THE RELATIONSHIP OF READABILITY
ON THE
SCIENCE ACHIEVEMENT TEST:
A STUDY OF 5TH GRADE ACHIEVEMENT PERFORMANCE

Zachary Scott Amos

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTER OF EDUCATION

May 2009

Committee:

Dr. Cynthia Diane Bertelsen, Advisor
Dr. Cindy Hendricks
Dr. Mohammed Darabie
ABSTRACT

Dr. Cynthia Bertelsen

This study examined the relationship between readability and test performance. This study investigated the historical, theoretical, and research views of high stakes testing, out-of-level testing, and comprehension. The Ohio Fifth Grade Science Achievement tests’ individual test questions were examined based upon the Flesch Kincaid readability formula. A revised version of the test was created with a lower reading level. A study was conducted on 34 students from a rural midwestern community. The tests were both administered, with two weeks of separation. The data were collected and analyzed through the StatCrunch application using T-Statistical and Summative data Analysis. From the results, their appeared to be a statistical difference in the test scores in favor of the revised version.
I dedicate this thesis to all of the people who assisted me throughout this process. This includes my chair: Dr. Cynthia Bertelsen, my committee members: Dr. Cindy Hendricks, Dr. Mohammed Darabie, and my personal support of my family, friends, and colleagues. Without all of these people and the exceptional resources provided by Bowling Green State University’s Jerome Library, I would not be where I am today!
ACKNOWLEDGEMENTS

I would first like to acknowledge Dr. Cynthia Bertelsen, Dr. Cindy Hendricks, and Dr. Mohammed Darabie for their kind guidance and wisdom. I would also like to thank “The King of Stats” Dr. Craig Mertler for his wonderful statistical support. Finally, I would also like to recognize my mother and father who have been there with constant support and encouragement throughout this adventure, and my many other adventures in life.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>CHAPTER I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Statement of the Problem</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Research Questions</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Rationale</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Definitions of Terms</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Limitations</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>7</td>
</tr>
<tr>
<td>II</td>
<td>CHAPTER II. REVIEW OF LITERATURE</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Historical Perspective</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Theoretical Perspective</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>New Information and Research</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Readability</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Formulas</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>27</td>
</tr>
<tr>
<td>III</td>
<td>CHAPTER III. METHODS AND PROCEDURES</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Methods</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Research Design</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Participants</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Instrumentation</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Procedures</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Data Collection</td>
<td>32</td>
</tr>
</tbody>
</table>
APPENDIX H. STUDENT LETTER .............................................................131

APPENDIX I. SUMMARY STATISTICS TABLE ...........................................133

APPENDIX J. EVALUATING QUESTIONS READABILITY, T-STAT

AND P-VALUE FOR TEST A AND TEST B ...............................................136
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 1A &amp; 1B</td>
</tr>
<tr>
<td>2</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 2A &amp; 2B</td>
</tr>
<tr>
<td>3</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 3A &amp; 3B</td>
</tr>
<tr>
<td>4</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 4A &amp; 4B</td>
</tr>
<tr>
<td>5</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 5A &amp; 5B</td>
</tr>
<tr>
<td>6</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 6A &amp; 6B</td>
</tr>
<tr>
<td>7</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 7A &amp; 7B</td>
</tr>
<tr>
<td>8</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 8A &amp; 8B</td>
</tr>
<tr>
<td>9</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 9A &amp; 9B</td>
</tr>
<tr>
<td>10</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 10A &amp; 10B</td>
</tr>
<tr>
<td>11</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 11A &amp; 11B</td>
</tr>
<tr>
<td>12</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 12A &amp; 12B</td>
</tr>
<tr>
<td>13</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 13A &amp; 13B</td>
</tr>
<tr>
<td>14</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 14A &amp; 14B</td>
</tr>
<tr>
<td>15</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 15A &amp; 15B</td>
</tr>
<tr>
<td>16</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 16A &amp; 16B</td>
</tr>
<tr>
<td>17</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 17A &amp; 17B</td>
</tr>
<tr>
<td>18</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 18A &amp; 18B</td>
</tr>
<tr>
<td>19</td>
<td>Paired T Statistics: Evaluating The Mean Of Test Scores For Test 19A &amp; 19B</td>
</tr>
<tr>
<td>20</td>
<td>Paired T Statistics: Evaluating The Sum Of Test Scores For Test A &amp; Test B</td>
</tr>
</tbody>
</table>
CHAPTER I. INTRODUCTION

As the importance of state and district assessments has increased, so have concerns about whether the tests that are being administered are appropriate for all students. According to the American Educational Research Association (1999), “many school districts are mandating tests to measure student performance and to hold individual schools and school systems accountable for that performance” (p. 1). Mandated tests are known as “Achievement Tests” and are used to determine advancement or graduation. Educational administrators and other officials use the results of the achievement tests to make very large and important decisions. The achievement tests are encouraged and mandated for schools at the national level. Furthermore, state departments of education have created and adapted similar tests to assess students’ performance on specific academic content standards.

In the state of Ohio, achievement tests are given to students in grades three through eight to assess their performance and abilities in reading, writing, mathematics, science, and social studies. Each achievement test assesses students on material or information they are expected to know by a certain grade level. The topics, skills, and information covered by the tests are grounded in the Ohio Academic Content Standards for each grade level. Despite the fact that some of these tests focus on content such as science and social studies, reading is a major component in all of the achievement tests.

Overall, mandatory state and national tests are being implemented in schools all over the United States to assess students’ abilities and levels of performance in a variety of content areas. Results from these tests may show a relationship that exists among students’ and their reading levels through their performance scores. As a result of looking at students’ performances on
different areas of state and national tests, teachers and test makers may become more aware of how to differentiate testing to meet the needs of students.

Statement of the Problem

In typical classrooms, there is a fluctuation of reading abilities that range from several grade levels below reading level to several grade levels above reading level. These students’ reading abilities have a direct impact on comprehension. If students cannot understand and/or comprehend the level of vocabulary then they will not be able to understand the test items.

The Ohio Fifth Grade Science Achievement Test is leveled at a 4.8 Flesch/Kincaid reading level. If students are two grade levels below reading level and are not on an Individual Education Plan then they have to take the test. However, students who are at lower reading levels will not be able to fully understand critical questions to answer questions that contain key science concepts. And without an alternative test, students’ will not be able to live up to their peak performance regarding these key concepts in the science curriculum that they have worked so long and hard to attain. For example, consider a student with a third grade reading level confronting a science test that entails substantial reading of text.

Presenting that student the test in the standard format will produce a biased indication of the student’s mastery of science concepts because the student’s difficulty deciphering the written page will depress his or her score. In other words, it won’t be clear if the problem is lack of content knowledge, or trouble with reading skills. However, it is possible one could obtain a more accurate appraisal by changing the mode or presentation, such as using a reading level that would be at a lower level. In light of the information already presented about important decisions being made on a grand scale for school districts this is a major concern for students, teachers, and administrators.
Research Questions

This study investigated the relationship that may be present among students’ test scores with an alternative science test and those with the traditional Ohio Achievement Test (OAT). The study investigated how the Ohio 5th Grade Science Achievement Test was constructed. This study also investigated differences and correlations between individual test questions and test performance. The driving research question that guided this investigation was, “How would an alternative 5th Grade Science Achievement Test with a lower reading level impact test scores?”

Rationale

Students are not going to be able to demonstrate their full potential if they are given material that is at a readability level that is much too high for their placement level. According to Adelman and Taylor (2006), a typical seventh grade classroom will have reading levels that range from second grade through beyond seventh grade. Furthermore, this tells us that with differentiated reading levels some students will not be able to understand the questions that are being asked of them. The students will be forced to make educated guesses rather than demonstrating their content knowledge. If students are given material that is at an appropriate level for them, they will be able to answer the required questions with full knowledge of what is being asked. This will in turn decrease student frustration and anxiety during the test. Since high-stakes-tests are given by every state in the United States, this study may also benefit students from other states.

Furthermore, not only the students will be the beneficiaries to the study, but the teachers of the students will benefit as well. When the reports from the state are sent back to the schools, many teachers are held accountable by the administration for their actions. According to the Ohio Department of Education (ODE) teachers whose students have low scores on the
proficiency tests for three or more years may be subjected to replacement (ODE, 2008). The results from these tests might be inaccurate and not reflect on teachers’ overall teaching of science content. In giving the students material that they will be able to understand, it will be a more accurate representation of their overall content knowledge. Teachers who may have not been praised in the past for their teaching would now be rewarded.

Additionally, schools are held accountable for making Adequately Yearly Progress (AYP). This is when the school is performing at the expected level. These are not three words that are taken lightly. If a school does not make AYP many things can happen. First, the administration after three years may be replaced (ODE, 2008). Secondly, if the school is publicly funded it may lose its funding or its students as a result of the OAT scores (ODE). For many struggling schools that are comprised of struggling readers, and this reality is directly correlated with funding.

Definition of Terms

Some terms that will be used in this study are identified below along with their descriptions and accompanying sources.

The Ohio Achievement Test (OAT)- are annual assessments that measure how well students have learned the reading and math concepts taught in grades 3 - 8. The achievement tests are designed specifically for Ohio students and are based on Ohio’s Academic Content Standards. They do not test a child’s intelligence or aptitude and are not meant to compare children’s abilities (ODE, 2008).

Content Standard: What students should know and be able to do; the overarching goal and theme. (ODE, 2008)
Standardized Test- “Tests that have statistics (or norms) for comparing performance of a student to a large sample of students. The test is given to each child in the norm sample, and norms are established to determine, for example how well the average 5th grader does on the test. These tests must be given with strict procedures in order for students to be compared accurately” (Jennings, Caldwell & Lerner, 2006, p. 114).

Readability level- “The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success of the group of readers. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting” (Chall & Dale, 1949, p. 23).

Comprehension- 1. “The reconstruction of the intended meaning of a communication; accurately understanding what is written or said.” 2. “The construction of the meaning of a written or spoken communication through a reciprocal, holistic interchange of ideas between the interpreter and the message in a particular communicative context” (Harris & Hodges, 1995, pp. 38-39).

Adequately Yearly Progress (AYP)- Under No Child Left Behind, each state has developed and implemented measurements for determining whether its schools and local educational agencies (LEAs) are making adequate yearly progress. AYP is an individual state's measure of progress toward the goal of 100% of students achieving to state academic standards in at least reading/language arts and math. It sets the minimum level of proficiency that the state, its school districts, and schools must achieve each year on annual tests and related academic indicators (U.S. Department of Education, 2008).
Limitations

Several limitations occurred with this study. The students in the study were at the 5.5 grade level when the study was conducted during the winter of 2009. The content that they were learning might change from a curriculum standpoint. The key concepts that were introduced into the science classroom may or may not be taught. The teaching staff might change as well as the lesson plans. Also, the content or standards tend to change after several years.

Additionally, students change from year to year. In upcoming years students’ reading levels and overall ability levels may progress or regress. Thus, the analysis of the study might be considerably different. For example, if the whole class is composed of extremely high readers who can interpret or comprehend all of the questions then the variances in between the two will be miniscule.

Furthermore, aside from the student changes, another change would be of the test itself. The 5th grade Ohio Achievement Test is reconstructed every year. In doing so the standards that are contained within the test may change as well as the numbering and organization of the test itself. This could possibly skew the data. The accessibility of the test can change as well. At present, copies are available to the public for free, and are easily accessible to the public on the Ohio Department of Education website (www.ode.state.oh.us). In the future, the tests may no longer be available.

Lastly, it was not possible for the researcher to create a revised test in which all questions fell within the same reading level. This was due, in part, to the number of specific content-related science terms that appeared within the questions.
Summary

Achievement testing has become a major concern within the classroom at almost every grade level. Students are being asked to all perform at the same normative level. However, research has shown that students’ reading levels vary greatly in the classroom. Numerous research studies conclude that students who are asked to read text that is not at their appropriately functioning reading level have problems with comprehension.

The 5th Grade Ohio Science Achievement Test is a measurement of the students’ grasp of content knowledge. A student who is a struggling reader will have a difficult time understanding/comprehending the questions on the achievement test. This led to the development of the research question that guided this investigation, “How would an alternative 5th Grade Science Achievement Test with a lower reading level impact test scores?” This investigation focused on the relationship that may exist between content knowledge, the readability levels of students, and the understanding of test questions on the Ohio 5th Grade Science Achievement Test.
CHAPTER II. REVIEW OF LITERATURE

The purpose of this investigation was to explore how an alternative 5th Grade Science Achievement Test with a lower reading level would impact their test scores. Before completing the study, a review of literature was conducted. The literature review focused on past and present literature in three main areas: “high stakes” testing, out-of-level testing, and the link between vocabulary and comprehension. All three of these topics were examined in full detail from a historical perspective, a theoretical perspective, and a researcher perspective.

Historical Perspectives

The need to know about students and teachers is part of the fabric of education in the United States. When the experiment known as mass education began in the United States in the mid-1800s, achievement testing was very limited in scope (Haladyna, Haas, & Allison, 1998). According to Heubert and Hauser (1999), “In the early 1900s, paper-pencil testing came into widespread use. At that time there was little discussion of what these tests actually measured” (p. 29). History provides striking examples of the actual or potential misuse of standardized tests to make decisions about individuals.

Unhappy with the increasing numbers of immigrants living in New York City, the president of Columbia University in 1917 embraced the use of the Thorndike Tests for Mental Alertness “to limit the number of Jewish students without a formal policy of restriction” (Heubert & Hauser, 1999, p. 21). Then in the early 1920s, the Stanford Achievement (SAT) was introduced, which marked a major milestone in educational achievement testing.

For more than three decades, under Title I of the Elementary and Secondary Education Act of 1965, program evaluation, through large-scale testing, has been an integral part in the federal support for the education of low-achieving children in poor neighborhoods. Heubert and
Hauser (1999) report, “The minimum competency testing movement, beginning in the 1970’s, gave large-scale, standardized achievement tests a visible and popular role in holding students (and sometimes schools) accountable” (p. 15). In addition to the increased prevalence of such tests and the escalation of consequences or “stakes” associated with them, the most notable change in the assessments themselves concerns their alignment with curriculum standards. Never before in U. S. history has the nation’s schools witnessed such an explosion of attention to standards.

The minimum competency testing movement of the 1980s was inaugurated by the *Nation at Risk* report (Goldberg & Harvey, 1983); this report has been overshadowed by a focus on high standards of achievement inspired by the nation's embrace of the national educational goals of the 1990s. According to Heubert and Hauser (1999), this momentum continued with Goals 2000, established by President George H.W Bush and was continued by President Clinton and Secretary of Education William Riley. Heubert and Hauser state, “By 1999, all fifty states had statewide assessments, although only 40 had developed standards in all core subjects... A total of 45 of the states had publicly rated the performance of the schools, and 16 had the power to close a school down” (p. 193).

In response to this, educators fired back with “out-of-level” testing. The term out-of-level testing typically means that a student who is in one grade is assessed using a level of a test developed for students in another grade. Below-grade-level testing is almost universally what is meant when the terms "out-of-level," "instructional-level," "off-level," or "functional-level" are used (National Center of Educational Outcomes [NCEO], 2008). Historically, the idea of using out-of-level testing was that it could evaluate student outcomes properly. This idea was used for
accountability standards for federal spending on educational type programs in the 1960s and 1970s (National Center of Educational Outcomes [NCEO], 2008).

Long, Schaffran and Kellogg (1977) reported that the Rhode Island State Department of Education, in 1971, authorized the testing of Title I students at their reading instructional level rather than their actual grade level. The final decision as to whether to test at grade level or instructional level was left to the discretion of the local educational agencies. Consequently, a variety of testing models evolved for Title I student selection and program assessment in reading. This “out-of-level testing led to a new set of problems related to data interpretation and program evaluation” (p. 203). It was perceived at the time that it would measure academic progress. This progress could be shown even below a student's attending grade level. This in turn demonstrated an intelligent and informed use of federal grant money. Then things changed with the standards educational reform in the 1990s, which pushed states to develop all encompassing assessments to measure all students' achievement toward on-grade level standards (NCEO, 2008). The results from these tests were used in states' accountability programs to meet the legal ramifications of the Individuals with Disabilities Education Act of 1997 and the Elementary and Secondary Education Act (ESEA) of 1994 (NCEO, 2008).

