READABILITY OF THE NEW YORK STATE REGENTS EXAM IN UNITED STATES HISTORY AND GOVERNMENT

by

Daniel E. Morton

A Master's Project
Submitted in Partial Fulfillment
of the Requirement for the Degree of
Master of Science in Education
Curriculum and Instruction Inclusive Education
Department of Education
State University of New York at Fredonia
Fredonia, New York

May 2015

Fredonia, The State University of New York
Department of Curriculum and Instruction

CERTIFICATION OF PROJECT WORK

We, the undersigned, certify that this project entitled REABILITY OF THE NEW YORK STATE REGENTS EXAM IN UNITED STATES HISTORY AND GOVERNMENT by Daniel E. Morton, candidate for the Degree of Masters of Science in Education, Curriculum and Instruction, is acceptable in form and content and demonstrates a satisfactory knowledge of the field covered by this project.

[redacted]

Master's Project Advisor
EDU 691 Course Instructor
Department of Curriculum and Instruction

4/30/15
Date

[redacted]

Department Chair Dr. Robert Dahlgren
Department of Curriculum and Instruction

5/4/2015
Date

[redacted]

Dean Christine Givner
College of Education
At Fredonia, The State University of New York

5/6/15
Date

Abstract

This study investigated the readability of the multiple-choice section on the New York State Regents Exam in United States History and Government. Every June Regents Exam, from 2014-1990, was analyzed for readability. The Homan, Hewitt, & Linder (1994) formula was utilized because this formula measures grade level readability for multiple-choice questions. Readability was determined by randomly selecting three multiple-choice questions from each exam to analyze. Readability was calculated for each question and averaged to determine the mean score for each exam. This study revealed that over time the NYS Regent Exam in United States History and Government has become easier to read. There are far-reaching implications with regard to teacher evaluations and test reliability and validity, as a result of this study.

Table of Contents

	Pages
Introduction	1
Statement of Problem	2
Literature Review	4
Methodology	13
Results	17
Discussion	21
References	27

Introduction

One graduation requirement for all students in the New York State educational system is to successfully complete the United States History and Government Regents Exam. This exam is a state generated cumulative test of United States history. The minimum passing score for this exam is 65 percent. There was a time that not every student had to take this examination. Students could pursue a Regents track or a general education track diploma in the State. This gave greater weight to achieving a Regents diploma. However, eventually it was deemed by the Board of Regents in the State of New York that every student would receive a Regents diploma in a wave of education reform in recent years past. Then like now, New York States and the country were experiencing another wave of educational reform.

Annual Professional Performance Review (APPR) and Common Core Learning Standards (Common Core) are this next wave of reform. APPR is designed to better evaluate teachers in their performance as educators and Common Core has increased the rigor of the standards students must meet in order graduate. One measure of an educator's effectiveness is the ability to show growth in their students' learning. Some districts achieve this by have students complete a baseline test at the beginning of the academic year which is then used to compare future performance against. For example, all my students completed a baseline test in United States History and Government. If my students show 20% growth from their baseline scores a very critical aspect of my APPR score has been fulfilled. Growth in determined by comparing baseline scores with the Regents exams at the end of the year. Debating the merits of this system is beyond the scope of this paper, that being said, there are numerous flaws in new evaluation system. What is important is the fact that student test score play a significant role in determining teacher performance and effectiveness.

In New York State all students are expected to take and complete these cumulative test even students with Individualized Educational Plans (IEPs) and/or 504 plans. Students who read at grade level, and students who read at or above a third grade level, all take these examinations. Again, debating the merits of this system is beyond this paper. Analysis is needed regarding the readability of the United States History and Government Regents Exam. Because student success and teacher accountability are so strongly link to this one exam the validity needs to be unquestioned. However, regardless of the level of readability of the students this examination should be constructed at an 11th grade reading level. This study seeks to investigate the readability of this exam.

Statement of Problem

To date, there has been no systematic readability analysis of the New York State History and Government Regents exam for 11th grade students. In recent years, a study of this caliber has become essential due in no small part to the ever increasing standards districts and educators across New York are facing. State wide Common Core testing standards and teacher Annual Professional Performance Review Plans (APPR) have ushered in a new era of accountability and scrutiny. Regents' assessment scores now play a critical role in determining the effectiveness of districts and educators. Districts that do not meet or exceed the new standards can be labeled as a district in need of an improvement plan. Educators who fail to meet APPR plans are given a Teacher Improvement Plan (TIP) and those who fail to improve can be removed from their position. This is an oversimplification of New York State education reform, however, a detailed explanation is beyond the scope of this paper. For high school educators that instruct courses that end in a Regents exam, how well their students perform factors into APPR scores, which determines teacher effectiveness.

In this new era, high stakes testing has even greater emphasis for all involved in the educational process. Much of the attention is on teacher performance in the classroom and testing outcomes, however, what if the assessments given to students were beyond their reading level? Hewitt and Homan (2004) measured the readability of state social studies assessments for 3rd, 4th, and 5th graders. The study analyzed only the multiple-choice section. It was determined that the higher the readability level the lower the percentage who answered the questions correctly. This was the only study discovered that analyzed the readability of state testing in the social studies. The implications from this study are significant. It is clear that further research is warranted in this area. If student performance on state testing decreases as readability increases what does this say about the validity of these exams? This study seeks to lay the foundation for further research into readability of New York State United States History and Government Regents exam. There is also a very practical need for such a study. Educators should know if the assessment that is determining their effectiveness has a readability beyond the intended grade level. This is important for both educators and policy makers. If there is to be an accurate measure of student performance readability must match the intended grade level. The United States History and Government exam is intended to assess content knowledge, not reading comprehension.

