COMPARING THE PERFORMANCE AND PREFERENCE OF STUDENTS EXPERIENCING A READING ALOUD ACCOMMODATION TO THOSE WHO DO NOT ON A VIRTUAL SCIENCE ASSESSMENT

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF EDUCATION

---

by
Angela Shelton
May 2012

Examining Committee

Diane Jass Ketelhut, Advisory Chair, University of Maryland, College of Education

Kristie Jones Newton, CITE Department

Joseph DuCette, Department of Psychological Studies in Education

John L. Pecore, CITE Department

Catherine C. Schifter, Department of Psychological Studies in Education

ABSTRACT

Many United States secondary students perform poorly on standardized summative science assessments. Situated Assessments using Virtual Environments (SAVE) Science is an innovative assessment project that seeks to capture students' science knowledge and understanding by contextualizing problems in a game-based virtual environment called *Scientopolis*. Within *Scientopolis*, students use an "avatar" to interact with non-player characters (NPCs), artifacts, embedded clues and "sci-tools" in order to help solve the problems of the townspeople. In an attempt to increase students' success on assessments, SAVE science places students in an environment where they can use their inquiry skills to solve problems instead of reading long passages which attempt to contextualize questions but ultimately cause construct-irrelevant variance. However, within these assessments reading is still required to access the test questions and character interactions. This dissertation explores how students' in-world performances differ when exposed to a Reading Aloud Accommodation (RAA) treatment in comparison to a control group. Student perceptions of the treatment are also evaluated. While a RAA is typically available for students with learning disabilities or English language learners, within this study, all students were randomly assigned to either the treatment or control, regardless of any demographic factors or learning barriers. The theories of Universal design for learning and brain-based learning advocate for multiple ways for students to engage, comprehend, and illustrate their content knowledge. Further, through providing more ways for students to interact with content, all students should benefit, not just those with learning disabilities. Students in the experimental group listened to the NPCs speak the dialogue that provides them with the problem, clues, and assessment questions, instead of relying on reading skills to gather the information. Overall, students in the treatment group statistically outperformed those in the control. Student perceptions of using the reading aloud accommodation were generally positive. Ideas for future research are presented to investigate the accommodation further.

ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family, below are my sentiments of appreciation to you.

I would like to express my deepest gratitude to my advisor, Dr. Diane Jass Ketelhut, for her excellent guidance, caring, and patience through this arduous process. You will forever be my research mom.

I would be remiss, not to also thank the rest of my dissertation committee and external reviewers who provided me with great feedback along the way and interesting ideas for future research.

Furthermore, I owe a sincere thank you to all of the researchers and teachers who are part of the SAVE Science team. Thank you for all of your time and help in making this research possible.

I would also like to thank my dad and my brothers. You were always supporting and encouraging me that I had it in me to finish this piece of work. To Arryn, you have always been more than just a sister to me. Furthermore, I would like to thank my grandmom, granddad, and Uncle Steve for being role models of the person I hope to be, once I finally grow up. To ALL of my Shelton family, you have provided me with all the love I needed to keep me smiling and laughing, which really is all a girl needs.

Finally, I would like to thank Mike and Dr. Walton for helping me push through my frustrations and make it to the other side of academia, with my sanity intact.

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Testing for Accountability

In 2001, the United States government passed the No Child Left Behind Act (NCLB), which provides guidelines for statewide measures of academic success and minimum levels of performance. Systems of high stakes assessments were developed to ensure that students were indeed learning and teachers were effectively educating; however, originally students with learning disabilities were often not tested or included in the aggregate samples (Mazzeo, Carlson, Voekl, & Lutkus, 2000). The Individuals with Disabilities Educational Act [IDEA] (1997, 2004) prohibits schools from excluding populations of exceptional students- English language learners or those with physical or learning disabilities.