Some states learned that large test instruments were not designed to be universally accessible to some students with disabilities (NCEO, 2008). Fourteen states currently permit out-of-level testing in their statewide testing programs. This number has risen dramatically since 1999, despite growing concern about the implications for students who take an out-of-level test instead of the on-grade level test (Bielinski, Thurlow, Minnema, & Scott, 2002). The National Research Council (NRC) (1999) for example, called for a research agenda that includes studies of “the need for particular types of accommodations and the adequacy and appropriateness of
accommodations applied to various categories of students” and “the validity of different types of accommodations” (pp. 110-111).

Furthermore, out-of-level testing made a come back in some states’ large assessment programs to resolve the problem of excluding too many students with disabilities from statewide testing. The reemergence of out-of-level testing in many different state assessment programs has decreased recently because of federal policies that call for assessment against grade-level academic standards (NCEO, 2008).

Furthermore, to delve further into testing, comprehension was studied. In 1917, E. L. Thorndike made a major contribution to, and gave new insights into, reading comprehension when he published a series of articles on reading as a thinking-reasoning process. He stated, “Understanding a paragraph is like solving a problem in mathematics. It consists in selecting the right elements of the situation and putting them together in the right relations, and also with the right amount of weight or influence or force for each” (as cited in Staufer, 1956, p. 328).

Thorndike also had many other views toward generating meaning from text. He stated, “The mind is assailed as it were by every word in the paragraph. It must select, repress, soften, emphasize, correlate, and organize, all under the influence of the right mental set or purpose or demand” (as cited in Staufer, 1956, p. 329). In 1932, Miles Tinker wrote that: “True comprehension, however, is quite different from such recall, for it involves understanding, selection, correlation, and organization, all of which are influenced by the mental set of the reader” (as cited in Staufer, 1956, p. 158).

Sadoski and Paivo (2007) report, “Scientific theories of the reading process can be traced to the 1960’s and 1970’s. The first actual volume that was dedicated explicitly to models and theories of reading was published less than 40 years ago by Singer & Ruddell” (p. 339). While
theories of cognition started slightly sooner, “modern cognitive psychology took shape in the
1950s and 1960s, and ever since has primarily adopted a piecemeal strategy toward research and
theory. That is, research and theory have been directed toward carefully limited aspects of
cognition” (p. 339).

Theoretical Perspective

The Classical test theory or CTT was the beginning structure for analyzing and
developing standardized tests (Crocker & Algina, 1986; Hambleton & Swaminathan, 1991). CTT
is based on the assumption that a test-taker has two scores: an observed score and a true score.
The observed score is the score shown of someone taking the test. It is an estimate of the true
scores of that test-taker plus/minus some unobservable measurement error (Crocker & Algina;
Hambleton & Swaminathan). One of the many advantages of the observed score is that it relies
on assumptions and is relatively easy to interpret. The main criticism of the true score is that it is
not an absolute characteristic of a test-taker since it depends on the content of the test. If there
are test-takers with different ability levels, a simple or more difficult test would result in different
scores. Another criticism of the true score is that the items’ difficulty would change depending
on the sample of test-takers who take a specific test (Hambleton, Robin, & Xing, 2000).
Therefore, it is difficult to compare the students’ tests results between different tests. In the end,
good techniques are needed to correct for errors of measurement (Hambleton, Robin, & Xing).

Second, the Item Response theory (IRT) is another form that has more recently been
established. It was originally developed to overcome the problems with CTT. This work was
produced in the 1960s (Birnbaum, 1968; Lord & Novick, 1968). One of the basic assumptions in
IRT is the possibility of a person taking the test is independent of the content of a test. The link
between the probability of answering an item correctly and the ability of a test-taker can be
modeled in different ways depending on the nature of the test (Hambleton & Swaminathan, 1991). It is common to assume that the items in a test measure one single dormant ability. According to IRT, test-takers with high ability should have a high probability of answering an item correctly. Another assumption is that it does not matter which items are used to estimate the test-takers’ ability. This assumption makes it possible to compare a test-takers’ result despite the fact that they have taken different versions of a test (Hambleton & Swaminathan).

Van der Linden and Glas (2000) report that IRT has been the preferred method in standardized testing since the development of computer programs. The computer programs can now perform the complicated calculations that IRT requires. Ralph Tyler (1950) is credited with advancing computer programs calculations in modern education. This idea is reflected in the practice of linking each individual test item to an objective. This “...link between test item and objective shows the crucial nature of ensuring correspondence between a curriculum and teaching, between a curriculum and testing, and between teaching and testing” (Haladyna, 2002, p. 87). Furthermore, the principle theories behind the achievement tests are to show the current status of student achievement, rather than to identify students in need of intervention or to determine appropriate instructional strategies (Olson, 2001).

Student comprehension was another area that was explored within the theoretical models to fully understand test construction. Automaticity theory, established by Laberge and Samuels (1974), demonstrates a model in which students’ process information automatically. Their idea was that “automatizing lower-level components frees attentional capacity, which can then be allocated to higher-level processing” (p. 133). This model helps students “organize their prosody, suggesting what should go with what, what should be emphasized, and what should be de-emphasized” (p. 133). Logan (1997) also believes that decoding problems lead to limited
cognitive processing. Logan states that, “Disfluent word-level processing disrupts text-level prosody” (p. 141). He continues by stating, “Reading for meaning is a complex activity that requires integration at all different word levels” (Logan, 1997, p. 141). Therefore, if students are not automatically processing the vocabulary in the text, they are not generating meaning, and cannot comprehend the material asked of them.

Furthermore, Just and Carpenter (1980) took a closer look into working memory with the eye fixation theory. Eye fixation theory is where a reader’s eye is fixated on a word due to difficulty with decoding a particular vocabulary word in a given text, which in turn causes cognitive overload in the working memory. To assess their theory, Just and Carpenter investigated the duration of eye movements in a sample of freshman college students. To do so, specialized instrumentation was used to measure each student’s duration of milliseconds, which was spent gazing at each individual word in a sample passage. The study concluded that students spent a longer duration of time fixating on unknown vocabulary. Therefore, the researchers concluded that the extra time spent would give the students a limited amount of working memory (Just & Carpenter). In other words, a person has a very limited working memory that can be exceeded in space (Just & Carpenter).

Furthermore, Miller (1956) also believes that comprehension is limited by decoding and that people have only a limited amount of working memory or short-term memory. Miller claims that a person can remember around seven items in his/her short-term memory. This item can be such things as a face, a number, a word, or anything that can be grouped together. By grouping the same or close to the same items into a collection, known as a “chunk”, a person’s short-term memory is expanded. Miller believes that one can store up to around seven or more words or pieces of words at a given time before a person’s working memory is filled. In other words, if a
student knew each of the seven words in a sentence that would be seven chunks, but if this individual needed to spend time decoding and chunking each letter, then he/she might use up all of his/her working memory on a single word.

Sweller's Cognitive Load Theory (1988) combines Miller's work and the schemata theory. Sweller claims a person’s short-term memory is limited in the number of elements it can contain at the same time. This theory treats schema, which is a combinations of elements, as the cognitive parts that make up an individual's knowledge base. Essentially, the schema or the combined elements when chunked together will expand a person’s memory.

Another model to be considered is Torgesen’s Verbal Efficiency theory (1986). Torgensen explains:

Verbal efficiency theory points out that the higher level thinking processes involved in skilled reading comprehension place heavy demands upon the attentional, or processing capacity of the reader. If lower level word identification, or decoding, processes also require a significant share of processing capacity, there will be few resources left to support the complex thinking operations that are required for good comprehension. One of the most important implications of this theory is that word recognition processes in reading must occur fluently (rapidly, effortlessly, accurately) before higher level comprehension processes can be fully developed.” (p.158)

New Information and Research

More group-administered achievement testing is taking place in states and local school districts than ever before. A survey published in Education Week revealed that every state has adopted mandatory tests at one or more grades at the elementary, middle, and high school levels, and 49 states have linked their academic standards to these tests (Achieve, Inc., 2002 as cited in
Olson, 2001). Moreover, Congress recently enacted legislation mandating annual testing in reading and mathematics for all children in grades 3 – 8 (Olson).

Despite the call for more accountability in our educational assessment system, public opinion regarding such practices does not always agree as shown by numerous polling data. For example, a poll taken by the American Association of School Administrators (AASA) reveals that “63 percent of registered voters disagreed that a student’s progress for a single school year be solely placed upon a single test, while 49% opposed students’ being held back a grade to achieve a passing score on an achievement test” (Olson 2000, p. 9). Only seven states provide extra funding for low-performing schools and just nine states allocate funds for remediation of failing students (Olson, 2000). Olson reports that by the close of 2000, graduation was contingent on test performance in 18 states and five states were starting to administer exit exams. By 2003, as many as 27 states will be able to withhold the diplomas of students who fail to pass their state accountability examinations (Olson, 2001).

Furthermore, the National Council of Teachers of English (2001) issued the following statement about the usage of high stakes testing:

The efforts to improve the quality of education, especially underachieving schools, are laudable, and the desire for accountability is understandable. However, high stakes tests often fail to assess accurately students’ knowledge, understanding, and capability. Raising test scores does not improve education. Therefore, the use of any single test in making decisions—such as graduation, promotion, funding of schools, or employment and compensation of administrators and teachers—is educationally unsound and unethical (p. 300).
On another note, the use of large-scale achievement tests as instruments of educational policy is growing. In particular, states and school districts are using such tests in making high-stakes decisions with important consequences for individual students. Heubert and Hauser (1999) claim, “These policies enjoy widespread public support and are increasingly seen as a means of raising academic standards, holding educators and students accountable for meeting those standards” (p. 1).

In 2003, Public Agenda, a nonpartisan opinion research and civic engagement organization, published a report (Johnson & Duffett, 2003) that included results from several different nation-wide surveys. This report revealed that most people in the United States support testing. According to the report, “71 percent of parents, teachers, and the general public said they support annual mandatory testing as a check on the performance of schools” (p. 23). The general public’s view toward the need of a large assessment has led to the beginning of a new era of testing known as out-of-level testing.

There are both benefits and drawbacks to out-of-level testing. The rationale for the use of out-of-level testing emerges from modern test theory, which demonstrates that the reliability with which a student’s performance varies as a function of where the test score falls within the full scoring (Minnema, Thurlow, Bielinski & Scott, 2001). Individuals who advocate for out-of-level testing assert that: (1) information gleaned from these assessments aligns more closely with the instruction a student is receiving, (2) measurement is more accurate, and (3) testing is less stressful for students (Minneman, Thurlow, Bielinski, & Scott).

Moreover, organizations supporting the use of out-of-level testing, such as the American Association of School Administrators, view three benefits of this type of testing for students with disabilities. One benefit is the reduced level of test frustration due to the use of appropriate texts. A second benefit is the improvement of the test measurement accuracy. Lastly, advocates claim
that the test items are better matched to students’ current educational goals and instructional level (Thurlow, Elliott, & Ysseldyke, 1999).

Those who oppose out-of-level testing question whether the benefits outweigh the drawbacks. Opponents cite the following concerns: Out-of-level testing (1) creates a lack of consistency between the test and the purpose of the test (Thurlow, Elliott, & Yesseldyke, 1999), (2) reflects low expectations for students (Thurlow, Elliot, & Yesseldyke), (3) reduces the number of students who are unable to meet the grade-level standards or curriculum., and (4) creates unfair expectations for all students.

Thurlow, Elliot and Yesseldyke (1999) propose that the “reporting of [out of level test] scores can become confused and convoluted” (p. 1). For example, if a fifth grade student is taking a test that is at the second grade level, and another student is taking a test at the appropriate fifth grade-level are the tests scored exactly the same? This practice varies across school districts. It has been said that some districts count them the same, while others disagree and weight them, and some do not take them into account at all (Thurlow, Elliott & Yesseldyke, 1999). According to Thurlow, Elliott and Yesseldyke, “What this suggests is that out-of-level testing contributes to problems in accurately assessing and communicating the academic achievement of all students--that is, unless the tests have been scaled and equated for out-of-level use” (p. 1). More importantly, peers may view students who take the out-of-level test differently and this perception may directly or indirectly affect them socially or academically.

Minnema, Thurlow and VanGetson (2004)’s study explored the drawbacks of out-of-level testing. They developed and administered a one page, double sided written survey that contained closed and open-ended items to describe educators’ perceptions and opinions about out-of-level testing. The survey was distributed in four states at state-level teacher and administrator meetings and training sessions. The research concluded, “…all states agreed
overall that out-of-level testing did reduce student frustration during test taking. With 60% in agreement for State 1 and 83% agreement for State 4, these two states overwhelmingly affirmed the use of out-of-level testing” (p. 33). Data for States 2 and 3 were inconclusive (Minnema, Thurlow & VanGetson, 2004). Student responses to this out-of-level test survey revealed the following:

Most students said that they liked taking the statewide test out of level (8 out of 10 students). Half of the students thought it was neither too hard nor too easy, although two students described the below grade-level test as “too babyish.” When asked about test rigor, seven students mentioned guessing at item responses, although six of these students indicated that they guessed only minimally and one student guessed frequently. Only one student indicated guessing at all throughout the out-of-level test. Two students did not mention guessing at test item responses. (Minnema, Thurlow & VanGetson, p. 12)

Yoshida (1976) completed a groundbreaking study on out-of-level testing. His study investigated 12 unified school districts by their district size, location, ethnic representation, and overall income status. Within each district, two groups were identified and randomly selected to be a part of the study. Most of the 359 student participants were identified as students with special needs who were “engaged in a full inclusion model” (p. 217). The teachers administered the MAT for all levels. The results revealed that the students from the sample did not appear to lower the reliability estimates. The conclusion from the study was that testing appeared to be successful in presenting test items to students with special needs in a way that appears to eliminate guessing, and “increases the likelihood of scores based upon how much students comprehend” (p. 217).
Furthermore, Plake and Hoover (1979) decided to delve deeper into the comparability of equal and raw scores obtained from in-level and out-of-level equivalent scores. They administered the Iowa Tests of Basic Skills (ITBS) as the achievement battery to fifth grade students. They also used three subtests from the ITBS test: Vocabulary, Reading Comprehension, and Mathematics Concepts. For the study, only level 11 was used for fifth grade. This process was used because it examined the difference between two lower grade and two higher grades (Plake & Hoover). Plake and Hoover concluded that “students being tested using the out-of-level tests had the same curriculum as the in-level students, and that the test questions were equally as difficult for the out-of level students as the in-level” (p. 276). They also concluded that out-of-level testing is “a reasonable attempt toward achievement information” and has “great potential” (p. 276). Additionally they found that “continued research is needed to resolve much of the problems associated with scoring and achievement level” (Plake & Hoover, p. 276).

The state of Ohio (ODE, 2008) developed an Alternate Assessment, which is administered to students with significant cognitive disabilities. The participation in this assessment is limited, and is not typically based upon disability condition, achievement level, school, attendance, or social/cultural factors. The catch is that “Federal regulations under the NCLB place a cap only on the total number of alternate assessment scores that may be counted as proficient in AYP accountability calculations for school districts. This cap has been established at an amount not to exceed one percent of the total tested population. Ohio policy also applies this cap to state accountability calculations for school districts” (ODE, p. 31). Furthermore “The one percent cap does NOT limit the number of students who may be assessed with the alternate
assessment, only the number of scores that can count as proficient in school and district accountability calculations” (ODE, p. 31).

Two Navy researchers, Duffy and U’Ren (1982), conducted a series of five studies involving United States Navy recruits on the relationship of readability, “readable writing” techniques, and comprehension. In the study, eight expository passages from a standardized reading test were revised by using word lists to simplify vocabulary and a restriction in syntactic structure to simplify the sentences. The passages were presented to recruits in five different conditions in which the reading task, the time allowed for reading, and the comprehension test format varied. The results concluded that a positive stance was indicated for low ability readers when the reading task was in a reading-to-learn format (Duffy & U’Ren). The results also indicated that readable revisions might facilitate comprehension under those circumstances. The results for the other recruits were inconclusive and did not show much or any variance.