As a social studies educator in New York State, I have often wondered from a historical perspective how the exam has changed over time. Analyzing this exam for readability will provide a greater depth of understanding into the frequency, caliber, and consistency of questions. This study will also shed light on the eras in United States History that received the greatest coverage.

Literature Review

Numerous researchers agree that determining the readability of textual information is an imperfect science (Armbruster, Osborn, & Davison, 1985; Burk & Greenberg, 2010; Crossley, Allen, & McNamara, 2011; Fry, 2002; McConnell, & Paden, 1983; Osborn, Jones, & Stein, 1985). A textbook could have multiple readability formulas applied to analyze the difficulty and each formula would provide a different conclusion as to the readability (Burk & Greenberg, 2010; Crossley, Allen, & McNamara, 2011; McConnell & Paden, 1983). Most readability formulas analyze three different 100 word sections of a text and average the sections to determine readability. In a hypothetical situation those three sections could be determined to have the following grade reading levels 3rd, 5th, and 10th. Therefore, in this hypothetical case the text would be 6th grade reading material. The implications of this scenario are staggering.

Readability Overview

What is the best methodology for determining readability and producing grade level textbooks and materials? What is the interplay of readability, comprehension, and student achievement? More importantly how does readability affect student achievement on standardized tests? In this era of accountability and standardization understanding readability in relation to state examinations is of critical importance not only to students but, teachers, school administrators, and parents. This review of the extant literature seeks to provide a better understanding about the implications of readability analysis of student learning materials (textbooks, materials, and tests) and more specifically the potential implications of readability on the New York State History and Government Regents exam.

Readability measures date back to the early 20th Century since then, numerous formulas have been developed for all different grade levels. Traditional readability formulas measure two

factors – sentence length for syntactic difficulty and word length measured in syllables for semantic difficulty. Readability formulas are an objective means of determining appropriate reading material for specific grade levels. Readability can be determined by hand method (counting) and computer programs (Fry, 2002). Fry comments: "Readability outside the classroom is used for a wide range of materials including such things as military training manuals, plain-language laws for insurance policies or loan contracts, and newspaper articles" (p. 289). According to Fry (2002), "readability has the strength of a large research base with many formal validity studies showing high correlations with reading comprehension [and] reading errors…readability [also] has the strength of objectivity and consistency that any person or computer will get the same score" (p. 291). Burk and Greenberg (2010) performed an exhaustive study analyzing numerous readability formulas and concluded that determining readability is a simplistic act not requiring expensive computer programs or extensive training. Fry (2002) writes with broad strokes as he summarizes and provides an overview of readability. This idea of consistency is however somewhat misleading.

The inconsistencies of readability formulas have been well documented (Armbruster, Osborn, & Davison, 1985; Burk & Greenberg, 2010; Crossley, Allen, & McNamara, 2011; Fry, 2002; McConnell, & Paden, 1983). What Fry (2002) means by consistency is that if one formula is applied to one specific reading passage whoever analyzes – it human or computer – will get the same result. McConnell & Paden (1983) provide data illustrating the inconsistency of multiple formulas applied to a single introductory level economic textbook. McConnell & Paden (1983) examined four readability analyses applied to a single textbook and ranked the scores from 1 to 9, with 1 representing the most difficult and 9 representing the least difficult. One textbook received the following scores Dale-Chall (8), Modified Dale-Chall (1), Fry (3-4),

and Flesch (8) (please note that the numerical values are preceded by the name of the readability formula). Armbruster, Osborn, & Davison (1985) analyzed four 100 word passages from a 5th grade textbook utilizing the simplistic Fry Graph. The four different passages received the following grade level scores 4th, 7th, 8th, and 11th grade. As stated above these inconsistencies are real and the implications stemming from these are significant. For educators to make an informed decision on course materials for developing readers multiple factors must be considered. Readability is just one of the factors. Textbook publishers should test materials with multiple readability formulas in order to provide educators with sound data so they can make informed decisions. Textbook publishers are, however, in the business of making money and need to provide accurate ratings for their educational materials if they want to remain relevant. There are other consequences besides readability score inconsistency.

Readability of Educational Materials

Armbruster, Osborn, & Davison (1985) state two major problems associated with selecting educational materials based on readability formulas. First, text construction and layout as well as author play important roles. Second, readability formulas do not take student factors such as motivation, resiliency, love of learning, and enthusiasm into account. Fry (2002) also states that student motivation and appropriateness of the materials are limitations of readability formulas. Another factor to consider and one in need of further research is the effect of taking complex texts and making them more readable.

Osborn, Jones, & Stein (1985) summarized earlier research that criticized taking complex texts and making them more readable. Doing so simply requires a publisher to decrease the rigor of vocabulary and shorten sentences.

> However, dividing long sentences into two or more shorter sentences may require a reader to make more inferences because the connection words that relate the parts of long sentence often get lost when the sentence is divided…[this] can seriously distort the logical relations between the parts of a text, disrupt the presentation of ideas, and make it difficult or impossible to convey the meaning of an original text (Osborn, Jones, & Stein, p. 14).