The underlying intent of NCLB is to provide all students with effective schools and propel them to a minimum level of academic success. Government officials designed this law as an attempt to eliminate the achievement gap that is seen consistently between white and minority students, as well as between students of differing socioeconomic statuses (Brown & Hunter, 2006; Packer, 2007). Students from low socioeconomic statuses and minority subgroups historically score lower in math, reading, writing, and science; a disparity that the United States federal government is attempting to eradicate through NCLB (United States Department of Education, 2006). Because students are presented with identical questions to answer, scores can be assigned objectively to the student, compared easily, and used as the primary indicators for successes in student learning outcomes (Chapman & Snyder, 2000). Further, these measurements can help the federal and state governments locate underperforming schools, investigate the source of the problem, and ensure these schools have the resources to remedy the situation. Both NCLB and IDEA have established polices that stimulated the creation of more accurate assessments via better questions or instituting accommodations to obtain a more valid measurement of student comprehension and abilities (Lehr & Thurlow, 2003).

Students are tested to determine their level of content proficiency as a direct measure of their educational success; however, generally, high stakes assessments are norm referenced for average learners, excluding students with learning disabilities and English language learners, which can therefore transform these tests into language proficiency assessments, rather than content specific ones (Abedi, 2002). Thus, policy makers, teachers, and parents may not be getting an accurate account of the true understanding of exceptional students, especially in science. Because science is based around inquiry, it is more difficult to assess knowledge and skills in narrow, multiple-choice questions. Frequently, lengthy passages are required to contextualize a problem or question in science. In order for students to express correctly their content knowledge and understanding on assessments, often they need skills outside of the tested construct, such as attention capabilities and reading abilities (Ketterlin-Geller, 2005). This chapter will focus on the issue of unintentionally tested constructs creating invalid score variations that disadvantage students with reading difficulties, the current prescriptive measures to fix the problem, and a proposed alternate solution.

Construct Irrelevant Variance

When an assessment measures skills and knowledge outside of a tested construct, it is referred to as construct irrelevant variance (Messick, 1989). Assessments that require reading as a prerequisite to demonstrating science knowledge have construct irrelevancy, because reading skills can act as a barrier to the demonstration of science understanding. According to Dolan, Hall, Banerjee, Chun, and Strangman (2005), this can exacerbate the difficulty of questions for students with reading difficulties including trouble with comprehension, fluency, phonemic awareness, word recognition, and language proficiency. Further, students with learning disabilities, defined as significant impediments to the acquisition or display of reading, writing, mathematical, speaking or listening skills (National Joint Committee on Learning Disabilities, 1998) have documented obstacles to learning and assessments that can be intensified with unintentionally tested constructs. Currently, learning-disabled students receive accommodations or modifications during testing, which are designed to increase the validity of the assessments for them, by alleviating these obstacles of construct-irrelevance (Crawford, 2007; Sireci, Scarpati, & Li, 2005; Elliot, Kratochwill, & Schulte, 1999).

Testing Accommodations

Because reading comprehension difficulty is the most commonly associated problem with learning disabilities (Gersten, Fuchs, Williams, & Baker, 2001), the most frequently used testing accommodation is having the test read aloud (Dolan & Hall, 2001) which helps students with difficulty in processing textual information, e.g., dyslexia. When the reading aloud of tests is specified as an accommodation for a student with a learning disability via their Individualized Educational Program (IEP), federal law (see NCLB, 2001 and IDEA, 2004) mandates that this accommodation be offered. Therefore, a teacher or aide must read every question on the assessment aloud to the student as a reading aloud accommodation (RAA); however, there are no guidelines for its administration. Furthermore, if there are multiple students taking identical assessments, personnel will often read it aloud to all the learning-disabled students simultaneously. While this makes the job easier and more economical, this process can negatively affect the student's performance. According to Dolan et al. (2005), if students do not understand the question or hear it properly, often they will not ask for a repeat of the question due to concerns of peer pressure or their public image. Therefore, while it appears that these students are receiving an accommodation, it may not be helping them access and express their content understanding knowledge. Moreover, teachers and administrators see that these students are receiving an accommodation per their IEP, which could potentially cause a misconception that the playing field is even, and that lack of understanding is the issue, not the learning disability. While testing accommodations are designed to provide students with learning disabilities a more equitable opportunity to express their understanding (Watts-Driscoll, 2007), examples of poorly administered accommodations illustrate that the adjustment may not be succeeding. Thus, designing the evaluations for maximum access of all students might be a better way to ensure equity.

