and portfolios are usually used as formative assessment tools in Singaporean schools. Students also take national standardized exams at the end of their final year of primary, secondary, and pre-university education (Mullis et al., 2012).

United States

In the United States, a decentralized curriculum is used rather than a national curriculum. Every state develops a curriculum framework and oversees the implementation of curriculum standards. School districts and sometimes individual schools decide which curriculum to use. The language of instruction is English, and school from kindergarten through 12th grade is publicly funded, but students do not have to go to school through to the end of the 12th grade; the requirement for compulsory education changes from state to state. Primary education usually takes five years, middle school takes three years, and high school takes four years in the United States (Mullis et al., 2012).

Even though the mathematics curriculum differs from state to state, emphasis is on mastering basic skills or procedures, understanding concepts or principles, and applying mathematics in real-life contexts. The Common Core Standards for Mathematical Practice suggests that eighth-grade teachers use instructional time to focus on three areas:

- Formulating and reasoning about expressions and equations, including solving linear equations;
- Grasping the concept of a function and using functions to describe quantitative relationships; and
- Analyzing two- and three-dimensional space and figures using distance, angle measure, similarity, and congruence. (Mullis et al., p. 983)

Mathematics curriculum frameworks include instructional benchmarks for each grade, learning strategies, and resources for instruction. Individual school districts or schools are responsible for providing instructional resources in schools. There are also no set textbooks in the United States. Multiple private companies produce textbooks in each state and states publish a list of eligible textbooks. School districts and schools choose textbooks from the state approved list. Besides local school districts and state agencies, there are national campaigns, such as Educate to Innovate, and private companies which provide funds and resources in order for students to improve their skills and motivation in STEM courses (Mullis et al., 2012).

Student assessment is conducted annually in the United States. The No Child Left Behind Act (NCLB) of 2001 requires states to assess students every year to see whether schools are making enough yearly progress toward proficiency benchmarks. Elementary school, middle school, and high school students take standardized tests in all states, but these tests do not have high stakes for students, as they are more important for schools and school districts. If students do not meet certain benchmarks, schools face interventions. There are some standardized tests that have high stakes for students, as well. These exams are the Scholastic Assessment Test (SAT) and the American College

Test (ACT). These exams are conducted throughout the United States, and results of these assessments are used in the process of undergraduate admissions (Mullis et al., 2012).

Turkey

The National Ministry of Education is the responsible agency for all educational activities in Turkey. The official language of instruction is Turkish and education from 1st grade through higher education is free. Turkey implements a centralized curriculum and textbooks where all schools use the same curriculum and all students have the same textbooks. The basic structure of education was primary education between 1st grade and eighth-grade, and secondary education between 9th grade and 12th grade, in 2011, and only primary education was compulsory (Mullis et al., 2012). However, the basic education structure of Turkey recently changed to four years of primary education, four years of middle school, and four years of high school education (also called 4+4+4). The coverage of compulsory education has been extended to cover high school as well in this recent change.

The goal of the mathematics curriculum is to educate students to use mathematics in their lives, solve problems, share their work, and enjoy learning mathematics. There are five content domains in the middle school mathematics curriculum, these being numbers, geometry, algebra, measurement, and probability and statistics. The length of each lesson is 40 minutes in middle school, and mathematics is taught in four lessons per week. Required resources for teaching and learning are provided by the state for free to all students (Mullis et al., 2012).

The National Ministry of Education conducts student assessment and students take two standardized tests before higher education. The first standardized test is the Student Achievement Level Examination, implemented at the end of middle school (eighth-grade). Students use the results of this test to enter high schools. Even though high school is mandatory and free, there are different kinds of high schools that eventually make it easier for students to get into a prestigious university. The other standardized exam is the university entrance exam that is offered to 12th grade students and high school graduates. This exam is only offered once in a year and it is very common for students to take the university entrance exam more than once. Students take many supplementary courses and buy resources to prepare for these tests (Mullis et al., 2012).

Variables

The study used teachers' testing frequency variable as the criterion variable (independent variable), three socioeconomic variables as control variables, and students' mathematics achievement scores (1st to 5th plausible values) as the dependent variable.

Testing Frequency Variable in TIMSS

Testing frequency data came from the teacher survey of the TIMSS 2011 administration. The teacher questionnaire consists of 30 main questions and more than 30 sub-questions. Teachers of TIMSS participant students were asked to complete the teacher questionnaire. It was estimated that it takes about 45 minutes for teachers to complete the questionnaire. Teacher data for this study came from sub-questions of the 19th question (TQM-19K). TIMSS used codes to identify each question, including sub-questions, and BTBM19K was used to identify teachers' quiz frequency variable. TIMSS

uses its own distinctive terminology for testing frequencies: every day or almost every day (daily quizzes), about half of the lessons (weekly quizzes), some lessons (monthly quizzes), and never (no quizzes). For the purposes of this study, I have in some cases replaced the TIMSS's testing frequency terms with the terms that are common in the literature.

Testing frequency variable is derived from the 19th teacher questionnaire: "In teaching mathematics to this class, how often do you ask students to take a written test or quiz?" Teachers needed to respond to this question with:

Every day or almost every day = 1

About half the lessons = 2

Some lessons = 3

Never = 4

The testing frequency variable was entered as a categorical variable for the IDB Analyzer and this categorical variable was dummy coded to enable comparison of the relationship between different testing frequencies and student achievement. The *Almost everyday* testing frequency was selected as the reference level (constant) in regression analyses.

SES Variables in the TIMSS

Socioeconomic status (SES) is probably the most widely used contextual variable in education research (Sirin, 2005). Though it is widely regarded as a "critical consideration" in education research, there is little consensus on its use or its definition (Chudgar et al., 2014; NCES, 2012). A panel was convened to reach a consensus on the definition of SES and the use of SES variables for The National Assessment of Educational Progress (NAEP) assessment program. Experts who participated in this panel

defined SES as access to financial, social, cultural, and human capital resources (NCES, 2012). Student socioeconomic status was found to be positively correlated with student achievement (Haveman & Wolfe, 2008). Moreover, SES was found to be the most important contextual factor that affects student achievement (Caponera & Losito, 2016). However, socioeconomic differences are still a challenge for many developed and underdeveloped countries (Bowden & Doughney, 2010). In international comparative assessment programs, students' SES are usually determined by their home background and resources – such as having computer at home, Internet connection, their own room, a study desk, and the number of books, as well as the spoken language at home – that affect student learning ((Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009).). Parents' education level and home resources are found to be strong predictors of student achievement (Caponera & Losito, 2016; Topcu, Erbilgin, & Arikan, 2016). This present study selected the following variables as indicators of family SES: home possession (number of books at home) and parents', specifically the fathers' education level. The study also used school areas' income levels as indicators of students' SES.

TIMSS 2011 student surveys and school surveys, filled out by school principals, included several questions meant to collect data on students' home and school backgrounds. These questions serve as indicators of students' SES. Caponera and Losito (2016) used students' home environments, including the parents' educational levels, home resources for study, and the number of books they have at home as variables to indicate students' SES. This study also used similar home background variables, as well as schools' SES indicator variables, as control variables. The study used two student

variables, retrieved from student questionnaires, and one school-based variable from the school questionnaires.

While both parents' education levels influence students' achievement, this study only used the fathers' education level because fathers' and mothers' education levels are correlated (see Appendix C) with each other, which may cause multicollinearity issues in the regression model. The study selected fathers' education level over the mothers' because the fathers' education level is a more important factor than the mothers' in students' achievement (Ermisch & Pronzato, 2010; Serafino and Tonkin, 2014). Ermisch & Pronzato (2010) note that it is possible that better educated mothers spend more time at their jobs, but less time with their kids. Furthermore, Wang (2007) has claimed that self-concept directly affects student achievement, and Janjetovic and Malinic (2004) found that a father's educational level is a more important factor than a mother's in improving students' self-concept in mathematics and science in particular. Moreover, the bigger impact of the fathers' education level, as opposed to the mothers' on student achievement, would be in the form of income inequality. Dickler (2016) has found that a man earns more money than a woman for doing the same job. Income levels of families have been found to affect students' achievement significantly because family income level determines the neighborhood they live in and the schools they attend, which each affect available resources for students succeed (Sirin, 2005).

The selected three SES variables were represented as following in TIMSS questionnaire;

1. About how many books are there in your home? (Do not count magazines, newspapers, or your schoolbooks). This question was the 4th question on the student

questionnaire and students had to pick one of the below choices.

None or very few (0–10 books) = 1

Enough to fill one shelf (11–25 books) = 2

Enough to fill one bookcase (26–100 books) = 3

Enough to fill two bookcases (101–200 books) = 4

Enough to fill three or more bookcases (more than 200) = 5

2. What is the highest level of education completed by your father or stepfather or male guardian? This question is derived from the 6th question of the student survey. Students needed to select one of the below options:

Some ISCED Level 1 or did not go to school = 1

ISCED Level 2 = 2

ISCED Level 3= 3

ISCED Level 4 = 4

ISCED Level 5B = 5

ISCED Level 5A, first degree = 6

Beyond ISCED Level 5A, first degree = 7

I don't know = 8

ISCED stands for International Standard Classification of Education, and Appendix D summarizes the ISCED school levels that are defined by UNESCO (2011).

3. Which best characterizes the average income level of the school's immediate area? This variable came from the 5th question of the school survey that was filled out by principals. School principals selected one of the below options:

High = 1

Medium = 2

Low = 3

The numbers on the right side of the equations represent the order of each sub-category. These numbers did not exist on the surveys, but were given to these categories by the IDB Analyzer to enable the system to compute statistical computations.

Procedure

The TIMSS 2011 utilizes a matrix-sampling design that involves creating an assessment pool of mathematics questions. Then 14 student achievement booklets are created from the assessment pool for eighth-grade mathematics, but each student completes only one booklet. TIMSS sorts test items as item blocks, in that each block includes 12-18 items at eighth-grade in the creation of test booklets. In each item block, content and cognitive domain items match the distribution across the item pool overall. Eight of the 14 booklets that were used in the TIMSS 2011 were adapted from the TIMSS 2007, and 6 more booklets were created specifically for the TIMSS 2011. As mentioned above, each student completes only one student test booklet, followed by a student questionnaire. Test booklets took about 90 minutes and student questionnaires about 30 minutes to complete for eighth-graders (Mullis, et al. 2012).

Country representatives worked very carefully to plan and document every procedure, cooperate with TIMSS, and deploy standardized procedures to ensure the quality and comparability of the data. To ensure high quality and comparability of TIMSS 2011 exam questions and questionnaires, the organization committee used a series of verification checks before implementing them. TIMSS 2011 also used detailed

documentation procedures to meet the sampling standards, and an ambitious quality assurance program to monitor the data collection activities (Johansone, 2016).

Data Analysis

All of the data for the eighth-grade mathematics students who were eligible to take the TIMSS 2011 exam, as well as those students' teachers' and schools' data, were used in this study. TIMSS 2011 eighth-grade student achievement scores, SES data, and teachers' quiz frequency data were obtained through IEA's International Database Analyzer (IDB Analyzer) software. IEA developed this software to simplify TIMSS data analysis procedures for users and enable users to analyze all variables for all participating education systems (NCES, 2016). The IDB Analyzer enables researchers to create appropriate datasets with the help of SPSS and then make accurate analyses. The IDB Analyzer has two modules: (1) the merge module and (2) the analyze module. In the merge module, countries and variables of interest were selected to create a base dataset for the analyze module. In the analyze module, the desired statistical method (regression, in this case) was selected first, and then independent, control, and dependent variables were moved to the designated box by using the base dataset that was created in the merge module. Then the analyze module created a new dataset that was ready to be analyzed through SPSS. For the first research question, the researcher used all countries in the data analysis; for the second research question, however, the researcher only used Korea, Singapore, Turkey, and the United States in the analysis.

All SES variables, as well as the testing frequency variable, were marked as categorical variables in TIMSS, but they were converted to dichotomous variables

through dummy coding by the researcher via the IDB Analyzer Analyses module. The IDB analyzer tool allows users to the reference category, or the group against users compare each of the other groups in a categorical variable and automatically creates dummy variables temporarily. Dummy variables get the same code as the original variable, plus _D#. D stands for Dummy coded and # represents the order number of the category. For example, I dummy coded testing frequency variable which has four categories (1 for everyday, 2 for half of the lessons, 3 for some lessons, and 4 for never) and I specified 1 as reference (constant) level. The IDB analyzer created BTBM19K_D2 (coded 1 for those with "half of the lessons", and zero otherwise), BTBM19K_D3 (coded 1 for those with "some lessons", and zero otherwise, and BTBM19K_D4 (coded 1 for those with "never", and zero otherwise). SPSS output window was checked to verify proper coding of categorical variables were achieved.

In the 1st control variable, "none or very few (0–10 books)", level 1 was selected as the reference level; in the 2nd control variable, "some ISCED Level 1 or did not go to school", level 1, was selected as the reference level; and in the last control variable, "low", level 3, was selected as the reference level.

Plausible values. As mentioned earlier, TIMSS creates 14 blocks of test questions and each student only take one block of questions. Plausible values are calculated to estimate students' academic performances based on their response patterns to one block of test questions and student questionnaire, as if students took all 14 blocks of test questions (Mullis & Martin, 2011). Plausible values are not used to obtain scores for each student, but rather to predict student performance based on responses to the test and student questionnaire through similar responses to the test and surveys in the sampled

population (Mullis & Martin, 2011). TIMSS generated five plausible value estimates through the IDB Analyzer plug-in for SPSS. Each plausible value is given a standard error by calculating the variability between them to indicate the level of error for each achievement mean (plausible value) for each student in the sample.

Mullis and Martin (2011) explain the benefit of using plausible values as assigning 14 blocks of questions to each student by giving them only one block of questions, which enables TIMSS to analyze more questions and reduce measurement error by calculated errors through five plausible values. Therefore, all five plausible values were used as dependent variables in this study. They were named 1st to 5th Plausible Values in the IDB Analyzer and entered as a continuous variable into the IDB Analyzer, and does not require dummy coding.

Organizing datasets: Datasets were created by using the IDB Analyzer and SPSS software packages. SPSS data files and the IDB Analyzer were downloaded from IEA's website. Using the merge module of the IDB Analyzer, country selection was made in the first step and then variables of the study were selected. Then, the created file was run through SPSS to create a base file for the analysis module of the IDB Analyzer. In the analyses module of the IDB Analyzer, math teacher weight (MTHWGT) was selected as the analysis type and regression analysis as the statistics type. Using plausible values option was also selected as this study used them as dependent variable. The final step was moving testing frequency variables and students' socioeconomic status variables into the independent variables box. International averages of student achievement scores (1st to 5th plausible values) were also moved to the dependent variable box. Categorical variables were dummy coded in the analysis module of the IDB Analyzer. After all steps were

completed, the new dataset was analyzed via SPSS to see if there was a significant relationship between testing frequency and the achievement scores of eighth-grade math students.

Research Question 1: To answer this research question, a multiple linear regression was conducted with all eighth-grade TIMSS 2011 participants with the testing frequency variable as the predictor variable, and the 1st to 5th plausible values (students' math achievement) variable as the outcome variables; the control variables were the number of books at home, fathers' education levels, and school areas' income levels.

Research Question 2: To answer this research question, a multiple linear regression was conducted for each of the following countries: Korea, Singapore, Turkey and the United States. The testing frequency variable was the predictor variable, the 1st to 5th plausible values (students' math achievement) variable was the dependent variable, and the control variables of the study were number of books at home, the fathers' education levels, and the school areas' income levels.

Statistical Assumptions for Multiple Regression

Several assumptions needed to be tested and satisfied in order to get correct results from the multiple regression analyses. Specifically, the following multiple regression assumptions: normality assumption, linearity assumption, homoscedasticity, and multicollinearity were checked and satisfied. Results of these assumption tests are provided in Appendix E.

Normality Assumptions: Multiple regressions analyses assume that variables in the model are normally distributed. Non-normally distributed variables may distort data and affect the relationship between variables and significance tests (Osborne & Waters,

2002). Normally distributed variables imply that residuals are normally distributed so that plots of values of residuals follow approximately a normal curve (Ballance, 2013). A normality assumption test can be done through statistical software packages by visual inspection of data plots, skew, kurtosis, and P-P plots. These plots provide information about the normal distribution of variables.

Linearity Assumption: The relationship between independent and dependent variables needs to be linear in order for multiple regressions to accurately estimate the relationship between them (Osborne & Waters, 2002). However, this requirement is not the case in every dataset, and especially in the social sciences. Therefore, researchers need to check data for a linear relationship between IVs and DVs. If the relationship between variables is not linear, regression analysis will not accurately estimate the relationship (Osborne & Waters, 2002). Most statistical packages test for linearity by providing residual plots and scatter plots.

Homoscedasticity: This assumption refers to a same variance of errors in all levels of independent variables (Osborne & Waters, 2002). If the variance of errors varies at different values of independent variables, then heteroscedasticity is observed in the data, which may distort analyses and increase the Type I error. An assumption check for homoscedasticity can be accomplished through visual inspection of a plot of the standardized residuals by the regression standardized predicted value (Osborne & Waters, 2002).

Multicollinearity: This assumption is also known as the collinearity assumption and it refers to having little or no correlation between independent (predictor) variables. If the multicollinearity assumption is violated, the result might be unusual and

misleading, standard errors will be inflated, and regression coefficients will have less power (Ballance, 2013). Tests of this assumption can be accomplished through statistical packages by selecting collinearity statistics or by running correlation analyses.

Assumption tests for this study were conducted through SPSS statistical software, as the IDB Analyzer does not provide assumption tests. Only the 1st plausible value variable was selected as a dependent variable, for only the purpose of assumption tests. Assumption test results indicated that the regression model of this study did not violate any of the multiple regression assumptions (see Appendix E).

Chapter IV: Results

This study investigated the relationship between testing frequency and eighth-grade students' mathematics achievement as measured by the TIMSS 2011 exam when controlled for students' socioeconomic status in all TIMSS 2011 participant countries. The testing frequency variable was derived from the TIMSS 2011 teacher questionnaire and the students' SES variables were derived from the TIMSS 2011 student and school questionnaires. The study also examined the relationship of testing frequency and the math achievement of 8th graders in four countries (Korea, Singapore, Turkey, and the United Sates) using the same variables.

Analyses

Multiple linear regression model was conducted to analyze the relationship between testing frequency and student achievement when controlling for students' home and school SES background variables. The IDB Analyzer in conjunction with SPSS tools was used to analyze the data. Five different regression analyses were conducted for this study; in the first regression analysis, all countries were selected in the dataset and the remaining four regression analyses were conducted for each of the four selected countries (Korea, Singapore, Turkey, and the United States).

It is important to note that SPSS alone was not enough to correctly analyze the data, as TIMSS data is constructed hierarchically, where students are nested in schools and schools are nested in countries, and linking codes and files are created to connect teachers with their students and schools. It is necessary for researchers to use the IDB Analyzer when analyzing TIMSS data because it is especially programmed to handle the TIMSS's hierarchical data structure.

It is also important to note that the SPSS output does not provide an ANOVA table when running datasets that are created through the IDB Analyzer in regression analyses (E. J. Gonzales, personal communication, October 27, 2016). Additionally, SPSS output also does not provide p-values in a coefficient table of regression analyses. However, significance tests can be calculated by using standard errors. IADB (2016) explains, "If the absolute value of the group difference divided by the standard error of the difference exceeds a t-value of 1.96, the result can be regarded as statistically significant on the 95% level" (p. 25). Therefore, this formula was used to test statistical significance.

Analyses for the First Research Question

Descriptive Statistics: Below is the descriptive data obtained from the regression analyses that included all TIMSS 2011 participating countries.

Table 4

Testing frequency variable's descriptive information of all TIMSS 2011 participant students

Testing Frequency	Number of Students	Percentages
Everyday or almost everyday	51,576	23%
About half the lessons	54,185	24%
Some lessons	116,417	52%
Never	1,646	.74%

Table 4 indicates that 52% of all students participating in the TIMSS 2011 received tests or quizzes as often as some lessons, while 23% of students received tests and quizzes almost every day. Fewer than 1% of students did not take any tests or quizzes, in all countries combined.

Table 5

Amount of books in your home variable's descriptive information of all TIMSS 2011 participant students

Number of Books	Number of Students	Percentages
0-10 books	45,083	20%
11-25 books	68,303	30%
26-100 books	59,232	26%
101-200 books	26,300	11%
More than 200 books	24,906	11%

Table 5 shows the percentages for the category of numbers of books that students have in their homes. Twenty percent of students had between 0-10 books, only 11% of students had more than 200 books in their home, and the biggest percentage (30%) of students had between 11-25 books in their homes.

Table 6

Fathers' educational level variable's descriptive information of all TIMSS 2011 participant students

Education Level	Number of Students	Percentages
ISCED level 1 or no school	22,233	10%
ISCED 2	26,833	11%
ISCED 3	47,417	21%
ISCED 4	18,814	8%
ISCED 5B	8,812	4%
ISCED 5A, first degree	27,466	12%
Beyond ISCED 5A, first degree	17,600	8%
I do not know	54,649	24%

Table 6 shows the students' fathers' highest educational levels. The largest portion, 24% of students, did not know their fathers' highest educational attainment, and 10% of students' fathers' educational level was primary education or below. Only 8% of students' fathers had bachelors or higher degrees, across all of the countries.

Table 7

Average income level of school area variable's descriptive of all TIMSS 2011 participant students

Income Level	Number of Students	Percentages
High	15,741	7%
Medium	133,070	59%
Low	75,013	34%

Table 7 indicates the average income levels of the schools' areas. While more than half of the students, 66%, lived in medium or high-income level areas, 34% of students went to school in low-income areas.

Regression results: A multiple regression analysis was computed to reveal the relationship between testing frequency and student achievement in eighth-grade mathematics when controlling for two student and one school SES variable in all countries combined. Table 8 indicates the results of the regression analyses. Testing frequency and SES variables explained 18% of the variance ($R^2 = .18$) in student achievement in all countries combined. Table 8 also shows coefficient estimates of testing frequency and SES variables. In all countries combined, the constant score was 394.34 when students took tests or quizzes almost everyday. Students' scores increased as testing frequency decreased, but the students' average dropped when testing was not used at all. Even though students' achievement increased by 3.06 points when testing was used about half of the lessons, and by 1.35 points when testing was used once in a month or several times in a semester, these increases were not statistically significant at the 95% level. Student achievement decreased by -7.39 points when testing frequency was never. This decrease is not a statistically significant decrease in student achievement (see Table 8).

Table 8

Multiple regression analyses predicting eighth-grade mathematics achievement from testing frequency and SES variables in all countries combined

Variable Name	Category Name	*B*	*SE B*	B	t-value
	Constant	394.34	3.52		112.13
Testing Frequency	About half the lessons	3.06	2.30	.01	1.33
	Some lessons	1.35	1.58	.00	.86
	Never	-7.39	4.69	.00	-1.58
Number of Books	11-25 books	19.36	.83	.10	23.33*
	26-100 books	44.41	.89	.23	49.73*
	101-200 books	56.10	1.13	.21	49.52*
	More than 200 books	53.72	1.33	.21	40.44*
Fathers' Educational Level	ISCED 2	9.87	2.91	.03	3.39*
	ISCED 3	25.39	2.86	.12	8.89*
	ISCED 4	39.41	3.26	.11	12.09*
	ISCED 5B	47.09	3.97	.11	11.87*
	ISCED 5A, first degree	57.01	2.95	.20	19.31*
	Beyond ISCED 5A, first degree	63.22	3.19	.17	19.83*
	I do not know	14.64	2.88	.08	5.09*
School Areas Average Income	Medium	16.93	1.26	.09	13.40*
	High	41.93	2.65	.09	15.80*

Note: R-Square = .18 (.00), Adjusted R-Square = .18 (.00)**
* Indicates statistical differences
** Numbers in parentheses are standard error of estimates

Table 8 also shows all the variables with sub-groups in the model. Results of the coefficient table show that, as the number of books at home increased, student achievement also increased. The same patterns were observed for fathers' educational levels and school areas' income levels. These increases were statistically significant for every sub-group of each of the SES variables.

Analyses for the Second Research Question

The following analyses were conducted through multiple regression analyses for each country separately, and results are reported for each country under different sections.

Korea

Descriptive Statistics: In Korea, 5,166 students representing a total of 593,779 eighth-grade Korean students took the TIMSS 2011 exam and filled out student questionnaires, but only 4,859 students' responses were valid for this study.

Table 9

Descriptive information of IV and control variables for Korean students

Variables	Valid N	Total Estimated N	Mean	Std. Deviation
Testing frequency	4,859	593,779	2.58	.62
Amount of books	4,859	593,779	3.62	1.26
Fathers' Edu. Level	4,859	593,779	5.37	2.06
Schools' income level	4,859	593,779	2.17	.66

Results indicate that the mean of students' test/quiz taking frequency was 2.58, which means students took tests/quizzes between *about half the lessons* and *some lessons*. On average, the amount of books that Korean students had in their homes was 3.61, which is between *26-100* and *101-200* books. The mean of Korean students' fathers' highest education level was 5.37, which is above the first stage of occupational education. Finally, the mean of the schools' average income levels is 2.17, which means slightly above a *medium* income level (see Table 9).

Table 10

Testing frequency variable's descriptive information of Korean students

Testing Frequency	Number of Students	Percentages
Everyday or almost everyday	338	7%
About half the lessons	1,401	29%
Some lessons	3,091	64%
Never	29	.60%

Table 10 indicates the percentages of students that received tests/quizzes at different

frequencies. More than half of the Korean students, 64%, took tests or quizzes rarely, in *some lessons*, 7% took them *almost everyday,* and only .60% (29) of students did not take tests or quizzes at all. The student-teacher ratio was 24:1 in Korea (see Table 10), which means that only about one teacher did not use tests or quizzes in his/her classroom. Because of the low number of students that fell into the *never* quiz frequency group, students in the *never* quiz frequency group were excluded from the results.

Table 11

Amount of books in your home variable's descriptive information of Korean students

Number of Books	Number of Students	Percentages
0-10 books	412	8%
11-25 books	496	10%
26-100 books	1,211	25%
101-200 books	1,166	24%
More than 200 books	1,574	32%

Table 11 shows the information about Korean students' book possession in their homes. The highest percentage of students, 32%, had *more than 200 books* in their home, while only 8% of students only had *0-10 books* in their homes.

Table 12

Fathers' educational level variable's descriptive information of Korean students

Education Level	Number of Students	Percentages
ISCED level 1 no school	66	1%
ISCED 2	113	2%
ISCED 3	1,516	31%
ISCED 4	0	0
ISCED 5B	301	6%
ISCED 5A, first degree	1,356	28%
Beyond ISCED 5A, first degree	420	9%
I do not know	1,087	22%

Table 12 presents information about the Korean fathers' highest education levels. Only 1% of students' fathers' highest education level was primary school or less, and 22% of

Korean students did not know their fathers' education level.