Therefore taking a complex textbook and making it more readable may actually make the material far more difficult to comprehend, thus leading to student frustration and decreased motivation (Armbruster, Osborn, & Davison, 1985; Osborn, Jones, and Stein,). McConnell (1985) observed that if an educator selects a textbook that is at or below the students' actual reading level, "students may fail to develop fully their reading and linguistic abilities" (p. 70). Crossley, Allen, & McNamara (2011) explain that publishers might take complex advanced textbooks and simplify them; making intermediate level textbooks, then further simplify them even further. Therefore, as the text moves through the stages of simplification meaning could become lost. Crossley, Allen & McNamara (2011) argue that intermediate level textbooks have the widest inaccuracies regarding readability because, "they [share] similarities with both beginning and advanced texts, many of which are not shared between beginning and advanced texts" (p. 97). This could be one explanation that many textbooks have wild fluctuations in readability levels.

Readability of Standardized Tests

Textbook readability research is extensive; however readability studies of standardized tests is limited in scope. New York like many other states requires students to complete cumulative standardized high stakes exams in order to graduate. What role does readability play

in student outcomes on these examinations? If a multiple-choice question is determined to have a high level of readability does this mean there is also a higher probability that a student will get it wrong?

One examination that every student must take is New York is the United States History and Government Regents exam. This exam consists of three parts: Part I includes 50 multiple-choice questions, Part II requires a thematic essay, and Part III includes document-based questions (DBQ), which are text-rich, and an essay based on those documents. The thematic essay is based on the various themes in United States History (e.g. foreign policy, domestic policy, social change, minority rights, presidential decisions, Constitutional principles, and Supreme Court decisions). Part III requires students to answer between 8-15 different document-based questions. The documents could be pictures, political cartoons, drawings, propaganda posters, or a variety of primary and secondary source materials. The majority of the documents are primary and secondary source materials. After answering the document-based questions, students weave them along with outside information into any essay. DBQ essay topics are much the same as the thematic essay topics.

This exam has a particularly forgiving scoring curve. For example, if a student answers every multiple-choice and document-based question correct he/she automatically starts off with an 80% - which is a mere five points away from mastery. The essays are graded on a 0-5 scale; therefore, utilizing this same example, if a student received a 1 on the thematic essay and a 2 on the DBQ essay he or she would receive roughly 90%. Debating the merits of this curve is beyond the scope of this paper however, knowing this information is vital for educators. This means it is extremely important for students to score well on the multiple-choice section and the

DBQs. Blackey (2009), an expert on multiple-choice questions and strategies, describes a multiple-choice question:

> Multiple-choice questions typically begin with a **stem**, which directs students to pick the correct answer from among four or five choices, or **options**. Ordinarily, each of the options is comparable in length to some or all of the remaining options for the question, so that no single option, and certainly not the correct one, stands out by virtue of its length or brevity. The wrong answers among the options are called **distracters**, whereas the correct answer is the **key.** (p. 54)

Blacky (2009) then goes on to describe the different types of questions for example, "analysis questions…interpretive questions…questions that include visuals" and finally provides numerous multiple-choice test taking strategies (p. 55). The article is extensive in its breadth and depth regarding the ins and outs of multiple-choice questions. However Blackey, who also acknowledges the fact that he writes questions for state assessments, does not mention even once readability of questions. Blackey (2009) states that, "questions are generally meant to be straightforward, not purposely complicated or confusing, and most multiple-choice exams include a variety of questions that run the gamut from fairly easy to especially challenging" (p. 63). One factor missing from his assessment of multiple-choice questioning is that of readability. Hewitt and Homan (2004) studied the readability of single-sentence items (multiple-choice questions) of state social studies examinations grades 3, 4, and 5. They concluded that, "mean scores progressively decreased as the readability level increased" on multiple-choice questions (p. 6). Hewitt and Homan (2004) went on to say that answering a question wrong might be a result of reading and comprehension difficulties rather than a deficiency in content knowledge. The possibility exists that students may be missing standardized test questions

because they are written poorly or not written at the appropriate grade level. Students may struggle to read and comprehend the questions, thus leading to poor test scores. The type and manner of instruction received also play a role in student outcomes on tests.

Bulgren, Marquis, Deshler, and Schumaker (2011) analyzed the impact of teacher-centered learning and student-centered learning on test performance. The group that received student-centered learning via an advanced graphic organizer scored significantly higher in testing as opposed to the group that received teacher-centered instruction via straight lecture and discussion. Therefore, quality of instruction is another factor that can impact student outcome on state assessments. A question may rate high on a readability scale but the content may not have been covered in class or the student could have just received poor instruction.

Other factors to consider are student motivation and study habits. Yonker (2011) researched the impact of study strategies on student performance on publisher-generated multiple-choice questions. Yonker determined via questionnaire that no matter the study strategy rote or authentic students did not perform well on complex multiple-choice questions. However, students that utilized authentic (deep) study strategies outperformed those who only utilized rote (surface) strategies on the fact-based multiple-choice questions. This study was conducted at the collegiate level; however, there was no mention of readability in the study. Perhaps the readability of the applied questions was beyond the average student's ability in the class, but this is mere speculation because this factor was not discussed in the study. This study did showcase the difficulty of assessing student learning beyond fact based questions. Many high school and college students struggle to correctly answer applied or complex multiple-choice questions which ask the students to synthesize or evaluate information to produce the answer.

Many question the validity of multiple-choice questions altogether. Can this type of question really assess student learning in a meaningful way?