Universal Design

Rather than trying to retrofit tests with accommodations (e.g. reading aloud of tests) for students with learning disabilities, Universal Design, a set of principles focusing on removing barriers of access from inception (Preiser & Ostroff, 2001) is now being applied to assessment. Universal design is an idea that started in architecture to design buildings for the ease of all users (Mace, 1998). Further, Mace, the creator of Universal Design, suggested that planning to allow access for all types of people from the start would benefit everyone, as opposed to the prior approach of retrofitting the design, which only benefits people with disabilities. For example, Dolan and Hall (2001) and Burgstahler (2007) both explain that curb cuts initially designed for people in wheelchairs actually help parents with strollers, people with walkers or canes, and an everyday walker who is just looking for easier access. Rather than having a machine that only lifts wheelchairs onto the curb, this design strategy removes a potential barrier for some and makes a building more accessible for all.

Even though it took decades, Universal Design has made its way into the field of education. According to the Center for Applied Special Technology (CAST; 2011), Universal Design for Learning allows for flexibility in how information is presented to students, how students participate in the learning process, and how they express their understanding and acquisition of skills and knowledge (Rose & Meyer, 2002). Moreover, removing or decreasing obstacles to learning in order to sustain high expectations for all students including those with disabilities or English language learners is a major tenet of the Universal Design for Learning theory. Instead of retrofitting lessons or tests via accommodations only for students with learning disabilities, all students have access to a variety of ways to express their knowledge and skills.

Primarily, the focus of Universal Design for Learning is to meet the diversity of needs of all learners, as really no two students learn or express knowledge in the exact same way. According to Dolan and Hall (2001), a learning situation designed via Universal Design for Learning affords "multiple means of recognition, expression, and engagement" (p.23). Multiple means of recognition includes the variety of ways of bringing the information to students, for example: audio support, printed or digital text, computer graphics, multimedia images, animations, and virtual worlds (Dolan & Hall, 2001). Multiple means of expression include the different ways students are allowed to demonstrate their content knowledge and skills such as typing, depicting, handwriting, recorded speech, or class presentations (Rose & Gravel, 2010). Multiple means of engagement provide diverse activities so there is a range of options for engaging with the material for varying learning styles, which can be done via a whole class demonstration, a laboratory exercise, a lecture, simulations, or a choice in topic. Each of these three paths towards a universally designed experience allows learners to customize their experience, become more invested and empowered in their learning, and experience the acquisition of knowledge and skills through the most effective method for each student.

Technology as a Vehicle

Technology is one potential pathway to achieving more universally designed lessons and assessments (Rose & Strangman, 2007). According to Dolan and Hall (2001), electronic media, such as computers, are the most effective and efficient way to meet the needs of most students. Computers offer a variety of ways to present information, engage a variety of students with different methods simultaneously, and allow for differences in expression of knowledge. Virtual Learning Environments, which facilitate student learning through a computer and possibly the internet, have been used in all subject matters in a variety of formats from web-based, like Blackboard and Moodle, to fully immersive environments like River City (Ketelhut, 2007; Shih-Wei & Cheing-Hung, 2005).

River City is an immersive virtual environment that transports students back in time to a city that is plagued with an infectious disease of unknown origin. Students must work together in collaborative teams to gather evidence and test hypotheses to determine a cause and help the town (Galas &Ketelhut, 2006). Immersive virtual environments, like River City, allow students the chance to experience environments, resources, and interactions that may be prohibited due to funding, safety, or other constraints such as time (Roussos et al., 1999). When used to assess student knowledge, immersive virtual environments, formerly used primarily in learning, allow for universal design and technology to combine for a more authentic assessment of knowledge. While the term "authentic assessment" is widely used, many connotations exist across various contexts. According to Jensen (2008) authentic assessments use brain based learning principles to examine the quality to which students understand and process both content and context, rather than how much material has been attained and successfully expressed.