Furthermore, in a correlational study conducted several years earlier, Marks, Doctorow and Wittrock (1974) studied the effects of word frequency and reading comprehension. They hypothesized that by varying the frequency of 15% of the words in elementary school reading materials there would be major gains in the comprehension of passages. To test the hypothesis, 222 sixth graders were randomly assigned to two reading treatments. These treatments were administered simultaneously in the same classroom. The differences between the two groups involved the frequency of words provided to each group. There was a 15% difference in the frequency of words provided to one group. “The results showed that the comprehension was significantly increased (p < .0001) with high frequency story passages. This indicated that the frequency in a small percentage of words enhanced story comprehension, while a few less familiar words inhibited comprehension” (p. 262).
In a similar study, Freebody and Anderson (1983) showed that the presence of difficult vocabulary does diminish comprehension. In the experiment, “72 sixth graders were given 300 word passages from the 5th Grade Scott Foresman Social Studies textbook” (p. 20). Three different passages were used in the study; one passage had easy vocabulary; one was medium in difficulty; one was hard in difficulty (Freebody & Anderson). They measured the comprehension by free recall, summarization, and sentence verification. The results from the study determined that there was a significant difference in only the sentence verification testing. On the recall measure, “the mean recall score was higher in the easy form than the difficult form, while the medium vocabulary’s measures were very inconsistent in eight of the nine passages” (Freebody & Anderson, p. 38).

Britton (1996) conducted a study that explored the improvement of instructional textbooks. In four separate studies, Britton tested two methods of revising instructional texts to improve student learning. There were 40 students in a high school biology class in the first study. A total of 30 ninth-grade students in English and history and 29 college freshmen were included in the second study, and 41 college students in the third. The methods tested included giving the original textbook excerpts without any changes to one group of students and the newly changed versions to a separate group, and then giving the same tests of learning to both groups. The findings revealed that one method “untangled the ‘cognitive knots’ in the text”, which made the text more understandable to read without the author’s own personal diction” (p. 10). The findings from the second method revealed that the "point" of the text was underlined. In every single instance, the groups who read the new version had significantly higher scores on the test than those who read the original textbook excerpts. These findings suggest that the revision techniques were effective in increasing learning (Britton).
Sweet and Snow (2003) believed the relationship between readability and comprehension has been a concept much overlooked. According to the authors, readability is an attribute of texts while comprehension is an attribute of readers. There is, therefore, a difference between the two concepts. Features of a text have a large effect on comprehension. Texts that are badly written or poorly structured are harder to understand. While reading, the reader constructs various representation of the text that is important for comprehension. These representations include for example, the “surface code (the exact wording of the text), the text base (idea units representing meaning), and a representation of the mental models embedded in the text” (p. 5).

According to William S. Gray (1937), reading is an inclusive process, and it assumes that the reader not only recognizes the essential facts or ideas presented, but also reflects on their significance by evaluating them critically, discovering relationships between them, and clarifying his understanding of the ideas apprehended. And the skills, abilities, and attitudes needed for problem solving are basically the same as those needed for comprehension in reading (cited in Stauffer, 1956, pp. 4-5).

According to Harrison, “It is common in the United States to assert that for independent study a book should be two years easier than the material used in class when the teacher is available to support and explicate” (p. 126). He continues by stating, “If questions are phrased in words which are less familiar than those which are used in the passage, readers will get fewer right than they would otherwise have done” (p. 130).

Sweet and Snow (2003) believe, “When text factors are poorly matched to a reader’s word knowledge and experience, the text may be too difficult for optimal comprehension to occur” (p. 7). In turn, “… some readers would become increasingly mystified and frustrated if comprehension did not come quickly” (p. 4). Fortunately, most classroom teachers know that
some students come to the task of comprehension better prepared than others (Snow & Sweet). Unfortunately, a child whose word recognition skills are inaccurate and/or labored will lose or have limited access to meaningful ideas in the text, and their comprehension will suffer.

According to Vellutino (2003), “The research to date has shown that a child who has difficulty identifying key words such as ‘dog,’ ‘cat,’ ‘play,’ ‘saw,’ or ‘catch’ will, no doubt, have difficulty identifying the central characters and/or central theme of this segment of the story” (pp. 52-53).

Clay and Imlach (1971) argue that “When students read aloud with awkward pauses, stops, and starts, they treat each word as a single entity it will limit the flow of the passage” (p. 53). At the same time, the slowness of their reading is a further barrier to meaning (Clay & Imlach). Clay and Imlach also state that, “their expressiveness in reading connected text is severely limited, shown by a lack of tone, inflection, and changes of rate or speed that enable the listener to fully interpret the text” (p. 135).

In response Graves, Slater, Roen, Redd-Boyd, Duin, Furniss, and Hazeltine (1988) believe that user-friendly texts could be easily assumed to reduce the demands on readers’ cognitive capacities and inferencing skills. They believe that “…in such texts, earlier content is restated when important for understanding a current section, connections that readers must make are clearly delineated, relevant background knowledge is explicitly presented, and distracting information (e.g. details, tangible facts) is kept to a minimum” (p. 244).

Furthermore, Marks et al. (1974) affirm that replacing 15% of the words in a reading text with low frequency words lead to a significant decrease in comprehension. Freebody’s and Anderson’s (1983) studies on the impact of placing low frequency words in important parts of the text as well as in the unimportant parts, reveal a decrease in the understanding of the whole text. Beck and McKeown (1991) suggest, “While interest in the topic remains keen, the amount
of research done on vocabulary size and other forms of measurement has declined significantly in recent decades” (p. 792).

**Readability**

On another note, in response to the ideas upon reading and comprehension many different readability formulas were discovered. As defined by DuBay (2004), “Readability is what makes some texts easier than others” (p. 3). Furthermore in a more broad or comprehensive manner Chall and Dale (1949), would identify readability as “the sum total of all those elements within a given piece of printed material that affect the success a group of readers have with it” (p. 23)

**Formulas**

In 1923, Lively and Pressey developed the first readability formula. The formula was developed after both Lively and Pressey were tired of having to teach science terms to students that were at too high a level. They were also tired of the chore of finding textbooks that would remedy this problem. Therefore, they developed a formula that measured the correlation per 1000 words that matched the Thorndike word list.

Rudolf Flesch created one of the most widely renowned readability formulas in 1948. The formula uses two different parts to gauge readability of a text. The first part of the formula involves identifying “the number of syllables and the number of sentences for each 100-word sample” (p. 223). The second part of his formula uses words that are more personal such as names, quotes, exclamations, pronouns and incomplete sentences (DuBay, 2004). In 1950, this formula was tested by Hayes, Jenkins, and Walker with experienced and inexperienced analyists. The conclusions from the study found that “reading ease with its components is analyzed quite reliably,” and “human interest is used with fair reliability” (p. 25).
Furthermore, to enhance Flesch’s formula, Kincaid, Fishburne, Rogers, and Chissom (1976) developed the Flesch-Kincaid Formula of Readability. The new formula was able to determine “the significant differences between modules less than one grade level apart using both comprehension scores and learning times” (p. 8).

Edgar Dale and Jeanne Chall (1949) developed the Dale-Chall formula for adults and children above the fourth grade level. This formula is similar to Flesch’s original formula except it uses different grade appropriate leveled words to compute the overall grade level of a text. According to Fry (1968) the “Dale-Chall gives two grade designations such as 5-6 or 7-8,” (p. 516). The word list differed also in that it did not agree with the Thorndike word list due to the authors’ concerns with the list.

Lastly, in 1977, Edward Fry developed the readability graph to show a quick visual way of determining the readability of school texts (Fry, 2002). To use the Fry Readability Graph, the user randomly selects three 100-word samples and counts the number of sentences in each sample. Then, the user counts the number of syllables in each of the three samples and averages the sentence count and syllable count and puts them into the graph to obtain a grade level (Fry, 2005).

There are many other formulas that are very well known, but are based on the above-mentioned formulas. These formulas are the “FORCAST”, “SMOG”, and ”Fog” readability formulas (DuBay, 2004, p. 55). These formulas, like the rest of the formulas, are used in an effort to help simplify textbooks and other materials.
Summary

There is a vast amount of information that supports as well as does not support high stakes testing. The amount of information weighing against high stakes testing is very widespread. According to Johnson and Duffet (2003), there appears to be a need to keep students and teachers accountable. The problem that exists from the accountability is that it leads to lack of funding and an inaccurate account of student performance. That is the reason that out-of-level tests were introduced.

Out-of-Level tests are a type of assessment that theoretically will adequately gauge a student’s overall skill ability. There are many different advantages and disadvantages associated with the concept of out-of-level testing. Unfortunately, statistical research about the effects of the use of out-of-level testing is very limited, and only a modest share of the relevant research examines the effects on student achievement. The need for further research and development is imperative toward student fair and equal achievement within the classroom. This led to examining the individual differences within actual text and comprehension. There is a significant amount of information that is available concerning the correlation between vocabulary and comprehension within a text.

Three main theories coincide with the idea. The theory of automaticity tells us that words need to be decoded automatically for comprehension or understanding to occur. Schema theory suggests that there can be a strong relationship between one single word and its background and a sentence. The eye fixation theory speculates that every word is looked at individually and not clustered, which spawned the theory of capacity and comprehension where a person’s working memory can only hold so much informational processes at a time. These theories and research studies conclude that there is a strong correlation between vocabulary and comprehension or understanding. Furthermore, identifying the need for tests that can be administered to a broader
range of students within the classroom for students to better demonstrate their own individual abilities of achievement is necessary.
CHAPTER III. METHODS AND PROCEDURES

Research has shown there is a significant difference that exists between student reading levels in our classrooms. Many students who are in the classroom are struggling readers. In some schools, the ratios are higher than others. The statistical data on the fifth grade Ohio Achievement Test shows that those who struggle on the reading proficiencies also have lower test scores on the Ohio Science Achievement test. With these findings, the current study focuses on students’ scores on specific standards that are rooted and assessed within the Ohio Fifth Grade Science Achievement Test. The purpose of the study was to examine the relationship between readability and test performance. The research question guiding this study was: “How would an alternative 5th Grade Science Achievement Test with a lower reading level impact test scores?” This chapter will explain the methods and procedures that were used to conduct the study.

Methods

*Research Design*

This study examined the relationships that exist between readability and science performance. To identify relationships that exist between these two variables, a correlational research design was employed. Students’ test scores from half of a 2007 fifth grade-science achievement test were collected and evaluated to determine the students’ overall performance. Students’ test scores from the normative test were then compared based on total test scores of the alternative Ohio Science Achievement Test with a lower reading level. The study also compared individual items on the test to identify differences.
Participants

The data, information, and scores that were used in the study were collected from one school district in Northwest Ohio. The regular Ohio Fifth Grade Achievement Test was given to a class of fifth graders in January 2009. The school that was chosen in the study is of lower socio-economic status in a rural community. The school was selected by convenience sampling. According to Mertler (2009), convenience sampling is the method of choosing arbitrarily and in an unstructured manner. It is the method most commonly employed in many practical situations (Mertler). Thus the school was chosen based on location, convenience, and availability. One fifth-grade classroom was randomly selected. A total of 34 students’ scores were collected and studied. Of the 34 students in this classroom 3% were of unknown ethnicity, 4% were Hispanic, and 93% were Caucasian.

Instrumentation

The Ohio Department of Education provides the public with half-length past proficiency tests (see Appendix A). The 2007 Ohio Fifth Grade Science Test was used to collect and analyze data. The test was also used to create an alternative Ohio Fifth Grade Science Test. Using each individual test question in order, and the Ohio Academic content standards of Science and Technology, Scientific Inquiry, Scientific Ways of Knowing, Earth and Space Sciences, Life Sciences, and Physical Sciences, the test was recreated. Using the Flesch-Kincaid readability formula and the third grade Dolch word list as guides, a revised test was created.

To answer the research question above, a revised version of the test was created to measure if reading had a statistically significant impact on test performance (see Appendix B). First, the standardized directions were copied and pasted into the revised version to keep the test as close to the original version as possible. The test needed to be a mirror image in content to
keep its content validity. Furthermore, each original individual question was typed into Microsoft Word, and each of the images for each question were then added. Next, all of the questions were analyzed one by one to check the readability levels.

The Ohio Academic Content Standards for 5th Grade were studied for each question to ensure that the content was kept the same throughout each individual question. Then each question was examined using the Dolch List. The Dolch List did not have much bearing on the readability levels on the test. Therefore, the three main factors that influenced the creation of the test were syllabication, word length, and sentence length. For example, in Question #1 the word “morning” was changed to the word “day.” The word “morning” is a 2-syllable word with 7 different letters, and the word “day” is only 1 syllable, and 3 letters. The changing of this one word did not have any bearing upon the test’s content, but it did influence the reading level of the test. Therefore, the word was changed and the readability level of the question was altered. This same type of process was done for questions #1-19 (see Appendix B). All of the questions for both tests are multiple-choice items except for questions #7, #11, and #19.

The time allotted for the full-length test is 2 hours and 30 minutes. For this purpose of this study, the time was altered from the original testing instructions to 1 hour and 15 minutes. An electronic clock was used to the alternative test (half the time of a regular achievement test). The directions from the Ohio Department of Education’s Achievement Test Administration guide were read exactly word for word. The only deviation was the time allotted to take the test (see Appendix C). A copy of the student answer document is located in Appendix D.

Procedures

Several steps and procedure were used in the planning and implementation of this study. First, one school was contacted to secure participants for the study. The principal was contacted
by a colleague who had worked in the building. Then a formal meeting was scheduled with a principal and a fifth grade teacher. This meeting was divided into four parts: purpose, testing, results, and questions/answers. The month of testing for the first test occurred in mid-January 2009. The second test was held on the first week in February 2009. The test was administered on the assigned dates.

The principal (see Appendix E), the teacher (see Appendix F), the parents (see Appendix G) and each of the students (see Appendix H) signed release of information waivers, which were approved by the BGSU Human Subjects and Review Board. Before the test was administered, the students were asked to create a pseudonym (first name only) of their choice to be used on both tests. This was done to conceal the students’ privacy/identity in the study. The test was timed 1 hour and 15 minutes and the students were given the exact directions based upon the Ohio Achievement Test Administration Testing booklet. The first test was given on the assigned date, and the testing data were entered in a database. The second test was given, and the information was added to the existing database.

Data Collection

Based on data collected from each of the tests’ total level of performance and individual item performance, the test scores were entered into StatCrunch. The StatCrunch program is an online computer program that calculates statistical analysis. The program is used to upload data files and analyze data using an extensive list of numerical and graphical procedures that StatCrunch offers. In addition, the user can export data analysis for easy access for the future. The program was used to determine the correlational aspects of the quantitative data from the raw scores of the achievement test.
Data Analysis

The statistical and graphical analyses of data were completed by running a series of statistical tests. All of the students’ unaltered science achievement tests were calculated by performing a paired T-Statistical formula and a summative statistical analysis. These data were further analyzed through the T-Statistical and Summary Stat feature in the report section of the Statcrunch database. The raw scores and mean scores were compared through statistical analysis within the database as well. These tests were performed to identify significant statistical differences that may appear.

Summary

The goal of this study was to identify correlations that exist between readability level and performance within the Ohio Fifth Grade Science Achievement Test. The test was conducted to determine whether there would be a correlation between the readability level of a content area test, and understanding or comprehending test questions. The students’ test scores were obtained from one school in northwest Ohio. The testing site was chosen due to convenience sampling. Each individual student’s score was compared between the unaltered sample test and the alternative test, to determine whether there would be any recognizable or relational trends.
CHAPTER IV. DATA ANALYSIS AND DISCUSSION OF RESULTS

The importance of state assessments is on the rise, and so should the judgments about whether tests that are being administered are appropriate for all students. According to the American Educational Research Association (1999), “many school districts are mandating tests to measure student performance and to hold individual schools and school systems accountable for that performance” (p. 1). In the state of Ohio, the 5th Grade Science Achievement tests are given to students in grades five through eight to assess student performance and ability level. Despite the fact that the test focuses on science content, reading is a major component of the achievement test even though reading has its own separate achievement test.