Douglas, Wilson, and Ennis (2012) studied the usefulness of multiple-choice questioning as a meaningful assessment and learning tool. The authors of this study evaluated student performance on online multiple-choice testing. Students took 3 pre-test and 3 post-tests. The pre-tests were online multiple-choice tests. Students had unlimited attempts to take the pre-test with instant feedback and the use of the class materials. For the post-tests students only had one attempt with limited time but still had access to class materials. The researchers discovered that over time students did better and better on pre and post-tests. Of course like most multiple-choice tests the questions were fact based questions. Douglas, Wilson, and Ennis concluded that multiple-choice test can promote and, "cultivate a climate of 'deep', reflective learning on the part of the participants" (p. 118).

Rothschild (2000) summarized the history of the DBQ. The DBQ on New York state history regents got its beginning in the Advanced Placement (AP) history examinations. The DBQ was introduced in 1973 on the AP United States History exam. From the exam's inception in 1955 through 1973 there were no DBQs. The push for DBQs came out of a need for the exam to remain relevant. In the 1970s college professors at all levels undergraduate and graduate were emphasizing social history in course curriculum, and more specifically the analysis of primary source documents. After three decades of the AP United States History exam utilizing the DBQ New York State also adopted the DBQ in 2001 (Rothschild, 2000). Since this adoption history educators in New York have made primary source analysis a regular part of the curriculum. The ability to evaluate bias, audience, overt meaning, historical context, and main ideas of primary

sources are skills now required in many if not all college level history classes. Therefore, in classrooms across New York since 2001 students are now exposed to this kind of analysis.

Rothschild (2000) highlighted one reason for the adoption of the DBQ is the attempt to promote social history or bottom up history. Social history implies the histories of the average American and that entails minority and marginalized groups. Swartz (2012) analyzed the DBQs on June New York States United States History and Government Regents exams from 2001-2009 for cultural responsiveness. Swartz concluded that, "for the most part, government-sanctioned document-based learning materials…are not culturally responsive. They often mask or marginalized cultures and groups" (p.151). This is fascinating because one motivator for incorporating the DBQ was to promote social history however it appears that the dominate culture is the only history that receives the lions share.

Implications

The implications of these findings are significant as they pertain to the readability of curricular materials. Educators need more than just a readability measure to determine if specific materials are best suited for their students and grade level. As an educator at the secondary level I have never considered the readability level of materials for my general education classes. After my first three years of teaching I can already tell if a particular reading passage – be it a primary or secondary source – is beyond my students. Those with experience know when students struggle to comprehend a particular passage of reading. Educators who wish to remain effective need to encourage students into tackling college-level texts at the secondary level; it requires patience, scaffolding, and perseverance. However, if all materials are beyond the reach of the high school students day in and day out they will become frustrated and fatigued. If students are

to have any hope of being successful at the collegiate level they must be exposed to rigorous reading materials long before they enter their freshman year of college.

There is a large body of research to validate and refute the use of readability as a measure of appropriate grade level material. This review of the research has yielded several takeaways when applying readability formulas. First, it is important to select the right formula. Researchers have concluded that no matter the readability formula all will accurately measure the same 100-word passage, but when measuring the readability of a multiple choice question has not been so straightforward. There are specific formulas for shorter passages. Second, when analyzing the readability of a textbook, it is important to select several different formulas to test the readability. This will result in a far more accurate assessment of the grade level of the material. Finally, teacher experience must also play a part when selecting appropriate grade level material.

Where this review intersects with my research is in the application of readability formulae on the New York State United States History and Government Regents examination. The section I analyzed in my study was the multiple-choice. It has come to light through this review of the extant research that readability plays a significant role in testing outcomes and student performance in general. Analyzing this section of the United States History and Government Regents exam will help other educational researchers and practitioners to better evaluate and understand the implications of readability on student, district, and state performance on standardized testing.

Methodology

What is the readability of the New York State United States History and Government Regents exam over time? This is my primary research question and what follows are my

secondary questions. How has the Regents Exam changed over time with regard to readability? What is the best formula or formulas to apply to a multiple-choice question?

There are numerous formulas to determine text readability, many of which require analyzing three or more 100 words passages: Dale-Chall, Modified Dale-Chall, New Dale-Chall, Lexile, ATOS, Fry, Coh-Metrix L2, and Flesch (Benjamin, 2012, Crossley; Allen, & McNamara, 2011; Fry, 2002; Hewitt & Homan, 2004; Homan, Hewitt, & Linder, 1994; McConnell, & Paden, 1983; Osborn, Jones, & Stein, 1985; Williamson, 2008). These are just a few of the formulas. In general, readability is calculated by examining two variables: sentence length and vocabulary (Fry, 2002). This is a very generic explanation, for every readability formula determines text complexity by manipulating these two variables. Some formulas require computer programs and others are performed manually. However, it does not matter the formula utilized, computer based or manual, both will determine readability, no one formula is better than another (Burk & Greenberg, 2010).

The disadvantage of all the formulas listed above, they do not provide analysis for single-sentence items for example, a multiple-choice question. The Homan-Hewitt Readability Formula provides single-sentence item readability analysis; their formula is as followed (Homan, Hewitt, & Linder, 1994).

$$Y = 1.76 + (.15 \times WNUM) + (.69 \times WUNF) - (.51 \times WLON)$$

This formula takes into account three variables, sentence length (WNUM), word difficulty (WUNF), and word length (WLON). The first variable (WNUM) of the formula was developed based Hunt (1965) T-Unit. Homan, Hewitt, and Linder (1994) define the T-Unit as the number of words in the signal-sentence item, the multiple-choice question. The second variable (WUNF) measures the number of difficult words as determine by Dale and O'Rourke's

(1981) *The Living Word Vocabulary*. This work was a 25 year study measuring the familiarity of 44,000 words from grades 4 through college. For this formula, a word is considered familiar if 80% of 4th graders know the word, all other words are considered unfamiliar. The final variable (WLON) measures the number of long words. Words longer than seven letters are identified as long according to the Homan-Hewitt Readability Formula (Homan, Hewitt, & Linder 1994). The following is an example of the formula. This sample question was pulled from the New York State Regents High School Examination, United States History and Government June 17, 2014 test (*Office of State Assessment*).