Immersive Virtual Environment Assessments

In lieu of answering multiple-choice questions, immersive virtual environment-based assessments provide students with a problem that they solve using the in-world tools and information so that evaluations can focus on both processes and answers (Clarke, 2009). Students can gather information via virtual observations, perform virtual measurements, or interact with non-player characters and digital objects (Nelson, Ketelhut, & Schifter, 2010). In using a virtual environment, teachers can administer an assessment that satisfies many of the requirements for the universal design for assessments because a variety of student needs can be satisfied simultaneously. Immersive virtual environment-based assessments can help diminish classroom constraints like long text passages for the contextualization of science problems, through allowing students to be immersed into the test questions as a virtual environment. Rather than reading long, elaborate passages, many of the details of the problem to be solved are provided as visual or auditory clues within the virtual environment. The teacher does not need to highlight important details, a common accommodation specified on IEPs, because there are visual cues above potential sources of evidence, in order to signify to students what characters, animals, or objects might be important. In addition, the teacher does not have to try to decipher messy handwriting or miss key pieces of information from the student's problem solving process, because all data from the student's experience (e.g. what they interacted with, any notes they took, what information they saved to view later) are saved in a database through students clicking and/or saving information.

While immersive virtual environment assessments may not be the best assessment for every student, part of universal design for assessment is allowing options. Some students will prefer paper-and-pencil exams and should have access to them if it is preferred. Stowell and Bennett (2010) found that students who typically experience a high level of test anxiety feel much more at ease with computer based assessments. In contrast, those with low anxiety on paper-and-pencil tests can experience high anxiety on computer-based assessments. This furthers the argument for Universal Design for Assessment because no two students will learn or express knowledge the exact same way. By allowing students choices, it increases the likelihood that students are less anxious and more likely to be able to stay engaged in the task and express the knowledge they have attained. Immersive virtual environments are a relatively new technology, especially in the realm of assessment, but already show a promising future as an integral piece in the assessment puzzle.

SAVE Science

Situated Assessment using Virtual Environments for Science Content and Inquiry (SAVE Science), a grant funded by the National Science Foundation (NSF), focuses on authentically assessing student science content knowledge and inquiry skills using an immersive virtual environment as the testing medium.   Science students use their content knowledge and science process skills to make and describe observations, use tools to extend these observations, make inferences, and generate hypotheses when charged with solving an in-world problem. This project hopes to eliminate barriers to testing as well as illuminate knowledge and skills not typically discovered through paper-and-pencil testing. Instead of students narrowing down their choices from a list of possible answers or simply guessing a solution to multiple choice test questions, students must actively participate in problem solving in order to apply their knowledge and synthesize an answer. Every virtual action a student makes is tracked and recorded in a database to show individual trajectories, allowing researchers to observe a holistic view of student's inquiry skills and science content knowledge. However, there is still a concern over the construct irrelevance of reading on science assessments. Currently, the project uses visual contextualization to help cue students as to what information is important; however, during the questioning sequence, students still must be able to read. In accordance with universal design for assessment principles, a read aloud accommodation should be offered to all students. This study investigated if the reading aloud accommodation can be placed within an immersive virtual environment, as a design structure to eliminate reading construct irrelevance, and if so, investigated its impact on students.

In order to determine if reading construct irrelevance variance is minimized by the computer-based reading aloud of text within a virtual environment, a randomly assigned group of students received this accommodation as a treatment, while those in a control group will not. As discussed more in depth in Chapter Two, according to universal design for assessment, by removing potential barriers for some students, i.e., English language learners and those with learning disabilities, all students will benefit. This design prevents students from being excluded from performing well on science assessments because they do not like to read, are poor readers, cannot read English, or because they cannot comprehend the passage.

Research Questions

More specifically, this dissertation is an attempt to address the issue of reading construct irrelevance in assessments, namely science evaluations based in immersive virtual environments. By eliminating the irrelevant construct, are students able to illustrate their science understanding and inquiry skills better? Is this construct irrelevance more crippling by demographic variables (e.g. gender or ethnicity) as indicated by the change in performance due to the RAA treatment? Can students perform better on virtual assessments, if the display of their knowledge is no longer contingent on reading? Do students like having an auditory affordance?