This study investigated the relationship that may be present among students’ test scores with an alternative science test and those with the traditional Ohio Achievement Test (OAT). This study also investigated differences and correlations between individual test questions and overall test performance. The driving research question that guided the investigation was, “How would an alternative 5th Grade Science Achievement Test with a lower reading level impact test scores?”

Data Analysis

Students’ scores from the Ohio Fifth Grade Science Achievement Test were collected and analyzed from one rural school to determine whether a significant relationship exists between readability and test performance. Scores from 34 students were examined (20 females, 14 males). For each question, each response was measured to examine if there was a difference between the standardized version of the Ohio 5th Grade Science Achievement Test and the revised version.

To answer the research question, a correlational research design was needed that involved the use of a Paired T Statistical Formula or t-test. The formula was calculated using the
StatCrunch database, which is found at www.statcrunch.com. The t-test was used to examine the significance between each individual question from the standardized test (test A) and revised version (test B). For example, the test of the questions 1A and 1B were analyzed together by evaluating the p-value. If the p-value was not under .05 then the question did not have statistical significance. Also, the t-test was used to identify if there was a statistically significant difference between the sum of test A and test B.

Secondly, to examine the test comparison as a whole, the test was analyzed through the Summary Statistical formula within StatCrunch. Furthermore, the formula dissected all of the questions on the test and displayed the mean and variance between each of the questions. The mean that is calculated is based upon the amount of students who correctly answered each of the questions, and the variance is the percentage that is different between each of those percentages. The subsequent paragraphs will address and describe the results that were obtained to answer the research question.

To answer the research question, the t-stat and the summary statistic were examined for each individual question (see Appendix I & J). If the mean was below 50% on any one question, then the Ohio Content Standard was identified. The summary statistics chart was a guide to examining each of the questions individually.

**Questions 1A and 1B**

The first question 1A and 1B had an original reading level of 1.4 and a revised readability level of 1.0. From the data that were collected, there was a mean of 1.0 for 1A and .94 for 1B. This means that 34/34 students or 100% were correct in 1A and 94% of students were correct on 1B. The difference was two questions. There was a total variance of .06 or 6% between the two. To determine statistical significance, t-tests were used. The T-Stat was 1.43 and the p-value was
Based on the results of the paired t-test as depicted in Table 1, there was not a statistical difference between the standard and revised version of scores for 1A and 1B.

Table 1
Paired T Statistics: Evaluating The Mean Of Test Scores For Test 1A &1B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A-1B</td>
<td>0.0588</td>
<td>0.0409594</td>
<td>33</td>
<td>1.436141</td>
<td>0.1604</td>
</tr>
</tbody>
</table>

Questions 2A and 2B

Questions 2A and 2B used the same procedures to interpret the results. The Flesch Kincaid formula revealed a readability score of 2.0 for the A test and a 1.4 for the B test. The content that was covered within this question was “5th Grade OACS: Scientific Inquiry #4 Identify one or two variables in a simple experiment” (Office of Curriculum and Instruction, 2003, p. 119). The results indicated a mean of .38 for 2A and .53 for 2B. This means that 38% of students were correct for 2A and 53% of students were correct on 2B. There was a total variance of .06 or 6% between the two. The T-Stat was -1.71 and the p-value was .10. Based on the results of the paired t-test as shown in Table 2, there was not a statistically significant relationship between the standard and revised version of scores for 2A and 2B.
Table 2

Paired T Statistics Evaluating The Mean Of Test Scores For Test 2A & 2B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A-3B</td>
<td>-0.14705883</td>
<td>0.08582568</td>
<td>33</td>
<td>-1.7134594</td>
<td>0.096</td>
</tr>
</tbody>
</table>

Questions 3A and 3B

The Flesch Kincaid readability score for Questions 3A and 3B was a 3.9 and a 2.0 respectively. The mean of .94 for 3A and .85 for 3B indicates 94% students were correct on test item 3A and 85% of students were correct on item 3B. There was a total variance of .09 or 9% between the two questions. The T-Stat was 1.79 and the p-value was .08. Based on the results of the paired t-test in Table 3, there was not a statistically significant relationship between the standard and revised version of scores for 3A and 3B.

Table 3

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 3A & 3B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3A-3B</td>
<td>0.0882353</td>
<td>0.049374826</td>
<td>33</td>
<td>1.7870502</td>
<td>0.0831</td>
</tr>
</tbody>
</table>

Questions 4A and 4B

For questions 4A and 4B, the t-test and summary statistics were used to interpret the results. The Flesch Kincaid formula was used to measure the readability. The Flesch Kencaid for
4A was 2.8 and for 4B the reading level was lowered to a 2.4. Through the collected data, there was a mean of .65 for question 4A and .71 for question 4B. Furthermore, 65% of students were correct on 4A and 71% of students were correct on 4B. There was a total variance of .05 or 5% between the two of test formats. The T-Stat was -.49 and the p-value was .62, as shown in Table 4 below. Based on the results of the paired t-test, it appears that there was not a statistically significant relationship between the standard and revised version of scores for 4A and 4B, but there was a total difference of 5%.

Table 4
Paired T Statistics: Evaluating The Mean Of Test Scores For Test 4A & 4B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>4A-4B</td>
<td>-0.05882353</td>
<td>0.118976444</td>
<td>33</td>
<td>-0.49441323</td>
<td>0.6243</td>
</tr>
</tbody>
</table>

Questions 5A and 5B

The t-test and summary statistics were used to interpret the results for questions 5A and 5B. The Flesch Kincaid formula was used again to determine the reading level. The reading levels were 8.5 for 5A and 6.7 for 5B. Through the data that were collected, there was a mean of .53 for question 5A and .65 for question 5B. A total of 53% of the students had correct answers for question 5A and 65% of students were correct on 5B. There was a variance of -0.11 or 11% between the two tests. The T-Stat was -0.94 and the p-value was 0.35. Based on the results of the paired t-test, there was no significant difference between the standard and revised version of scores for 5A and 5B, but there was a notable difference of 11%.
Table 5

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 5A & 5B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5A-5B</td>
<td>-0.11764706</td>
<td>0.12499344</td>
<td>33</td>
<td>-0.9412258</td>
<td>0.3534</td>
</tr>
</tbody>
</table>

*Question 6A and 6B*

In the questions 6A and 6B the t-test and summary statistics were used to interpret the results. Through the data that were collected, there was a mean of .12 for 6A and .21 for 6B. This means that 12% students were correct on question 6A and 21% of students were correct on 6B. The readability level of question 6A was 3.2 and 2.9 for question 6B. There was a total variance of -0.09 or 9% between the two tests. The content standard assessed in this questions was the “5th Grade OACS: Physical Sciences: #2 Trace how thermal energy can transfer from one object to another by conduction” (Office of Curriculum and Instruction, 2003, p. 118). The T-Stat was -1.0 and the p-value was 0.32 (see Table 6). Based on the results of the paired t-test, there was not a statistically significant relationship between the standard and revised version of scores for 6A and 6B, but there was a difference of 9%.
Table 6
Paired T Statistics: Evaluating The Mean Of Test Scores For Test 6A & 6B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>6A-6B</td>
<td>-0.0882353</td>
<td>0.0882353</td>
<td>33</td>
<td>-1.000</td>
<td>0.3246</td>
</tr>
</tbody>
</table>

Questions 7A and 7B

The t-test and summary statistics were used to interpret the results. Through the data collected, there was a mean of .97 for question 7A and 1.12 for question 7B. These data were based upon the student’s exact scores being overall averaged. There was a total variance of -0.14 or 14% between the two of questions. The readability level for question 7A and 7B was 5.3. The T-Stat was -1.71 and the p-value was 0.10. Based on the results of the paired t-test, there was no statistically significant relationship between the standard and revised version of scores for 7A and 7B, but there was a difference of 14%.

Table 7
Paired T Statistics: Evaluating The Mean Of Test Scores For Test 7A & 7B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7A-7B</td>
<td>-0.14705883</td>
<td>0.08582568</td>
<td>33</td>
<td>-1.7134594</td>
<td>0.096</td>
</tr>
</tbody>
</table>
Questions 8A and 8B

For questions 8A and 8B, the t-test and summary statistics were used to interpret the results. The Flesch Kincaid readability level was a 5.3 for the standard version and was lowered to a 4.7 on the revised test. Through the collected, data there was a mean of .91 for question 8A and .97 for question 8B. There was a total variance of -0.06 or 6% between the two. The T-Stat was -1.0 and the p-value was 0.32. Based on the results of the paired t-test, there was not a statistically significant relationship between the standard and revised version of scores.

Table 8.
Paired T Statistics: Evaluating The Mean Of Test Scores For Test 8A & 8B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>8A-8B</td>
<td>-0.05882353</td>
<td>0.05882353</td>
<td>33</td>
<td>-1.00</td>
<td>0.3246</td>
</tr>
</tbody>
</table>

Questions 9A and 9B

The Flesch Kinkaid readability level for questions 9A and 9B were 5.8 and 4.7 respectively. The t-test and summary statistics were used to interpret the results. Through the data collected, there was a mean of .56 for question 9A and .62 for question 9B. There was a total variance of -0.06 or 6% between the two. The T-Stat was -0.63 and the p-value was 0.54. Based on the results of the paired t-test, there was not a significant relationship between the standard and revised version of scores.
Table 9

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 9A & 9B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>9A-9B</td>
<td>-0.05882353</td>
<td>0.09384989</td>
<td>33</td>
<td>-0.6267832</td>
<td>0.5351</td>
</tr>
</tbody>
</table>

*Questions 10A and 10B*

A t-test and summary statistics were used to interpret the results for test questions 10A and 10B. The Flesch Kincaid formula was used to determine the readability level. The readability for question 10A was an 8.5 and 7.2 for question 10B. Through the data collected, there was a mean of .79 for question 10A and .73 for question 10B. There was a total variance of -0.06 or 6% between the two. The T-Stat was 0.70 and the p-value was 0.49. Based on the results of the paired t-test, there was not a statistical relationship between the standard and revised version of scores as depicted in Table 10 below.

Table 10

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 10A & 10B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10A-10B</td>
<td>0.05882353</td>
<td>0.083816886</td>
<td>33</td>
<td>0.70181</td>
<td>0.4877</td>
</tr>
</tbody>
</table>
Questions 11A and 11B

The readability level for question 11A was a 5.0 for test A and a 2.3 for test B. There was a mean of .71 for question 11A and .97 for question 11B based upon the averages of all of the scores. There was a total variance of -0.26 or 26% between the two. The T-Stat was -2.72 and the p-value was 0.01 (see Table 11). Based on the results of the paired t-test, there is a statistically significant relationship between the standard and revised version of scores.

Table 11
Paired T Statistics: Evaluating The Mean Of Test Scores For Test 11A & 11B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>11A-11B</td>
<td>-0.2647059</td>
<td>0.09727869</td>
<td>33</td>
<td>-2.7211087</td>
<td>0.0103</td>
</tr>
</tbody>
</table>

Questions 12A and 12B

The Flesch Kincaid formula, t-stat, and summary statistics were used for questions 12A and 12B. The readability level for question 12A was 2.8 and a 2.6 for question 12B. There was a mean of .79 for question 12A and .85 for question 12B based upon the averages of all of the scores. There was a total variance of -0.06 or 6% between the two. The T-Stat was -0.81 and the p-value was 0.42 (see Table 12). Based on the results of the paired t-test, there is not a statistical significance relationship between the standard and revised version of scores.
Table 12

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 12A & 12B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>12A-12B</td>
<td>-0.05882353</td>
<td>0.07240676</td>
<td>33</td>
<td>-0.81240386</td>
<td>0.4224</td>
</tr>
</tbody>
</table>

*Questions 13A and 13B*

The Flesch Kincaid readability level for question 13A was a 3.6 and 2.6 for question 13B. There was a mean of .65 for 13A and .65 for 13B based upon the averages of all of the scores. There was a “0” or no variance, because they both were the same. The T-Stat was “0” and the p-value was 1.00 (as shown in Table 13). Because there was an exact same mean score of the summary stat, there was not a statistical relationship between the standard and revised version of scores.

Table 13

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 13A & 13B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>13A-13B</td>
<td>0.00</td>
<td>0.07312724</td>
<td>33</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
**Questions 14A and 14B**

Questions 14A and 14B were analyzed using the same method. The readability for question 14A was a 4.7 and a 3.8 for question 14B. The data collected demonstrated a mean of .65 for question 14A and .38 for question 14B. There was a total variance of 0.26 or 26% between the two. The T-Stat was 3.02 and the p-value was 0.004 (as depicted in Table 14 below). This question represented the Content Standard, “Physical Sciences #3 Describe that electrical current in a circuit can produce thermal energy, light, sound and/or magnetic forces (Office of Curriculum and Instruction, 2003, p.118). Based on the results of the paired t-test, it there was a statistically significant relationship between the standard and revised version of scores.

Table 14
Paired T Statistics: Evaluating The Mean Of Test Scores For Test 14A & 14B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>14A-14B</td>
<td>0.2647059</td>
<td>0.0876391</td>
<td>33</td>
<td>3.0204084</td>
<td>0.0048</td>
</tr>
</tbody>
</table>

**Questions 15A and 15B**

The readability level for question 15A was a 5.3 and a 4.7 for question 15B. The data revealed a mean score of .35 for question 15A and .26 for question 15B. There was a total variance of 0.26 or 26% between the two of them. The T-Stat was 1.00 and the p-value was 0.32 (see Table 15). The content for this question met the “Earth and Space Sciences: C. Describe Earth’s resources including rocks, soil, water, air, animals, and plants and the ways that they can
be conserved” Content Standard (Office of Curriculum and Instruction, 2003, p. 108). Based on the results of the paired t-test, there was no statistical relationship between the standard and revised version of the test scores.

Table 15
Paired T Statistics: Evaluating The Mean Of Test Scores For Test 15A & 15B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>15A-15B</td>
<td>0.0882353</td>
<td>0.0882353</td>
<td>33</td>
<td>1.00</td>
<td>0.3246</td>
</tr>
</tbody>
</table>

*Questions 16A and 16B*

The readability level for question 16A was a 10.9 and was a 9.6 for question 16B. The data revealed a mean of .44 for question 16A and .35 for question 16B. There was a total variance of 0.09 or 9% between the two. The T-Stat was 0.76 and the p-value was 0.45 (see Table 16). This question was developed to address the content standard “4th grade Earth and Space Sciences #4 Describe weather by measurable quantities such as temperature, wind direction, wind speed, precipitation, and barometric pressure” (Office of Curriculum and Instruction, 2003, p. 114). Based on the results of the paired t-test and summary statistics below, it appears that there was not a statistically significant relationship between the standard and revised version of the test scores.
Table 16

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 16a & 16b.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>16A-16B</td>
<td>0.0882353</td>
<td>0.11459956</td>
<td>33</td>
<td>0.76994437</td>
<td>0.4468</td>
</tr>
</tbody>
</table>

*Question 17A and 17B*

The reading levels from the Flesch Kincaid formula for question 17A was a 5.8 and a 4.2 for question 17B. The data showed there was a mean of .32 for 17A and .24 for 17B. There was a total variance of 0.09 or 9% between the both of them. The T-Stat was 0.90 and the p-value was 0.37 as depicted in Table 17. The content that was portrayed in the question was the “5th Grade OACS: Physical Sciences: #2 Trace how thermal energy can transfer from one object to another by conduction” (Office of Curriculum and Instruction, 2003, p. 118). Based on the results of the paired t-test and summary statistics, there was not a statistically significant relationship between the standard and revised version of the test scores.

Table 17

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 17A & 17B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>17A-17B</td>
<td>0.0882353</td>
<td>0.09781615</td>
<td>33</td>
<td>0.90205246</td>
<td>0.3736</td>
</tr>
</tbody>
</table>
Question 18A and 18B

The readability level for question 18A was calculated to a 1.4 and 1.1 for question 18B. The mean for question 18A was .82 and .85 for question 18B. There was a total variance of -0.02 or 2% difference between both items. The T-Stat was -0.44 and the p-value was 0.66 (see Table 18). Based on the paired t-test and summary statistics, there was not a statistically significant relationship between the standard and revised version of the test scores.