Table 13

Average income level of school area variable's descriptive information of Korean students

Income Level	Number of Students	Percentages
High	689	14%
Medium	2,663	55%
Low	1,507	31%

Table 13 shows information about Korean students' school areas' average income levels. While more than half of the students (55%) went to school in medium-income level areas, 31% of students went to school in low-income areas, and 14% of students went to school in high-income area schools.

Regression results: A multiple regression analysis was computed to reveal the relationship between testing frequency and student achievement in eighth-grade math courses when controlling for SES variables in Korea. Table 14 indicates the results of the regression analyses. Testing frequency and SES variables explained 22% of the variance ($R^2 = .22$) in student achievement in Korea. Table 14 also shows coefficient estimates of testing frequency and SES variables. In Korea, the mathematics achievement constant (intercept) was 514.79, which indicates the score for when students take tests or quizzes *everyday or almost everyday*. Students' scores increased as testing frequency decreased in Korea. Students' achievement increased by 10.63 points when testing was used in *about half of the lessons* and 11.50 points when testing was used once in a month or several times in a semester. Despite these high achievement differences between testing frequencies, none of the differences are statistically significant at a 95% level.

Table 14

Multiple regression analyses predicting eighth-grade mathematics achievement from testing frequency and SES variables in Korea

Variable Name	Category Name	*B*	*SE B*	B	t-value
	Constant	514.79	14.77		
Testing Frequency	About half the lessons	10.63	9.60	.05	1.11
	Some lessons	11.50	9.81	.06	1.17
Number of Books	11-25 books	9.31	6.81	.03	. 1.37
	26-100 books	43.46	5.29	.21	8.22*
	101-200 books	57.03	5.30	.32	12.65*
	More than 200 books	85.30	5.49	.44	15.55*
Fathers' Educational Level	ISCED 2	2.40	14.31	.00	.17
	ISCED 3	15.23	10.49	.08	1.45
	ISCED 5B	23.22	11.96	.06	1.94
	ISCED 5A, first degree	49.78	10.87	.25	4.58*
	Beyond ISCED 5A, first degree	40.19	12.52	.12	3.21*
	I do not know	3.34	11.57	.02	.29
School Areas Average Income	Medium	8.40	6.03	.12	4.98*
	High	30.01	3.70	.05	2.27*

Note: R-Square = .22 (.02)**, Adjusted R-Square = .22 (.02)**
* Indicates statistical differences
** Numbers in parentheses are standard error of estimates

Table 14 also shows all the SES variables with sub-groups in the model. Results of the coefficient table show that, as the number of books at home increased, student achievement also increased. All the differences in sub-groups of the number of books are statistically significant except in the *11-25 books* sub-group. The same pattern was observed for fathers' educational levels and school areas' income levels. Only *ISCED Level 5B* and *ISCED Level 5A* have statistically significant differences with the reference level in the fathers' highest educational level variable. Regarding the school areas' income levels variable, only high-income level school areas have statistically significant differences with the reference level, the low-income level.

Singapore

Descriptive Statistics: Descriptive data in Tables 15-Table 20 was obtained from regression analyses that included only Singaporean students' data. In Singapore, 5,927 students representing a total of 47,764 eighth-grade Singaporean students participated in the TIMSS 2011, taking an exam and filling out the student questionnaire. The number of valid students was 5,641 for this study.

Table 15

Descriptive information of IV and control variables for Singaporean students

Variables	Valid N	Total Estimated N	Mean	Std. Deviation
Testing frequency	5,641	47,764	2.67	.54
Amount of books	5,641	47,764	2.82	1.20
Fathers' Edu. Level	5,641	47,764	5.24	2.44
Schools' income level	5,641	47,764	2.12	.52

Results indicate that the mean of students' test/quiz-taking frequency was 2.67 in Singapore, which is between about half the lessons and some lessons, but closer to some lessons in frequency. The number of books that the average Singaporean house had was between 26 and 200 books. The average eighth-grade Singaporean students' fathers' highest education level was 5.24, which is just above first stage of occupational education, and the mean of schools' average income levels was 2.12, which is just above the medium-income area schools (see Table 15).

Table 16

Testing Frequency variable's descriptive information of Singaporean students

Testing Frequency	Number of Students	Percentages
Everyday or almost everyday	177	3%
About half the lessons	1,492	26%
Some lessons	3,955	70%
Never	17	.30%

Table 16 represents the percentages of students who received tests/quizzes at different frequencies. A significant percentage of students, 70%, took tests or quizzes very rarely, in *some lessons*, only 3% took them *almost everyday,* about a quarter of them, 26%, took them in *about half the lessons,* and only 17 (.30%) students did not take any tests or quizzes. The student-teacher ratio was 17 in Singapore in 2011(see Table 16), which indicates that only one teacher did not use tests or quizzes in his/her classroom. Because of the low number of students falling into the *never* quiz frequency group, this group was excluded from further analysis.

Table 17

Amount of books in your home variable's descriptive information of Singaporean students

Number of Books	Number of Students	Percentages
0-10 books	908	16%
11-25 books	1,429	25%
26-100 books	1,853	33%
101-200 books	797	14%
More than 200 books	654	12%

Table 17 shows the information regarding Singaporean students' book possession in their homes. The highest percentage of students, 33%, had *more than 200 books* in their home, and 12% of students only had *0-10 books* in their homes.

Table 18

Fathers' educational level variable's descriptive information of Singaporean students

Education Level	Number of Students	Percentages
ISCED level 1 or no school	523	9%
ISCED 2	332	6%
ISCED 3	979	17%
ISCED 4	491	9%
ISCED 5B	443	8%
ISCED 5A, first degree	593	11%
Beyond ISCED 5A, first degree	498	9%
I do not know	1,782	32%

Table 18 presents information on the Singaporean students' fathers' highest education levels. Almost one out of three (32%) students did not know their fathers' educational level in Singapore, and 9% of students' fathers' highest education reported a primary education level or lower.

Table 19

Average income level of school area variable's descriptive information of Singaporean students

Income Level	Number of Students	Percentages
High	472	8%
Medium	3,951	70%
Low	1,218	22%

Table 19 represents information about students' school areas' average income levels in Singapore. Almost three quarters of the students, 70%, attended schools in medium-income level areas, and almost a quarter of them, 22%, attended in low-income school areas. Only 8% of students attended high-income area schools in Singapore.

Regression Results: An additional multiple regression analysis was computed to see the relationship between testing frequency and student achievement in eighth-grade mathematics courses when controlling for SES variables in Singapore. Table 20 indicates

the results of the regression analyses. Testing frequency and SES variables explained 19% of the variance ($R^2 = .19$) in student achievement in Singapore. Table 20 presents the coefficient estimates of testing frequency and SES variables in Singapore. The mathematics achievement constant (intercept) was 534.04, which indicates the score when students took tests or quizzes *everyday or almost everyday.* There was no linear relationship between testing frequency and student scores in Singapore. Students' scores decreased by -7.35 points when testing frequency became *about half the lessons*, but student scores increased by 5.64 points when testing frequency became *some lessons*. These increases or decreases by change in testing frequency were not statistically significant differences at a 95% level.

Table 20

Multiple regression analyses predicting eighth-grade mathematics achievement from testing frequency and SES variables in Singapore

Variable Name	Category Name	*B*	*SE B*	B	t-value
	Constant	534.04	17.05		
Testing Frequency	About half the lessons	-7.35	13.57	-.04	-.54
	Some lessons	5.64	13.84	.03	.41
Number of Books	11-25 books	25.63	4.71	.13	5.44*
	26-100 books	51.49	5.53	.29	9.31*
	101-200 books	53.96	6.20	.23	8.70*
	More than 200 books	56.80	6.61	.22	8.59*
Fathers' Educational Level	ISCED 2	-6.43	6.32	-.02	-1.02
	ISCED 3	13.45	4.76	.06	2.82*
	ISCED 4	10.53	5.06	.04	2.08*
	ISCED 5B	27.53	5.52	.09	4.99*
	ISCED 5A, first degree	33.01	6.85	.12	4.82*
	Beyond ISCED 5A, first degree	42.64	6.75	.15	6.32*
	I do not know	-3.29	5.03	-.02	-.65
School Areas Average Income	Medium	28.57	9.55	.15	5.43*
	High	59.01	10.87	.20	2.99*

Note: R-Square = .19 (.02)**, Adjusted R-Square = .19 (.02)**
* Indicates statistical differences
* *Numbers in parentheses are standard error of estimates

Table 20 also shows all the variables with sub-groups in the model. Results of the coefficient table show that, as the number of books at home increased, students' achievement also increased, and all the differences in sub-groups regarding the number of books were statistically significant. The same pattern was observed for school areas' income levels, and these differences were also statistically significant. However, the fathers' educational level variable did not have a perfect linear relationship with student achievement. Student scores decreased by -6.43 points when fathers' education level increased to *ISCED level 2*, and by -3.29 points when students *do not know* their fathers' education level, but both of these decreases were not statistically significant. Student achievement increased when fathers' education level get to *ISCED levels 3,4,5B, 5A,* and *beyond 5A*, and all of these increases were statistically significant. Student achievement also increased by the increase of school areas' income levels, and these increases were statistically significant increases.

Turkey

Descriptive Statistics: Results in the following tables were obtained from regression analyses that included only Turkish students' data. Table 21 indicates that 6,625 students' responses were valid for this study, but the total number was 6,928 to represent a total of 1,108,775 eighth-grade Turkish students who took the TIMSS 2011 exam and filled out student surveys.

Table 21

Descriptive information of IV and control variables for Turkish students

Variables	Valid N	Total Estimated N	Mean	Std. Deviation
Testing frequency	6,625	1,108,775	2.11	.84
Amount of books	6,625	1,108,775	2.48	1.11
Fathers' Edu. Level	6,625	1,108,775	2.36	1.91
Schools' income level	6,625	1,108,775	2.50	.52

The mean of students' test/quiz-taking frequency was 2.11 in Turkey, which means that Turkish students took tests or quizzes in about half of the lessons. The mean for the number of books Turkish students have at home was 2.48, which is between 11-25 books and 26-100 books. The mean for Turkish students' fathers' highest education levels was 2.36 in 2011. It is just above ISCED Level 2, which is above middle school, and the mean of the schools' average income levels was 2.50, which is in the middle of medium and low-income area schools.

Table 22

Testing Frequency variable's descriptive information for Turkish students

Testing Frequency	Number of Students	Percentages
Everyday or almost everyday	2,048	31%
About half the lessons	1,875	29%
Some lessons	2,702	41%
Never	0	0

Table 22 represents the percentages of Turkish students who received tests/quizzes at different frequencies. Almost one out of three students took tests or quizzes *almost everyday*, 29% took them in *about half the lessons*, 41% of students took them in *some lessons*, and there was no student who did not take any tests or quizzes at all in 2011. Therefore, the *never* testing frequency group was excluded from the results.

Table 23

Amount of books in your home variable's descriptive information for Turkish students

Number of Books	Number of Students	Percentages
0-10 books	1,251	19%
11-25 books	2,476	37%
26-100 books	1,815	27%
101-200 books	672	10%
More than 200 books	411	6%

Table 23 shows information about the average number of books that Turkish students had at home. More than half of the Turkish students (56%) had less than 25 books, 27% had between *26-100 books*, and only 6% of Turkish students had *more than 200 books* at their home.

Table 24

Fathers' educational level variable's descriptive information for Turkish students

Education Level	Number of Students	Percentages
ISCED level 1 or no school	3,383	51%
ISCED 2	940	14%
ISCED 3	1,355	20%
ISCED 4	0	0
ISCED 5B	279	4%
ISCED 5A, first degree	331	5%
Beyond ISCED 5A, first degree	91	1%
I do not know	246	4%

Table 24 indicates information on Turkish students' fathers' highest education levels. Half of the Turkish students fathers' highest education levels was listed as *ISCED Level 1* or below. The fathers' highest education level in this group was either lower secondary education or primary education. ISCED Level 1 also includes no school level. Only 5% of students' fathers had their bachelor and graduate school degree, and 4% of Turkish students did not know their fathers' education level.

Table 25

Average income level of school area variable's descriptive information for Turkish students

Income Level	Number of Students	Percentages
High	55	1%
Medium	3,070	46%
Low	3,500	53%

Table 25 presents data on Turkish students' school areas' average income levels. More than half, 53%, of students attended schools in low-income area schools, 46% attended in medium-income area schools, and only 1% of Turkish students attended schools in high-income areas.

Regression results: A multiple regression analysis was conducted with only Turkish students' data to analyze the relationship between testing frequency and student achievement in eighth-grade mathematics courses when controlling for SES variables. Table 26 indicates the model statistics of the regression analysis. The model statistics table indicates that testing frequency and SES variables explain 25% of the variance (R^2 = .25) in eighth-grade students' mathematics achievement in Turkey. Regression coefficient estimates for testing frequency and SES variables are presented in Table 26. The mathematics achievement constant (intercept) was 387.56, which indicates the score when students took tests or quizzes *every day or almost every day.* Students' achievement scores decreased as testing frequency decreased, which indicates a good linear relationship between testing frequency and math achievement in Turkey. Students' scores decreased by -12.27 points when testing frequency became *about half the lessons* and by -10.76 points when testing frequency became *some lessons.* These decreases by change in testing frequency are not statistically significant differences at a 95% level.

Table 26

Multiple regression analyses predicting eighth-grade mathematics achievement from testing frequency and SES variables in Turkey

Variable Name	Category Name	*B*	*SE B*	β	t-value
	Constant	387.56	7.29		
Testing Frequency	About half the lessons	-12.27	8.31	-.05	-1.48
	Some lessons	-10.76	7.86	-.05	-1.37
Number of Books	11-25 books	38.91	4.46	.17	8.72*
	26-100 books	70.88	5.22	.28	13.57*
	101-200 books	87.44	6.47	.24	13.51*
	More than 200 books	94.66	10.60	.21	8.93*
Fathers' Educational Level	ISCED 2	-4.15	4.03	-.01	-1.03
	ISCED 3	26.28	3.70	.09	7.09*
	ISCED 5B	66.68	8.63	.12	7.72*
	ISCED 5A, first degree	89.92	8.50	.18	10.58*
	Beyond ISCED 5A, first degree	113.46	26.11	.13	4.34*
	I do not know	-48.50	8.33	-.08	-5.82*
School Areas Average Income	Medium	27.53	5.20	.12	5.30*
	High	36.24	19.61	.03	1.85

Note: R-Square = .25 (.03)**, Adjusted R-Square = .25 (.03)**
* Indicates statistical differences
* *Numbers in parentheses are standard error of estimates

Table 26 also shows all coefficient estimates for SES variables and their sub-groups in the model. Results of the coefficient table show that, as the number of books at home increased, students' achievement also increased, and all of the differences in sub-groups regarding the number of books at home are statistically significant in Turkey. Similar results are shown in fathers' highest educational levels. Student achievement usually increased by the fathers' educational level, except for in the *ISCED Level 2* and *I do not know* groups; student achievement decreased for these two groups, and all of these increases and decreases are statistically significant except for the *ISCED Level 2*. Student achievement also increased by school areas' income levels. Student achievement

increased by 27.53 points when school areas income levels reached medium level, and this increase is statistically significant. Students' math achievement increased by 36.24 points when school areas income levels became high, but this increase is not statistically significant at a 95% level.

United States

Descriptive Statistics: Descriptive data in the below tables was obtained from the final regression analyses that included only American students' data. In the United States, 10,477 students representing a total of 2,229,781 eighth-grade American students took the TIMSS 2011 exam and filled out student questionnaires, but only 7,025 students' responses were valid for this study.

Table 27

Descriptive information for IV and control variables of American students

Variables	Valid N	Total Estimated N	Mean	Std. Deviation
Testing frequency	7,025	2,229,781	2.43	.75
Amount of books	7,025	2,229,781	2.95	1.30
Fathers' Edu. Level	7,025	2,229,781	5.69	2.26
Schools' income level	7,025	2,229,781	2.30	.60

The mean of students' test/quiz-taking frequency was 2.43 in the United Sates, which is between about half of the lessons and some lessons in terms of testing frequencies. The mean for the number of books American students had at home was 2.95, which is almost 26-100 books. The mean for American students' fathers' highest education levels was 5.69, which is about the first stage of the occupational education level. The mean for

school areas' average income levels was 2.30, which is between medium and high-income levels, but closer to the medium-income level (see Table 27).

Table 28

Testing frequency variable's descriptive information of American Students

Testing Frequency	Number of Students	Percentages
Everyday or almost everyday	1,163	16%
About half the lessons	1,733	25%
Some lessons	4,104	58%
Never	25	.36%

Table 28 shows the information regarding American students' test or quiz receiving frequency in numbers and percentages. More than half, 58% of the students, took tests or quizzes in *some lessons,* 16% of them took tests or quizzes *almost every day*, every one out of four students took them in *about half the lessons*, and only 25 (.36%) students did not take any tests or quizzes in the United States. The student-teacher ratio was 14 in the United States in 2011(see Table 28), which indicates that only about 1-2 teachers did not use tests or quizzes in their classrooms. Because of the low number of students in the *never* quiz frequency group, the *never* testing frequency group was excluded from the results analyses.

Table 29

Amount of books in your home variable's descriptive information of American students

Number of Books	Number of Students	Percentages
0-10 books	1,111	16%
11-25 books	1,605	23%
26-100 books	2,017	29%
101-200 books	1,187	17%
More than 200 books	1,105	16%

Table 29 indicates the information about American students' possession of books in their homes. Almost half, 49%, of American students had less than 25 books, about one out of

three students, 29%, had *26-100 books,* and 16% of students had *more than 200 books* in their homes.

Table 30

Fathers' educational level variable's descriptive information of American Students

Education Level	Number of Students	Percentages
ISCED level 1 or no school	175	2%
ISCED 2	534	8%
ISCED 3	1,379	20%
ISCED 4	307	4%
ISCED 5B	220	3%
ISCED 5A, first degree	1,227	17%
Beyond ISCED 5A, first degree	793	11%
I do not know	2,390	34%

Table 30 presents data on American students' fathers' highest education levels. While only 2% of students' fathers' highest education level was *ISCED Level 1*, 34% of students did not know their fathers' education level, and this percentage is the highest percentage among the other levels of the fathers' educational level variable.

Table 31

Average income level of school area variable's descriptive information of American students

Income Level	Number of Students	Percentages
High	581	8%
Medium	3,753	53%
Low	2,691	38%

Table 31 presents data on American students' school areas' average income levels. More than half, 53%, of students attended schools in medium-income area schools, 38% of students attended in low-income area schools, and 8% of American students attended schools in high-income areas.

Regression results: The final multiple regression analysis was conducted with

only American students' data to analyze the relationship between testing frequency and student achievement in eighth-grade mathematics courses when controlling for SES variables. Table 32 indicates that the regression model explained 19% of the variance (R^2 = .19) in math achievement of eighth-grade American students. The regression coefficient estimates for testing frequency and SES variables for the United States are presented in Table 32. The mathematics achievement constant (intercept) was 449.82, which indicates the score when students took tests or quizzes *every day or almost every day*. Students' achievement scores decreased as testing frequency decreased, which indicates a good linear relationship between testing frequency and math achievement in the United States. Average scores decreased by -5.68 points when testing was *used about half the lessons* and by -.22 points when tests were used in *some lessons*. These differences were not statistically significant at a 95% level.

Table 32

Multiple regression analyses predicting eighth-grade mathematics achievement from testing frequency and SES variables in United States

Variable Name	Category Name	*B*	*SE B*	β	t-value
Math Achievement	Intercept (Constant)	449.82	9.45		
Testing Frequency	About half the lessons	-5.68	7.33	-.03	-.78
	Some lessons	-.22	6.47	.00	-.03
Number of Books	11-25 books	11.49	3.54	.06	3.25*
	26-100 books	35.51	4.25	.22	8.35*
	101-200 books	52.82	4.67	.27	11.32*
	More than 200 books	58.18	4.95	.29	11.75*
Fathers' Educational Level	ISCED 2	10.04	7.42	.04	1.35
	ISCED 3	22.28	6.87	.12	3.24*
	ISCED 4	23.83	7.59	.07	3.14*
	ISCED 5B	17.65	7.91	.12	2.23*
	ISCED 5A, first degree	40.45	6.98	.21	5.79*
	Beyond ISCED 5A, first degree	46.21	7.39	.20	6.25*
	I do not know	12.26	6.55	.08	1.87
School Areas Average Income	Medium	15.94	5.36	.11	2.39*
	High	22.95	9.59	.08	2.97*

Note: R-Square = .19 (.02)**, Adjusted R-Square = .19 (.02)**
* Indicates statistical differences
* *Numbers in parentheses are standard error of estimates

The coefficient model also includes SES variables' and their sub-groups' coefficient estimates for the United States in Table 32. Results of the coefficient table show that, as the number of books at home increased, students' achievement also increased, and all the differences in sub-groups are statistically significant. The same results were shown for fathers' highest educational levels. Student achievement increased by their fathers' educational levels and all of these increases are significant except for in the *ISCED Level 2* and *I do not know* groups. Student achievement also increased by school areas' income levels, and these increases are statistically significant at a 95% level.

Findings

Results of the multiple regression analyses show that there was no statistically

significant relationship between testing frequency and eighth-grade student achievement in mathematics in all countries combined, or in any of the four selected countries. Even though there were not significant differences between testing frequencies, however, students who took daily tests or quizzes (*almost everyday*) scored 3.90 points more than students who *never* took quizzes, students who took weekly quizzes (*about half the lessons*) scored 7.37 points more than students who *never* took tests or quizzes, and students who took monthly tests or quizzes (in *some lessons)* scored 5.74 points more than students who *never* took tests or quizzes in all countries combined (see Table 8). Results also indicate that almost all students took at least one test or quiz in all TIMSS 2011 participating countries. Even though about a quarter of the students took tests or quizzes every day, the testing frequency trend, 52%, was for monthly tests or quizzes according to the TIMSS 2011 results.

This study also analyzed the testing frequency and student achievement relationship in individual countries. The study selected four countries: Korea (1^{st}), Singapore (2^{nd}), the United States (9^{th}), and Turkey (24^{th}). The first three of these countries were high performing countries and Turkey was the only low performing country chosen. Multiple regression analyses were performed for each country separately, and results show that high performing countries' testing frequency practices are similar to each other, but different than Turkey. Students performed best when tests or quizzes were used *weekly* in all countries combined*,* performed best when tests or quizzes were implemented *daily* in the United States and Turkey, and students performed best when the testing frequency was *monthly* in the top two performing countries, Korea and Singapore. Testing frequency practices varied from country to country. While 23% of

students took daily quizzes in all of the countries combined, 7% in Korea, 3% in Singapore, 16% in the United States, and 31% in Turkey took daily quizzes. The highest percentage of students took monthly quizzes in all individual countries, as well as in all countries combined. In all countries combined, the testing frequency was *monthly quizzes* for 52% of students, 64% for Korean students, 70% for Singaporean students, 58% for American students, and 41% for Turkish students. Interestingly enough, in general, as countries' rankings increased, the percent of students who took monthly quizzes also increased, but the percent of students who took daily quizzes decreased. A similar trend was observed in the mean scores of testing frequency for individual countries. The testing frequency mean was 2.58 in Korea, 2.67 in Singapore, 2.43 in the United States, and 2.11 in Turkey. The mean scores indicate that an average Turkish student took a test or quiz almost *every week* while students in high performing countries took a test between once every week and once in a month.

Results also indicate significant differences in terms of SES variables between high performing countries and Turkey. All selected SES variables significantly affect student achievement in all countries combined and in all individual countries. Results show that the number of books at home is a significant predictor of student achievement everywhere. Student achievement linearly increased as the number of books students had at home increased in all countries combined, as well as in individual countries. The percent of students who had less than 25 books in their home was 50% in all countries combined, and the other 50% of students had more than 25 books. When looking at individual countries for the number of books at home, the percentage of students who had less than 25 books decreased in high performing countries (36% in Korea, 41% in

Singapore, 39% in the United States), but increased, 56%, in the low performing country, Turkey.

The same pattern seen in regard to the possession of books was observed for fathers' highest educational levels. It also had a positive linear relationship with student achievement and is a significant predictor of 8th students' math achievement in all countries combined, as well as in individual countries. Students whose fathers' highest education was beyond college level performed best, and students whose fathers' highest education was below middle school performed lowest. These results are to be expected, as parents' education levels are perceived as a significant factor in student success. In all countries combined, 21% of students' fathers' highest education level was lower than *ISCED Level 2* (middle school). When it comes to individual countries, this percentage increased in Turkey (65%) but decreased in high performing countries, amounting to 3% in Korea, 15% in Singapore, and 10% in the United States. Another interesting finding of this study is the high percentage of students who do not know their fathers' educational level, and especially in high performing countries. The percentage of students who did not know their fathers' education level was 24% in all countries combined, 22% in Korea, 32% in Singapore, and 34% in the United States, but only 4% in Turkey. Students who selected their fathers' education level as *I do not know* did not score significantly differently than below the middle school group, except in Turkey, where they scored significantly lower, which may mean that students whose fathers' education level was low preferred to select *I do not know* for their fathers' education level. Buckley (2009) suggests that students who are in low SES may give responses to survey questions that

may not represent their actual socioeconomic status or they will give answers to represent themselves as being in a better SES.

A similar trend to other SES variables was also observed for school areas' income levels. School areas' income levels were a significant predictor of student achievement and had a positive linear relationship with student achievement. In general, students who attended schools in high-income areas scored best, and students who attended schools in low-income areas scored lowest. The percentage of students that attended schools in a low-income area was 34% in all countries combined, though it increased in Turkey (53%) and decreased in high performing countries, except in the United Sates (38%); the associated percentages are 31% in Korea and 22% in Singapore. Even though the percent of students who attended schools in low-income areas was higher in the United States than in all countries combined, American students scored higher than the international average on the TIMSS 2011.

Hypotheses Tests

The study tested whether there was a significant relationship between testing frequency and mathematics achievement scores of eighth-grade students in all countries combined and in four specific countries when controlled for SES variables.

a. The results showed that there was no statistically significant relationship between testing frequency and eighth-grade student achievement in mathematics in all counties combined.
b. The results also showed there was no statistically significant relationship between testing frequency and eighth-grade student achievement in mathematics in any of the four selected countries.

Therefore, the study failed to reject its hypothesis.