Dissertation Overview

Through a review of literature found in Chapter Two, and data collected through the quantitative methodology described in Chapter Three, this dissertation attempts to answer the aforementioned questions. Student performances on in-class assessments, immersive virtual environment-based assessments, and survey responses are compared using statistical analyses. The results of quantitative analyses are presented within Chapter Four to answer the research questions. The focus of the final chapter is examining the results and interpreting how the findings impact the field of science assessment.

If the auditory affordance is found to help students express their science content knowledge and process skills, it will be provided as an option for all students in future implementations. Currently there is little research into using virtual environments as assessments, but it is quickly becoming a blossoming field. Determining whether auditory support is a necessary option for universal design on these assessments is a key factor in its design and implementation. Findings showing this design strategy as successful could help student expression of knowledge, better inform assessment creators, and ultimately reduce science anxiety of students. In addition, if individualized auditory affordances make the difference due to the elimination of reading construct irrelevance, this will be important to more than just the field of science education. In the wake of high stakes testing, accurately assessing knowledge is a necessity in ensuring that no student is left behind and all are getting the support they need to succeed in school.

CHAPTER TWO

LITERATURE REVIEW

Introduction

This chapter provides a review of the literature to ground this dissertation. The theory of Brain-based learning, which combines educational implications with findings from neuroscientific research detailing how the brain functions, provides the broadest lens that supports the following ideas. All students learn differently because individual brains develop according to unique personal experiences. In order to maximize instruction, teachers should use multiple methodologies in instructional practices so that students experience a multitude of conceptual and sensory experiences. By universally designing instruction and assessment, teachers integrate numerous ways for students to observe, acquire, and communicate their understanding of knowledge so that the classroom is free of learning barriers, allowing for a greater number of students to succeed than would with more traditional pedagogy. Technology is an efficient and effective way of designing and integrating universal design and brain-based principles within the classroom. Immersive virtual environments are a potential technology that is being researched as to their effectiveness in integrating multiple methodologies in learning and assessment. Within immersive virtual environment-based research projects like the Virtual Performance Assessment project and the Situated Assessment using Virtual Environments for Science Content and Inquiry (SAVE) Science project, students are part of a problem and solve it virtually, rather than reading about it on a paper and pencil test. While these projects are integrating brain-based methodologies by immersing students into assessments and providing students the opportunity to process information actively with potentially less stress than a paper and pencil assessment, is this enough to assess student content knowledge more accurately?

These assessments are designed to provide students multiple pathways to solve in-world problems; however, as they are currently designed they do not provide students multiple ways to access the problem, rather students can only obtain the details of in-world problem through visual means like reading and observing. When assessing student knowledge, it is important to eliminate the assessment of unintended constructs, (e.g. reading on a science test), that may prohibit students from exhibiting their content understanding. In order to reduce the measurement of construct irrelevant variance on paper-and-pencil assessments for students with learning disabilities, accommodations are provided. Accommodations are specifically designed for students depending on their learning disabilities. Because reading comprehension difficulties are one of the most common areas of deficiency in learning disabled students, the read aloud accommodation is widely used. This accommodation provides an audio version of an assessment, usually performed by a human reader, to students with learning difficulties, so that reading proficiency is no longer a requirement to perform on an assessment. Embedding this accommodation into virtual environment assessments would potentially make it more aligned to a brain-based perspective and provide the ability for students to process the tested information while expressing their content understanding, regardless of their reading proficiency.

Theoretical Framework-Brain-based Learning

In the middle of his first presidency, George H. W. Bush proclaimed that the last decade of the 1900s would be the “decade of the brain” in order to increase initiatives for funding, programs, and general awareness of findings in brain research (Bush, 1990).

While articles and books had been published about the brain in previous decades, the field of neuroscience began using magnetic resonance imaging (MRI) and Positron Emission Topography (PET) scans during learning and memory exercises to map out areas of the brain and understand how human brains function (Jensen, 2008; Weiss, 2000). These findings about the brain as an organ and its inner workings, were applied to the field of education to produce brain-based instruction. According to Jensen (2008), brain based learning is a multifaceted approach designed to maximize learning consistent with what research suggests best supports brain functionality. Caine and Caine (1991) outlined twelve principles for brain-based learning in order to explain how the human brain functions.