Question #9:

The term *judicial review* is best defined as the

(1) right of a defendant to appeal the verdict of a jury

(2) ability of congress to create new federal courts

(3) authority of the Senate to confirm Supreme Court justices

(4) power of the Supreme Court to determine the constitutionality of law

Choice (1) $1.76 + (.15 \times 20) + (.69 \times 14) - (.51 \times 4) = 12.38$

Choice (2) $1.76 + (.15 \times 17) + (.69 \times 13) - (.51 \times 5) = 10.73$

Choice (3) $1.76 + (.15 \times 18) + (.69 \times 14) - (.51 \times 6) = 11.06$

Choice (4) $1.76 + (.15 \times 20) + (.69 \times 14) - (.51 \times 5) = 11.87$

The mean score for this test question is 11.51 therefore; this question receives a readability score of 11.51. Scores calculated with this formula indicate grade level for example, this exam is an 11th grade test therefore scores should fall between 11.00 and 11.99 according to the Homan-Hewitt Readability Formula (Homan, Hewitt, & Linder, 1994).

For this study I analyzed the readability of the multiple-choice sections from 1990 to the 2014 June Regents Exam. Due to time constraints readability data was collected in the following manner. I put fifty slips of paper in a container each slip labeled with numbers one through fifty. For each exam I randomly select 3 slips of paper and analyze that item (question) that corresponds with that number. If the question selected has a word that was not in the *Living Word Vocabulary,* I selected another slip from the container. Three multiple-choice questions were analyzed from each June exam, from 1990 through 2014. Mean score were calculated for each test.

There were several limiting factors throughout the data collecting phase. First, as stated above, if a question contained a word that was not in the *Living Word Vocabulary* that question could not be analyzed for this study. This proved to be disconcerting because there were numerous questions that were excluded. For example, all names, ethnicities, countries, and continents, were excluded. Therefore, if a question contained a president's name I could not analyze that question because not counting that word would have manipulated the readability score. Further, if a question contained the words Spanish, European, Germany, or Asia, I could not analyze it because, again, they were not in the *Living Word Vocabulary*. Therefore, just selecting three suitable questions to analyze for each exam proved tedious.

Second, numbers were excluded from the analysis of each question. For example, if a question said "in the early 1800s" or "the late 19th Century" I did not count the numbers in these questions toward readability.

Third, the *Living Word Vocabulary* is quite dated at this point. The only version of this work was published in 1981. A revised version would have given greater strength to my data. Dale and O' Rourke's (1981) study was extremely expensive and was conducted over many

years therefore, updating this work would be a monumental undertaking. However, such an endeavor would be beneficial to future studies such as this and would provide all those analyzing and discussing readability an appropriately comprehensive resource.

Finally, this study was extremely time consuming. My initial intent with this study was to analyze every multiple choice question on each exam. However, to analyze just one question would take rough 20-40 minutes to analyze, depending on the length. Time was a major limiting factor. Once I began collecting the data I was forced to scale back the scope of my very ambitious goal. Table 1 shows all the data collected over the past several months. Because I could only analyze three questions per exam the overall strength of my data is diminished.

Results

Tables 1, 2, and 3 present the results of the data collected. Table 1 shows the tests from the most recent June Regents Exam back to June 1990. The table displays the Regents Exam year, questions numbers, readability of those questions, and overall exam readability. As stated in the methodology, the selection of three questions at times was frustrating because a number of questions on each exam had to be eliminated because a particular word in the question was not in the *Living Word Vocabulary*. Questions that were analyzed are arranged in numerical order for each exam and the corresponding readability follows in Table 1. Some exams have wild fluctuations is scores per question. For example, June 2011 question number 50 had a readability of 17.2 which is college level, while question number 38 scored 6th grade reading level. The most consistent of all the exam questions was 2008. Each of the questions analyzed scored in the 9th grade reading level.

Table 1
Readability NYS U.S. History and Government Regents Exam Utilizing Homan-Hewitt Readability Formula

Exam	Question Numbers	Question Readability	Exam Readability
2014	17	6.2	8.4
	23	9	
	35	10	
2013	27	10.3	9
	44	7.9	
	36	8.7	
2012	28	9.3	9.7
	12	8.7	
	21	11.2	
2011	38	6.2	10.9
	49	9.3	
	50	17.2	
2010	6	7	8.3
	19	10.2	
	46	7.8	
2009	9	12.3	9.5
	34	6.1	
	41	10.1	
2008	13	9.3	9.4
	22	9.4	
	47	9.6	
2007	6	9	8.6
	4	8.7	
	20	8.1	
2006	10	7.5	8.3
	15	8.1	
	46	9.3	
2005	15	8.7	9.3
	21	8.9	
	22	10.2	
2004	3	9.6	11.9
	20	9.4	
	45	16.7	
2003	1	10.8	8.8
	14	8.1	
	24	7.6	
2002	10	11.2	10.2