This study also investigated which testing frequency results in better achievement scores in all countries combined and in four selected countries when controlled for SES. The coefficient table showed that using daily tests or quizzes resulted in the highest achievement scores in all countries combined, and that not using quizzes or tests at all (*never*) resulted in the lowest achievement scores. The testing frequency relationship was different in each individual country. In Korea and Singapore, students performed best when tests or quizzes were implemented on a monthly basis, but students performed best when tests or quizzes were implemented daily in Turkey and the United States. These differences were not statistically significant differences in any of the four countries.

Summary of Results

The regression model explains the 18% variance in all countries combined, 19% in Singapore and United States, 22% in Korea, and 25% in Turkey. The significance tests of the model statistics are not known, as the IDB Analyzer does not produce an ANOVA table in SPSS output. However, coefficient tables showed that there was no statistically significant difference between testing frequencies with relationship to eighth-grade students' math achievement in all countries combined, or in any of the four selected countries. Students who took tests or quizzes almost every day scored 3.90 points higher than students who never took tests or quizzes, but this difference was not statistically significant. The difference between other testing frequencies goes even higher in comparison to the never testing frequency in all countries combined, but these differences were not significant neither.

When it comes to individual countries, no significant differences were detected between testing frequencies. However, the never testing frequency was excluded from the regression model, as there were not enough students in this testing frequency group in individual countries. The mean of the testing frequency was lowest, 2.11, in Turkey, and highest in Singapore, at 2.67. It was 2.43 for American students and 2.58 for Korean students. This means that Turkish students received tests or quizzes most frequently and that Singaporean students took tests or quizzes least frequently among the four selected countries. Coefficient tables showed that the most beneficial testing frequency was for monthly quizzes in Korea and Singapore, while the daily testing frequency was the most beneficial for Turkish and American students.

Results also indicated that SES variables are very important factors in students' mathematics' achievement scores. Students' mathematics scores consistently and significantly increased as their family and school background conditions got better in all countries combined and in all individual countries.

Chapter V: Conclusion

Formative assessment is a crucial part of instructional activities because it provides feedback, informs policy, and informs all stakeholders in education about teaching and learning practices. While there are several formative assessment strategies, quizzing is one of the most commonly used methods of formative assessment. Frequent quizzing is encouraged by The Test Enhanced Learning in Classroom model (TELC) of Roediger, it being found to increase student learning. It is also understood that frequent quizzing is good practice. It is underpinned by learning theories. It shows gaps in student knowledge, motivates students to study, organizes learning materials, engages students, reduces test anxiety, helps transfer knowledge to new topics, and provides feedback for students and teachers (Roediger, et al. 2011; Shirvani, 2009). Ideally, quizzing should always provide feedback (Brame & Beil, 2015). This brings the student to a higher level of understanding through interacting with the teacher (Hein, 1991). Many studies (Shirvani, 2009; Gholami & Moghaddam, 2013; & Zgraggen, 2009) were conducted to measure the relationship of testing frequency and student achievement in various subjects. Some of these studies confirmed the argument that frequent testing increases student achievement (Shirvani, 2009; Roediger, et al. 2009; & Gholami & Moghaddam, 2013), but some of them found no significant differences between testing frequencies (Basol & Johanson, 2009). There are even some studies which found that infrequent testing might be better than frequent testing to increase student achievement in mathematics (Zgraggen, 2009). Despite these findings, frequent testing is underpinned by learning theories as an effective strategy to enhance student learning. Frequent testing is a retrieval practice and an active learning strategy (Francisco, 2014; Karpicke & Grimaldi,

2012). Frequent testing activates learners' prior knowledge and so helps them to construct their own learning (Pelech, 2015).

Summary of the Purpose

Previous studies found different relationships between testing frequency and student achievement, which prevents any consensus on optimal testing frequency. The primary purpose of this study was to identify whether there is a significant relationship between testing frequency and student achievement in eighth-grade mathematics. Furthermore, the study investigated the optimal relationship between testing/quiz frequency and student achievement in eighth-grade mathematics. Another purpose of this study was to determine if the optimal relationship between quiz frequency and student achievement differs from high performing countries to low performing countries. This study also sought to clarify teachers' practices of tests or quizzes in order to see if teachers in high performing countries and low performing countries use quizzes at different frequencies. Thus, educators can see the relationship of student achievement in mathematics and testing frequency in different countries and learn other countries' test or quiz implementing practices. It is especially important for low performing countries to analyze high performing countries' testing frequency practices so that they may implement testing at the same or similar frequencies.

Summary of the Procedure

This study explored TIMSS 2011 data, where about 250,000 students from more than 40 countries participated, to test the relationship between testing frequency and student achievement in all participant countries, as well as in a small number of pre-selected high and low performing countries (Korea, Singapore, Turkey, and the United

States). Teachers' quiz frequency data was retrieved from the TIMSS 2011 teacher questionnaire, and student achievement data was retrieved from the TIMSS 2011 exam results as obtained from the TIMSS database. While previous research studies regarding the effect of quiz frequency on student achievement were conducted in a single or in several institutions, with a limited number of participants, this study was a wider and deeper study with a huge number of participants from more than 40 countries, which enables international comparison of quiz frequency practices.

Additionally, three student socioeconomic status variables (number of books at home, father's highest education level, and school areas' income level) were used as control variables in order to accurately measure the relationship between testing frequency and student achievement. Several multiple regression analyses were utilized to determine the ideal quiz frequency in general and in pre-selected countries. Multiple regression analyses also revealed quiz frequency practices in different countries.

Findings Related to Literature

Results of multiple regression analyses indicated that there is no significant relationship between testing frequency and student achievement in the participant countries in general or in any of the four selected countries. These results match a meta-analysis of the prior testing frequency effect on student achievement studies that was conducted in 2009 by Basol and Johanson, but contradict findings of test enhanced learning in classroom (TELC) studies. Basol and Johanson (2009) analyzed testing frequency studies and found some improvements by testing frequency, but no significant differences were found between different testing frequencies. Furthermore, only a few

variables significantly affect mathematics achievement once variables are controlled for students' and schools' socioeconomic status (Caponera & Losito, 2016).

Additionally, the study sought to identify the optimal testing frequency. The study found that different testing frequencies have different relationship effects in each country and in all countries. All testing frequencies (daily, weekly, and monthly) had a better relationship with achievement scores in comparison to the *never* testing frequency in all countries. Student scores were highest when testing frequency was *weekly* in all countries combined, which indicates that the optimal testing frequency is weekly testing, in general. These results run parallel to Gholami and Moghaddam's (2013) study, where they found weekly tests to result in better student achievement in comparison to a no testing group. This result is also similar to Bangert et al.'s (1998) findings. They found that better student performance was associated with frequent testing, but the improvement in student achievement diminished as the number of tests increased. This conclusion also matches the beliefs of opponents of frequent testing who claim that extreme use of tests or quizzes may not be very beneficial for student learning because frequent testing reduces instructional time, leads students to score better on the tests, and emphasizes tests rather than learning (Gholami & Moghaddam, 2013).

One of the primary purposes of this study was to find the relationship of testing frequency and student achievement in mathematics in pre-selected high and low performing countries. The study found that there are differences in using tests or quizzes in individual countries, and the relationship varies from country to country. Student achievement was highest when students took weekly tests or quizzes in Korea and Singapore, which were the top two performing countries in the TIMSS 2011

administration, but student achievement was highest when students took daily tests or quizzes in the United States (9th in TIMSS 2011) and in Turkey (24th in TIMSS 2011). Shirvani (2009) also found that daily quizzes are better than weekly quizzes at improving student achievement in the United States. It can be concluded that the optimal quiz frequency is *daily* quizzes in the United States and Turkey, but that having monthly quizzes is the optimal frequency in Korea and Singapore. However, significant improvements should not be expected from simply changing testing frequency because there is not a significant relationship between testing frequency and students' mathematics achievement.

This study has also provided some important information about overall practices of testing or quizzing around the world. It has found that, globally, tests/quizzes are significant parts of instructional activities, with almost all students taking at least several quizzes in a year. This confirms the value of frequent testing as a teaching strategy for the enhancement of learning, through informing students about their progress and teachers about the effectiveness or otherwise of their classroom practices and lesson planning.

Teachers' practices of implementing tests or quizzes is another important finding of this study. The study found that teachers in four pre-selected countries utilized quizzes at different frequencies. The quiz frequency was between weekly and monthly in three high performing countries (Korea, Singapore, and the United States), but close to weekly in the low performing country (Turkey). However, it was a common practice to utilize quizzes in all selected countries. The study found that only about one teacher did not implement quizzes at all in four selected countries, as based on the TIMSS 2011 teacher questionnaire. While there is no weighted data on overall countries' practices of testing

frequency, this study found that tests or quizzes were usually used more frequently in low performing countries than in high performing countries (see Appendix F). None of the high performing countries implemented tests on a weekly basis or more often, but a lot of low performing countries, and especially middle eastern countries, implemented tests on a weekly basis or even more often. These results suggest that low performing countries may be spending too much time on implementing tests or quizzes, such that it reduces instructional time.

While this study found no significant relationship between quiz frequency and student achievement, it found that students' SES factors (number of books students have at home, highest education level of their fathers, and school areas' income levels) and student achievement are statistically significantly related to each other in general and in all pre-selected high and low performing countries. Students' achievement scores significantly and constantly increased as their socioeconomic status improved. These results confirm previous studies' findings regarding the positive relationship between socioeconomic status and student achievement (Aikens & Barbarin, 2008; Haveman & Wolfe, 2008).

Results suggest that teachers should implement at least several tests or quizzes, supported with feedback, in a semester, but that teachers should not use tests or quizzes too often in order to avoid over-emphasizing test scores at the expense of instructional time.

Implications

This study found that testing frequency has varying relationships with students' mathematics achievement in eighth-grade. These relationships also differ from country to

country. Students in the top two performing countries benefitted most from *monthly* quizzes, but students in the United States and Turkey benefitted most from *daily* quizzes, while overall students benefitted most from taking *weekly* quizzes. These results make it harder to draw clear conclusions from the study, but we can easily conclude that teachers in any country will find it beneficial to their students to implement tests or quizzes at some frequency. However, this study demonstrates that the best testing frequency varies from country to country. It is also integral to note that an important caveat is that student achievement will not increase or decrease radically simply through changing testing frequency, because there is not a significant relationship between testing frequency and student achievement. These results are akin to findings of Patnam's (2013) and O'Dwyer, Wang, and Shields's (2015) TIMSS studies, where they investigated the relationship between teaching strategies and student achievement. They found very little relationship between student achievement and teaching methods.

This study also investigated the relationship between student achievement in math and students' socioeconomic statuses. The main finding was that students' socioeconomic statuses are more important predictors of student achievement than testing frequency. The study found that students who were in a higher SES showed significantly better achievement scores than students who were in a low SES. We can conclude that student achievement is related to both instructional and non-instructional factors. These non-instructional factors might be out of the control of teachers, and radical politic changes and time are needed to increase parents' educational levels and improve the income of school populations, as well as the countries' entire populations. However, an awareness of the impact of these factors on the achievement of low SES students and the

benefits of frequent testing or quizzing could inform pre-service teacher education, lesson planning, and classroom practice in the vitally important areas of student engagement and motivation to succeed. Aikens and Barbarin (2008) suggest that schools which have a high number of low SES students should distribute these students to multiple schools in order to reduce their burdens and decrease the number of students that require access to necessary resources at schools. Obviously, this is not a permanent solution to diminish the effect of SES on student achievement, but it can at least reduce the gap between high and low SES students.

Limitations

TIMSS population samples are created to represent a country's students, and only those students' teachers are entitled to complete the survey. It follows then that teachers who complete the teacher surveys may not represent the main body of teachers in a country. Therefore, results and conclusions cannot be generalized to the totality of teachers in a country. However, the high numbers of teachers who participated in TIMSS studies make it possible to say that teachers' testing frequency practices represent the majority of the teachers.

There are always limitations to what can be learned from a survey because self-reporting may not reflect actual practices. By design, the relationship of teachers' responses to survey questions and students' test scores may not reveal the total relationship of a phenomenon. Students' SES are also determined through student responses to survey questions, but since this involves self-reporting, the results may not represent actual socio-economic statuses of students because students from a low socio-economic status or from countries that have a low gross domestic product may give

responses in such a way as to represent themselves as being from a higher SES than their actual SES would indicate (Buckley, 2009).

Another limitation of this study is the use of plausible values. Plausible values were used to estimate average student achievement in mathematics and science. Even though the plausible values are highly reliable, it is not the same as all students taking the entire exam. Carstens and Hastedt (2010) say researchers should use caution when conducting analyses with plausible values because using the plausible values incorrectly, such as in using only one plausible value or using only statistical software, may indicate statistically significant differences when there are no differences, or may not show significant differences when there are actual statistical differences (Carstens & Hastedt, 2010). This study would indicate that researchers should use IEA software (IDB) that is specifically designed for analyses with plausible values, and especially for large-scale dataset analyses, such as with TIMSS, PIRLS, and PISA.

This study was limited to written tests and quizzes, but teachers may also implement online quizzes and publishers' pre-created quizzes in addition to written quizzes.

Future Research Suggestions

This study investigated the relationship of testing frequency and general mathematics achievement scores. Besides general mathematics achievement, TIMSS also provides content-specific domain achievements in mathematics. TIMSS included Numbers, Algebra, Geometry, and Data and Chance content domains in the eighth-grade math exam, which enables researchers to look into each content domain. TIMSS also measures students in their cognitive skills levels, such as knowing, applying, and

reasoning. Future studies can focus on the relationship between testing frequency and student achievement in each content and domain to see if testing frequency has different effects on different subjects in mathematics. Thus, teachers would then know the best testing frequency for each content domain. Studies can also focus on the relationship of testing frequency and cognitive domains to investigate whether a given testing frequency works better for one cognitive level than the others so that teachers can create tests based on the relevant findings.

This study was limited to the relationship of written tests or quizzes with achievement scores, but did not consider the issue of online testing. Written tests have some disadvantages when compared to online tests. One of the biggest disadvantages is in procedures for providing feedback. Roediger, Putnam, and Smith (2011) point out the significance of providing feedback to improve student learning on their TELC model, but it takes a significant amount of time for teachers to read all papers and provide feedback for each question to each student in the traditional testing method. This challenge may prevent teachers from providing feedback on frequent tests. However, there are some online testing tools that can provide feedback and grade papers automatically, so that teachers do not have to devote much time to testing and providing feedback. To measure the relationship between online testing practices and student achievement, future research based on large-scale studies such as the TIMSS, PISA, and NAEP could focus on the relationship between frequent online testing and student achievement.

References

Agarwal, P. K. (2016). Retrieval practice. Retrieved from http://www.retrievalpractice.org/

Agarwal, P. K., Bain, P., & Chamberlain, R. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24*(3), 437-448. Retrieved from http://www.jstor.org/stable/43546801

Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, McDermott, K. B, & McDaniel, M.A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition, 3,* 131-139.

Aikens, N. L., & Barbarin, O. (2008). Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. *Journal of Educational Psychology, 100*, 235-251.

Al-Rukban, M. O. (2006). Guidelines for the construction of multiple choice questions tests. *Journal of Family & Community Medicine*, *13*(3), 125–133.

Amador, J. M., & Soule, T. (March 01, 2015). Girls Build Excitement for Math from Scratch. *Mathematics Teaching in the Middle School, 20,* 7, 408-415.

Anderson, M. (2015). 6 facts about Americans and their smartphones. Retrieved from http://www.pewresearch.org/fact-tank/2015/04/01/6-facts-about-americans-and-their-smartphones/

Anthony, M. (2016). Cognitive development in 11-13 year olds. Retrieved from http://www.scholastic.com/parents/resources/article/stages-milestones/cognitive-development-11-13-year-olds

Armstrong, P. (2016). Bloom's Taxonomy. Retrieved from https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/

Badger, E., Thomas, B., & ERIC Clearinghouse on Tests, M. D. (1992). *Open-Ended Questions in Reading. ERIC/TM Digest*.

Ballance, D. L. (2013). Assumptions in Multiple Regression: A Tutorial. Retrieved from http://www.dianneballanceportfolio.com/uploads/1/2/8/2/12825938/assumptions_in_multiple_regression.pdf

Bangert-Drowns, R. L. Kulik, J. A. & Kulik, G. L. C. (1991). The effects of frequent classroom testing. Journal of Educational Research, 85, 89-99.

Barker, T. (2011). An Automated Individual Feedback and Marking System: An Empirical Study. *Electronic Journal of E-Learning*, *9*(1), 1-14.

Basol, G. & Johanson, G. (2009). Effectiveness of frequent testing over achievement: A meta analysis study. International Journal of Human Sciences, 6(2), 99–121.

Bilican, S., Demirtasli, R. N., & Kilmen, S. (2011). The Attitudes and opinions of the students towards mathematics course: The comparison of TIMSS 1999 and TIMSS 2007. *Educational Sciences: Theory and Practice, 11,* 3, 1277-1283.

Bofah, E. A., & Hannula, M. S. (December 01, 2015). TIMSS data in an African comparative perspective: Investigating the factors influencing achievement in mathematics and their psychometric properties. Large-scale Assessments in Education: an Iea-Ets Research Institute Journal, 3, 1, 1-36.

Bonwell, C. C., Eison, J. A., Association for the Study of Higher, E., ERIC Clearinghouse on Higher Education, W. D., & George Washington Univ., W. D. (1991). Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports.

Bowden, M., & Doughney, J. (2010). SocioEconomic status, cultural diversity and the aspirations of secondary students in the Western Suburbs of Melbourne, Australia. *Higher Education, 59*(1), 115-129. Retrieved from http://www.jstor.org/stable/25622168

Brame, C. J. & Biel, R. (2015). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. CBE—Life Sciences Education 14, 1-12.

Brooks, J. G., & Brooks, M. (1993). *In search of understanding: The case for constructivist classrooms*. Alexandria, VA: ASCD.

BSU, (2015). Ordering questions. Retrieved from. https://bbcrm.edusupportcenter.com/link/portal/8197/8382/Article/5536/Ordering-Questions-Instructor

Buckley, J. (2009). Cross-national response styles in international educational assessments: Evidence from PISA 2006. Retrieved from https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Burk, M. J. (1987). The effect of practice testing to learning on the achievement and attitude of geometry students. Unpublished master's thesis, Glassboro State College, Glassboro, NJ.

Butler A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 1118–1133.

Butler A. C., Godbole N., Marsh E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. Journal of Educational Psychology, 105, 290–298.

Campbell, L. (2008). Beginning with what students know: The role of prior knowledge in learning. Retrieved from https://www.corwin.com/sites/default/files/upm-binaries/25914_081222_Campbell_Ch1_excerpt.pdf

Caponera, E., & Losito, B. (2016). Context factors and student achievement in the IEA studies: evidence from TIMSS. *Large-scale Assessments in Education: an Iea-Ets Research Institute Journal, 4,* 1, 1-22.

Carstens, R., & Hastedt, D. (2010). The effect of not using plausible values when they should be: An illustration using TIMSS 2007 grade 8 mathematics data. Paper presented at the 4th International Association for the Evaluation of Educational Achievement (IEA) International Conference, Gothenburg, Sweden.

Cassady, J. C. & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27,* pp. 270–295.

Chacos, B. (2017). 12 powerful websites that can replace your PC's desktop software Retrieved from http://www.pcworld.com/article/2459671/websites/12-powerful-websites-that-can-replace-your-desktop-software.html

CERI (2008). Assessment for learning formative assessment. Retrieved from http://www.oecd.org/site/educeri21st/40600533.pdf

Chickering, A. W., and Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *American Association of Higher Education Bulletin, 39(7),* 3-7

Chudgar, A., Luschei, T. F., and Fagioli, L. (2014). A call for consensus in the use of student socioeconomic status measures in cross-national research using the trends in international mathematics and science study (TIMSS). Teachers College Record.

CITL (2015). Assessing student learning. Retrieved from https://citl.indiana.edu/teaching-resources/assessing-student-learning/

CITT (2016). Gagne's 9 events of instruction. Retrieved from http://citt.ufl.edu/tools/gagnes-9-events-of-instruction/

Clay, B. & Root, E. (2001). Is this a trick question: A short guide to writing effective test questions. Retrieved from http://www.k-state.edu/ksde/alp/resources/Handout-Module6.pdf

Coffey, H. (2009). Zone of proximal development. Retrieved from http://www.learnnc.org/lp/pages/5075

Connor-Greene, P. A. (2000). Assessing and promoting student learning: Blurring the line between teaching and testing. *Teaching of Psychology, 27,* 2, 84-88.

Dempster, F. N. (1992). Using tests to promote learning: A neglected classroom resource. Journal of Research and Development in Education, 25(4), 213-217.

Dickler, J. (2016). Men still earn more than women with the same jobs. Retrieved from http://www.cnbc.com/2016/12/05/men-still-earn-more-than-women-with-the-same-jobs.html

Dineen, P., Taylor, J., & Stephens, L. (1989). The effect of testing frequently upon the achievement of students in high school mathematics courses. School Science Mathematics, 89(3), 197-200.

Dindyal, J. (2008). An overview of the gender factor in mathematics in TIMSS-2003 for the Asia-Pacific region. *Zdm: the International Journal on Mathematics Education, 40,* 6, 993-1005.

Doe, S. R., Gingerich, K. J., & Richards, T. L. (2013). An evaluation of grading and instructional feedback skills of graduate teaching assistants in introductory psychology. *Teaching Of Psychology*, *40*(4), 274-280.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research & Evaluation*, *14* (7). Retrieved from http://pareonline.net/getvn.asp?v=14&n=7

Dustin, D. S. (1971). Some effects of exam frequency. The Psychological Record, 21, 409-414.

Eberly Center (2015). What is the difference between formative and summative assessment? Retrieved from https://www.cmu.edu/teaching/assessment/basics/formative-summative.html

Eison, J. (2015). Using active learning instructional strategies to create excitement and enhance learning. Retrieved from https://www.cte.cornell.edu/documents/presentations/Eisen-Handout.pdf

Ermisch, J & Pronzato, C. (2010). Causal effects of parents' education on children's education. Retrieved from https://www.iser.essex.ac.uk/files/iser_working_papers/2010-16.pdf

Evans, J. A. (2015). *Gender, self-efficacy, and mathematics achievement: An analysis of fourth grade and eighth-grade TIMSS data from the united states* (Order No. 3723105). Available from ProQuest Dissertations & Theses Full Text. (1728883955). Retrieved from http://search.proquest.com/docview/1728883955?accountid=2837

FairTest (2007). The dangerous consequences of high-stakes standardized testing. Retrieved from http://www.fairtest.org/dangerous-consequences-highstakes-standardized-tes

Fitch, M. L., Drucker, A. J., & Norton, J. A. J. (January 01, 1951). Frequent testing as a motivating factor in large lecture classes. *Journal of Educational Psychology, 42,* 1, 1-20

Francisco, A. (2014). Ask the cognitive scientist: Retrieval practice. Retrieved from http://digitalpromise.org/2014/10/15/ask-the-cognitive-scientist-retrieval-practice/

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafore, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning improves student performance in science, engineering, and mathematics. Proceedings of the *National Academy of the Sciences, 111*(23), 8410-8415.

Fulton, B. A. (2016). The relationship between test anxiety and standardized test scores. Retrieved from

http://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=3361&context=dissertations

Garrison, C. & Ehringhaus, M. (2016). Formative and summative assessment in the classroom. Retrieved from https://www.amle.org/BrowsebyTopic/WhatsNew/WNDet/TabId/270/ArtMID/888/ArticleID/286/Formative-and-Summative-Assessments-in-the-Classroom.aspx

Gholami, V. & Moghaddam, M. M. (2013). The effect of weekly quizzes on students′ final achievement score. *International Journal of Modern Education and Computer Science, 5,* 1, 36-41.

Gray, A. (1997). Constructivist teaching and learning. Retrieved from http://www.saskschoolboards.ca/old/ResearchAndDevelopment/ResearchReports/Instruction/97-07.htm

Halpern, D. F., & Hakel, M. D. (2003). Applying the Science of Learning to the University and Beyond: Teaching for Long-Term Retention and Transfer. *Change*, *35*(4), 36–41. Retrieved from http://www.jstor.org/stable/40165500

Harris, D. K., & Changas, P. S. (1994). Revision of Palmore's Second Facts on Aging Quiz from a True-False to a Multiple-Choice Format. *Educational Gerontology*, *20*(8), 741-54.

Haveman, R., & Wolfe, B. (2008). The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature, 33*(4), 1829-1878.

Hein, G. E. (1991). Constructivist learning theory. Retrieved from https://www.exploratorium.edu/education/ifi/constructivist-learning

Hertzberg, O. E., Heilman, J. D. and Leuenberger, H. W. (1932). The value of objective tests as teaching devices in educational psychology. *Journal of Educational Psychology, 33*, 371-380.

Horst, H. T. & Martens, R. (2016). Closed-ended questions. Retrieved from https://www.utwente.nl/ces/toetsing/Docenten/testing-grading/job-aid/closed-ended-questions/

Hutchings, M. & Kazmi, N. (2015). Exam factories: The impact of accountability measures on children and young people. Retrieved from https://www.teachers.org.uk/sites/default/files2014/exam-factories_0.pdf

IADB (2016). PRIDI user guide. Retrieved from http://www.iadb.org/education/pridi/database/PRIDI_User_Guide.pdf

ICEF (2014). Global demand for STEM training a growing factor in overall enrolment trends. Retrieved from http://monitor.icef.com/2014/03/global-demand-for-stem-training-a-growing-factor-in-overall-enrolment-trends/

IEA (2016). TIMSS 2015. Retrieved from http://www.iea.nl/timss_2015.html

Jancarík, A., & Kostelecká, Y. (2015). The Scoring of Matching Questions Tests: A Closer Look. *Electronic Journal Of E-Learning* 13 (4), 270-276.

Janjetovic, D. & Malinic, D. (2004). Family variables as predictors of mathematics and science self-concept of students. *1st IEA International Research Conference.* Nicosia, Cyprus.

Johansone, I. (2016). Operations and quality assurance. Retrieved from http://timssandpirls.bc.edu/methods/pdf/TP_Operations_Quality_Assurance.pdf

Johnson, P.E. (2006) Effect of frequent testing on learning mathematics. *International Journal of Mathematical Education in Science and Technology, 21*(5), 733-737, doi: 10.1080/0020739900210507

Karpicke, J. D. & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science (New York, N.Y.), 331,*6018, 772-5.

Karpicke, J. D. & Grimaldi, P. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review, 24*(3), 401-418. Retrieved from http://www.jstor.org/stable/43546799

Kerkman, D. D., & Johnson, A. T. (2014). Challenging Multiple-Choice Questions to Engage Critical Thinking. *Insight: A Journal of Scholarly Teaching*, *9,* 92-97.

Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology, 25,* 6, 427-436.

Kharbach, M. (2014). 10 useful web tools for creating online quizzes and polls. Retrieved from http://www.educatorstechnology.com/2014/02/10-useful-web-tools-for-creating-online.html

Kika, F. M., & Mclaughlin, T. F., & Dixon, J. (1992). Effects of frequent testing of secondary algebra students. Journal of Educational Research, 85(3), 159-62

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice, 41,* 4, 212-18.

Kulp, D. H. II (1933). Weekly tests for graduate students? *School and Society, 38,* 157-159.

Lay, Y. F., Ng, K. T., & Chong, P. S. (March 01, 2015). Analyzing affective factors related to eighth-grade learners' science and mathematics achievement in TIMSS 2007. *The Asia-Pacific Education Researcher, 24*(1), 103-110.

Livingston, J. A. (1997). Metacognition: An overview. Retrieved from http://gse.buffalo.edu/fas/shuell/cep564/Metacog.htm

Lowe, T. W. (2015). Online quizzes for distance learning of mathematics. *Teaching Mathematics Applications,* 34*,* 138-148.

Ma, X. (1995). The effect of informal oral testing frequency upon mathematics learning of high school students in China. Journal of Classroom Interaction, 30(1), 17-20.

Marden, N. Y., Ulman, L. G., Wilson, F. S., & Velan, G. M. (2013). Online feedback assessments in physiology: effects on students' learning experiences and outcomes. *Advances in Physiology Education,* 37(2)*,* 192-200. doi**:**10.1152/advan.00092.2012

Marshall, B. (2007). A crisis for efficacy? *Education Review, 20*(1), 29-35.

Martin, M. O. & Kelly, D. L. (1996). TIMSS technical report: Design and development. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., and Mullis, I. V.S. (2012). Methods and procedures in TIMSS and PIRLS 2011. Retrieved from http://timss.bc.edu/methods/

Martin, R. R., & Srikameswaran, K. (1974). Correlation between frequent testing and student performance. *Journal of Chemical Education. 51*(7), 485-486. Retrieved September 20,2008, from the ERIC database, No. EJ101605.

Mawhinney, V. T., Bostow, D. E., Laws, D. R., Blumenfeld, G. J., & Hopkins, B. L. (1972). A comparison of students studying-behavior produced by daily, weekly, and three-week testing schedules. *Journal of Applied Behavior Analysis, 4*(4), 257-264.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399-414. doi:10.1037/a0021782

Mcdaniel, M. A., Roediger, H. L., & Mcdermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 2, 200-206.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L., III. (2013). Quizzing in middle school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*, 360–372.

Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education,30*(4), 159-167. doi:10.1152/advan.00053.2006

Mullis, I.V.S. & Martin, M.O. (Eds.). (2011). TIMSS 2011: TIMSS and PIRLS achievement scaling methodology (pp. 1-11). Chestnut Hill, MA: Boston College. Retrieved from http://timss.bc.edu/methods/pdf/TP11_Scaling_Methodology.pdf

Mullis, I. V. S., Martin, M. O., and Foy, P. (2005). IEA's TIMSS 2003 International report on achievement in the mathematics cognitive domains: Findings from a developmental project. Retrieved from http://timssandpirls.bc.edu/PDF/t03_download/T03MCOGDRPT.pdf

Mullis, I. V. S., Martin, M. O., Foy, P., and Arora, A. (2012). TIMSS 2011 international results in mathematics. Retrieved from http://timssandpirls.bc.edu/timss2011/downloads/T11_IR_Mathematics_FullBook.pdf

Mullis, I. V. S., Martin, M. O., Minnich, C. A., Stanco, G. M., Arora, A., Centurino, V. A., & Boston College, T. C. (2012). *TIMSS 2011 encyclopedia: Education policy and curriculum in mathematics and science. Volume 1: A-K*. International Association for the Evaluation of Educational Achievement.

Mullis, I.V.S., Martin, M.O., Foy, P., Ruddock, G. J., O'Sullivan, C. Y, & Preuschoff, C (2009). TIMSS 2011 assessment frameworks. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., and Preuschoff, C. (2009). TIMSS 2011 assessment frameworks. Retrieved from http://timssandpirls.bc.edu/timss2011/downloads/TIMSS2011_Frameworks.pdf

Murayama, K. (2009). Objective test items. Retrieved from http://www.education.com/reference/article/objective-test-items/

NCTE (2013). Formative assessment that truly informs instruction. Retrieved from http://www.ncte.org/positions/statements/formative-assessment/formative-assessment_full

Nicol, D. (2011). Developing students' ability to construct feedback. *The Quality Assurance Agency for Higher Education*. Retrieved from http://www.enhancementthemes.ac.uk/docs/publications/developing-students-ability-to-construct-feedback.pdf?sfvrsn=30

Nunes, L. D., & Karpicke, J. D. (2015). Retrieval-based learning: research at the interface between cognitive science and education. In R. A. Scott, & S. M. Kosslyn (Eds.), Emerging trends in the social and behavioral sciences (pp. 1-16). John Wiley & Sons, Inc.

O' Dwyer, L. M., Wang, Y., & Shields, K. A. (2015). Teaching for conceptual understanding: A cross-national comparison of the relationship between teachers' instructional practices and student achievement in mathematics. *Large-scale Assessments in Education: an Iea-Ets Research Institute Journal, 3,* 1, 1-30.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research & Evaluation, 8(2). Retrieved from http://pareonline.net/getvn.asp?v=8&n=2

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing instructions and study to improve student learning. Retrieved from http://psych.wustl.edu/memory/TELC/Improving_Student_Learning.pdf

Pastötter, B. & Bauml, K. T. (2014). Retrieval practice enhances new learning: the forward effect of testing. Retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00286/full

Patnam, V. S. (2013). *Factors related to student achievement in mathematics and comparison of the U.S. with other countries: A study based on TIMSS 2007 report* (Order No. 3591696). Available from ProQuest Dissertations & Theses Full Text. (1434876028). Retrieved from http://search.proquest.com/docview/1434876028?accountid=2837

Pelech, J. R. (2016). Comparing the effectiveness of closed-notes quizzes with open-notes quizzes: Blending constructivist principles with action research to improve student learning. *i.e.: inquiry in education, 8*(1)*,* 1-21. Retrieved from: http://digitalcommons.nl.edu/ie/vol8/iss1/5

Polesel, J., Dulfer, N., & Turnbull, M. (2012). The experience of education: The impacts of high stakes testing on school students and their families. Retrieved from https://www.whitlam.org/__data/assets/pdf_file/0008/276191/High_Stakes_Testing_Literature_Review.pdf

Prince, M. (2004). Does active learning work? A Review of the Research. *Journal of Engineering Education*, *93*(3), 223-231.

Rodriguez, M. C. (2010) The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17*(1)*,* 1-24, doi: 10.1207/ s15324818ame1701_1

Rodriguez, M. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17*, 1-24.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011b). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology. Applied, 17*(4), 382-95.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-7.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 1*(3), 181-210.

Roediger H. L., III, Putnam A. L., Smith M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation, 44,* 1–36.

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248.

Rojas, S. A. (2014). Towards automatic recognition of irregular, short-open answers in Fill-in-the-blank tests. *Tecnura*, *18*(39), 47-61.

Salas-Morera, L, Arauzo-Azofra, A & García-Hernández, L. 2012. Analysis of online quizzes as a teaching and assessment tool. *Journal of Technology and Science Education 2*(1)*,* 39-45.

Sanchez, W. B. (2013). Open-Ended Questions and the Process Standards. *Mathematics Teacher*, *107*(3), 206-211.

Scriven, M. (1967). The methodology of evaluation. Retrieved from http://www.comp.dit.ie/dgordon/Courses/ILT/ILT0005/TheMethodologyOfEvaluation.pdf

Serafino, P. and Tonkin, R. (2014). Intergenerational predictors of poverty in the UK and EU, in IARIW Conference, the Netherlands.

Shilo, G. (2015). Formulating Good Open-Ended Questions in Assessment. *Educational Research Quarterly*, *38*(4), 3-30.

Shirvani, H. (2009). Examining an assessment strategy on high school mathematics achievement: Daily quizzes vs. weekly tests. *American Secondary Education, 38*(1), 34-45.

Sirin. S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3)*,* 417-453.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*, 784-802.

Socrates Programme (2009). Improving quality of science teacher training in European cooperation. Retrieved from http://www.iqst.upol.cz/e-learning/m1/e-learning-m1-u2.php

Spector, J. (2015). Common Core tests giving kids anxiety, psychologists say. Retrieved from http://www.lohud.com/story/news/education/2015/11/20/common-core-anxiety/76114566/

Spielberger, C. D. & Vagg, P. R. (1995). *Test anxiety: Theory, assessment, and treatment.* Washington, DC: Taylor & Francis

Stiggins, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. Retrieved from http://bibliotecadigital.academia.cl/jspui/bitstream/123456789/586/1/Rick%20Stiggins.pdf

Suda, K. J., Bell, G. C., & Franks, A. S. (2011). Faculty and student perceptions of effective study strategies and materials. *American Journal of Pharmaceutical Education*, *75*(10), 201. http://doi.org/10.5688/ajpe7510201

Techopedia (2015). Mobile application. Retrieved from https://www.techopedia.com/definition/2953/mobile-application-mobile-app

TELC (2011). Test-enhanced learning in the classroom. Retrieved from http://psych.wustl.edu/memory/TELC/

TIMSS 1995 (1996). Third International Mathematics and Science Study. Retrieved from http://timssandpirls.bc.edu/timss1995.html

TIMMS 2007 (2007). About TIMSS. Retrieved from http://timss.bc.edu/TIMSS2007/about.html

TIMSS & PIRLS 1999 (1999). Third International Mathematics and Science Study Repeat – TIMSS 1999. Retrieved from http://timssandpirls.bc.edu/timss1999.html

TIMSS & PIRLS 2011 (2011). About TIMMS 2011. Retrieved from http://timssandpirls.bc.edu/timss2011/index.html

TIMSS 2015 (2016). About TIMSS 2015. Retrieved from http://timssandpirls.bc.edu/timss2015/

Topcu, M.S., Erbilgin, E., & Arikan, S. (2016). Factors predicting Turkish and Korean students' science and mathematics achievement in TIMSS 2011. *Eurasia Journal of Mathematics, Science & Technology Education, 12*(7)*,* 1711-1737.

Townsend, N. R., & Wheatley, G. H. (1975). Analysis of frequency of tests and varying feedback delays in college mathematics achievement. *College Student Journal, 9*(1), 32-35.

Trifoni, A. & Shahini, M. (2011). How does exam anxiety affect the performance of university students? *Mediterranean Journal of Social Sciences, 2*(2)*,* 93-100

Turner, C. (2014). U.S. tests teens a lot, but worldwide, exam stakes are higher. Retrieved from http://www.npr.org/2014/04/30/308057862/u-s-tests-teens-a-lot-but-worldwide-exam-stakes-are-higher

Turney, A. H. (1931). The effect of frequent short objective tests upon the achievement of college students in educational psychology. *School and Society, 33*, 760-762.

UNESCO (2011). International standard classification of education. Retrieved from http://www.uis.unesco.org/Education/Documents/isced-2011-en.pdf

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* Cambridge, Mass.: Harvard University Press.

Wang, J. (2007). A trend study of self-concept and mathematics achievement in a cross-cultural context. *Mathematics Education Research Journal, 19*(3), 33-47.

WATgreen Project (2015). Advantages and disadvantages of open and closed questions. Retrieved from http://environment.uwaterloo.ca/research/watgreen/projects/library/1020/ocq.html

Webopedia (2015). Application software. Retrieved from http://www.webopedia.com/TERM/A/application.html

Weimer, M. (2015). Advantages and disadvantages of different types of test questions. Retrieved from http://www.facultyfocus.com/articles/educational-assessment/advantages-and-disadvantages-of-different-types-of-test-questions/

Wilder, D. A., Flood, W. A., & Stromsnes, W. (2001). The use of random extra credit quizzes to increase student attendance. Journal of Instructional Psychology. Retrieved March 7, 2010, from http://findarticles.com/p/articles/mi

William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1)*,* 3-14. http://dx.doi.org/10.1016/j.stueduc.2011.03.001.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012): How and when do students use flashcards?, *Memory, 20*(6)*,* 568-579.

Wood, W. C. (1998). Linked multiple-choice questions: The tradeoff between measurement accuracy and grading time. *Journal Of Education For Business*, *74*(2), 83-86.

Wooten, M. M., Cool, A. M., Prather, E. E., & Tanner, K. D. (2014). Comparison of Performance on Multiple-Choice Questions and Open-Ended Questions in an Introductory Astronomy Laboratory. *Physical Review Special Topics - Physics Education Research*, *10*(2), 020103. https://dx.doi.org/10.1103/PhysRevSTPER.10.020103

Zarei, A. A. (2008). On the learnability of three categories of idioms by Iranian EFL learners. *Journal of Humanities of the University of Kerman, 2*(2), 82-100

Zhao, Y. (2016). Stop copying others: TIMSS lessons for America. Retrieved from http://zhaolearning.com/2016/11/30/stop-copying-others-timss-lessons-for-america/

Zgraggen, F. D. (2009). *The effects of frequent testing in the mathematics' classroom*. Doctoral dissertation. University of Wisconsin-Stout.

Appendices

Appendix A: Review of Online Test/quiz Tools

Given the advantages of integrating teacher-generated online quizzes, it is important that teachers are aware of the special tools (web pages, apps, or software) designed t to create this form of assessment. The following sections will identify a sample range of currently available tools and discuss their strength and weaknesses.

Web Sites

A website is composed of connected web pages about a topic or related topics on the Internet and they are created by individuals or organizations for various purposes. Web sites aimed at helping educators generate their own quizzes are usually created by organizations. Below, some websites that enable teachers to create their own quizzes will be discussed.

ProProf Quiz Maker: This website offers various tools for educators including quiz generator tool for a monthly fee. It provides six types of quizzes (multiple choice, matching, ordering, fill in the blank, essay, and check box) for teachers to generate their own quizzes. It also provides automatic grading and statistics about quizzes. However, it fails to give immediate feedback when students select a wrong answer.

Testmoz Test Generator: Testmoz is specifically designed to create tests for free. However, it provides for only three types of quizzes (multiple choice, multiple response, and fill in the blank) to be created. It does have an automatic grading function with correct-incorrect feedback, but it does not let teachers insert explanation feedback for each choice.

QuizWorks: Quiz Works enables teachers to create multiple-choice and true/false type of online quizzes for an annual fee, but it can provide other types of quizzes for an additional fee Quiz Works' multiple-choice type quizzes provide test statistics for teachers and feedback for students. However, students will see feedback at the end of the quiz instead of after each attempt.

Google Forms: A Google account is required for this web tool. Even though it is designed to create surveys, Google Forms can be used to create interactive online tests. Since it is not developed to enable educators to create quizzes or tests, only limited question types (multiple-choice and true/false) can be generated through Google Forms. Google Forms are free, provide tests statistics, and work very well with other Google tools. However, providing immediate feedback through Google Forms requires teachers to be familiar with this tool and do some tweaking around questions and choices to provide feedback for each choice.

QuizStar: QuizStar is a part of the 4teachers.org website that provides various online tools for teachers. It is free, but an account is required. Teachers can generate quizzes with multiple-choice, multiple-select, true/false, and short answer type questions through this tool. QuizStar can enable teachers to insert feedback for every question, but not for every choice, and does automatic grading for every quiz type except short answer questions. It also enables teachers to use other teachers' tests and let them customize tests based on their needs by using timers and selecting the number of attempts.

The above-mentioned tools are intended to serve as examples of the numerous web sites that enable educators to create online quizzes. Generally, quiz generator websites are free, but they require an account to use their web sites. These web sites

usually provide correct/incorrect feedback and automatic grading for multiple-choice and true/false tests. There are also some advanced web sites that usually require the payment of a fee. These web sites include more question types, more detailed statistics about results, and explanation feedback rather than just correct/incorrect feedback.

Mobile Applications

Mobile applications, also known as apps, are special software designed to run on handheld devices like a smartphone or a tablet computer. Mobile applications are usually developed to provide similar services to those that users access through computers (Techopedia, 2015). There are several mobile application software environments but IOS and Android are the two most commonly used app development environments. These apps are essential for smartphones and tablet computers. As of 2015, 64% of adult Americans own a smartphone and they use their phones for conducting research about their health condition, educational activities, and job and employment opportunities (Anderson, 2015). It is obvious that mobile applications are a significant part of life and education. However, there are very limited numbers of mobile apps designed for educational assessments. Most of the quiz generating apps that are available in IOS and Android store are created for gaming purposes rather than as an assessment tool. There are limited numbers of apps in IOS and Android stores created for assessing students, but most of them are not available in both stores. This creates problems for classroom use, since not all students in a classroom use the same mobile operating system. The other drawback of current apps is cost. Many quiz apps allow users to download the app for free, but users are then restricted to creating either a limited number of quizzes or a limited number of questions in a quiz. Teachers and students need to pay a fee to get full

functional app in both Apple and Google Play stores. In the following sections, some mobile applications that enable teachers to create quizzes will be discussed.

Quiz and Flashcard Maker: This quiz maker app is only available for IOS users for a fee. It allows users to create their own flashcards and multiple-choice type quizzes that can be sent out to students. These questions can be graded automatically. Providing explicit feedback is not available for this app, and the non-existence of an Android app makes it difficult for teachers to use in a classroom.

mTestTaker: This app is available for a fee in App Store for iPhone, iPad and iPod Touch users, but it is not available for Android users. This app only enables teachers to create multiple choice question quizzes. Giving feedback to test takers is not possible. However, automatic grading is available.

Quiz Maker: Quiz Maker is a game and an assessment tool that is available only for devices that have IOS operating system. It is free to download, but then the user has to pay a fee to be able to get full functionality. Quiz maker app enables teachers to use sound, pictures, and texts when creating a quiz. However, only multiple-choice test can be created through this app and providing feedback other than correct/incorrect feedback is not possible.

Quiz Creator: This app is only available for Android users and offers both a free and a paid version. The free version gives users limited access, but, if users upgrade it to the paid version, they can have full access to the app. This app allows teachers to create quizzes by using variety of question types but it is designed as a gaming tool rather than assessment tool. Teachers cannot send their quizzes to individual students, rather several teams need to compete on created quizzes.

Create a Quiz: This app is also only available for Android users and has free and paid versions. It offers only multiple-choice and standard question types and does not allow teachers to insert feedback into questions or choices.

Revision Quiz Maker: This app is the only free app that is available in both IOS and Android platforms. This app lets users generate four types of questions in their quizzes: multiple-choice, ordering, fill in the blank, and matching. True/false quiz type also can be generated in multiple-choice format. However, there is no option for users to send out quizzes to other users. This make it useless for teachers who like to create their own quizzes and send it out to students, but students can use this app as studying aid tool.

The review of test/quiz generating mobile applications showed that there is not a quality quiz generating app that is available in both IOS or Android operating systems. Available apps are either available for only IOS users or Android users. This is a critical problem preventing teachers from integrating quiz generating technologies because not all students in one classroom use the same operating system. Even though apps are available for both operating systems users, they either fail to provide feedback or they only enable teachers to create quizzes with limited types of questions. Another disadvantage of these tools is their price because not all teachers and students want to pay to integrate quiz generating systems.

Application Software

The term application software, also known as end user programs, is defined as a single program, or group of programs, developed for the end user and which must be downloaded on to a computer (Webopedia, 2015). Chrome web browser and Dropbox cloud storage program are some examples of millions of applications software that we

use very often. Software used to be a preferable tool over websites because of their powerful features, but advancements in web browsers enabled the creation of powerful websites that can replace software (Chacos, 2017). In this section, applications software specifically designed to empower teachers to create assessment methods (tests, exams, and quizzes) will be discussed. There are many applications software (especially learning management systems - Blackboard and Moodle for example) that can be used to create tests, but need to be downloaded as a whole package if teachers wish to use their test creating programs, as these cannot be purchased separately. The expense involved is a deterrent for many teachers and schools.

iSpring Solutions: This application software is free, but the paid version of quiz maker program enables teachers to create quizzes with a variety of question types like multiple-choice, multiple correct, true/false, and essay question. It automatically grades quizzes and give detailed reports to users. However, the free version does not allow teachers to provide detailed feedback for students, other than correct/incorrect feedback.

Articulate Quiz Maker: This software is a part of course development software that consists of several applications, but users could buy quiz maker application as a standalone tool. This is a very powerful software that enables teachers to create quizzes with many question types, and enables the use of multimedia materials (images, sounds, and videos) in questions and answers. It also allows teachers to provide custom feedback for students for every question and choice. This software is also very advanced in terms of providing test analytics to the teachers.

There are many applications software choices for quiz creating purposes. However, these tools' prices are beyond a teacher's budget. These applications are

usually designed for schools or institutions to buy rather than an individual and most of this applications software include a variety of tools needed to develop a whole course rather than just developing assessment tools. The features that this software provide for teachers to create tests are almost limitless, but all these tools require teachers to get some training or watch tutorials in order to generate effective assessment tools.

To conclude, teachers need better online tools and training in their use to create their own tests/quizzes and provide feedback in order to increase student learning through taking quizzes. All available tools have some drawbacks that prevent teachers generating their own effective quizzes. The most powerful tools are applications software for quiz creating, but their price and difficulties of using them militate against their adoption by teachers. Available mobile applications are either free or cheap and easy to use, but most of these tools are not available in cross platforms. Additionally, they do not enable teachers to integrate all kinds of questions in a quiz and they fail to provide immediate explanation feedback. On the other hand, web sites are free and as powerful as application software (Chacos, 2017) and easier to use than software but harder to use than a mobile app. The available web sites provide many effective quiz features (variety of question types, correct/incorrect feedback, automatic grading, and test reports) but they do not enable teachers to insert custom feedback for questions and choices. Therefore, a web site that offers all of the above-mentioned functions and is synchronized with mobile apps that are available in both App Store and Google Play will be a great addition for teachers' tool kit for developing their preferred forms of assessment.

Appendix B: Review of Quiz Types

Multiple-Choice Quizzes

Multiple-choice tests are a form of assessment in which test takers are asked to select the best possible answer out of multiple choices. Multiple-choice tests are known as the most widely used and objective type of test method (Al-Rukban, 2006). Multiple choice quizzes are frequently implemented in every level of education especially in large classes, despite the argument that they prevent students from thinking critically and emphasize memorization rather than comprehension (Kerkman & Johnson, 2014).

Proponents of multiple-choice tests claim, however, that if these tests are well constructed and well written they can be used to measure high order skills as well as basic skills (Steve, 1997). Al-Rukban (2006) argues that multiple-choice tests can be used to measure various levels of learning, are objective, easy to score, reliable, and time-efficient. These features are even enhanced by the utilization of online tools and mobile devices. Automatic grading and providing timely feedback are now made possible by many websites, applications, and software for multiple-choice tests that help teachers save significant amount of time and effort (Online Testing Tools for Teachers, 2015).

Fill in the Blank Quizzes

Fill in the blank questions include a phrase, sentence, or paragraph that has a blank space indicating a missing word or words (answer) that students need to complete (ITS, 2014). Fill in the blank tests are usually more challenging than multiple-choice tests because they require students to recall concepts and rather than memorizing (Rojas, 2014). Since this question type does not offer choices, it prevents students from finding the correct answer by guessing or finding the correct answer by eliminating wrong

choices (Classroom Assessment, 2015). However, fill in the blank questions are limited in terms of measuring various learning levels (Murayama, 2009).

Matching Quizzes

Matching quiz type has a content area in one side and a list of names or statements is on another side, which must be correctly matched (Murayama, 2009). Matching questions are typically two columns of a test where a column on the left side of a page presents stimuli and another column on the right side of the page presents responses and test takers need to match the response associated with a given stimulus (Jancarik & Kostelecka, 2015). Countries (stimuli) on one side and capitals (responses) on the other side is typical matching question type that students need to match capitals with countries. Providing more choices (responses) than stimuli reduces the possibility of finding the correct answer through guesswork (Jancarik & Kostelecka, 2015). CITL (2015) lists the advantages of matching quizzes as following. They

- need less time for reading and response, which allow teachers to cover more content.
- provide objective measurement of student achievement
- provide reliable test scores
- enable teachers to grade effectively and accurately.

Matching quiz types have disadvantages as well. For example, it is hard to measure high level of learning outcomes and they are difficult to create because of the need to find a set of stimuli and responses (CITL, 2015). Clearly, matching quizzes cannot be used for every topic but they can be utilized to measure simple recall of information in many courses.

Ordering Quizzes

Ordering question is another type of question that teachers can use to create online and paper-based quizzes. Students need to order a list of items in a manner to find the correct answer and teachers can use labels and images that let students create the desired order through drag and drop actions through online tools. Teachers need to consider partial credits when grading but quizzes should be low stakes for students to work as retrieval practice (Pelech, 2015). Ordering questions are very useful for measuring students' knowledge of chronological events or ordering numbers from small to big or big to small (BSU, 2015). Ordering questions are also useful to logically order a list of events or stories.

True/False Quizzes

True/False questions are presentations of statements to students and students indicate in some manner whether the statement is true or false and true/false questions are commonly used to measure simple recall of knowledge (Clay & Root, 2001). One of the disadvantages of true/false quizzes is that they allow respondents to guess and respondents have a 50% chance of being right without knowing the correct answer and an educated good guess that knows how to use clues can increase chance of being right in true/false question types (Harris & Changas, 1994). Despite these disadvantages, true/false tests have their own positive features that lead educators to implement this type of tests, such as easy to score, written quickly, objectively scored, and reliable (Clay & Root, 2001). University of Minnesota's Measurement Services (2015) further describes these advantages as ability to measure a variety of learning outcomes, accuracy and economy in grading, reliability and being amenable for item analysis to inform

instruction. Similar to most quiz types, these tests can be generated as online and paper-based by teachers.

A project submitted to The WATgreen Project (2016), an educational initiative at the University of Waterloo to determine, among other issues, the advantages and disadvantages of questions types in tests listed the following advantages of the closed-ended questions type:

- it is easy for test takers to respond
- it takes less time to answer for respondents
- it takes less time to score the answers
- teachers can easily compare students' scores
- it is easier to transfer data to statistical tools and easier to analyze them
- choices can help to clarify the question for better understanding
- test takers are more likely to answer even sensitive issues
- there are fewer irrelevant or confused answers to questions
- conducting replication of these tests is easier

Another advantage of closed-ended tests is that they enable educators to create online version of these tests that allow teachers to provide immediate feedback (Kharbach, 2014). Some advance tools even provide explanation feedback in addition to correct-incorrect feedback (Online Testing Tools for Teachers, 2015).