1. The brain is a parallel processor.
2. Learning engages the entire physiology.
3. The search for meaning is innate.
4. The search for meaning occurs through "patterning."
5. Emotions are critical to patterning.
6. The brain processes parts and wholes simultaneously.
7. Learning involves both focus attention and peripheral perception.
8. Learning always involves conscious and unconscious processes.
9. We have at least two different types of memory: A spatial memory system and a set of systems for rote learning.
10. We understand and remember best when facts and skills are embedded in natural spatial memory.
11. Learning is enhanced by challenged and inhibited by threat.
12. Each brain is unique.
    (Pgs. 79-87)

An overall theme of these tenets is that the brain processes multiple things concurrently, (e.g. visual and auditory or parts and whole).Therefore presenting materials in multiple modalities creates more potential pathways for retrieval while allowing the brain to process information more naturally. Further, the brain processes and uses both focused attention and sensory information as well as conscious and unconscious processes to encode an experience. Negative emotions acquired through any of the aforementioned mental processes, even peripherally, can cause students to have unfavorable perceptions about learning, as emotions play a key part in learning and memory. Because affective and cognitive processes cannot be separated, negative emotions can inhibit learning. Thus the classroom environment should always be supportive and non-threatening. Supportive not only means nurturing, but also the nourishment of learning through challenging students. Because the brain processes the familiar while simultaneously attending to new stimuli, a combination of the two allows for a comfortable challenge for students. Learning is optimal when students are challenged but the perceived threat is low. One way to challenge students but provide a supportive environment is by embedding learning within real life experiences. This enhances factual recall and procedural knowledge and while facilitating natural learning and the transfer of knowledge. Because all brains are constructed individually depending on experiences, no two people learn the same and therefore, multifaceted, brain based instruction provides a greater opportunity for all students to learn and acquire new information (Lyons, 2003).

Infusing the theory of brain-based learning into classroom instruction requires a significant paradigm modification in teaching practices (Jensen, 2008). In order to move away from rote memorization and towards more significant, natural and permanent learning, three conditions are necessary for lessons: “relaxed alertness, orchestrated immersion, and active processing” (Caine & Caine, 1990). Relaxed alertness, one of the twelve principles, occurs when students experience a high level of challenge but a low level of threat, meaning that the challenge is not so overwhelming that it is intimidating, yet it is not so easy that there is no intellectual stimulation. Further, these authors express that because all learning is experiential to some degree, teachers must orchestrate the immersion of students within content experiences. The school experience should provide a context that situates all student learning within multiple, complex experiences that are authentic. This provides students with multiple pathways to absorb, comprehend, and store information. Moreover, because each brain is unique students make meaning through processing these experiences, teachers should provide opportunities for active processing. Similar to metacognition, active processing is where students must think about how they are acquiring information because it helps students develop personalized self-monitoring and learning strategies (Flavell, 1979). Within the brain, reflexive learning is immediate and natural in order to navigate one’s world safely and overcome dangers or obstacles, while reflective thinking is not necessarily innate nor does it have the same urgency (Slywester, 2010). Thus, reflexive learning is not naturally used or developed in children, which is why metacognitive strategies must be taught and practiced. According to Crawford (2003) in accordance with brain-based learning, assessments should be designed to help students learn how to self-evaluate their understanding and develop reflective, metacognitive skills. Learning processes like knowing how to learn and acquire new skills are what is required for success in the global economy, not the rote, memorization skills that were necessary in the industrialized era (Gabriel, 1999; Stevens & Goldberg, 2001). Therefore, in order to prepare students for the current workplace, instruction and assessment must change towards processes rather than memorization and reflect more brain-based learning strategies (Bellah et al., 2008).

Brain-based instruction and assessment moves beyond one method of information transmission and assessment, in order to provide more opportunities for students to interact with and remember material. The human brain processes images, sounds, and other information through multiple modalities (i.e. auditory, visual, and kinesthetic) when neural pathways are activated to respond via stimuli (Block & Parris, 2008); however, personal preference and previous learning experiences determine student's dominant learning style (Clemons, 2005).