Table 18

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 18A & 18B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>18A-18B</td>
<td>-0.029411765</td>
<td>0.066559106</td>
<td>33</td>
<td>-0.44188944</td>
<td>0.6615</td>
</tr>
</tbody>
</table>

Question 19A and 19B

The Flesch Kincaid readability level for question 19A was a 3.5, while the readability level for question 19B decreased to 2.9. The mean for question 19A was 2.06 and 2.94 for question 19B. There was a variance shown of -0.88 or 88% difference between both items (see Table 19). The T-Stat was -4.82 and the p-value was <0.0001. Based on the paired t-test below and summary statistics, there was a strong statistically significant relationship between the standard and revised version of the test scores.
Table 19

Paired T Statistics: Evaluating The Mean Of Test Scores For Test 19A & 19B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>19A-19B</td>
<td>-0.88235295</td>
<td>0.18288979</td>
<td>33</td>
<td>-4.8245063</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

*Summary of Test A (SUM A) and Test B (SUM B)*

Lastly, the paired t-test and the summary stat were used to test the sum of the standard achievement test (Sum A), and the sum of the revised achievement test (Sum B). The test was first measured through the summary stat that calculated Sum A with a mean of 13.65, and Sum B with a mean of 14.82. Then, the paired t-test was calculated and revealed the difference between the means of -1.17. The t-stat was -2.16 and the p-Value was at 0.04 (see Table 20). Since the p-value was less than 0.05, the difference between the sums A & B was statistically significant.

Table 20

Paired T Statistics: Evaluating The Sum Of Test Scores For Test A & Test B.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Difference</th>
<th>Standard Error</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>20A-20B</td>
<td>-1.1764706</td>
<td>0.5447376</td>
<td>33</td>
<td>-2.1597016</td>
<td>0.0382</td>
</tr>
</tbody>
</table>
Discussion of Results

The research question from this study was “How would an alternative 5th Grade Science Achievement Test with a lower reading level impact test scores?” The statistical findings from this study appear to indicate that there is a significant relationship between the readability of a test and of student performance in accord with the correlation summative data. What this means is that the total test scores from each of the tests showed large fluctuation. The students who took the revised test scored higher than the unaltered test by a total of 40 points.

The data from this study suggest that there was not a statistically significant relationship between all of the test questions, and even in some cases, the revised test created a contradictory decline in test performance. The one test question that showed a statistically significant decline was question 14, but this could have been the result of a test construction error since the test was designed to change the other terms that were not science related. The deviation that appeared to cause this fluctuation of scores was through shortening the word “electricity” to the word “electric”. The reason that this question is significantly different is that students most likely learn the full science term “electricity” instead of a condensed version such as “electric.”

Furthermore, the test showed scattered results in data ranging throughout the test items. Deviation from the remaining negative questions averaged to be a difference of 2-3 students. This could have been simply a judgmental error or just a change of mind toward the subject, or possibly the content was taught before the first test was given. The main areas of overall weakness were in the areas of questions 2, 6, 11, and 17. These areas did not show a significant amount of difference, and remained below 50% even with the revised version. The content standards were then addressed, and the rationale was the content being the main attributing
factor. Therefore, each of these is related to the content or skills and the curriculum, and not through readability levels.

Summary

It appears that on the Ohio 5th Grade Science Achievement test, the reading level does significantly affect test performance in favor of revised, easier-to-read questions. The Statistical T-Stat formula and the Summary Statistics show that there is a significant difference. There was also the finding that through the t-statistics, not all of the questions were equal, but the item differences were not significant enough to have an overall bearing upon the test itself. The test had an error in construction on item 14, which after a close examination, was due to a single word, which appeared to impact content knowledge. Even though there was not a statistically significant relationship upon all of the questions, this did show a low trend for the majority of the test questions toward the revised test. The overall results of the percentages can be used to help guide curriculum instruction within the school to help better differentiate instruction and allot more time for the areas of weakness.
CHAPTER V. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Achievement testing has become a major concern within the classroom at almost every grade level. Students are asked to perform at the same level. However, research has shown that students’ reading levels vary greatly in the classroom. Guzetti (1990) said, “Researchers have estimated that instructional reading levels in most content classrooms may range from three grades above students’ actual grade placement to three grades below” (p. 17). Numerous research studies conclude that students who are asked to read text that is not at their appropriately functioning reading level have problems with comprehension.

The purpose of the current study was to determine if readability was a major contributing factor for performance on the Ohio 5th Grade Science Achievement Test. The implications of the study could lead to a major change in how achievement tests are constructed, and how achievement tests are viewed in the United States. In this chapter, a summary of the study, along with a discussion of its conclusions and recommendations will be offered.

Summary

Achievement tests first emerged in the mid 1800s as a mass educational reform. Widespread acceptance of these tests spanned the United States in the 1900s. Achievement tests have evolved into an all-encompassing reality as a result of the No Child Left Behind Act. These achievement tests appear to be the government’s solution to educational reform. The implications that have resulted from these tests are that all students, teachers, and administrators are held accountable for the test scores. The heavy weight that is placed upon these tests has led the researcher to question if these tests are fully assessing student content knowledge. Many struggling readers within our schools may have problems decoding the text in the achievement tests questions. According to Adelman and Taylor (2006), a typical seventh grade classroom will
have reading levels that range from second grade through seventh grade and beyond. Furthermore, this implies that some students may not be able to understand the questions that are being asked of them when taking tests.

The purpose of this study was to determine whether readability was a contributing factor in student performance on the Ohio Fifth Grade Science Achievement Test. To do so, a revised version of the test was designed to mirror the academic content standards, while attempting to lower the readability of each test question. Both tests were then given to a group of 34 fifth grade students within a two-week period. The data were analyzed statistically through the Flesch Kincaid Readability formula, StatCrunch’s T-Statistical Formula, and Summative of data function. The data were then analyzed to identify whether there was a relationship between test questions and performance. Furthermore, the results of the test showed that the test as a whole was statistically significant, which means that it was a valid instrument in assessing students’ content knowledge, which proved that readability does have an impact on test performance.

Conclusions

It was found in the creation of the test itself that content related terms have a significant effect on the readability of test questions. The content related terms make lowering the reading level of a question a daunting task. With some questions, the readability could be dropped to the primer level, while it was not possible to get below the sixth grade level with others. Furthermore, not every sentence could be reduced to an established standard of reading level across the board, and there is nothing that can otherwise be done without disturbing content related terms and concepts.

The creation of the revised test demonstrates that it is possible to reduce the readability on a form of a standard Science Achievement Test. This was beneficial to students by providing an independent level reading, which is particularly important for struggling readers. According to the data in the study, there was a significant difference between the regular and revised
achievement test. The results of this study suggest that by lowering the readability on the test, the score students receive on the test may be a better predictor of their overall science knowledge, rather than their reading level vocabulary. Freebody’s and Anderson’s (1983) study also showed that the presence of difficult vocabulary does diminish comprehension, which leads to a decline in performance. On another note, Chall and Dale (1949) felt that a reading level was “The sum total of all those elements within a given piece of printed material that affect the success a group of readers have with it” (p. 23).

It was also noted that there are weaknesses to the current tests themselves. If the current tests are not assessing the curricular knowledge that students have acquired, then is our current testing system flawed? This is an area that should be further researched. NCEO (2008) asserts that states should be pushed to develop all encompassing assessments to measure all students' achievement toward on-grade level standards. Therefore, if we are currently using such a flawed system, are we portraying to the public a skewed vision of our students’ actual academic performance?

Recommendations

The results from this study suggest that our reliance on achievement tests should be questioned. Although these tests appear to be a central focus in today’s educational system, there is a great cost for students, teachers, and administrators. Understanding that achievement tests have both a positive and negative effect on student achievement could lead to further research in finding alternative ways to measure student achievement.

Future Research

This study involved 34 students from a rural community that was predominately Caucasian. Further research needs to address a larger sample size that involves a more diverse
population. A larger sample size will help to validate these findings and to justify the need for
different measures that could be taken when creating achievement tests. In addition,
consideration needs to be taken to prepare and administer tests that are free of bias. Examining
culturally responsive teaching practices could shed light on ways to assess students from various
eythnic, racial, socio-economic, and cultural backgrounds. Since students from diverse
populations did not participate in this study, it would be important to conduct future studies that
include a more inclusive student population.

Teachere

National and State achievement tests are used to ensure that teachers are providing
quality research proven instruction, and students are learning essential concepts and skills that
are developmentally appropriate. Teachers need to embrace the fact that not all students are on
the same reading level. The materials available within today’s classrooms need to be
differentiated to meet students’ individual needs. Many students may spend a majority of their
time decoding the reading materials they are assigned. Word length of all reading materials can
be a cause of reading difficulty for many students. If the students have a lower level of
vocabulary knowledge, they could become overwhelmed cognitively since they spend most of
their mental energy decoding words, thus reducing their ability to comprehend.

In many classrooms, teachers see comprehension as an area of concern for many
students. The findings from this study suggest that the readability of a text has a relationship with
test performance. A main factor related to readability that emerged from this study was the use of
specific science concepts and/or terms. These concepts and/or terms need to be identified and
explicitly taught to students prior to reading science texts and science-related materials. Thus, it
is important that teachers are cognizant of the specific content-specific terms and concepts that
are included in the district and/or state adopted materials. Teachers should also be aware of the numerous word recognition, vocabulary, and comprehension strategies that could assist students in their learning.

**Teacher Educators**

Teacher candidates who are being prepared for tomorrow’s classrooms have a heavy burden placed upon them. Today, teachers are being judged by society through the means of achievement tests. The No Child Left Behind Act that is currently in place mandates the administration of achievement tests as a way to measure student learning. In the present educational climate, the achievement tests are the main assessment measure used to assess students’ level of performance and achievement. The current educational and accountability system does not appear to be changing in the near future. Thus, it is essential that teacher educators continue to examine and share educational materials that represent a range of readability levels with teacher candidates. The teacher educator could then model and demonstrate how to use these varied materials within the classroom setting.

Science educators need to identify and teach students essential science concepts and terminology prior to reading science texts. Future educators should be using different reading strategies when delving into a science text, especially in the area of word recognition and vocabulary. In addition, teacher educators need to share with teacher candidates the current assessment system used in schools today may be flawed. As a result, teacher educators and teacher candidates can work together to examine and develop more appropriate classroom-based assessment systems.
Educational Policy Makers

It is important that educational policy makers understand that the current assessment system has both benefits and drawbacks. They must be ready to respond to critics who find this type of assessment system problematic. They should be researching alternative ways to measure student achievement and teacher accountability. One recommendation is to encourage educational policy makers at the local, state, and federal levels to fund large-scale evaluation studies that investigate student achievement across several different dimensions. These studies should include investigations of individual students’ academic progress and success across several different types of formal and informal assessment protocols.

The use of Value-Added Assessments could also supplement the “high stakes” assessments and give students, parents, teachers, administrators, and policymakers additional data to review, discuss, and use when making informed educational, instructional, and budgetary decisions. According to the Office of Curriculum and Instruction (2003), “The ODE is developing a 2% modified alternate assessment in reading and mathematics for grades 5-8, and it has obtained federal funding to assist in their development” (p. 31). The Office of Curriculum and Instruction adds, “ODE will pilot and field test modified grade-level assessments for certain students with disabilities during the 2008-2009 and 2009-2010 school years, and it plans to introduce the modified alternate assessments as fully operational during the 2010-2011 school year” (p. 31). The need for alternate assessment is only being done within the mathematics and reading areas. As shown by the results in this study, future research needs to be aimed toward science achievement tests that reflect all students’ full potential.
Summary

There is a vast amount of pressure that has been put upon the shoulders of students’ measuring their skills and overall abilities. This is in place to alert teachers, administrators, governmental officials, parents, and the general public about how our nation is doing in our schools. These scores are used to make large-scale decisions about our educational system. Throughout Ohio, students are given the Ohio 5th Grade Science Achievement Test in March every year. This test emphasizes on the noted specific content standards for the state of Ohio that focus on students’ specific skills.

The students’ performances highlighted in this study show that reading level does in fact have an effect on test performance. There is little current research that has been provided on the effect of readability, and state achievement tests. This is an area that is in great need of exploration for the sake of our students and teachers. It is achievement tests that are making the big decisions in thousands of lives every year, and these tests should be questioned.
REFERENCES


Exceptional Children, 53, 157-162.


APPENDIX A

REGULAR ACHIEVEMENT TEST
Student Name: ________________________________

Ohio Achievement Tests

Grade 5

Science

Student Test Booklet

Half-Length Practice Tests

Copyright © 2006 by Ohio Department of Education. All rights reserved.
Directions:

Today you will be taking the Ohio Grade 5 Science Practice Test. Three different types of questions appear on this test: multiple choice, short answer and extended response.

There are several important things to remember:

1. Read each question carefully. Think about what is being asked. Look carefully at graphs or diagrams because they will help you understand the question.

2. For short-answer and extended-response questions, use a pencil to write your answers neatly and clearly in the space provided in the answer document. Any answers you write in the Student Test Booklet will not be scored.

3. Short-answer questions are worth two points. Extended-response questions are worth four points. Point values are printed near each question in your Student Test Booklet. The amount of space provided for your answers is the same for two- and four-point questions.

4. For multiple-choice questions, shade in the circle next to your choice in the answer document for the test question. Mark only one choice for each question. Darken completely the circles on the answer document. If you change an answer, make sure that you erase your old answer completely.

5. Do not spend too much time on one question. Go on to the next question and return to the question skipped after answering the remaining questions.

6. Check over your work when you are finished.
Science

1. It rained early in the morning. A student sees a puddle of water on the sidewalk when she travels to school. The water in the puddle is gone when she travels home.

   What happened to the water in the puddle?

   A. It froze.
   B. It melted.
   C. It condensed.
   D. It evaporated.

2. Students have two blocks the same size. They drop each block into a beaker of water.

   Why does block 1 float and block 2 sink?

   A. Block 1 is a different material than block 2.
   B. Block 1 absorbs more light than block 2.
   C. Block 2 repels more water than block 1.
   D. Block 2 weighs less than block 1.
Use the picture to answer question 3.

Owl Butterfly

3. The owl butterfly has patterns on its wings that look like large eyes.

   How does this help the butterfly survive?

   A. It helps the butterfly fly faster.
   B. It helps the butterfly see better.
   C. It helps the butterfly scare enemies.
   D. It helps the butterfly absorb sunlight.

4. Butterflies get food from the flowers of a plant. They also lay their eggs on the leaves of the plant. As the caterpillars develop, they eat the leaves of the plant.

   How does the plant benefit from butterflies?

   A. Butterflies help the plant grow larger flowers.
   B. Butterflies’ eggs help the leaves to fall off the plant.
   C. Butterflies help pollinate flowers so that seeds can form.
   D. Butterflies help add nutrients to the nectar of the flowers.
5. Which diagram shows the flagpole at 5:00 p.m. on a summer day?

A. West
   North
   East
   South

B. West
   North
   East
   South

C. West
   North
   East
   South

D. West
   North
   East
   South

Go to next page
6. People wear hats when outside in the winter.

   How do hats help people stay warm?

   A. Hats stop thermal energy from leaving their heads.
   B. Hats slow down the thermal energy leaving their heads.
   C. Hats stop cold from entering their bodies through their heads.
   D. Hats slow down cold from entering their bodies through their heads.

7. A student asks, “Does the size of the wheels affect how far toy cars roll on the floor?”

   The student hypothesizes that toy cars with large wheels roll farther. The student wants to make sure that the force that starts the cars moving is always the same.

   In your Answer Document, describe or draw how the student can set up an investigation of his hypothesis.

   Then, describe or draw how the student can collect data to support his hypothesis. (2 points)
8. Students bump into each other when they turn the corner in the hallway shown. They plan to place a mirror in the hall so that they can see one another before reaching the corner.

Overhead View of Hallway

Where should they place the mirror?

A. position A  
B. position B  
C. position C  
D. position D
Science

Use the following picture to answer questions 9-11.