Disadvantages of closed-ended quizzes as listed by the WATgreen Project (2015) are as follows:

- they can limit students' ideas
- students who do not know the answer can respond by guessing the answer

- students might be frustrated if they think the best answer is not a choice
- questions may confuse students if there are many distractors in choices
- teachers may not be able to notice misinterpretation of a question
- students may mark wrong choice even though they know the answer
- these tests force students to give simple answers to complex issues
- they push respondents to choose an answer that they would not choose in real life

These tests are also criticized for not being able to measure all levels of learning outcomes and students who do not know the concept may get high scores by guessing and utilizing some test taking techniques (Weimer, 2015).

Open-Ended Quizzes

Open-ended questions are questions that have multiple solutions or explanations, which have potential to reflect students thinking process, understanding, and misconceptions. Teachers who implemented open-ended quizzes found that teachers get a great amount of information about what students know and do not know on a topic (Sanchez, 2013). Shilo (2015) points out the importance of asking right questions and says that formulating appropriate questions can enable teachers to examine student' knowledge, improve thought processes and enhance student's learning abilities. Teachers need to choose open-ended quizzes over multiple-choice quizzes if they want to analyze student learning because multiple-choice quizzes cannot reveal students' complete understanding. Teachers generally prefer to use multiple-choice questions because are a lot easier to score than open-ended quizzes (Wooten, Cool, Prather, and Tanner, 2014). According to the University of Waterloo WATgreen Project (2015) the main advantages of open-ended quizzes are as following:

- they enable students to provide various answers to a question
- students can explain and clarify their response in many ways
- teacher may see another approach to a question
- students can go as deep as they want to answer to complex issues
- respondents are free to be creative and express themselves in details
- teachers can see students' logic and thinking processes in their answers

Another and perhaps the most important advantage of open-ended exam types, essays, problems, graphic reading, and such, is being able to measure all levels of learning outcomes in Bloom's Taxonomy (Wood, 1998).

The disadvantages of open-ended quizzes are identified as following in the WATgreen Project (2015):

- students may provide various degrees of details in their answers
- student responses may go to irrelevant directions in useless details
- conducting statistical procedures are difficult
- students may lose correct answer if questions are too general
- students need more time, show more effort and, more thought is required to answer the questions
- questions may intimidate students
- answers take more space in the answer sheet
- grading requires more time and effort in open-ended exams (Badger, Thomas, & ERIC Clearinghouse on tests, 1992; Schinske, 2011).

- grading accuracy and reliability is another disadvantage of open-ended items because different graders may score differently since responses might be very different from each other (Doe, Gingerich, & Richards, 2013).

To conclude, open-ended tests and closed-ended tests have their own advantages and disadvantages. Being able to measure all levels of learning is an important advantage of open-ended tests. However, Wood (1998) indicates that well written and well-thought close-ended items such as multiple-choice question can measure all levels of learning taxonomy too. Closed-ended questions are preferred by educators because of their features like, grading accuracy and consistency, allowing for implementation in large groups, time efficiency, amenable to conduct statistical analyses (Horst & Martens, 2016), and through recent advancement in mobile technology educators can now create quizzes with immediate feedback through their mobile devices.

Appendix C: Correlation between SES Variables

Table 33

Correlation between socioeconomic status variables and dependent variable

	Number of books	Mothers' edu. level	Fathers' edu. level	School income level	Math achievement
Num. of books	1.00	.14	.14	-.14	.28
Mothers' edu. level	.14	1.00	.58	-.11	.10
Fathers' edu. level	.14	.58	1.00	-.11	.10
School Income level	-.14	-.11	-.11	1.00	-.19
Math achievement	.28	.10	.10	-.19	1.00

Appendix D: UNESCO's school levels

Table 34

UNESCO's ISCED school levels

UNESCO Levels	School Levels
ISCED Level 0	Early childhood education
ISCED Level 1	Primary education
ISCED Level 2	Lower secondary education
ISCED Level 3	Upper secondary education
ISCED Level 4	Post-secondary non-tertiary education
ISCED Level 5	Short-cycle tertiary education
ISCED Level 6	Bachelor's or equivalent level.
ISCED Level 7	Master's or equivalent level
ISCED Level 8	Doctoral or equivalent level.

Appendix E: Multiple Regression Assumptions Tests

Assumption tests for the first research question

a. Normality Assumptions: Normality assumption is tested through explore function of SPSS with 1st plausible values (DV) and four levels of predictor variable; testing frequency. Histograms in Figures 5, 6, 7, and 8 indicates that normality assumption is not violated for any level of the predictor variable.

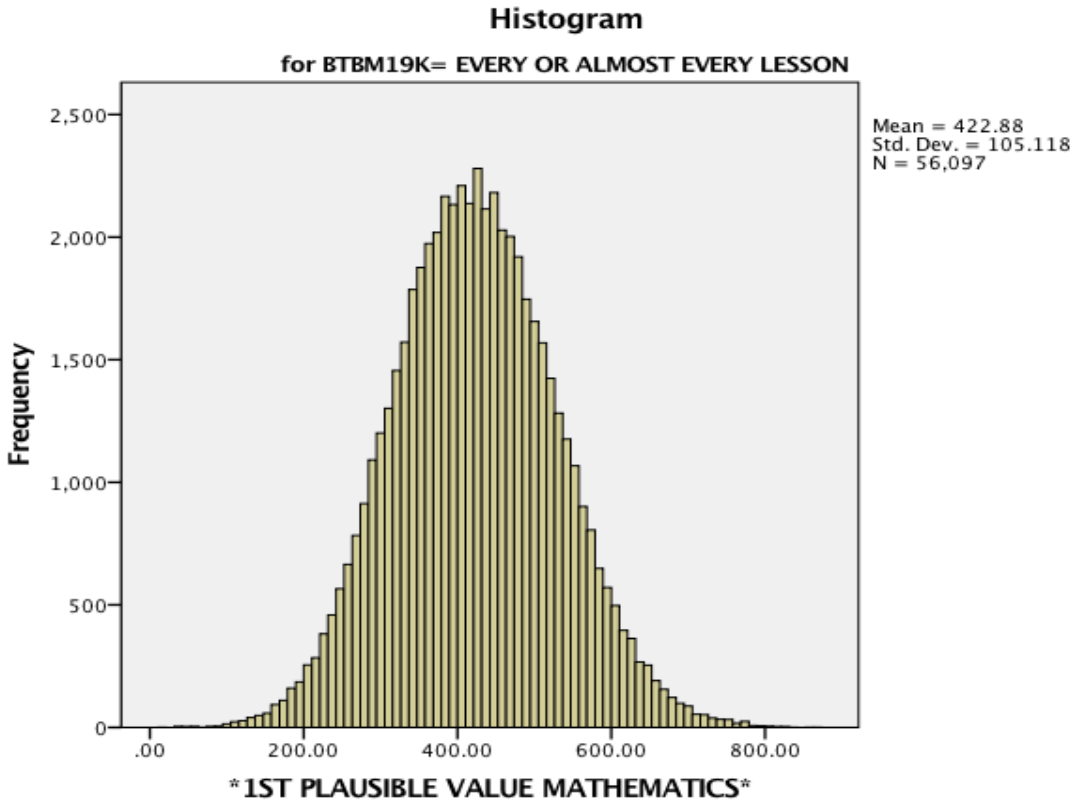

Figure 1: Distribution of achievement scores for every or almost every lesson testing frequency.

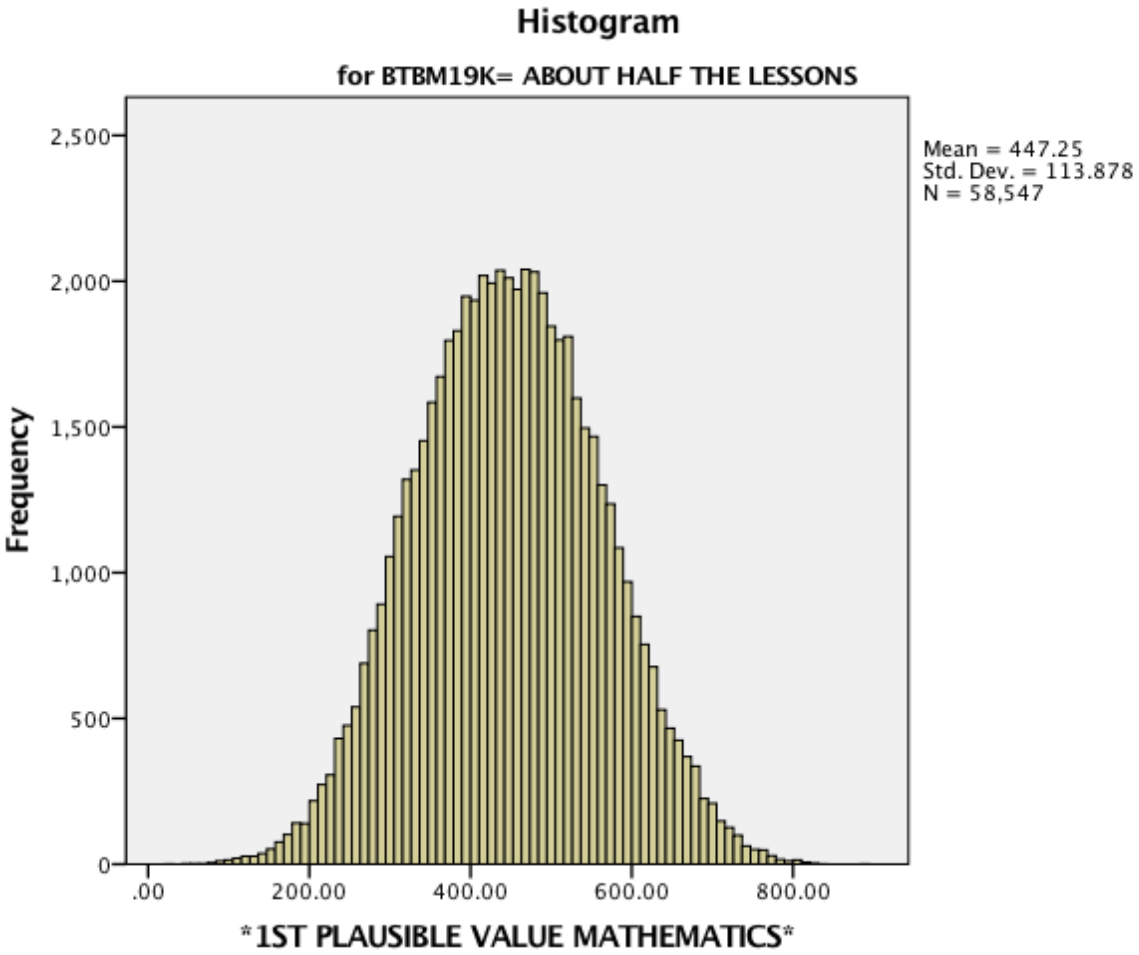

Figure 2: Distribution of achievement scores for about half the lessons testing frequency.

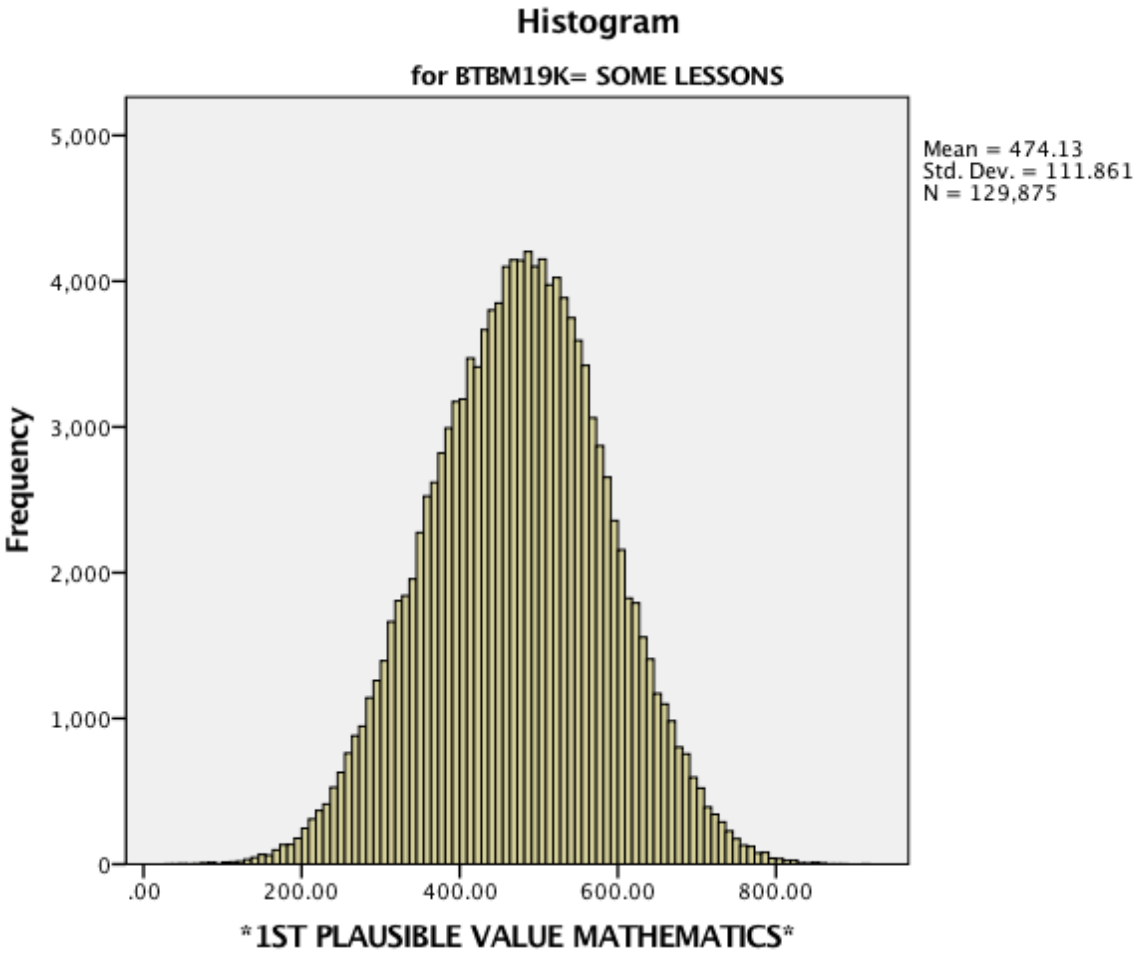

Figure 3: Distribution of achievement scores for some lessons testing frequency.

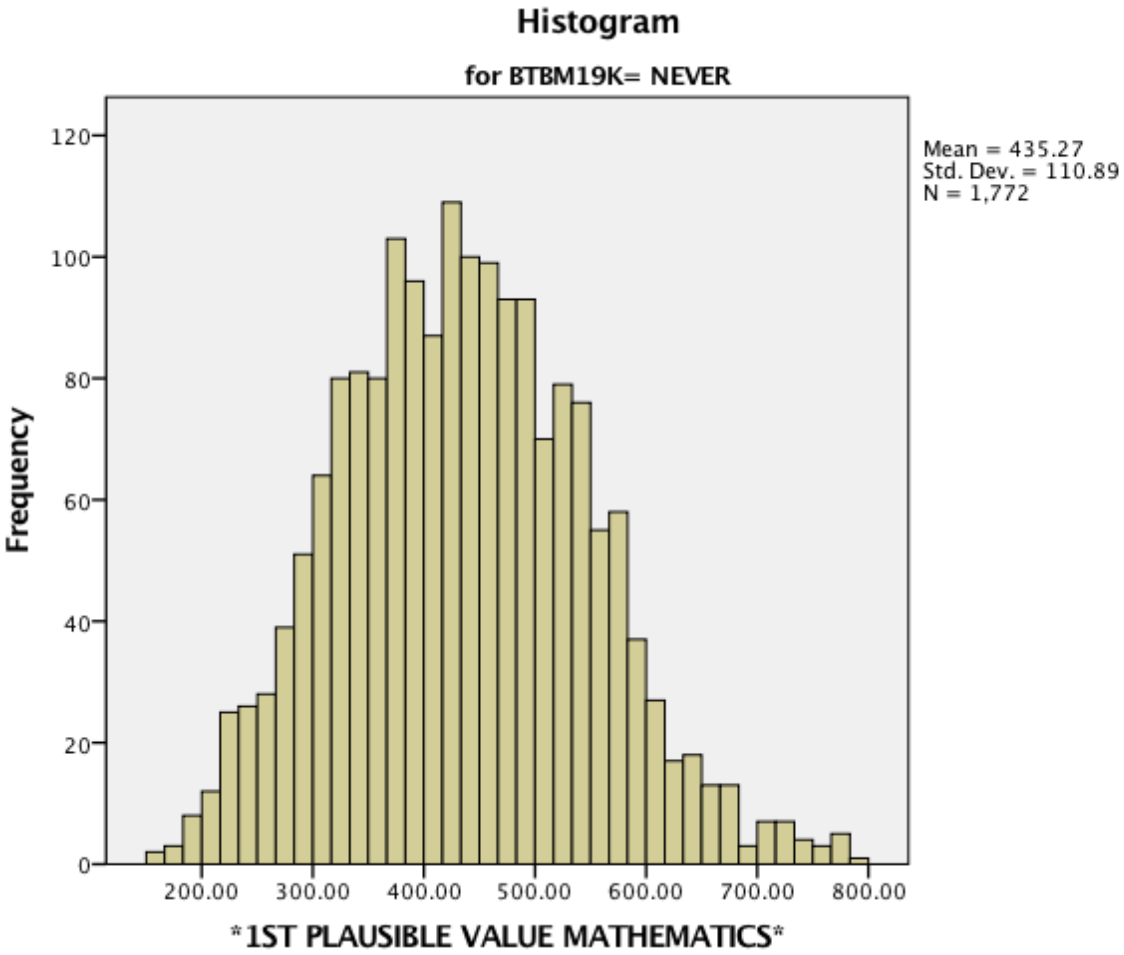

Figure 4: Distribution of achievement scores for never lesson testing frequency.

b. Linear Relationship: Linearity assumption between predictor and dependent variables are tested through linear regression function of the SPSS. Figure 9 is a scatterplot representation of standardized residuals by predicted values. It can be seen that values are scattered around a horizontal line and there is not much departure from linearity. Therefore, linearity assumption is met in the model.

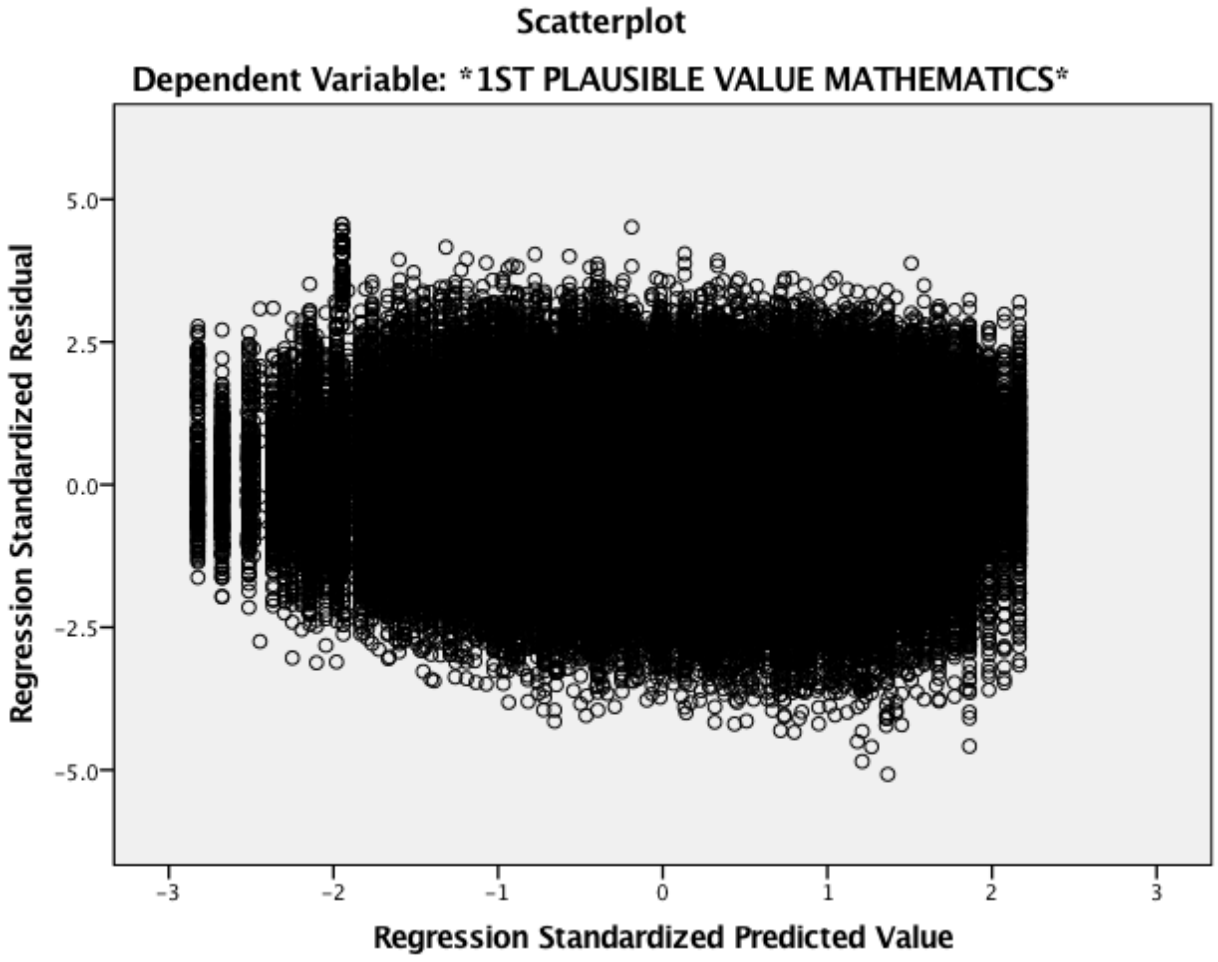

Figure 5: Scatterplot indicates linear relationship with standardized residuals by predicted values

c. Homoscedasticity: Homoscedasticity assumption test was also tested through linear regression function of the SPSS. Figure 10 indicates variances of errors are similar across the predicted values. Residuals are equally scattered around the horizontal line. It can be concluded that homoscedasticity assumption is not violated in the regression model.

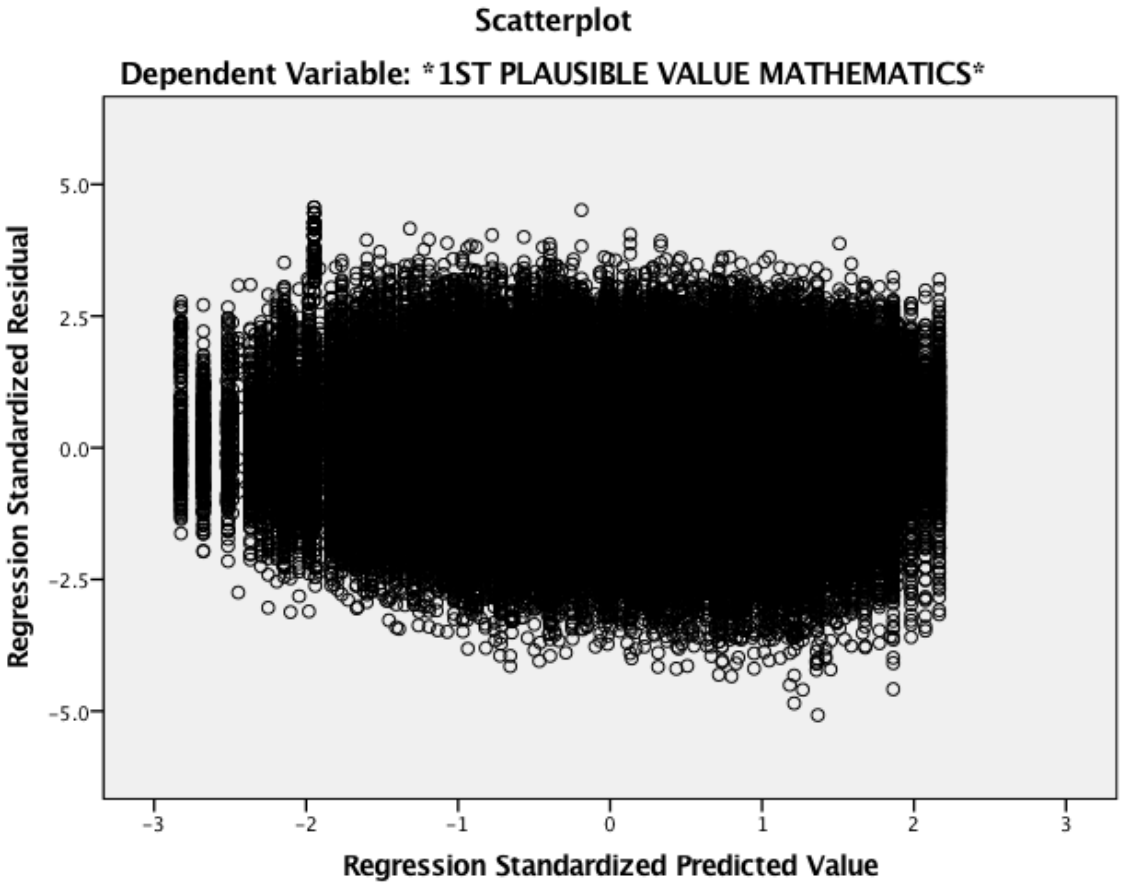

Figure 6: Scatterplot indicates homoscedasticity with standardized residuals by predicted values.

d. Multicollinearity: Multicollinearity assumption was tested through correlation function of SPSS between all independent variables in the model. The results are represented in the Table 35 and it indicates low correlation between independent variables. Therefore, there is no multicollinearity problem in the model.

Table 35

Correlation between independent variables

Independent Variables	Testing frequency	Number of books at home	Fathers' highest level of education	School areas' income level
Testing frequency	1.00	.12	.11	-.04
Number of books at home	.12	1.00	.20	-.21
Fathers' highest level of education	.11	.20	1.00	-.17
School areas' income level	-.04	-.21	-.17	1.00

Assumption tests for the second research question

1. Korea

a. Normality Assumptions: Histograms in Figures 11, 12, and 13 indicates that normality assumption is not violated for any level of the predictor variable in the model, except *never* testing frequency level as there were very small number of students in this group. Because of low number of students in never quiz frequency group, this quiz frequency group was excluded from data analysis.