	14	9.7	
	45	9.6	
2001	6	9.6	9.6
	29	7.4	
	34	11.9	
2000	4	11	9.8
	29	8.9	
	44	9.5	
1999	5	8.2	9.1
	22	9.7	
	20	9.5	
1998	4	9	10.2
	42	11.5	
	43	10	
1997	8	9.2	7.2
	40	5.2	
	44	7.1	
1996	3	12.3	9.4
	8	9.3	
	10	6.7	
1995	4	5.2	8.2
	43	8.8	
	44	10.5	
1994	3	16.5	12.4
	17	9.1	
	42	11.7	
1993	16	12.1	9.8
	27	7.7	
	36	9.5	
1992	27	8.9	10.9
	41	12.6	
	44	11.2	
1991	4	9.7	8.6
	36	11	
	39	5.1	
1990	11	9.3	11.4
	35	11.1	
	37	13.7	

My primary research question was to investigate the readability of the New York State Regents Exam in United States History and Government over time. Table 2 presents a

scatter plot diagram connected with a line showing readability over time of the June Regents Exams. The first exam plotted June 1990 had a readability score of 11.4 and the last point plotted June 2014 had a readability of 8.4. Table 2 shows a wide range of scores over time however, there is a clear downward trend associated with the data.

Table 2
Readability Over Time

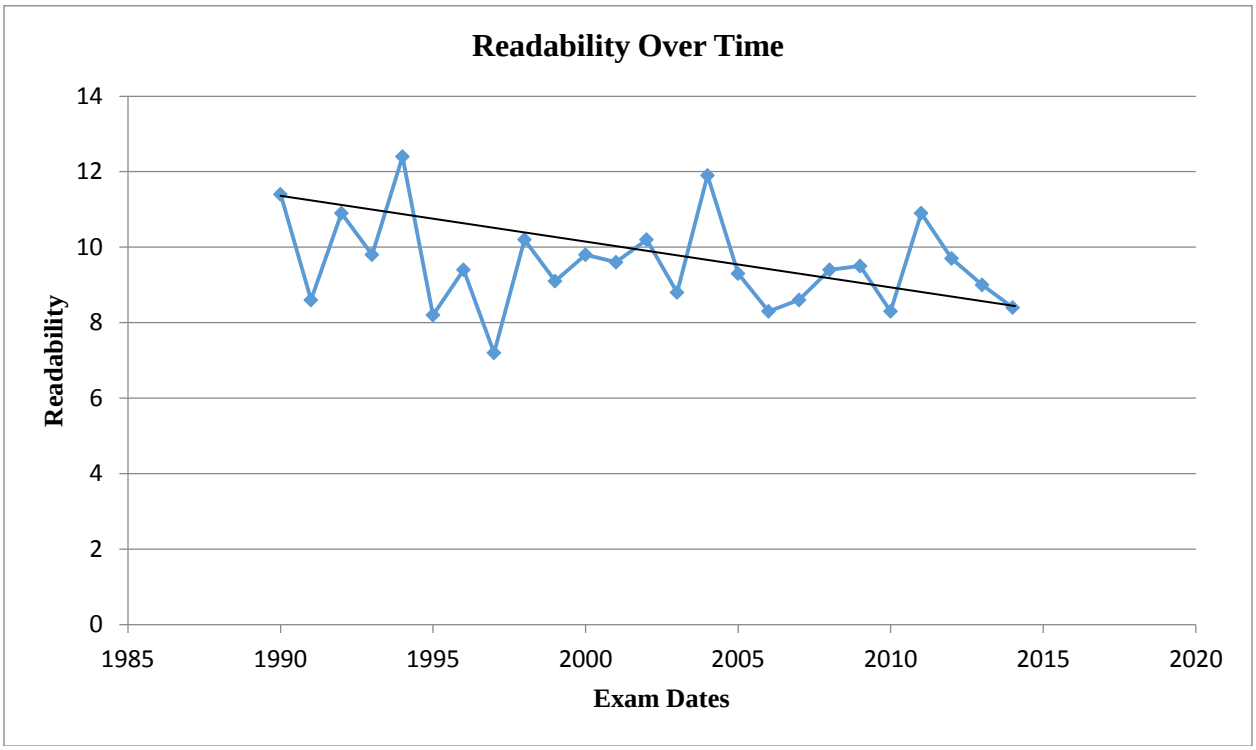

Table 3 displays statistical information regarding the mean, median, mode, and range of the exams readability. The mean exam readability is 9.6; a median of 9.4; there are several modes 8.3, 8.6, 9.4, 9.8, 10.2, and 10.9; and a range score of 5.2. There were 6 mode scores, each had two tests with these numbers. There was a total of 25 exams measured for readability, one exam scored a readability in the 7th grade range (7.0 - 7.9), seven exams scored in the 8th grade range (8.0 - 8.9), ten scored in the 9th grade range (9.0 - 9.9), four in the 10th grade range

(10.0 - 10.9), two in the 11th grade range (11.0 - 11.9), and one exam scored in the 12th grade range (12.0 - 12.9).

Table 3

June Regents Exams Readability Statistics

Mean	9.6
Median	9.4
Mode	8.3, 8.6, 9.4, 9.8, 10.2, 10.9
Range	5.2

Discussion

I initially hypothesized that the readability of the NYS U.S. History and Government Regents Exam would decrease over time. This hypothesis has largely been confirmed. Table 2 reveals a gradual downward trend. I anticipated that the decrease in readability would be far more pronounced as opposed to a gradual trend. I also was anticipating that there would have been a sharp decrease in the readability level between years 2000 – 2001. The reason for this assumption was that starting in 2001 every student in New York State was mandated to take and pass the Regents Exam as a graduation requirement. However, there was no dramatic fall in readability in 2001. What I did not anticipate but what has been revealed by the data is that readability appears to have been falling since June of 1990. If time was not a constraint for this study it is plausible to presume that exams dating back to 1980, may have revealed exam scores with even higher readability levels.