Canyon and River

9. Some students plan to learn more about the soil found near the river. Students collect soil samples from several spots.

Which tool would help them investigate the ability of the soil to hold water?

A. thermometer
B. rock hammer
C. magnifying glass
D. graduated cylinder
10. A geologist wrote many books on how rivers affect land. The geologist described detailed observations made over a long time.

Why do scientists record details about scientific observations?

A. to prove that scientists work hard
B. to make science books more interesting
C. to make people want to read about science
D. to provide evidence that supports conclusions

11. The picture shows evidence that different natural processes shape the canyon over time.

In your Answer Document, identify one natural process that could have helped shape the canyon in the picture and describe evidence of this process. (2 points)
Science

12. A student sets up an investigation with two identical plants. He uses containers and soil that are the same. On day 1, he adds fertilizer to plant B. Each plant gets 10 mL of water every day.

The student provides the pictures of the plants that he took on day 1 and day 10.

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant A</td>
<td><img src="image" alt="Plant A Day 1" /></td>
<td><img src="image" alt="Plant A Day 10" /></td>
</tr>
<tr>
<td>Plant B</td>
<td><img src="image" alt="Plant B Day 1" /></td>
<td><img src="image" alt="Plant B Day 10" /></td>
</tr>
</tbody>
</table>

Fertilizer

Which question does his investigation answer?

A. How does water affect plant growth?
B. How do soil types affect plant growth?
C. How does fertilizer affect plant growth?
D. How fast do different types of plants grow?
13. A student took three ice cubes from the freezer and put them in a glass of freshly squeezed orange juice. After 10 minutes, the student tried to take the ice cubes out of the juice. There was no ice left.

What type of change took place?

A. physical, because the ice cubes evaporated
B. physical, because the ice cubes changed into liquid
C. chemical, because the ice cubes’ energy became heat
D. chemical, because the ice cubes became a new substance

14. A copper wire with a plastic coating is placed near a compass, as shown in figure 1. When both ends of the wire are connected to a battery as shown in figure 2, the compass needle moves.

Why does the compass needle move?

A. Electricity flows from the wire to the compass.
B. Magnetic force flows from the battery to the wire.
C. Thermal energy flows through the wire to the compass.
D. Electricity flows through the wire, producing magnetic force.
15. Soil in an empty field blows away during a strong wind.
Which activity slows the erosion of this field over time?

A. watering the field
B. plowing the field in rows
C. planting grass in the field
D. building an electric fence

16. Measurements from a barometer help predict changes in weather.

What does a barometer measure?

A. humidity
B. air pressure
C. wind speed
D. temperature
17. Three identical blocks are pushed together. The starting temperature of each is shown.

\[
\begin{array}{ccc}
X & Y & Z \\
100^\circ C & 20^\circ C & 60^\circ C \\
\end{array}
\]

Which traces the transfer of thermal energy among the blocks?

A. X ← Y → Z  
B. X → Y → Z  
C. X → Y ← Z  
D. X ← Y ← Z

18. In many ways, Earth is like other planets in the solar system.

In which way is Earth different?

A. Earth has a moon.  
B. Earth orbits the sun.  
C. Earth has mountains.  
D. Earth has lots of water.
19. A class observed grasshoppers, frogs, mice, snakes, and owls in a grassy field. They are all part of the same food web.

Students combined their observations of what the organisms eat in the table.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Food Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>grasshopper</td>
<td>grass</td>
</tr>
<tr>
<td>mouse</td>
<td>grass, grasshoppers</td>
</tr>
<tr>
<td>frog</td>
<td>grasshoppers</td>
</tr>
<tr>
<td>snake</td>
<td>grasshoppers, mice, frogs</td>
</tr>
<tr>
<td>owl</td>
<td>grasshoppers, mice, frogs, snakes</td>
</tr>
</tbody>
</table>

In your Answer Document, draw a food web with four of these organisms.

When drawing the food web, be sure to use the names of the four organisms and draw arrows to trace the energy flow among the organisms. (4 points)
APPENDIX B

REVISED ACHIEVEMENT TEST
Revised test *

Student Name: ____________________________

Achievement Test

Science

Student Test Booklet

Note: This copy of the State of Ohio’s Ohio Achievement Test was revised as part of this study. The revisions made are italicized. Readability level and the specific Content Standards are noted parenthetically.
Directions:
Today you will be taking the 5th Grade Science Achievement Test. Three different types of questions appear on this test: multiple choice, short answer, and extended response.

There are several important things to remember:

1. Read each question carefully. Think about what is being asked. Look carefully at graphs or diagrams because they will help you understand the question.

2. For short-answer and extended-response questions, use a pencil to write your answers neatly and clearly in the space provided in the answer document. Any answers you write in the Student Test Booklet will not be scored.

3. Short-answer questions are worth two points. Extended-response questions are worth four points. Point values are printed near each question in your Student Test Booklet. The amount of space provided for your answers is the same for two- and four-point questions.

4. For multiple-choice questions, shade in the circle next to your choice in the answer document for the test question. Mark only one choice for each question. Darken completely the circles on the answer document. If you change an answer, make sure that you erase your old answer completely.

5. Do not spend too much time on one question. Go on to the next question and return to the question skipped after answering the remaining questions.

6. Check over your work when you are finished.
1. It rained early in the day. A student sees a pool of water on the sidewalk when she travels to school. The water in the pool is gone when she travels home. What happened to the pool of water?

A. It froze.
B. It melted.
C. It condensed.
D. It evaporated.

2. Students have 2 blocks the same size. They drop each block into a beaker of water.

Why does block 1 float and block 2 sink?

A. Block 1 is not the same material as block 2.
B. Block 1 takes in more light than block 2.
C. Block 2 repels more water than block 1.
D. Block 2 weighs less than block 1.
Use the picture to answer question 3.

8. The owl butterfly has patterns on its wings that look like large eyes.

How does this help the butterfly survive?

A. It helps it fly faster.
B. It helps it see better.
C. It helps it scare enemies.
D. It helps it absorb sunlight.

4. Butterflies get food from plants flowers. They lay eggs on the leaves of the plant. As the caterpillars change, they eat the leaves of the plant.

How does the plant get help from a butterfly?

A. Butterflies help the plant grow larger flowers.
B. Butterflies’ eggs help the leaves to fall off the plant.
C. Butterflies help pollinate flowers so that seeds can form.
D. Butterflies help add nutrients to the nectar of the flower.

Go to next page
Which shows a flag at 5:00 p.m. on a summer day?

A. North

B. West

C. South

D. East
People wear hats in the winter cold.

How do hats help keep people warm?

A. Hats stop thermal energy from leaving their heads.
B. Hats slow down the thermal energy from leaving.
C. Hats stop the cold from coming into their bodies.
D. Hats slow down the cold from entering their bodies.

A student asks, “Do the size of the tires change how far toy cars roll on the floor?” The student hypothesizes that toy cars with large tires roll farther. The student keeps the force that makes the cars move the same.

In your Answer Book, tell or draw how the student can set up an investigation of his hypothesis. Then, tell or draw how the student can collect data to support his hypothesis.

(2 points)
Students run into each other when they turn the corner in the hallway as shown. They plan to place a mirror in the hall so that they can see each other before reaching the corner.

Overhead View of Hallway

Where should they place the mirror?

A. Position A
B. Position B
C. Position C
D. Position D
Use the following picture to answer questions 9–11.

**Canyon and River**

- Plants
- Sandbar
- Rocks

---

Some students plan to learn more about the soil found near the river. Students collect soil from several spots.

Which tool would help them find the ability of the soil to hold water?

A. thermometer
B. rock hammer
C. magnifying glass
D. graduated cylinder

---

Go to next page
10. A geologist wrote many books on how rivers change land. The geologist gave detailed observations made over time.

   Why do scientists record details about observations?

   A. to prove they work hard
   B. to make science books more interesting
   C. to make people want to read about science
   D. to give support to conclusions

11. The picture above shows different natural processes that shape the canyon over time.

   In your Answer Sheet book, pick 1 process that could help shape the canyon in the picture and tell why. (2 points)
A student investigates two of the same plants. He uses pots and soil that are the same. On day 1, he adds fertilizer to plant B. Each plant gets 10 mL of water every day.

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant A</td>
<td><img src="plant_a_day1.png" alt="Image" /></td>
<td><img src="plant_a_day10.png" alt="Image" /></td>
</tr>
<tr>
<td>Plant B</td>
<td><img src="plant_b_day1.png" alt="Image" /></td>
<td><img src="plant_b_day10.png" alt="Image" /></td>
</tr>
</tbody>
</table>

These are the pictures he took on day 1 and day 10.

Which question does his investigation answer?

A. How does water change plant growth?
B. How do soil types change plant growth?
C. How does fertilizer change plant growth?
D. How fast do different plants grow?
A student took 8 ice cubes from the freezer and put them in a glass of fresh orange juice. After 10 minutes, the student tried to take the ice cubes out of the juice. There was no ice left.

What type of change took place?

A. physical, due to the ice cubes evaporated.
B. physical, due to the ice cubes changed into liquid.
C. chemical, due to the ice cubes’ energy became heat.
D. chemical, due to the ice cubes became a new substance.

A copper wire with a plastic coating is placed near a compass, as shown in figure 1. When both ends of the wire are put onto a battery as shown in figure 2, the compass needle moves.

Why does the compass needle move?

A. Electric flows from the wire to the compass.
B. Magnetic force flows from the battery to the wire.
C. Thermal energy flows through the wire to the compass.
D. Electric flows from the wire, producing magnetic force.
15. Soil in a clear field blows away during a strong wind.

Which item slows the erosion of the field over time?

A. water the field  
B. plow the field in rows  
C. plant grass in the field  
D. build an electric fence

16. A barometer helps predict changes in weather.

What does a barometer measure?

A. humidity  
B. air pressure  
C. wind speed  
D. temperature
17. Three blocks are pushed together. The starting temperature of each is shown.

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100°C</td>
<td>20°C</td>
<td>60°C</td>
</tr>
</tbody>
</table>

Which shows the transfer of thermal energy among the blocks?

A. \(X \leftrightarrow Y \rightarrow Z\)
B. \(X \rightarrow Y \rightarrow Z\)
C. \(X \rightarrow Y \leftrightarrow Z\)
D. \(X \leftrightarrow Y \leftrightarrow Z\)

18. Earth is like other planets in the solar system.

In which way is Earth different?

A. Earth has a moon.
B. Earth orbits the sun.
C. Earth has mountains.
D. Earth has lots of water.
A class looked at grasshoppers, frogs, mice, snakes, and owls in a grassy field. They are part of the same food web. Students combined their observations of what the organisms eat in the table.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Food Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>grasshopper</td>
<td>grass</td>
</tr>
<tr>
<td>mouse</td>
<td>grass, grasshoppers</td>
</tr>
<tr>
<td>frog</td>
<td>grasshoppers</td>
</tr>
<tr>
<td>snake</td>
<td>grasshoppers, mice, frogs</td>
</tr>
<tr>
<td>owl</td>
<td>grasshoppers, mice, frogs, snakes</td>
</tr>
</tbody>
</table>

In your Answer sheet, draw a food web with 4 of these organisms. When drawing the food web, be sure to use the names of the four organisms and draw arrows to show the energy flow among the organisms. (4 points)
References:

Ohio Achievement Tests
Grade 5
Science and Social Studies Half-Length Practice Tests
Administration Manual
Script

You are now going to take the Ohio Grade 5 Science Half-Length Practice Test. It is important that you do your best work on this test. Otherwise, it will not really show how well you can do in science. You can make sure your test scores give a true picture of what you know and what you can do by doing your best on this test. Remember to do your own work. You are not to copy or share work with anyone.

You have a book to read or some work that you can do at your desk in case you finish early. I will refer to this book or work as your silent work. Place your silent work where you can find it easily, but you are not to take it out at any time during the test. After you have completed the test and I have collected your test booklet and answer document, you can take out your book or other silent work.

Make sure that you have a pencil on your desk. If you need another pencil during the test, raise your hand and I will give you one. If you need to change an answer, make sure that you completely erase the answer you do not want.

Are there any questions?

Answer any questions.

I will hand out the Student Test Booklets and answer documents now. Do not open your Student Test Booklets or answer documents until I tell you to do so.

Make sure that each student receives a Student Test Booklet and answer document and has a pencil. Hold up a Student Test Booklet and answer document. Point to the place on the covers where students are to print their names.

Print your first and last name carefully on the line provided on the covers of the Student Test Booklet and the answer document.

Allow time for students to print their names on the covers.

Turn to page 1 in your test booklet and read along as I read the directions.

Directions:

Today you will be taking the Ohio Grade 5 Science Practice Test. Three different types of questions appear on this test: multiple choice, short answer and extended response.

There are several important things to remember:

1. Read each question carefully. Think about what is being asked. Look carefully at graphs or diagrams because they will help you understand the question.

2. For short-answer and extended-response questions, use a pencil to write your answers neatly and clearly in the space provided in the answer document. Any answers you write in the Student Test Booklet will not be scored.
3. Short-answer questions are worth two points. Extended-response questions are worth four points. Point values are printed near each question in your Student Test Booklet. The amount of space provided for your answers is the same for two- and four-point questions.

4. For multiple-choice questions, shade in the circle next to your choice in the answer document for the test question. Mark only one choice for each question. Darken completely the circles on the answer document. If you change an answer, make sure that you erase your old answer completely.

5. Do not spend too much time on one question. Go on to the next question and return to the question skipped after answering the remaining questions.

6. Check over your work when you are finished.

Are there any questions?

Pause for students’ questions.

We are now ready to begin. You may turn to page 2 of your Student Test Booklets and page 2 of your answer documents and start working.

Using the Test Results

Please review these test results with your students. The focus of this review should be on the format and types of test questions that appear on the practice test. This review will also help prepare them for the Grade 5 Science Achievement Test by alleviating their anxiety.
Grade 5 Science Half-Length Practice Test

Scoring Key

<table>
<thead>
<tr>
<th>Question</th>
<th>Scoring Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D. It evaporated</td>
</tr>
<tr>
<td>2</td>
<td>A. Block 1 is a different material than block 2.</td>
</tr>
<tr>
<td>3</td>
<td>C. It helps the butterfly scare enemies.</td>
</tr>
<tr>
<td>4</td>
<td>C. Butterflies help pollinate flowers so that seeds can form.</td>
</tr>
<tr>
<td>5</td>
<td>A.</td>
</tr>
</tbody>
</table>

6. B. Hats slow down the thermal energy leaving their heads.

7. See scoring guidelines on page 11.

8. B. position B

9. D. graduated cylinder

10. D. to provide evidence that supports conclusions

11. See scoring guidelines on page 15.

12. C. How does fertilizer affect plant growth?

13. B. physical, because the ice cubes changed into liquid

14. D. Electricity flows through the wire, producing magnetic force.

15. C. planting grass in the field

16. B. air pressure

17. C. $X \rightarrow Y \leftarrow Z$

18. D. Earth has lots of water.

19. See scoring guidelines on page 18.

Examples of correct responses for the constructed-response items are included in the scoring guidelines. These examples are illustrations of common responses and do not represent the full range of strategies and possible solutions for the question.
Scoring Guidelines and Sample Responses for Constructed-Response Questions

Question 7

7. A student asks, "Does the size of the wheels affect how far toy cars roll on the floor?"

The student hypothesizes that toy cars with large wheels roll farther. The student wants to make sure that the force that starts the cars moving is always the same.

In your Answer Document, describe or draw how the student can set up an investigation of his hypothesis.

Then, describe or draw how the student can collect data to support his hypothesis. (2 points)

Standard and Benchmark Assessed

Standard: Scientific Inquiry

Benchmark: Develop, design and safely conduct scientific investigations and communicate the results.

Rationale: This question asks students to show how to set up and collect data during an investigation designed to determine whether the size of the wheels affects how far a toy car will travel.

The response will receive full credit if the response draws or describes a way to set up an investigation so that the force starting the motion is constant AND draws or describes a method for collecting data to support the hypothesis.