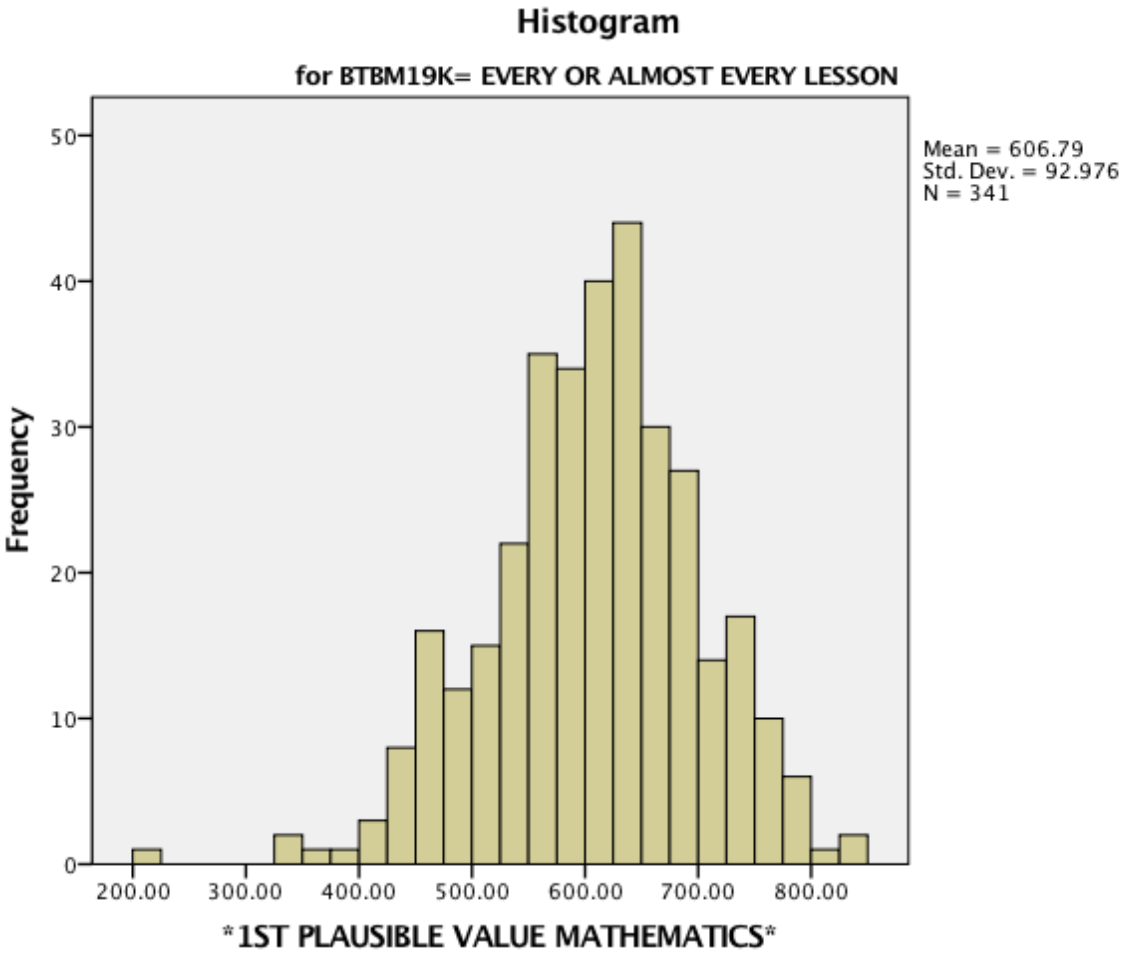

Figure 7: Distribution of achievement scores for every or almost every lesson testing frequency

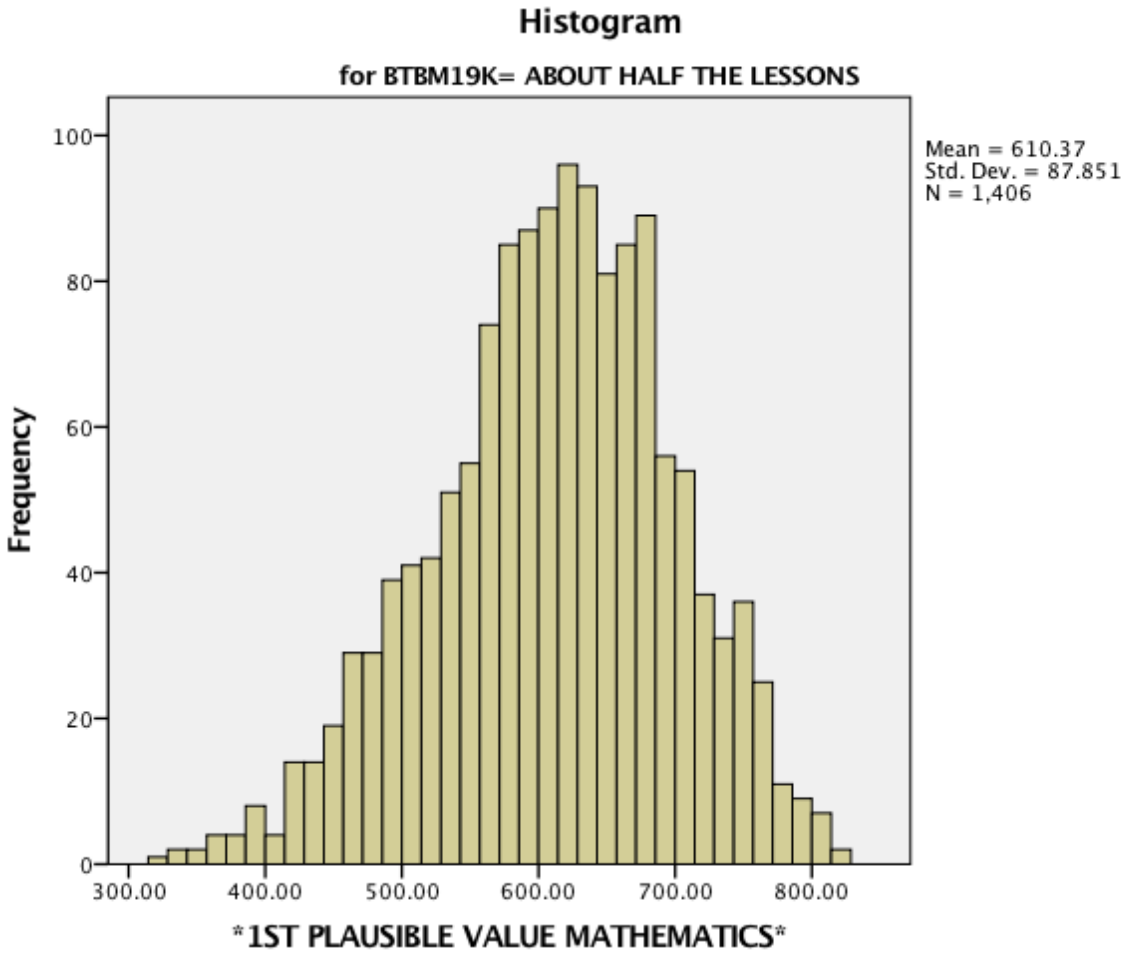

Figure 8: Distribution of achievement scores for about half the lessons testing frequency.

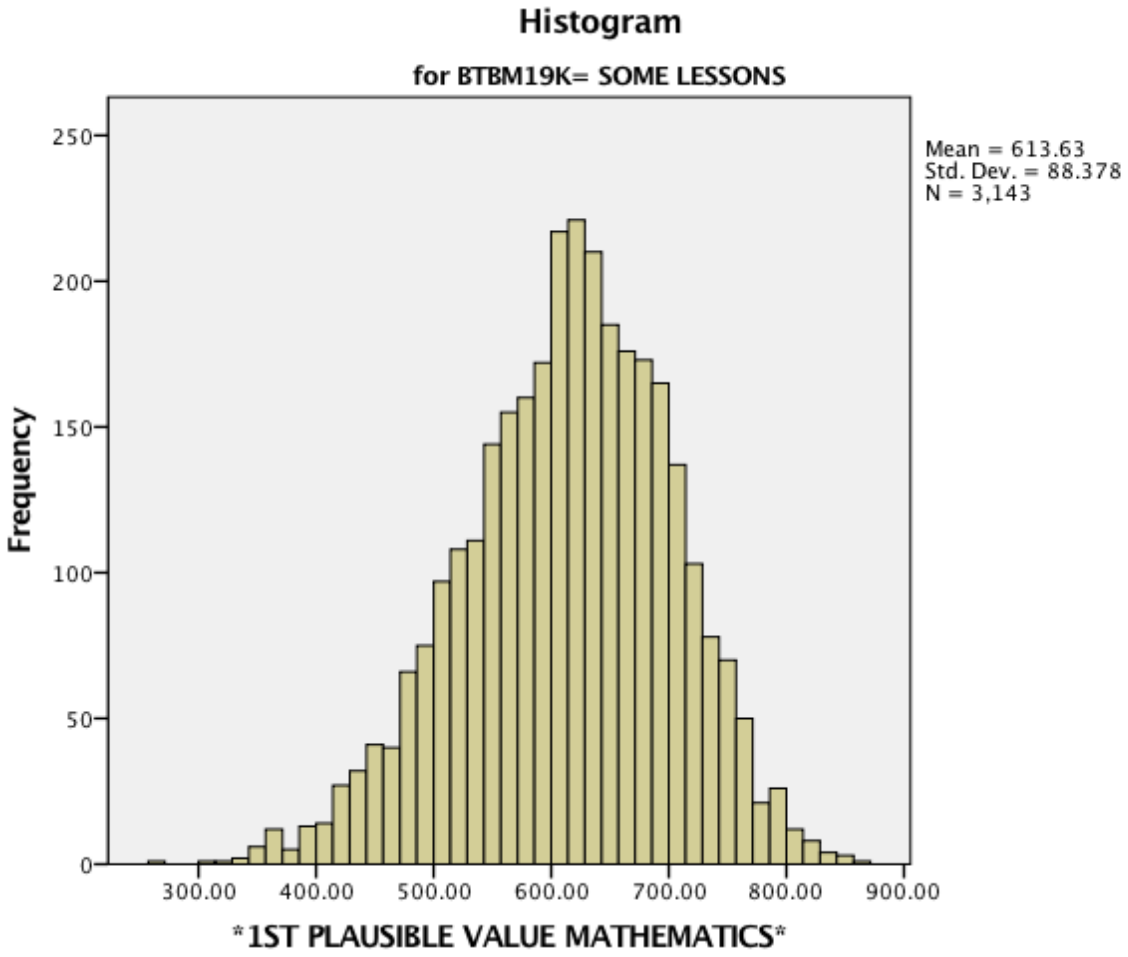

Figure 9: Distribution of achievement scores for some lessons testing frequency.

b. Linear Relationship: Linearity assumption between predictor and dependent variables are tested through linear regression function of the SPSS. Figure 14 is a scatterplot representation of standardized residuals by predicted values. It can be seen that values are scattered around a horizontal line and there is not much departure from linearity. Therefore, linearity assumption is met in the model.

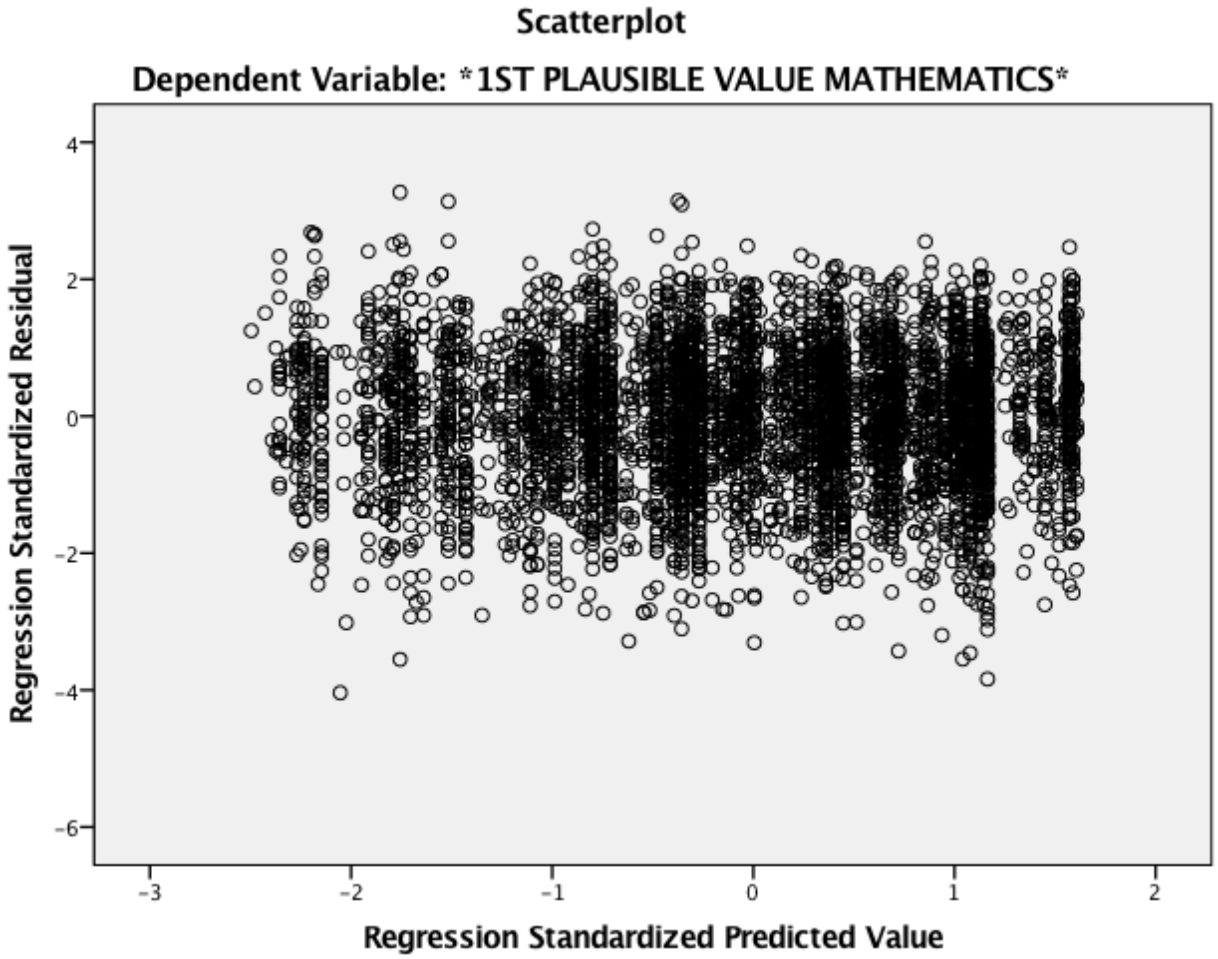

Figure 10: Scatterplot indicates linear relationship with standardized residuals by predicted values

c. Homoscedasticity: Homoscedasticity assumption test was also tested through linear regression function of the SPSS. Figure 15 indicates variances of errors are similar across the predicted values. Residuals are equally scattered around the horizontal line. It can be concluded that homoscedasticity assumption is not violated in the regression model.

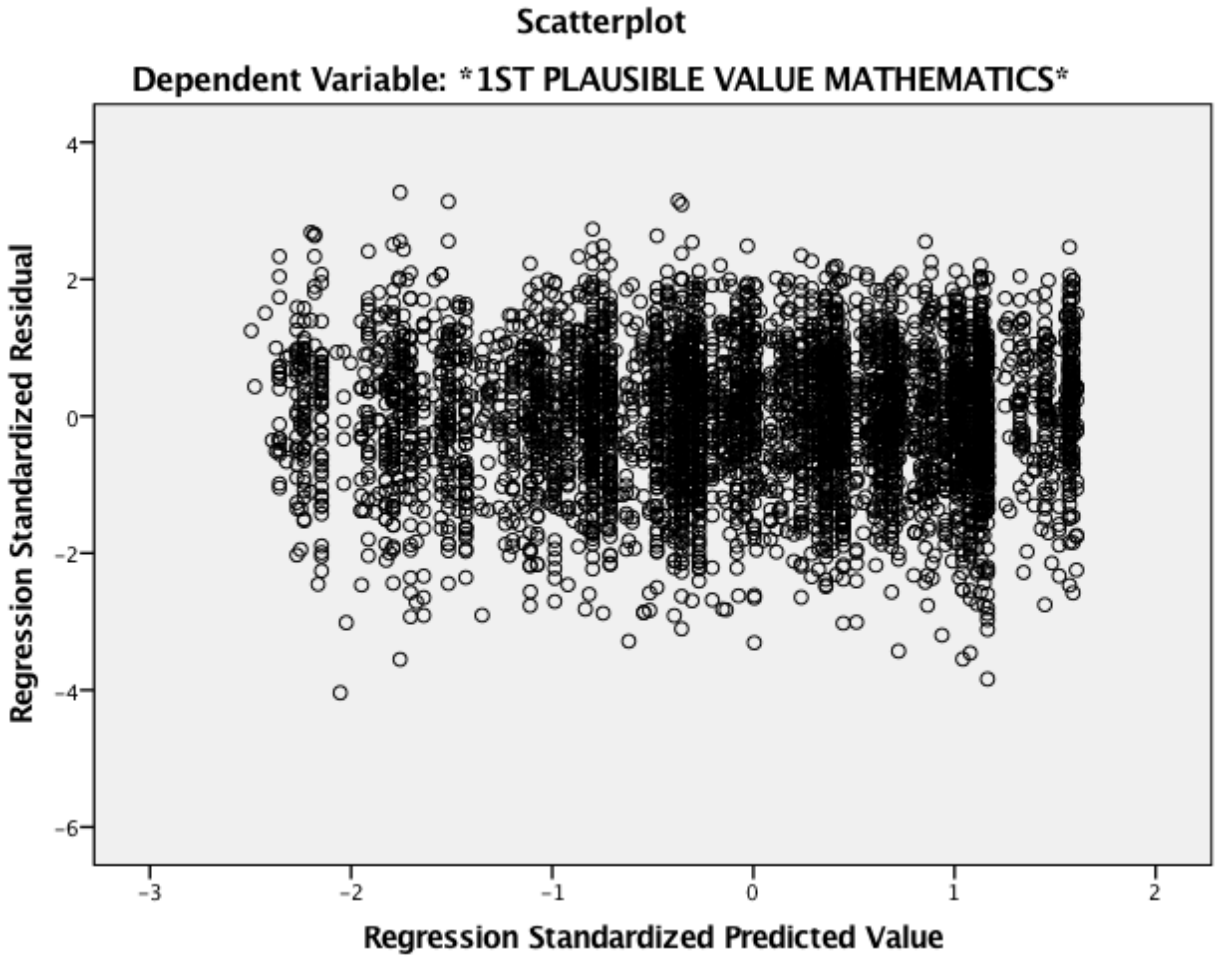

Figure 11: Scatterplot indicates homoscedasticity with standardized residuals by predicted values.

d. Multicollinearity: Multicollinearity assumption was tested through correlation function of SPSS between all independent variables in the model. The results are represented in the Table 36 and it indicates low correlation between independent variables. Therefore, there is no multicollinearity problem in the model.

Table 36

Correlation between independent variables in Korea

Independent Variables	Testing frequency	Number of books at home	Fathers' highest level of education	School areas' income level
Testing frequency	1	.02	.01	.09
Number of books at home	.02	1	.10	-.16
Fathers' highest level of education	.01	.10	1	-.08
School areas' income level	.09	-.16	-.08	1

2. Singapore

a. Normality: Histograms in Figures 16, 17, and 18 indicates that normality assumption is not violated for any level of the predictor variable in the model, except *never* testing frequency level as there were very small number of students in this group. Because of low number of students in never quiz frequency group, this quiz frequency group was excluded from data analysis.

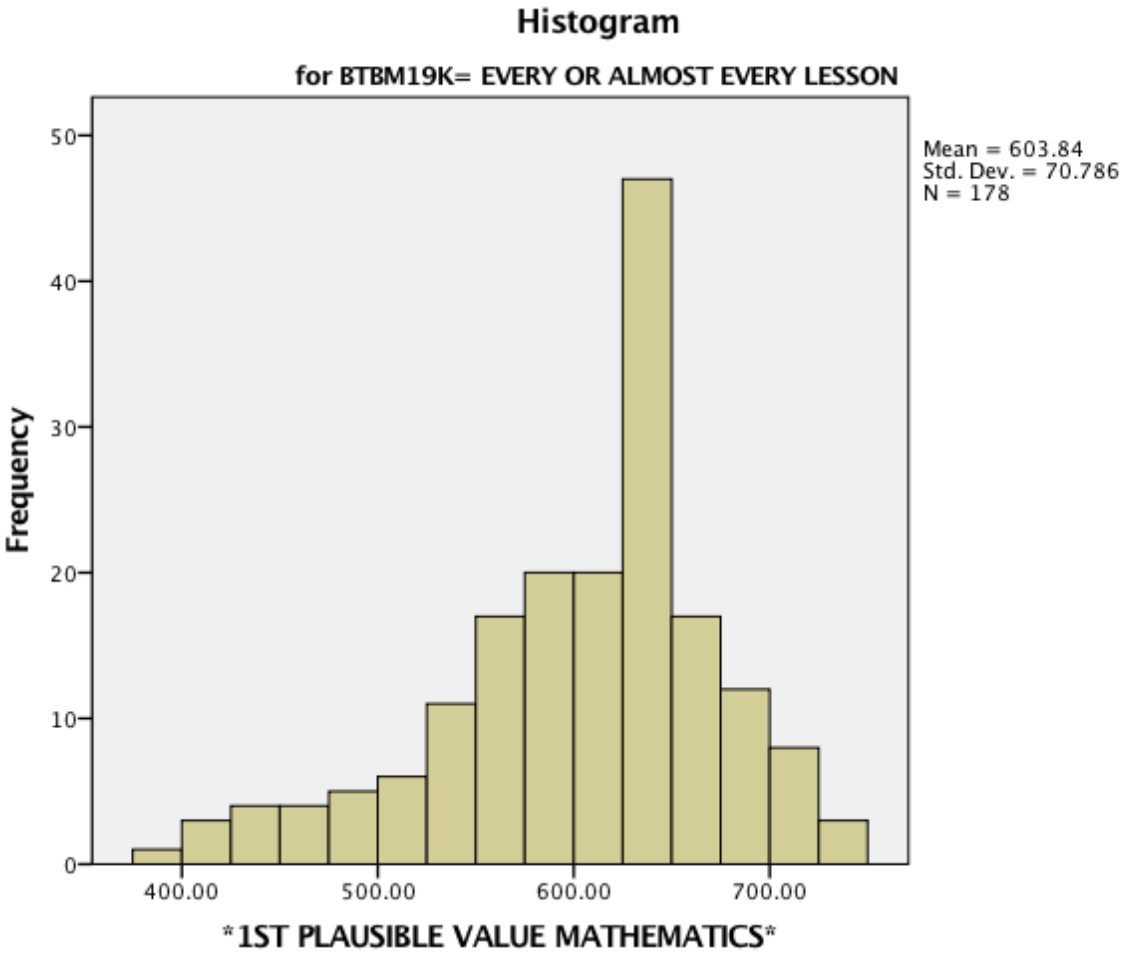

Figure 12: Distribution of achievement scores for every or almost every lesson testing frequency.

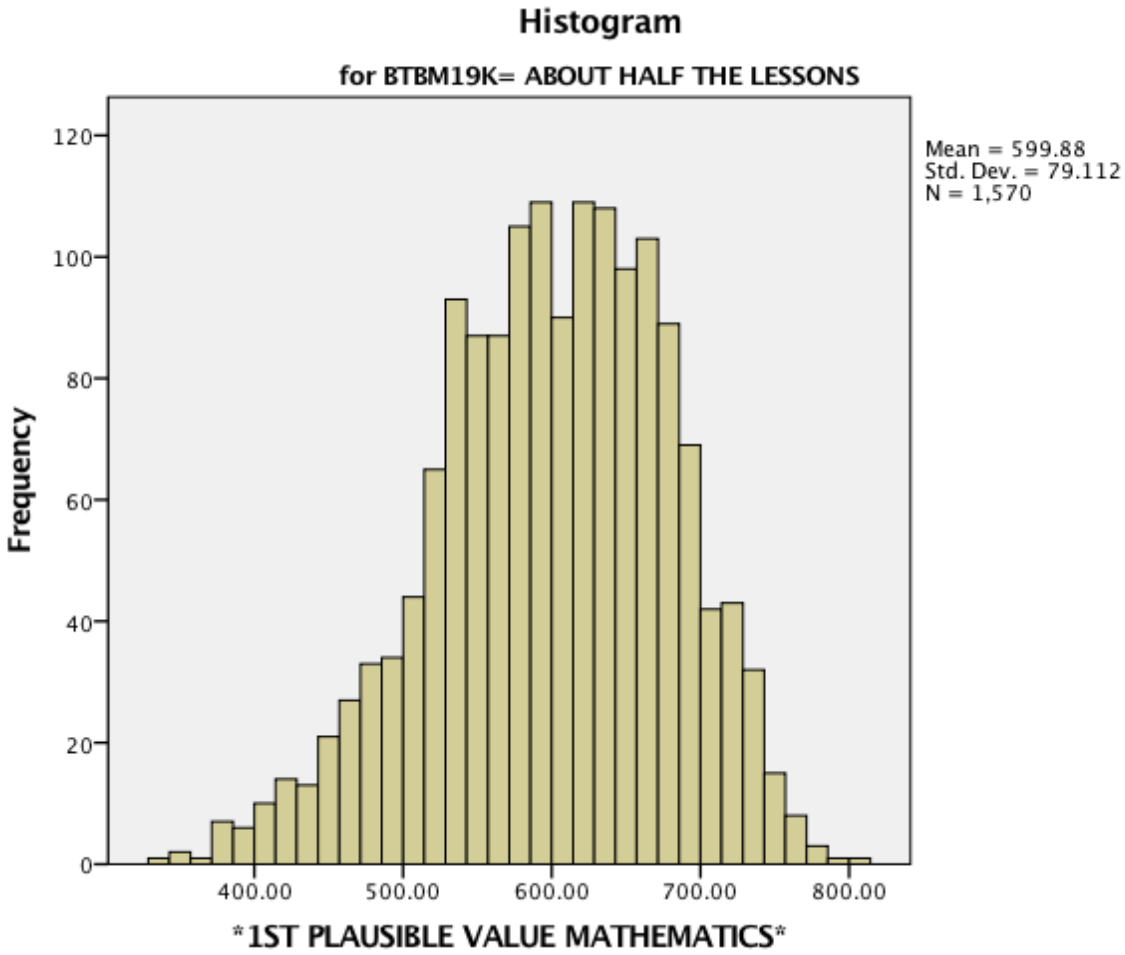

Figure 13: Distribution of achievement scores for about half the lessons testing frequency.