This study was only an analysis of the multiple choice section of the examination. The Regents exam contains three other sections a thematic essay, document based questions (DBQ), and a document based question essay. These sections were not included in this study. This is significant for evaluating the exams overall readability because the document section of the exam contain a significant amount of reading and analysis. A single reading passage associated

with a document could potentially have 10 – 200 words per passage. These numbers are rough estimates, a passage could have fewer or more words. Future research is warranted to analyze the DBQ section of this exam for readability. The essay sections, specifically the DBQ analysis, plays a prominent role in overall student outcomes and readability of this exam.

Individual question analysis revealed much regarding each examination's readability. The only exam that had consistent readability numbers was June 2008. Each of the questions analyzed fell within the 9th grade reading range. The exam with the greatest discrepancy in readability was June 2011. Question number 38 received a score of 6.2 while question number 50 had a readability of 17.2, this is problematic. In theory, question number 50 was outside the reading ability of the many high school students who took the exam that year. What would be fascinating to study is the percentage of students across New York who got that question correct and measure that against those that got the question incorrect. This was a major curiosity at the genesis of my research. What if a student was adequately instructed in preparation for the state examination? That student had a highly qualified educator, who diversified instructions, delivered the curriculum to address all different kinds of learners, and yet; what if the question was beyond the reading ability of a student? Is that a valid question? I would argue no, it is not.

What about high needs students? There are many students across New York State who have Individualized Educational Plans (IEPs), which allow for the modification of curriculum, and a "low pass" score among other accommodations. A "low pass" means that a student can receive a 55% percent on the Regents Exam and still "pass" the exam for graduation credit. All students regardless of the intellectual abilities had to pass a series of Regents Exams in order to graduate high school. For example, if a student has undergone a series of achievement testing, and that student has been identified as having an IQ of 68, they would be borderline

intellectually disabled. Average intelligence ranges around 100. It would be safe to assume that this hypothetical student has a 3rd or 4th grade reading ability. If this were true, every question on the United States History and Government Regents Exam would be outside of that student's abilities. I instruct several students that fall into this category, which beg the question; is this exam an adequate tool to measure their knowledge of U.S. history? Is this an adequate tool to measure my abilities as an educator?

The mean readability of the examination for years 1990 - 2014 is 9.6. Why is the mean readability at the 9th grade reading level? I have three hypotheses. First, the exam's readability over time is 9.6 because this ensures that for the majority of students taking it will be assessed on their knowledge of United States History and not their reading ability. There are numerous students who do not read at grade level, have IEPs which allow for "low pass" scores, and many students who experience extreme test anxiety. Many students have told me over the years that their brains just went blank once they get the test in front of them. Many students understand the implications of not performing well on these state Regents Exams. Failure means taking the exam again in August, potentially a stint in summer school, enrollment in a Reagents Preparation course the following academic year, or even postponing or not completing graduation. Students know and understand this and contribute to low testing scores and poor performance. All of these are factors for the State to consider when creating these questions for high-stakes exit exams for high school juniors across the state. I can only assume that State officials want to make the exam rigorous, but at the same time accessible and within the limits of testing time constraints. On all State testing exams students are only allotted 3 hours to complete the exam. Students with IEPs can receive extended time to complete the exam however, all the time in the world cannot assist a student who cannot read and interpret the questions before them.

Second, the examination's readability is at a 9th grade level because there is a desire within state government to boost test scores. If the readability of the test is well below the intended grade level, in theory this should increase overall scores across the State. It is hard to imagine a state board of regents, legislature, or governors who would go before the public and admit that they want to decrease the rigor of state Regents Exams, lower teaching standards, reward schools for low graduation rates, and at the same time increase school funding. In my review of the current research there has been no systematic study of the readability of the U.S. History and Government Regents Exam. This topic is currently not open for public debate presumably because the public is ill-informed or unaware of the significance of this topic on test validity and student performance. Because this topic is not discussed it is very easy to create questions for which readability falls below the intended grade level while at the same time appearing to be a rigorous assessment of the curriculum.

Another factor which determines a student's Regents score is a generous grading curve. For example, on the most recent Regents Exam June 2014 if a student answered all the multiple choice and document based questions correct that student would, according to the conversion chart, receive an 81% on the exam (*Office of State Assessment*). That same hypothetical student could write two essays only receiving a score of 1 out of a potential 5 on each and receive an 88% on the Regents. That is a mastery score. If a student gets full credit on all the multiple choice questions and gets just one document based question correct that student receives a passing grade of 65%. This means a student can complete half of the exam and earn a passing score. Debating any further the merits of the conversion chart is beyond the scope of this paper, however these brief talking points help to emphasize the importance of the multiple choice section. Any high school history teacher across New York will likely agree that for students

who struggle in school, struggle on testing, or who have IEPs, the key to their success lies in the multiple choice section. If one of those borderline students can get at least thirty of the multiple choice questions correct, all the document questions correct, and can achieve scores of 2s for both of the essays a student will receive a 70% on the exam.

Finally, there is perhaps no consideration give to the readability of the questions. This final thought seems very unlikely. It would be against the State's interest to produce an examination that is not a valid and reliable measure of the United States History curriculum. It would be a disservice to the State, school districts, teachers, and students to produce an examination that is beyond the reading level intended for the test. Better to produce an exam with a readability at or below the intended level.