☑ Scoring Guidelines

<table>
<thead>
<tr>
<th>Points</th>
<th>Student Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>The response draws or describes a way to set up this investigation so that the force starting the motion is constant AND draws or describes a method for collecting data to support the hypothesis.</td>
</tr>
<tr>
<td>1</td>
<td>The response draws or describes a way to set up this investigation so that the force starting the motion is constant; OR draws or describes a method for collecting data to support the hypothesis.</td>
</tr>
<tr>
<td>0</td>
<td>The response fails to demonstrate any understanding of how the student can test the hypothesis or collect data. The response does not meet the criteria required to earn 1 point. The response indicates inadequate or no understanding of the task and/or the idea or concept needed to answer the question.</td>
</tr>
</tbody>
</table>
SCORE POINT 2

Student Sample

7. A student asks, “Does the size of the wheels affect how far toy cars roll on the floor?”

The student hypothesizes that toy cars with large wheels roll farther. The student wants to make sure that the force that starts the cars moving is always the same.

In your Answer Document, describe or draw how the student can set up an investigation of his hypothesis.

Then, describe or draw how the student can collect data to support his hypothesis. (2 points)

Score Rationale

The response draws or describes a way to set up this investigation so that the force starting the motion is constant (“... set up a ramp with the car on top. At the bottom would be a track for the car to run on.”) AND draws or describes a method for collecting data to support the hypothesis (“... run the same car with different sized wheels to see how far it goes each time ... next to the track would be a tape measure to measure how far it will go.”) (also has a drawing to help show the ramp and track set-up to be sure cars get the same starting force and to show the location of the measuring tape for collecting data).
SCORE POINT 1

Student Sample

7. A student asks, "Does the size of the wheels affect how far toy cars roll on the floor?"

The student hypothesizes that toy cars with large wheels roll farther. The student wants to make sure that the force that starts the cars moving is always the same.

In your Answer Document, describe or draw how the student can set up an investigation of his hypothesis.

Then, describe or draw how the student can collect data to support his hypothesis. (2 points)

![Image of a table with columns labeled Size of wheels and Length traveled.]

Score Rationale

The response fails to draw or describe a way to set up this investigation so that the force starting the motion is constant (The response does not show evidence of a set-up) BUT describes a method for collecting data to support the hypothesis ("Take the same car with different size wheels and measure the length it goes.") (a chart was also included that shows how the data collected during an investigation would be recorded).
SCORE POINT 0

Student Sample

7. A student asks, “Does the size of the wheels affect how far toy cars roll on the floor?”

The student hypothesizes that toy cars with large wheels roll farther. The student wants to make sure that the force that starts the cars moving is always the same.

In your Answer Document, describe or draw how the student can set up an investigation of his hypothesis.

Then, describe or draw how the student can collect data to support his hypothesis. (2 points)

He could have other people vote then bring in some toy cars as an example. Then add up the data.

Score Rationale

The response fails to demonstrate any understanding of how the student can test the hypothesis or collect data. The response fails to draw or describe a way to set up the investigation so that the force starting the motion is constant (“He could have other people vote then bring in some toy cars as an example.”) AND fails to describe a method for collecting data to support the hypothesis (“Then add up the data.”) (does not describe the measurement of different distances).
Question 11

11. The picture shows evidence that different natural processes shape the canyon over time.

In your Answer Document, identify one natural process that could have helped shape the canyon in the picture and describe evidence of this process. (2 points)

Standard and Benchmark Assessed

Standard: Earth and Space Sciences

Benchmark: Summarize the processes that shape Earth’s surface and describe evidence of those processes.

Rationale: This question asks students to identify one natural process that changes Earth’s surface AND correctly describe evidence of that process from the picture of the canyon and the river.

The response will receive full credit if the response correctly identifies a natural process that helped shape the canyon shown and described evidence of that process from the picture.

Scoring Guidelines

<table>
<thead>
<tr>
<th>Points</th>
<th>Student Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>The response correctly identifies one natural process that could have helped shape the canyon AND correctly describes evidence of that process from the picture of the canyon and river.</td>
</tr>
<tr>
<td>1</td>
<td>The response only demonstrates partial understanding of how natural processes change Earth’s surface. The response correctly identifies one process but does not describe evidence from the picture related to that process; OR correctly describes evidence of a change but does not identify a process that could have caused the change; OR correctly identifies a natural process that can form canyons with a description of evidence of the process that is not specific to the canyon shown.</td>
</tr>
<tr>
<td>0</td>
<td>The response fails to demonstrate any understanding of how natural processes change Earth’s surface. The response does not meet the criteria required to earn 1 point. The response indicates inadequate or no understanding of the task and/or the idea or concept needed to answer the question.</td>
</tr>
</tbody>
</table>
SCORE POINT 2

Student Sample

11. The picture shows evidence that different natural processes shape the canyon over time.

In your Answer Document, identify one natural process that could have helped shape the canyon in the picture and describe evidence of this process. (2 points)

I think one natural process was a landslide because of the curved slope by the rocks.

The rock could have been in the rockside

Score Rationale

The response correctly identifies one natural process that could have helped shape the canyon (landslide) AND correctly describes evidence of that process from the picture of the canyon and river ("... because of the curved slope by the rocks.")).
SCORE POINT 1

Student Sample

11. The picture shows evidence that different natural processes shape the canyon over time.

   In your Answer Document, identify one natural process that could have helped shape the
canyon in the picture and describe evidence of this process. (2 points)

   I could have been wind if blew very
   hard.

Score Rationale

The response correctly identifies one natural process that could have helped shape the canyon ("I
could have been wind if blew very hard.") (wind erosion) BUT fails to correctly describe evidence of
that process from the picture of the canyon and river.

SCORE POINT 0

Student Sample

11. The picture shows evidence that different natural processes shape the canyon over time.

   In your Answer Document, identify one natural process that could have helped shape the
canyon in the picture and describe evidence of this process. (2 points)

   The rock help out a lot with the
   process became with the rock
   the water goes thought the rive;
   Smoother.

Score Rationale

The response does not demonstrate any understanding of how natural processes change Earth’s
surface. The response fails to correctly identify one natural process that could have helped shape the
canyon (rock) AND fails to correctly describe evidence of that process from the picture of the canyon
and river ("... became with the rock the water goes though the river smoother.").
Question 19

A class observed grasshoppers, frogs, mice, snakes, and owls in a grassy field. They are all part of the same food web. Students combined their observations of what the organisms eat in the table.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Food Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>grasshopper</td>
<td>grass</td>
</tr>
<tr>
<td>mouse</td>
<td>grass, grasshoppers</td>
</tr>
<tr>
<td>frog</td>
<td>grasshoppers</td>
</tr>
<tr>
<td>snake</td>
<td>grasshoppers, mice, frogs</td>
</tr>
<tr>
<td>owl</td>
<td>grasshoppers, mice, frogs, snakes</td>
</tr>
</tbody>
</table>

In your Answer Document, draw a food web with four of these organisms.

When drawing the food web, be sure to use the names of the four organisms and draw arrows to trace the energy flow among the organisms. (4 points)

Standard and Benchmark Assessed

Standard: Life Science

Benchmark: Analyze plant and animal structures and functions needed for survival and describe the flow of energy through a system that all organisms use to survive.

Rationale: This question asks students to trace and describe the flow of energy through a food web consisting of four organisms.

The response will receive full credit if the response represents a complete food web showing the interrelationships of four named organisms. The arrows' positions and directions trace the flow of energy through the food web.

☑ Scoring Guidelines

<table>
<thead>
<tr>
<th>Points</th>
<th>Student Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>The response provides a complete food web drawing, which includes four combinations of organisms. The food web uses the names of four organisms and shows all relationships among the organisms identified. The arrows are drawn in the correct direction to show the organization and flow of energy among the organisms.</td>
</tr>
</tbody>
</table>

(continued on following page)
<table>
<thead>
<tr>
<th>Points</th>
<th>Student Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>The response shows an essential understanding of the task by showing food web organism names for all four organisms, but it includes one category of mistake. The categories of mistakes include: • omission of arrows; • arrow(s) pointing in the wrong direction; • arrow(s) between two organisms that do not share a feeding relationship; • conflicting arrow (e.g., two-headed arrows, two arrows pointing in both directions between two organisms).</td>
</tr>
<tr>
<td>2</td>
<td>The response shows a partial understanding of the task by providing some evidence of understanding two food chains. The response names four organisms, but it includes an omission and one other category of mistake OR provides a perfectly correct food chain composed of more than two organisms.</td>
</tr>
<tr>
<td>1</td>
<td>The response omits significant aspects of the food web, showing a limited understanding of the task by showing one food chain in the web. The response provides a correct feeding relationship between two organisms (the feeding relationship does not have to begin with a producer) OR an incorrect food chain with three or more organisms.</td>
</tr>
<tr>
<td>0</td>
<td>The response fails to demonstrate any understanding of the feeding relationships among organisms. The response does not meet the criteria required to earn 1 point. The response indicates inadequate or no understanding of the task and/or the idea or concept needed to answer the item.</td>
</tr>
</tbody>
</table>
19. A class observed grasshoppers, frogs, mice, snakes, and owls in a grassy field. They are all part of the same food web. Students combined their observations of what the organisms eat in the table.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Food Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>grasshopper</td>
<td>grass</td>
</tr>
<tr>
<td>mouse</td>
<td>grass, grasshoppers</td>
</tr>
<tr>
<td>frog</td>
<td>grasshoppers</td>
</tr>
<tr>
<td>snake</td>
<td>grasshoppers, mice, frogs</td>
</tr>
<tr>
<td>owl</td>
<td>grasshoppers, mice, frogs, snakes</td>
</tr>
</tbody>
</table>

In your Answer Document, draw a food web with four of these organisms.

When drawing the food web, be sure to use the names of the four organisms and draw arrows to trace the energy flow among the organisms. (4 points)

Score Rationale

The response provides a complete food web drawing, which includes at least four combinations of organisms. The food web uses the names of at least four organisms, including a producer (grass), and shows all relationships among the organisms identified. The arrows are drawn in the correct direction to show the organization and flow of energy among the organisms.
19. A class observed grasshoppers, frogs, mice, snakes, and owls in a grassy field. They are all part of the same food web. Students combined their observations of what the organisms eat in the table.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Food Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>grasshopper</td>
<td>grass</td>
</tr>
<tr>
<td>mouse</td>
<td>grass, grasshoppers</td>
</tr>
<tr>
<td>frog</td>
<td>grasshoppers</td>
</tr>
<tr>
<td>snake</td>
<td>grasshoppers, mice, frogs</td>
</tr>
<tr>
<td>owl</td>
<td>grasshoppers, mice, frogs, snakes</td>
</tr>
</tbody>
</table>

In your Answer Document, draw a food web with four of these organisms.

When drawing the food web, be sure to use the names of the four organisms and draw arrows to trace the energy flow among the organisms. (4 points)

Score Rationale
The response provides a food web drawing, which includes at least four combinations of organisms. The food web uses the names of at least four organisms, including a producer (grass), and shows all relationships among the organisms identified. The response contains one category of mistake (the arrows are pointing in the wrong direction).
SCORE POINT 2
Student Sample

19. A class observed grasshoppers, frogs, mice, snakes, and owls in a grassy field. They are all part of the same food web. Students combined their observations of what the organisms eat in the table.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Food Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>grasshopper</td>
<td>grass</td>
</tr>
<tr>
<td>mouse</td>
<td>grass, grasshoppers</td>
</tr>
<tr>
<td>frog</td>
<td>grasshoppers</td>
</tr>
<tr>
<td>snake</td>
<td>grasshoppers, mice, frogs</td>
</tr>
<tr>
<td>owl</td>
<td>grasshoppers, mice, frogs, snakes</td>
</tr>
</tbody>
</table>

In your Answer Document, draw a food web with four of these organisms.

When drawing the food web, be sure to use the names of the four organisms and draw arrows to trace the energy flow among the organisms. (4 points)

Grass→Grasshopper→Mouse→Snake→Owl.

Score Rationale

The response demonstrates a partial understanding of the task by providing some evidence of understanding a food chain. The response provides a producer (grass) but fails to provide an interrelationship required for it to be a food web. The response provides a correct drawing of a food chain with no other category of mistake.
19. A class observed grasshoppers, frogs, mice, snakes, and owls in a grassy field. They are all part of the same food web. Students combined their observations of what the organisms eat in the table.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Food Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>grasshopper</td>
<td>grass</td>
</tr>
<tr>
<td>mouse</td>
<td>grass, grasshoppers</td>
</tr>
<tr>
<td>frog</td>
<td>grasshoppers</td>
</tr>
<tr>
<td>snake</td>
<td>grasshoppers, mice, frogs</td>
</tr>
<tr>
<td>owl</td>
<td>grasshoppers, mice, frogs, snakes</td>
</tr>
</tbody>
</table>

In your Answer Document, draw a food web with four of these organisms.

When drawing the food web, be sure to use the names of the four organisms and draw arrows to trace the energy flow among the organisms. (4 points)

\[ \text{grasshopper} \rightarrow \text{mouse} \rightarrow \text{frog} \rightarrow \text{snake} \rightarrow \text{owl} \]

Score Rationale

The response demonstrates a partial understanding of the task by providing some evidence of understanding a food chain. The response provides a correct drawing of a food chain but includes one other category of mistake (wrong interrelationship; mouse → frog).
19. A class observed grasshoppers, frogs, mice, snakes, and owls in a grassy field. They are all part of the same food web. Students combined their observations of what the organisms eat in the table.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Food Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>grasshopper</td>
<td>grass</td>
</tr>
<tr>
<td>mouse</td>
<td>grass, grasshoppers</td>
</tr>
<tr>
<td>frog</td>
<td>grasshoppers</td>
</tr>
<tr>
<td>snake</td>
<td>grasshoppers, mice, frogs</td>
</tr>
<tr>
<td>owl</td>
<td>grasshoppers, mice, frogs, snakes</td>
</tr>
</tbody>
</table>

In your Answer Document, draw a food web with four of these organisms.

When drawing the food web, be sure to use the names of the four organisms and draw arrows to trace the energy flow among the organisms. (4 points)

owl → snake → frog → mouse → grasshopper

Score Rationale

The response fails to demonstrate any understanding of the feeding relationship among organisms. The response fails to provide a drawing of the food web. A worded response is not given credit.
APPENDIX D

STUDENT ANSWER DOCUMENT
Student Name: ________________________________

Ohio Achievement Tests

Grade 5

Science

Answer Document
Half-Length Practice Tests

Copyright © 2006 by Ohio Department of Education. All rights reserved.
1. A B C D
2. A B C D
3. A B C D
4. A B C D
5. A B C D
6. A B C D

7. Write your response to question 7 in the space below.
8. A B C C
9. A B C C
10. A B C C

11. Write your response to question 11 in the space below.
Science  Use Pencil Please

12.  A  B  C  D
13.  A  B  C  D
14.  A  B  C  D
15.  A  B  C  D
16.  A  B  C  D
17.  A  B  C  D
18.  A  B  C  D
19. Write your response to question 19 in the space below.
APPENDIX E

ADMINISTRATOR LETTER
January 2009

Dear Administrator,

Hello, my name is Zachary S. Amos, and I am a student at Bowling Green State University in the Graduate Reading Program. I became interested in literacy and reading education because I was coming in contact with elementary schools students who were moving into the middle grades with inadequate reading abilities. I often heard students complaining about reading the words and trying to make meaning of the sentences in the text. This led me to be concerned about how the students would do if they were tested in other areas aside from reading. “Would they be able to understand questions in the Achievement tests that they were required to take?” I am currently doing a research study on the 5th Grade Science Achievement Test. The purpose of my study is to find out if there is a difference between test performance and readability. I want to see if the current test is valid in assessing a students’ knowledge of Science concepts. If there is a valid difference in data, this study may act as a catalyst toward further research, which could impact students, teachers, and administrators substantially for the better.

In order to give this study I would need the help of your students. I would like to give half of a past 5th Grade Science Achievement Test to your class. The test is voluntary and exactly 19 questions in length ranging from multiple choice to short answer questions. Your child would have approximately one hour and fifteen minutes to complete the Achievement test during class time. The test would be followed two weeks later with a revised 5th Grade Science Achievement test, but with the newly revised test the reading level would be at a 3.4 grade level compared to the 4.8 grade from the first one. Basically, the Science concepts will be exactly the same, but the wording will be less difficult for your child. The second means of showing Achievement’s procedure will be exactly the same as the first. Afterwards, I will be comparing the scores from both of the tests. The only people that will be viewing the data from the tests would be myself, Dr. Cynthia Bertelson, and ________.