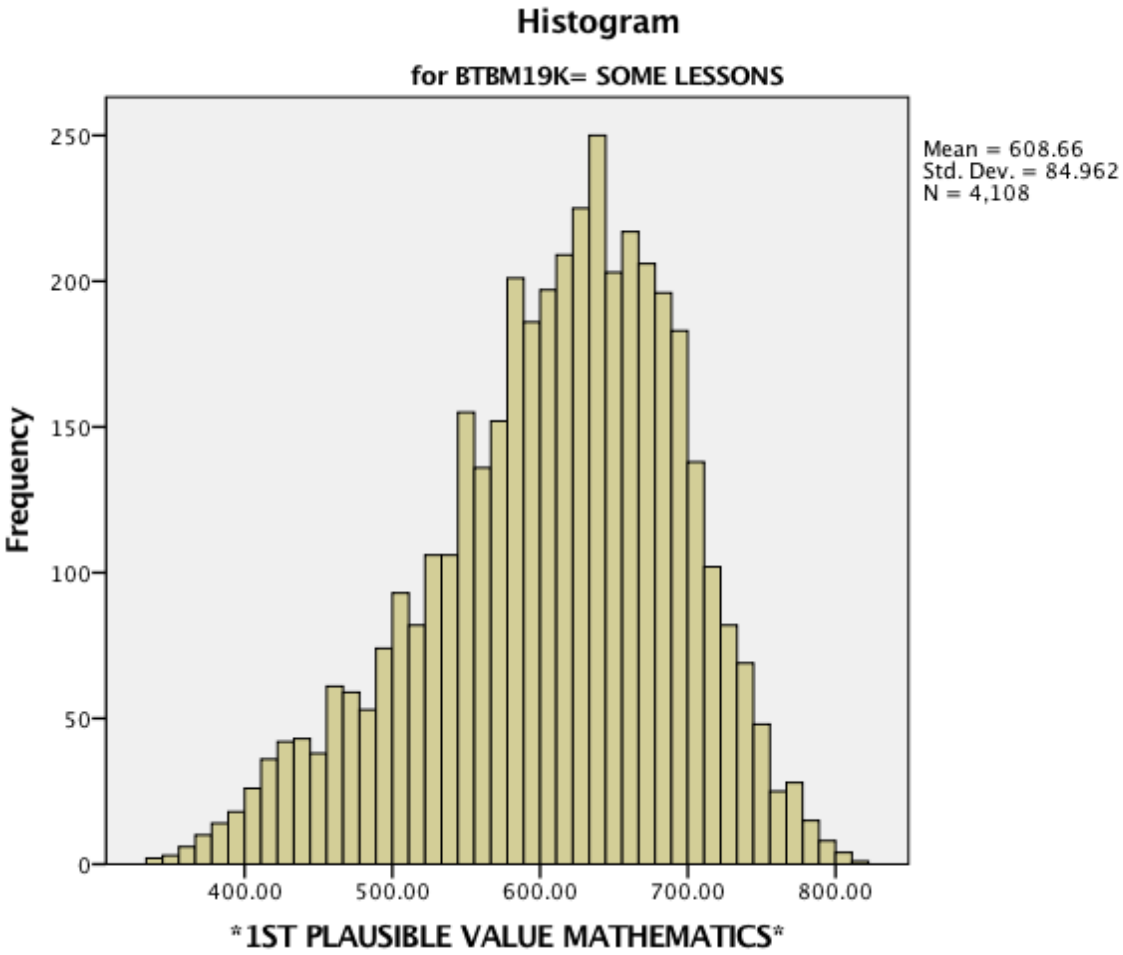

Figure 14: Distribution of achievement scores for some lessons testing frequency.

b. Linearity assumption: Linearity assumption between predictor and dependent variables are tested through linear regression function of the SPSS. Figure 19 is a scatterplot representation of standardized residuals by predicted values. It can be seen that values are scattered around a horizontal line and there is not much departure from linearity. Therefore, linearity assumption is met in the model.

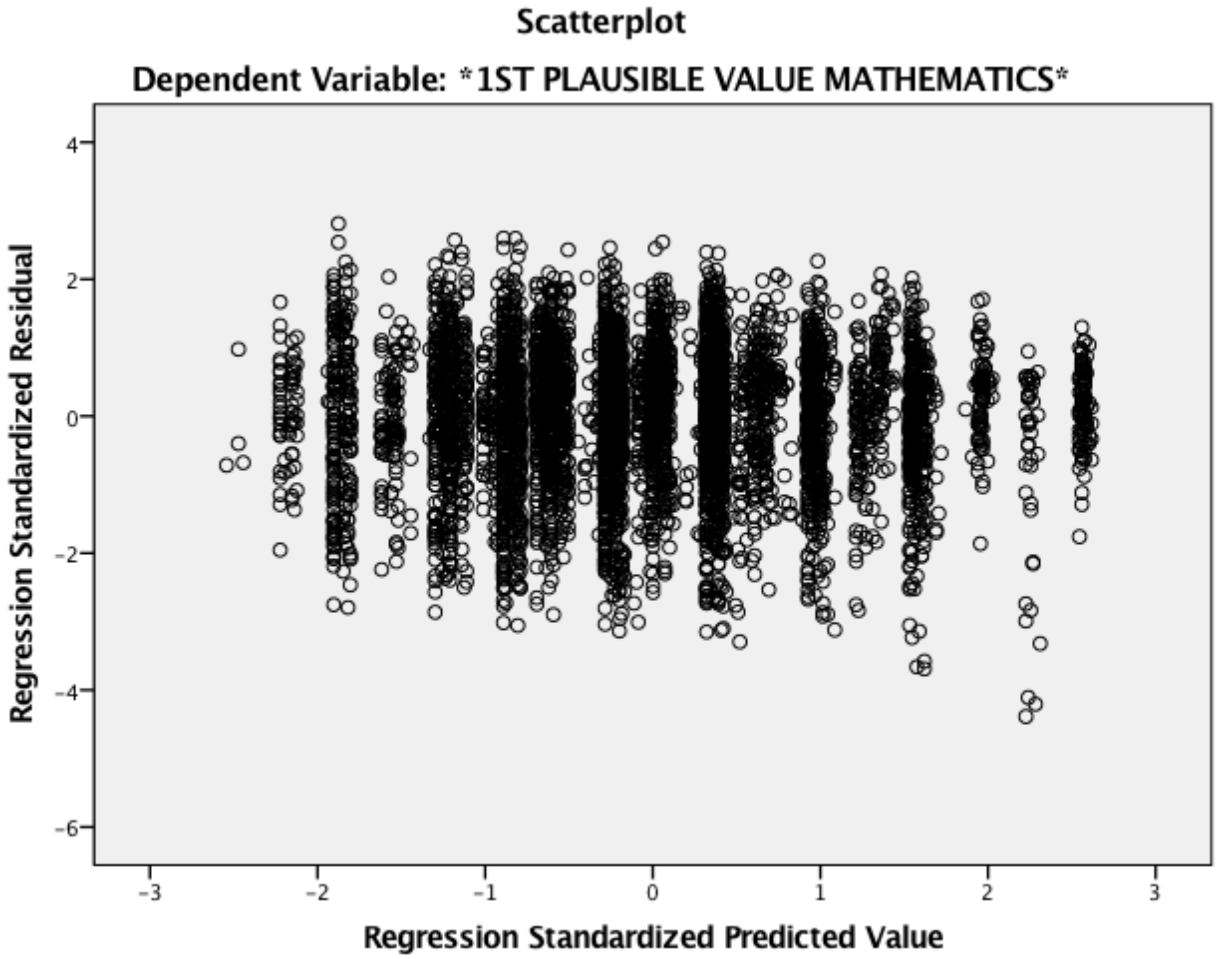

Figure 15: Scatterplot indicates linear relationship with standardized residuals by predicted values

c. Homoscedasticity: Homoscedasticity assumption test was also tested through linear regression function of the SPSS. Figure 20 indicates variances of errors are similar across the predicted values. Residuals are equally scattered around the horizontal line. It can be concluded that homoscedasticity assumption is not violated in the regression model.

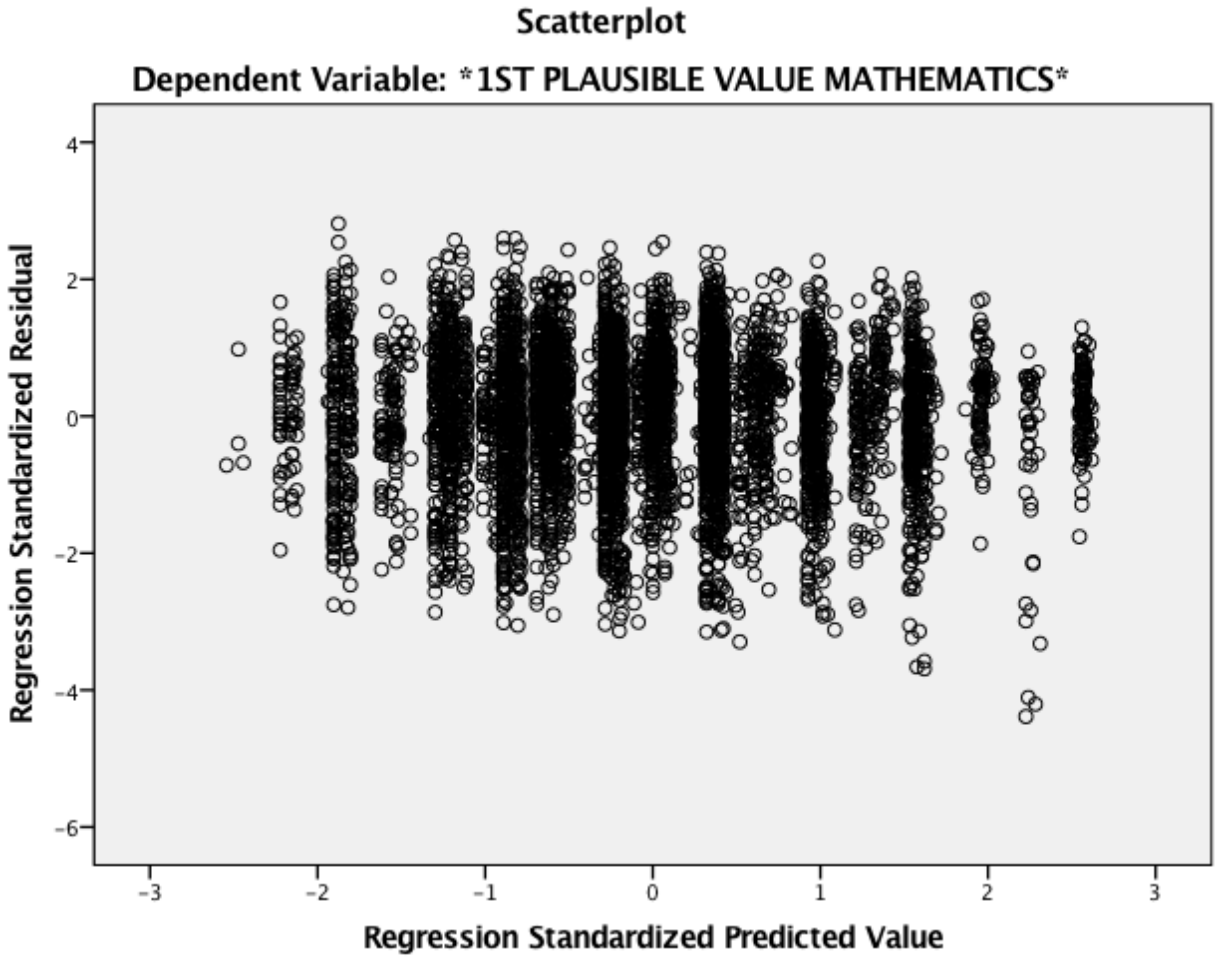

Figure 16: Scatterplot indicates homoscedasticity with standardized residuals by predicted values.

d. Multicollinearity: Multicollinearity assumption was tested through correlation function of SPSS between all independent variables in the model. The results are represented in the Table 37 and it indicates low correlation between independent variables. Therefore, there is no multicollinearity problem in the model.

Table 37

Correlation between independent variables in Singapore

Independent Variables	Testing frequency	Number of books at home	Fathers' highest level of education	School areas' income level
Testing frequency	1	.04	.03	.09
Number of books at home	.04	1	.13	-.20
Fathers' highest level of education	.03	.13	1	-.08
School areas' income level	.09	-.20	-.08	1

3. Turkey

a. Normality Assumption: Histograms in Figures 21, 22, and 23 indicates that normality assumption is not violated for any level of the predictor variable in the model, except *never* testing frequency level as there were no students in this group. Because of that, this quiz frequency group was excluded from data analysis.

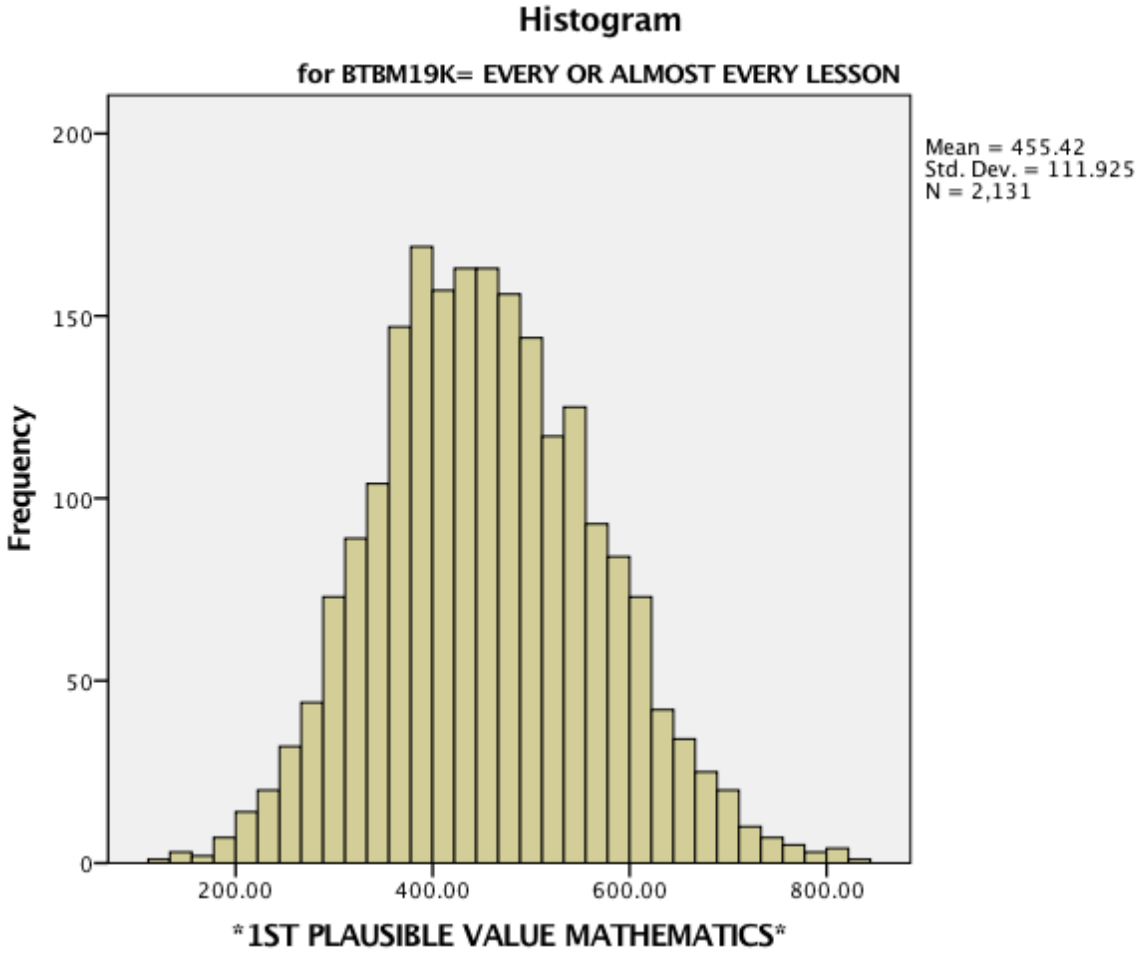

Figure 17: Distribution of achievement scores for every or almost every lesson testing frequency.

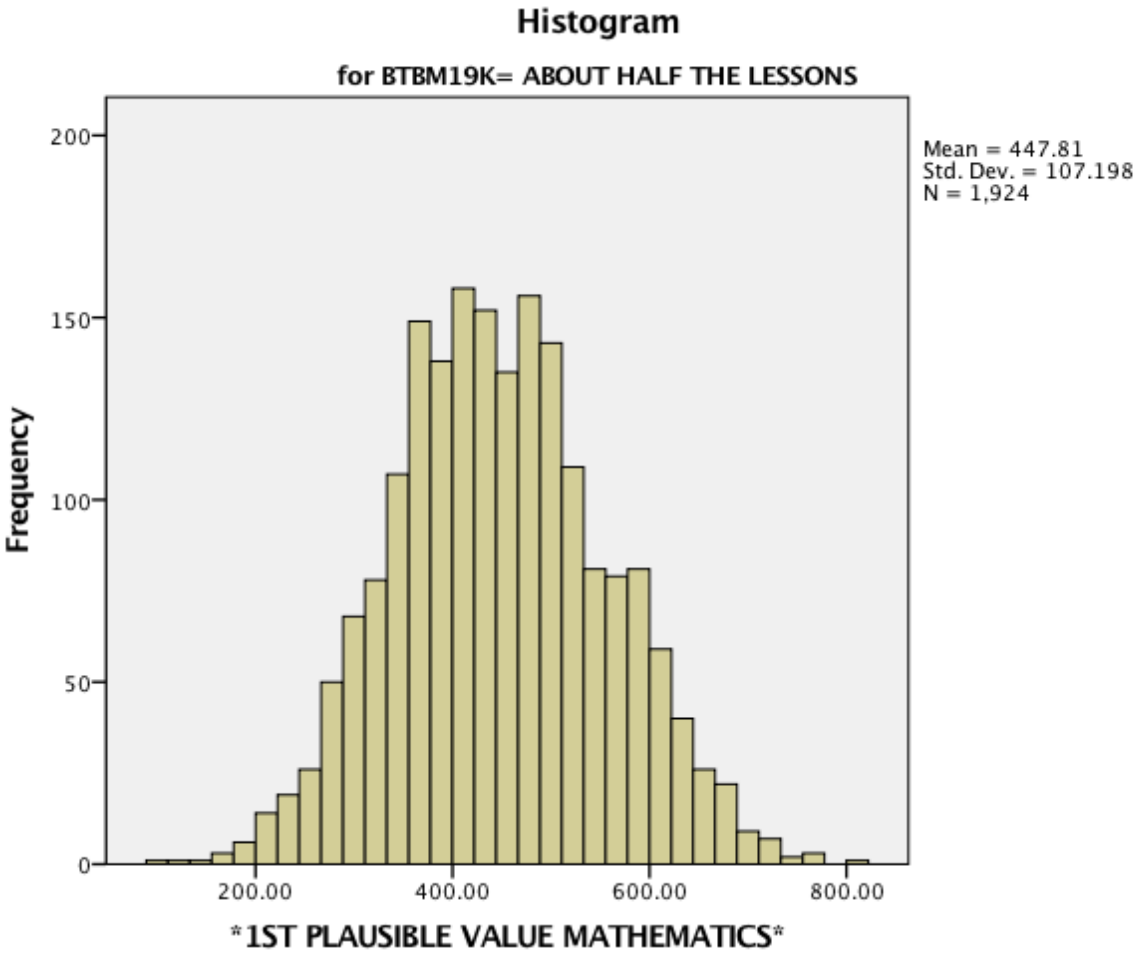

Figure 18: Distribution of achievement scores for about half the lessons testing frequency.

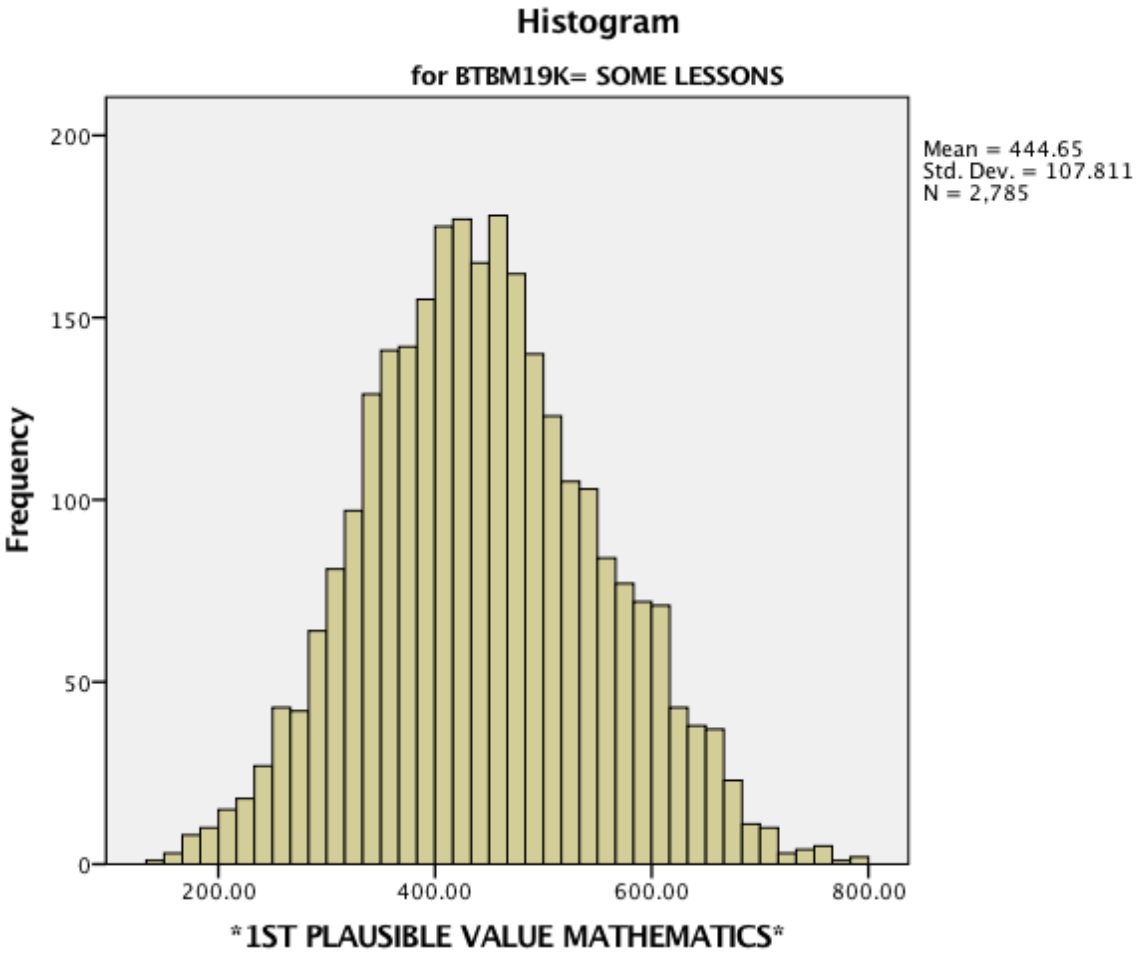

Figure 19: Distribution of achievement scores for some lessons testing frequency.

b. Linearity assumption: Linearity assumption between predictor and dependent variables are tested through linear regression function of the SPSS. Figure 24 is a scatterplot representation of standardized residuals by predicted values. It can be seen that values are scattered around a horizontal line and there is not much departure from linearity. Therefore, linearity assumption is met in the model.

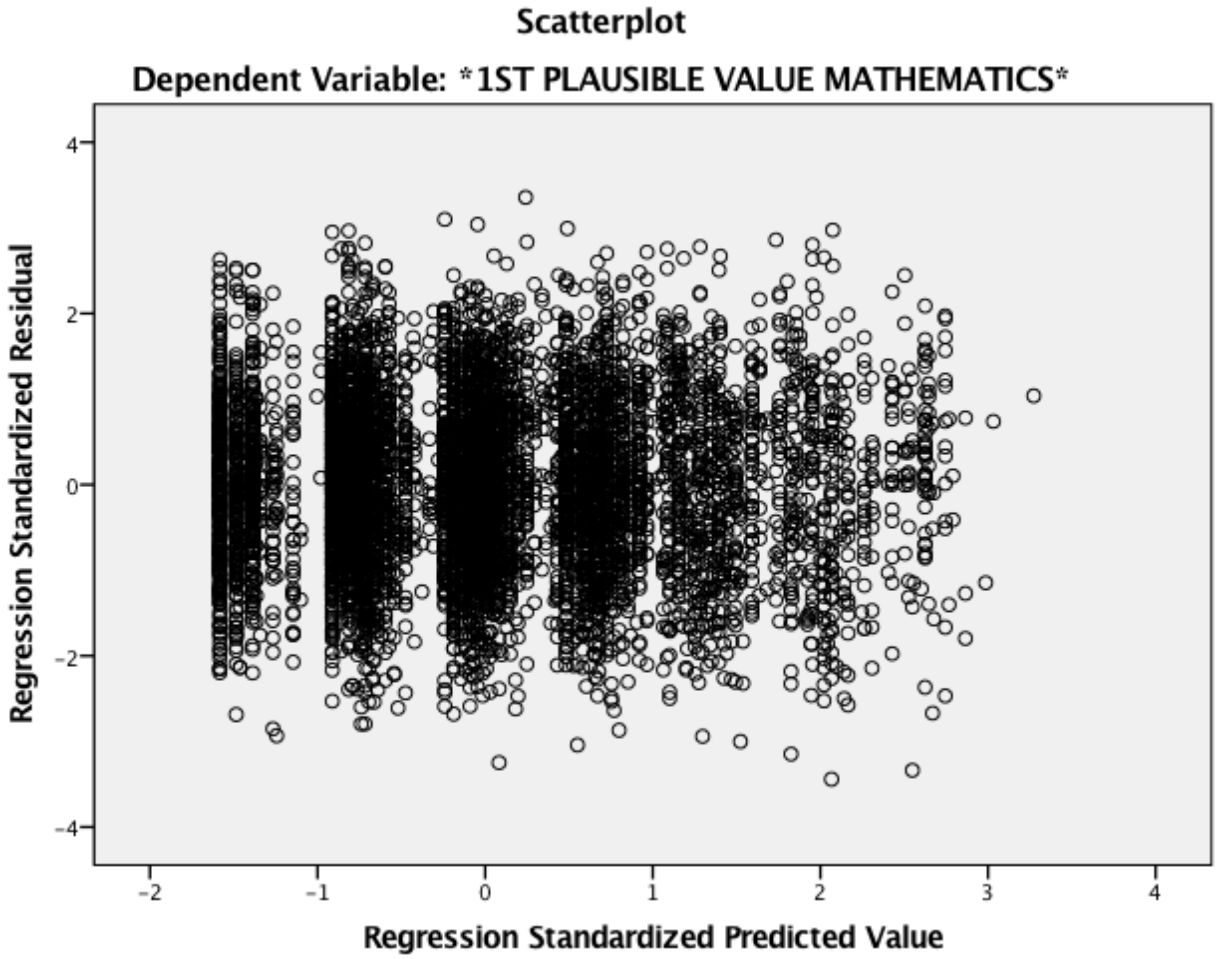

Figure 20: Scatterplot indicates linear relationship with standardized residuals by predicted values

c. Homoscedasticity: Homoscedasticity assumption test was also tested through linear regression function of the SPSS. Figure 25 indicates variances of errors are similar across the predicted values. Residuals are equally scattered around the horizontal line. It can be concluded that homoscedasticity assumption is not violated in the regression model.