Dale and O'Rourke's (1981) *Living Word Vocabulary* was a limiting factor in this study. This work was produced in 1981 with no revisions because of this it is safe to assume that it is on the verge of being an outdated resource. Thiers was a 25-year study measuring the familiarity of 44,000 word meanings. Hundreds of thousands of students were sampled to create this work which in itself in an incredible feat. This undertaking took considerable time and money to create which I can only assume is why it has not been undated or reproduced. However, it would be reckless to not highlight the fact that many of the words in the vocabulary are no longer in use for example, argot (meaning: jargon, slang, slang of thieves) or are in need of a revised familiarity score. Take for example the word American. According to the *Living Word Vocabulary* only 69% of 4th graders are familiar with this word. I used this word as an example because I looked this up numerous times while collecting my data and each time it shocked me that over the twenty-five years of data collect it was determined that a mere 69% of 4th graders were familiar with the word American. Once can only speculate what an updated version of this

work would uncover - maybe 80% are familiar with this word or maybe 60%. What is safe to assume is that the familiarity scores of these 44,000 words would change - to what degree I know not. This limitation certainly dilutes the strength of my data. However, all readability formulas are subject to scrutiny. If another single-sentence readability formula were created tomorrow it would almost certainly yield different results than the data I collected. That being said, the more readability formulas applied to any reading passage or multiple choice question will help ensure a more holistic score for educators and administrators to analyze and assess.

There is certainly a need for further research in the area of readability in relation to the NYS Regents Exam in United States History and Government. I would argue a more exhaustive study is warranted, not just of the multiple choice questions, but also of the document-based questions. New York State, along with the nation, is currently in an era of sweeping educational reform. There are titanic pressures on students, teachers, districts, and state governments. Students need to develop the skills required for the 21st Century workforce. The first couple of steps in that journey are high school and then usually college or technical school. Districts and specifically teachers have been pressured like no other time in our nation's history to educate and prepare that workforce for success, and states want to flaunt the educational outcomes of their student populations to ensure public support and federal funding. However, what if the measures by which teachers, districts, and students are being evaluated were invalid due to unrealistic readability levels?

References

Armbruster, B. B., Osborn, J. H., & Davison, A. L. (1985). Readability formulas may be dangerous to your textbooks. *Educational Leadership, 42*(7), 18. Retrieved from ERIC database .

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1), 63-88. doi:10.1007/s10648-011-9181-8

Blackey, R. (2009). So many choices, so little time: Strategies for understanding and taking multiple-choice exams in history. *History Teacher, 43*(1), 53-66. Retrieved from ERIC Database.

Bulgren, J. A., Marquis, J. G., Lenz, B., Deshler, D. D., & Schumaker, J. B. (2011). The effectiveness of a question-exploration routine for enhancing the content learning of secondary students. *Journal of Educational Psychology, 103*(3), 578-593. doi:10.1037/a0023930.

Burke, V., & Greenberg, D. (2010). Determining readability: How to select and apply easy-to-use readability formulas to assess the difficulty of adult literacy materials. *Adult Basic Education and Literacy Journal, 4*(1), 34-42. Retrieved from ERIC database.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84-101. Retrieved from ERIC database.

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary.* Chicago: World Book International.

Douglas, M., Wilson, J., & Ennis, S. (2012). Multiple-choice question tests: a convenient,

flexible and effective learning tool? A case study. *Innovations in Education & Teaching International, 49*(2), 111-121. doi:10.1080/14703297.2012.677596.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher, 56*(3), 286-291. Retrieved from ERIC database.

Homan, S., Hewitt, M., & Linder, J. (1994). The development and validation of a formula For measuring single-sentence test item readability. *Journal of Educational Measurement, 31* (4), 349-358.

Hewitt, M. A., & Homan, S. P. (2004). Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction, 43*(2), 1-16. Retrieved from ERIC database.

Hunt, K. (1965). *Differences in grammatical structures written at three grade levels* (NCTE Research Report No. 3). Urbana, IL: National Council for Teachers of English.

Maestri, M. (2006). The myth of a multicultural curriculum: An analysis of New York State U.S. history regents. *The History Teacher, 39*(3), 381-402. Retrieved from ERIC database.

McConnell, C., & Paden, D. W. (1983). Readability: Blind faith in numbers?. *Journal of Economic Education, 14*(1), 65-71. Retrieved from ERIC database.

Office of State Assessment (2015). United States History and Government Regents Examination Retrieved from http://www.nysedregents.org/USHistoryGov/home.html

Osborn, J. H., Jones, B., & Stein, M. (1985). The case for improving textbooks. *Educational Leadership, 42*(7), 9. Retrieved from ERIC database.

Reich, G. A. (2011). Testing collective memory: Representing the Soviet Union on multiple-

choice questions. *Journal of Curriculum Studies, 43*(4), 507-532. doi:10.1080/00220272.2011.578665.

Rothschild, E. (2000). The impact of the document-based question on the teaching of United States history. *History Teacher, 33*(4), 495-500. Retrieved from ERIC database.

Swartz, E. E. (2012). Distinguishing themes of cultural responsiveness: A study of document-based learning. *Journal of Social Studies Research, 36*(2), 135-167. Retrieved from ERIC database.

Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal Of Advanced Academics, 19*(4), 602-632. Retrieved from ERIC database.

Yonker, J. E. (2011). The relationship of deep and surface study approaches on factual and applied test-bank multiple-choice question performance. *Assessment & Evaluation in Higher Education, 36*(6), 673-686. doi:10.1080/02602938.2010.481041

.