Overall, the risk is nothing greater than what your students would experience in everyday life. Prior to giving each of the tests your child will create a pseudoname/ fake name to put on the test. The tests will also be under lock and key in my ________________ at Bowling Green State University. Furthermore, your child will have the opportunity to withdraw from this study at any time. If any student withdraws from the study, no information from the test will be used nor will this impact the relationship for the school. Participating in the study will not impact grades/ class standing/ relationship to the institution. The benefit to participating in this study is the opportunity to help identify a problem that could be potentially causing harm to many schools, teachers, administrators, and students across Ohio.

If there are any questions or concerns in regards to this research study, please feel free to contact me, Zachary S. Amos at __________________. You may contact my Bowling Green State University advisor, Cynthia Bertelson, __________________. You can also contact the Chair, Human Subjects Review Board, Bowling Green State University at 419-372-7716 or by e-mail at hsrb@bgsu.edu, with questions or concerns about participant rights.

Thank you for your consideration,
Zachary S. Amos

Graduate Student

I, ________________________________ grant Zachary S. Amos permission to conduct his study with my students and to use information gathered for his study.
APPENDIX F

TEACHER LETTER
January 2009

Dear Teacher,

Hello, my name is Zachary S. Amos, and I am a student at Bowling Green State University in the Graduate Reading Program. I became interested in literacy and reading education because I was coming in contact with elementary schools students who were moving into the middle grades with inadequate reading abilities. I often heard students complaining about reading the words and trying to make meaning of the sentences in the text. This led me to be concerned about how the students would do if they were tested in other areas aside from reading. “Would they be able to understand questions in the Achievement tests that they were required to take?” I am currently doing a research study on the 5th Grade Science Achievement Test. The purpose of my study is to find out if there is a difference between test performance and readability. I want to see if the current test is valid in assessing a students’ knowledge of Science concepts.

In order to give this study I would need the help of your students. I would like to give half of a past 5th Grade Science Achievement Test to your class. The test is voluntary and exactly 19 questions in length ranging from multiple choice to short answer questions. Your child would have approximately one hour and fifteen minutes to complete the Achievement test during class time. The test would be followed two weeks later with a revised 5th Grade Science Achievement test, but with the newly revised test the reading level would be at a 3.4 grade level compared to the 4.8 grade from the first one. Basically, the Science concepts will be exactly the same, but the wording will be less difficult for your child. The second means of showing Achievement’s procedure will be exactly the same as the first. Afterwards, I will be comparing the scores from both of the tests. The only people that will be viewing the information from the tests would be myself, Dr. Cynthia Bertelson, and _______. Overall, the risk is nothing greater than what your students would experience in everyday life. Prior to giving each of the tests your child will create a pseudoname/ fake name to put on the test. The tests will also be under lock and key in the _______ at Bowling Green State University. Furthermore, your child will have the opportunity to withdraw from this study at any time. If any student withdraws from the study, no information from the test will be used nor will this impact the relationship for the school. Participating in the study will not impact grades/ class standing/ relationship to the institution. The benefit to participating in this study is the opportunity to help identify a problem that could be potentially causing harm to many schools, teachers, administrators, and students across Ohio.

If there are any questions or concerns in regards to this research study, please feel free to contact me, Zachary S. Amos at ___________. You may contact my Bowling Green State University advisor, Cynthia Bertelson, at ___________. You can also contact the Chair, Human Subjects Review Board, Bowling Green State University at 419-372-7716 or by e-mail at hsrb@bgsu.edu, with questions or concerns about participant rights.

Thank you for your consideration,

Zachary S. Amos
Graduate Student

I, __________________________ grant Zachary S. Amos permission to conduct his study with my child and to use information gathered for his study.
APPENDIX G

PARENT LETTER
January 2009
Dear Parent/Guardian,
Hello, my name is Zachary S. Amos, and I am a student at Bowling Green State University in the Graduate Reading Program. I became interested in literacy and reading education because I was coming in contact with elementary schools students who were moving into the middle grades with inadequate reading abilities. I often heard students complaining about reading the words and trying to make meaning of the sentences in the text. This led me to be concerned about how the students would do if they were tested in other areas aside from reading. “Would they be able to understand questions in the Achievement tests that they were required to take?” I am currently doing a research study on the 5th Grade Science Achievement Test. The purpose of my study is to find out if there is a difference between test performance and readability. I want to see if the current test is valid in assessing a student’s knowledge of Science concepts.
In order to give this study I would need the help of your child. I would like to give half of a past 5th Grade Science Achievement Test to your child. The test is voluntary and exactly 19 questions in length ranging from multiple choice to short answer questions. Your child would have approximately one hour and fifteen minutes to complete the Achievement test during class time. The test would be followed two weeks later with a revised 5th Grade Science Achievement test, but with the newly revised test the reading level would be at a 3.4 grade level compared to the 4.8 grade from the first one. Basically, the Science concepts will be exactly the same, but the wording will be less difficult for your child. The second means of showing Achievement’s procedure will be exactly the same as the first. Afterwards, I will be comparing the scores from both of the tests. The only people that will be viewing the information from the tests would be myself, Dr. Cynthia Bertelson, and ______.
There will be no risk in your child’s identity in giving and collecting the data. Prior to giving each of the tests your child will create a pseudonym/ fake name to put on the test. The tests will also be under lock and key in __________________ at Bowling Green State University. Furthermore, your child will have the opportunity to withdraw from this study at any time. If your child withdraws from the study, no information from the test will be used nor will this impact the relationship for the school. Participating in the study will not impact grades/ class standing/ relationship to the institution. The benefit to participating in this study is the opportunity to help identify a problem that could be potentially causing harm to many schools, teachers, administrators, and students across Ohio. If there are any question or concerns in regards to this research study, please feel free to contact me, Zachary S. Amos at ______. You may contact my Bowling Green State University advisor, ______. You can also contact the Chair, Human Subjects Review Board, Bowling Green State University at 419-372-7716 or by e-mail at hsrb@bgsu.edu, with questions or concerns about participant rights.
Thank you for your consideration,

Zachary S. Amos
Graduate Student

I, ____________________________ grant Zachary S. Amos permission to conduct his study with my child and to use information gathered for his study.
APPENDIX H

STUDENT LETTER
January 2009

Dear Student,

My name is Zachary Amos and I am a graduate student from Bowling Green State University. I would like to do a research study on you that could help many students, teachers, and administrators. What I would be asking of you is that you take an Achievement test for me two different times. The test is completely voluntary. The test would also be good practice for you when taking your test in the Spring. The test itself will be timed and you will have an hour to complete each one. It will not be graded at all, and will not change anything for you in ____ class. If you feel that it is causing you discomfort you can stop at anytime. It will not cause you anymore stress or risk than everyday life. If you would be willing to help please sign below.

Sincerely,
Zachary S. Amos

Graduate Student

I ______________________ am willing to be a part of the research study.
APPENDIX I

SUMMARY STATISTICS TABLE
### Summary Statistics Table

<table>
<thead>
<tr>
<th>Column</th>
<th>n</th>
<th>Mean</th>
<th>Variance</th>
<th>Std. Dev.</th>
<th>Std. Err.</th>
<th>Median</th>
<th>Range</th>
<th>Min</th>
<th>Max</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>34</td>
<td>1</td>
<td>0</td>
<td>0.23883258</td>
<td>0.040959448</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1B</td>
<td>34</td>
<td>0.9411765</td>
<td>0.057040997</td>
<td>0.084595144</td>
<td>0.040959448</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2A</td>
<td>34</td>
<td>0.38235295</td>
<td>0.2433155</td>
<td>0.49327022</td>
<td>0.084595144</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2B</td>
<td>34</td>
<td>0.5294118</td>
<td>0.25668448</td>
<td>0.5066404</td>
<td>0.08688811</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3A</td>
<td>34</td>
<td>0.9411765</td>
<td>0.057040997</td>
<td>0.23883258</td>
<td>0.040959448</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3B</td>
<td>34</td>
<td>0.85294116</td>
<td>0.35949063</td>
<td>0.061652135</td>
<td>0.08318903</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4A</td>
<td>34</td>
<td>0.64705884</td>
<td>0.23529412</td>
<td>0.48507124</td>
<td>0.08318903</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4B</td>
<td>34</td>
<td>0.7058824</td>
<td>0.21390374</td>
<td>0.4624973</td>
<td>0.07931763</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5A</td>
<td>34</td>
<td>0.5294118</td>
<td>0.25668448</td>
<td>0.5066404</td>
<td>0.08688811</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5B</td>
<td>34</td>
<td>0.64705884</td>
<td>0.23529412</td>
<td>0.48507124</td>
<td>0.08318903</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6A</td>
<td>34</td>
<td>0.11764706</td>
<td>0.32703498</td>
<td>0.56086034</td>
<td>0.070387416</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6B</td>
<td>34</td>
<td>0.20588236</td>
<td>0.41042563</td>
<td>0.070387416</td>
<td>0.070387416</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7A</td>
<td>34</td>
<td>0.9705882</td>
<td>0.6269355</td>
<td>0.107518554</td>
<td>0.070387416</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7B</td>
<td>34</td>
<td>1.117647</td>
<td>0.6402985</td>
<td>0.109810285</td>
<td>0.070387416</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8A</td>
<td>34</td>
<td>0.9117647</td>
<td>0.28790224</td>
<td>0.49374826</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8B</td>
<td>34</td>
<td>0.9705882</td>
<td>0.17149858</td>
<td>0.029411765</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9A</td>
<td>34</td>
<td>0.5588235</td>
<td>0.50399476</td>
<td>0.08643439</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9B</td>
<td>34</td>
<td>0.61764705</td>
<td>0.49327022</td>
<td>0.084595144</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10A</td>
<td>34</td>
<td>0.7941176</td>
<td>0.41042563</td>
<td>0.070387416</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10B</td>
<td>34</td>
<td>0.7352941</td>
<td>0.44781107</td>
<td>0.07679897</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>11A</td>
<td>34</td>
<td>0.7058824</td>
<td>0.67552054</td>
<td>0.11585081</td>
<td>0.070387416</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>11B</td>
<td>34</td>
<td>0.9705882</td>
<td>0.7581994</td>
<td>0.13003013</td>
<td>0.070387416</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>12A</td>
<td>34</td>
<td>0.7941176</td>
<td>0.41042563</td>
<td>0.070387416</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12B</td>
<td>34</td>
<td>0.85294116</td>
<td>0.35949063</td>
<td>0.061652135</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>13A</td>
<td>34</td>
<td>0.64705884</td>
<td>0.48507124</td>
<td>0.08318903</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>13B</td>
<td>34</td>
<td>0.64705884</td>
<td>0.48507124</td>
<td>0.08318903</td>
<td>0.070387416</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14A</td>
<td>34</td>
<td>0.64705884</td>
<td>0.23529412</td>
<td>0.48507124</td>
<td>0.08318903</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14B</td>
<td>34</td>
<td>0.38235295</td>
<td>0.2433155</td>
<td>0.49327022</td>
<td>0.084595144</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15A</td>
<td>34</td>
<td>0.3529412</td>
<td>0.23529412</td>
<td>0.48507124</td>
<td>0.08318903</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15B</td>
<td>34</td>
<td>0.2647059</td>
<td>0.20053476</td>
<td>0.44781107</td>
<td>0.07679897</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16A</td>
<td>34</td>
<td>0.44117647</td>
<td>0.2540107</td>
<td>0.50399476</td>
<td>0.08643439</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16B</td>
<td>34</td>
<td>0.3529412</td>
<td>0.23529412</td>
<td>0.48507124</td>
<td>0.08318903</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>17A</td>
<td>34</td>
<td>0.32352942</td>
<td>0.2254902</td>
<td>0.47485808</td>
<td>0.08143749</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>17B</td>
<td>34</td>
<td>0.23529412</td>
<td>0.18538325</td>
<td>0.43056154</td>
<td>0.07384069</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>18A</td>
<td>34</td>
<td>0.8235294</td>
<td>0.14973262</td>
<td>0.386953</td>
<td>0.06636189</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18B</td>
<td>34</td>
<td>0.85294116</td>
<td>0.12923351</td>
<td>0.35949063</td>
<td>0.061652135</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>19A</td>
<td>34</td>
<td>2.0588236</td>
<td>0.8449198</td>
<td>0.9191952</td>
<td>0.15764068</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>19B</td>
<td>34</td>
<td>2.9411764</td>
<td>1.2691622</td>
<td>1.1265709</td>
<td>0.19320533</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>SUM A</td>
<td>34</td>
<td>13.6470585</td>
<td>12.417112</td>
<td>3.5237923</td>
<td>0.6043254</td>
<td>14</td>
<td>17</td>
<td>3</td>
<td>20</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>SUM B</td>
<td>34</td>
<td>14.823529</td>
<td>15.301248</td>
<td>3.911681</td>
<td>0.6708478</td>
<td>15</td>
<td>18</td>
<td>6</td>
<td>24</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>
APPENDIX J

EVALUATING QUESTION READABILITY, T-STAT, AND P-VALUE FOR TEST A AND TEST B
## Evaluation Question Readability, T-Stat, and P-Value for Test A & Test B

<table>
<thead>
<tr>
<th>Question</th>
<th>Original Readability Score (Test A)</th>
<th>Revised/New Readability Score (Test B)</th>
<th>t-Stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.4</td>
<td>1.0</td>
<td>1.436141</td>
<td>0.1604</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>1.4</td>
<td>-1.71346</td>
<td>0.096</td>
</tr>
<tr>
<td>3</td>
<td>3.9</td>
<td>2.0</td>
<td>1.78705</td>
<td>0.0831</td>
</tr>
<tr>
<td>4</td>
<td>2.8</td>
<td>2.4</td>
<td>-0.49441</td>
<td>0.6243</td>
</tr>
<tr>
<td>5</td>
<td>8.5</td>
<td>6.7</td>
<td>-0.94123</td>
<td>0.3534</td>
</tr>
<tr>
<td>6</td>
<td>3.2</td>
<td>2.9</td>
<td>-1.0</td>
<td>0.3246</td>
</tr>
<tr>
<td>7</td>
<td>5.3</td>
<td>5.3</td>
<td>-1.71346</td>
<td>0.096</td>
</tr>
<tr>
<td>8</td>
<td>5.3</td>
<td>4.7</td>
<td>-1.0</td>
<td>0.3246</td>
</tr>
<tr>
<td>9</td>
<td>5.8</td>
<td>4.7</td>
<td>-0.62678</td>
<td>0.5351</td>
</tr>
<tr>
<td>10</td>
<td>8.5</td>
<td>7.2</td>
<td>0.70181</td>
<td>0.4877</td>
</tr>
<tr>
<td>11</td>
<td>5.0</td>
<td>2.3</td>
<td>-2.72111</td>
<td>0.0103</td>
</tr>
<tr>
<td>12</td>
<td>2.8</td>
<td>2.6</td>
<td>-0.8124</td>
<td>0.4224</td>
</tr>
<tr>
<td>13</td>
<td>3.6</td>
<td>2.6</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>14</td>
<td>4.7</td>
<td>3.8</td>
<td>3.020408</td>
<td>0.0048</td>
</tr>
<tr>
<td>15</td>
<td>5.3</td>
<td>4.7</td>
<td>1.0</td>
<td>0.3246</td>
</tr>
<tr>
<td>16</td>
<td>10.9</td>
<td>9.6</td>
<td>0.769944</td>
<td>0.4468</td>
</tr>
<tr>
<td>17</td>
<td>5.8</td>
<td>4.2</td>
<td>0.902052</td>
<td>0.3736</td>
</tr>
<tr>
<td>18</td>
<td>1.4</td>
<td>1.1</td>
<td>-0.44189</td>
<td>0.6615</td>
</tr>
<tr>
<td>19</td>
<td>3.5</td>
<td>2.9</td>
<td>-4.82451</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>