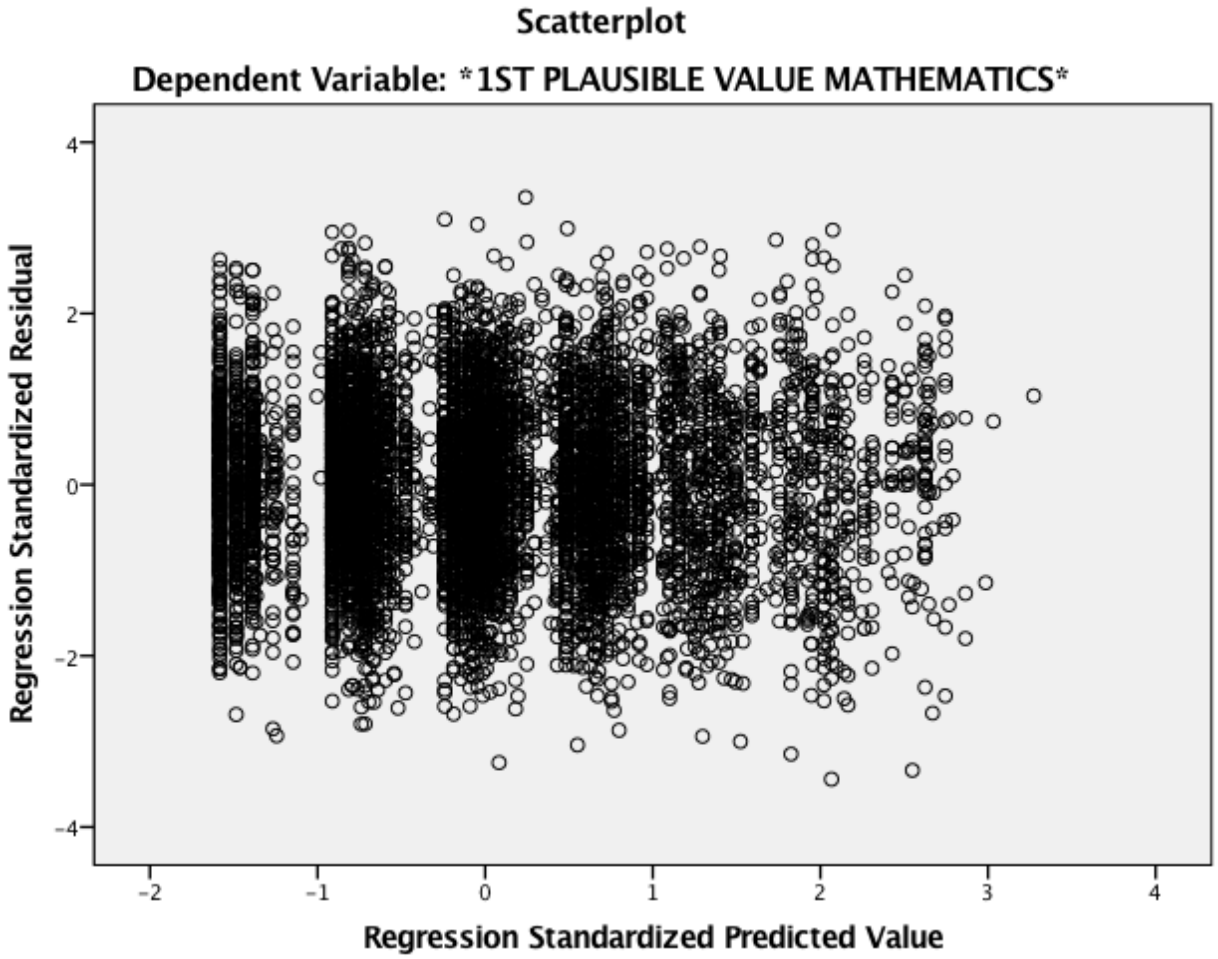

Figure 21: Scatterplot indicates homoscedasticity with standardized residuals by predicted values.

d. Multicollinearity: Multicollinearity assumption was tested through correlation function of SPSS between all independent variables in the model. The results are represented in the Table 38 and it indicates low correlation between independent variables. Therefore, there is no multicollinearity problem in the model.

Table 38

Correlation between independent variables in Turkey

Independent Variables	Testing frequency	Number of books at home	Fathers' highest level of education	School areas' income level
Testing frequency	1	-.03	-.03	-.02
Number of books at home	-.03	1	.27	-.24
Fathers' highest level of education	-.03	.27	1	-.22
School areas' income level	-.02	-.24	-.22	1

4. United States

a. Normality Assumption: Histograms in Figures 26, 27, and 28 indicates that normality assumption is not violated for any level of the predictor variable in the model, except *never* testing frequency level as there were very small number of students in this group. Because of low number of students in never quiz frequency group, this quiz frequency group was excluded from data analysis.

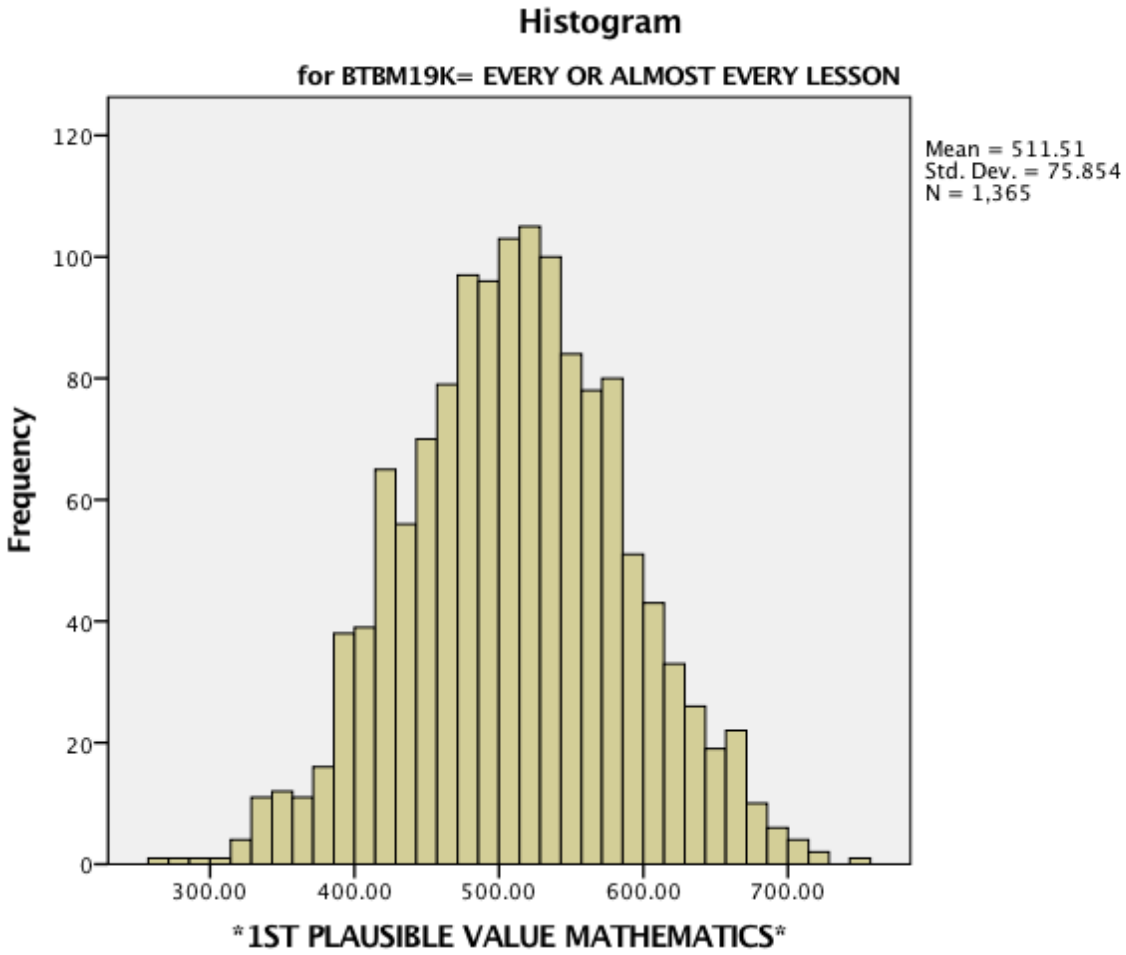

Figure 22: Distribution of achievement scores for every or almost every lesson testing frequency.

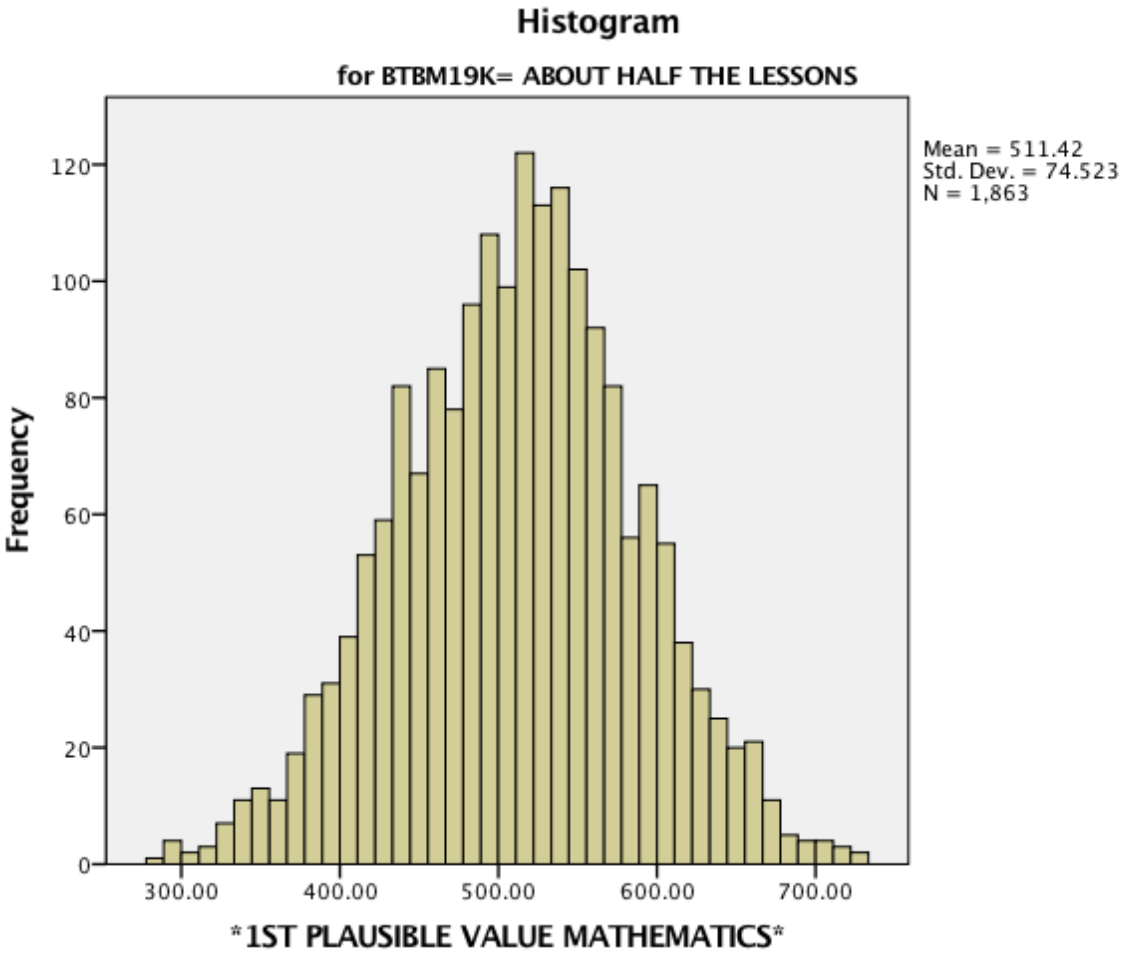

Figure 23: Distribution of achievement scores for about half the lessons testing frequency.

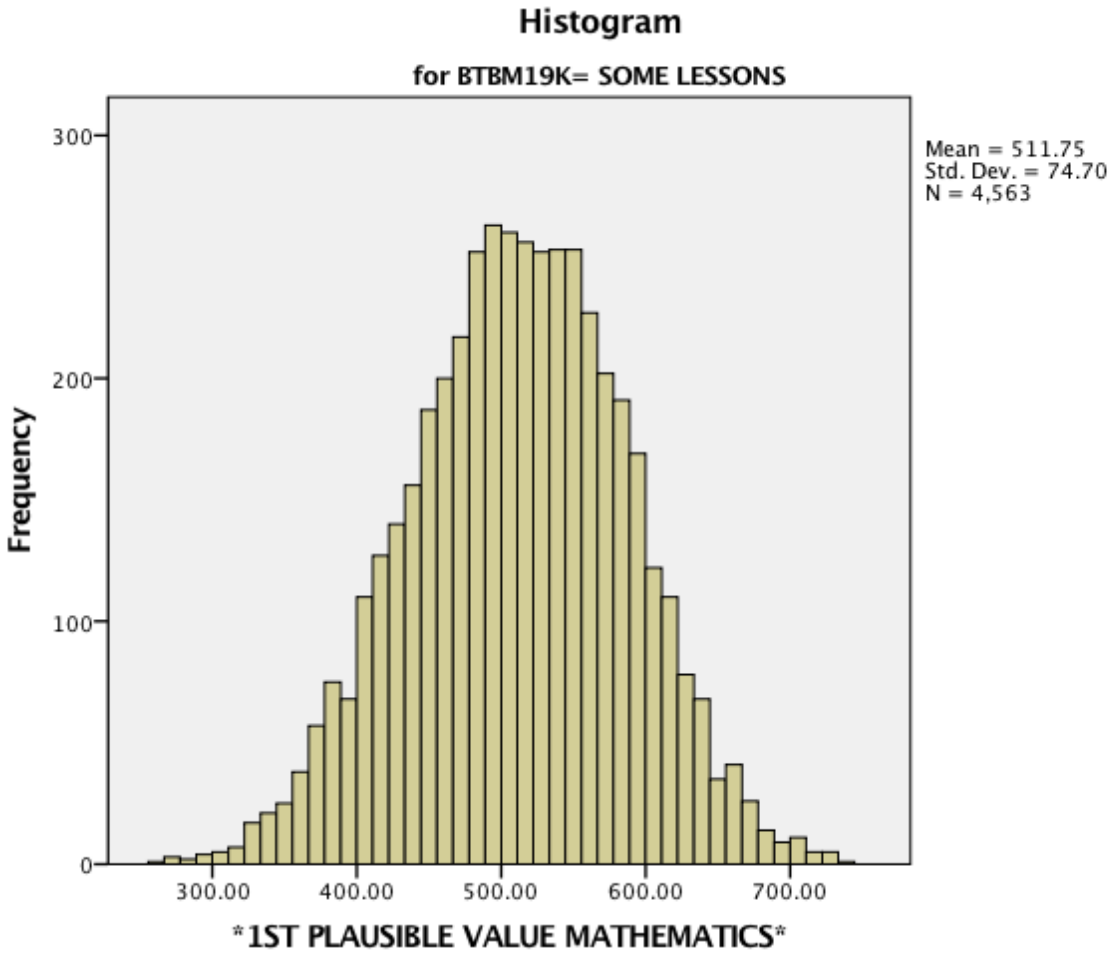

Figure 24: Distribution of achievement scores for some lessons testing frequency.

b. Linearity assumption: Linearity assumption between predictor and dependent variables are tested through linear regression function of the SPSS. Figure 29 is a scatterplot representation of standardized residuals by predicted values. It can be seen that values are scattered around a horizontal line and there is not much departure from linearity. Therefore, linearity assumption is met in the model.

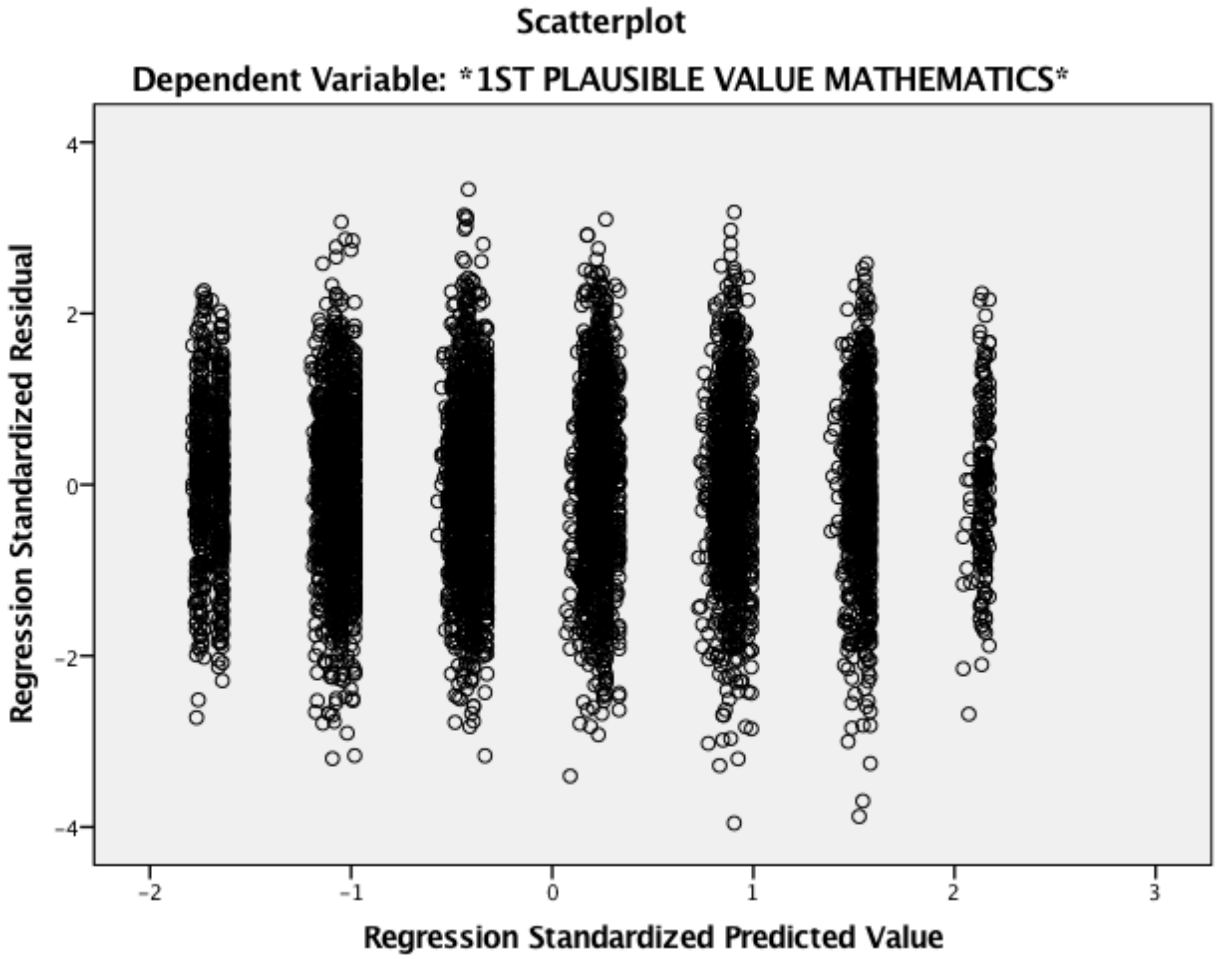

Figure 25: Scatterplot indicates linear relationship with standardized residuals by predicted values

c. Homoscedasticity: Homoscedasticity assumption test was also tested through linear regression function of the SPSS. Figure 30 indicates variances of errors are similar across the predicted values. Residuals are equally scattered around the horizontal line. It can be concluded that homoscedasticity assumption is not violated in the regression model.

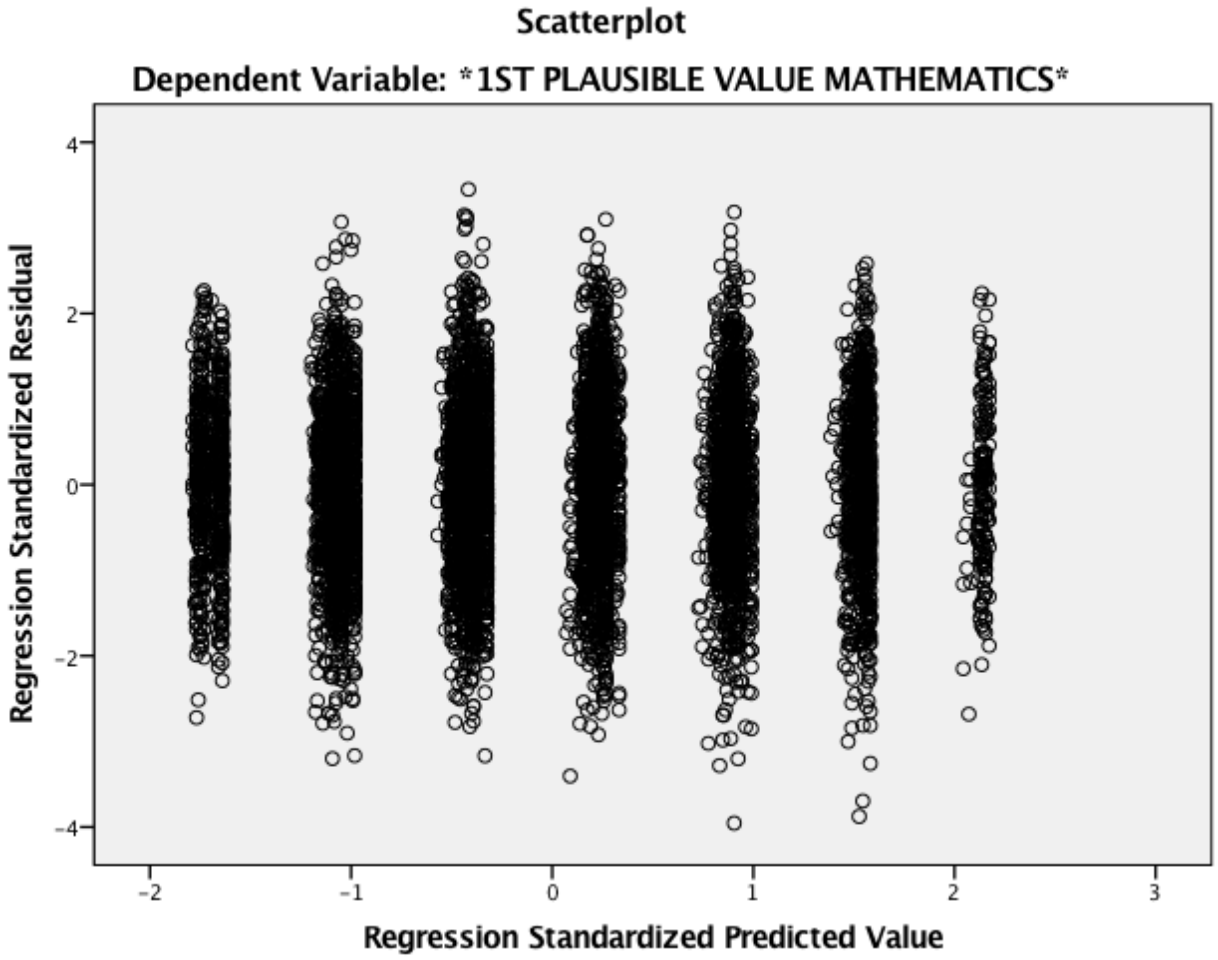

Figure 26: Scatterplot indicates homoscedasticity with standardized residuals by predicted values.

d. Multicollinearity: Multicollinearity assumption was tested through correlation function of SPSS between all independent variables in the model. The results are represented in the Table 39 and it indicates low correlation between independent variables. Therefore, there is no multicollinearity problem in the model.

Table 39

Correlation between independent variables in United States

Independent Variables	Testing frequency	Number of books at home	Fathers' highest level of education	School areas' income level
Testing frequency	1	.04	.03	.02
Number of books at home	.04	1	.12	-.23
Fathers' highest level of education	.03	.12	1	-.13
School areas' income level	.02	-.23	-.13	1

Appendix F: Testing Frequency Mean Scores for All Countries

Table 40

TIMSS 2011 participating countries' average achievement scores and their testing frequency means

Countries	Testing Frequency Mean	eighth-grade Math Achievement
Korea	2.59	613
Singapore	2.67	611
Chinese Tapei	2.72	609
Hong Kong	2.54	586
Japan	2.49	570
Russia	2.37	539
Israel	2.50	516
Finland	3.01	514
United States	2.43	509
England	2.97	507
Hungary	2.58	505
Australia	2.79	505
Slovenia	2.96	505
Lithuania	2.88	502
TIMSS Scale Centerpoint		500
Italy	2.59	498
New Zealand	2.79	488
Kazakhstan	2.26	487
Sweden	2.86	484
Ukraine	2.50	479
Armenia	2.86	467
Romania	2.16	458
United Arab Emirates	1.99	456
Turkey	2.11	452
Lebanon	1.53	449
Malaysia	2.31	440
Georgia	2.34	431
Thailand	2.02	427
Macedonia	2.66	426
Tunisia	2.02	425
Chili	2.33	416
Iran	1.56	415
Qatar	1.97	410
Bahrain	2.00	409
Jordan	2.59	406
Palestinian National Authority	1.87	404
Botswana	2.62	397
Saudi Arabia	1.78	394
Syria	2.39	386
Indonesia	1.83	380
Morocco	1.83	371
Oman	2.00	366
Honduras	2.04	352
South Africa	2.36	338
Ghana	1.80	331