Research studies into the effectiveness of integrated brain-based instructional principles indicate that students experiencing multiple methodologies for information processing are outperforming those in control groups. Ali, Hukamdad, Ghazi, Shahzad, and Khan (2010) found students who experienced brain based learning statistically learned more than those in a control group, through a pre/post test design, when examining secondary physics students in Pakistan. The experimental group experienced a colorful, decorated, scented classroom with group discussions and experiments designed to facilitate the social construction of knowledge, while the control group experienced more traditional classroom instruction within a white room. By providing multiple modalities, not only are neural pathways strengthened and more accessible, but a larger number of students experience their preferred learning style. Similarly, in a study of Turkish pre-service teachers, Duman (2010) found that the experimental group, which learned through brain-based strategies, had a statistically significant ($p<.001$) higher achievement than the control group. Specifically, the control group only experienced lecture or question and answer, while those in the treatment group experienced group work, videos, stress reducing techniques, and the ability to walk around the room to collaborate with others. This difference in instructional method had a positive effect on attitude and motivation for the experimental group as well, regardless of their dominant learning style. According to Tomlinson and Kalbfleisch (1998) in order to educate students most effectively, teachers must use brain-based strategies like differentiating instruction by providing multiple pathways for constructing and expressing knowledge. One way to accomplish a more brain-based classroom is through integration of universal design.

Universal Design for Assessments

Universal design is a theory that began in architecture in order to make buildings more accessible to all, without a multitude of retrofits designed to fit small portions of the society, such as people in wheelchairs (Mace, Hardie, & Place, 1991). The theoretical principles addressed in order to provide access for the widest variety of people are summarized as providing equitable, flexible, intuitive, and simplistic functionality; having tolerance for mistakes; requiring low physical effort; and allowing unlimited access regardless of mobility or size (Connell et al., 1997). According to Welch (1995), universal design goes beyond specifications for individual users and focuses on creating a product more accessible and functional for all users from inception. In the 1990s, the Center for Assistive Special Technology (CAST) adopted this theoretical framework and applied it to learning (see Rose, Meyer, Strangman, & Rappolt, 2002 for a review of universal design for learning). Thompson, Johnstone, and Thurlow (2002) then applied this idea of providing a greater accessibility to testing and developed universal design for assessment. They modified the original seven principles to pertain to the evaluation of student knowledge and outlined the following:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amendable to accommodations

5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

(pgs. 1-2)

Through these principles, universal design for assessment ensures greater accessibility for all students on evaluations of understanding, especially those with learning disabilities and language barriers, while simultaneously providing more valid interpretations of performance (McGuire, Scott & Shaw, 2006). Rather than trying to evaluate all students in an identical way, universal design for assessment provides students and test proctors (usually teachers) a multitude of options (e.g. a reading aloud accommodation or definitions provided when students mouse over options on a computerized test) from the beginning so less external accommodations have to be made (Rose & Meyer, 2002). While the majority of universal design for assessment research articles are theoretical, Johnstone, Thompson, Moen, Bolt, and Kato (2005) used differential item function analysis, item to total correlation, and item rankings of difficulty to determine which questions were problematic by performing secondary analysis on data from fourth and eighth grade high stakes math assessments. Because they looked at items with respect to all disabilities and administered accommodations, their sample sizes became very small along with the effect sizes, while the number of problematic questions was very large. They advised that states should analyze their questions using item rank, item correlations, and differential item functioning to find their problematic questions. However, they recommended using focus groups and experts to determine why the questions are problematic in order to fix the issues of questions that are low in accessibility standards for any subgroup, which they did not complete within their study.

Rather than seven design principles or ways to analyze already implemented questions, the Center for Assistive Special Technology provides three categories that creators of universally designed assessments should focus on from inception to implementation: multiple means of presentation, expression and engagement (CAST, 2011), which are congruent with brain based learning ideals. In reference to testing, multiple means of representation equates to providing graphics, text, and/or audio, not just simply text similar to what many paper-and-pencil tests supply (Dolan & Hall, 2001). This idea is compatible with one of the main tenets of brain-based learning explains that the brain processes acquired information concurrently and in parallel (Caine & Caine, 1991). Multiple means of expression could entail allowing students to audio record their responses or use a speech to text program as well as the option to draw or type their answers rather than handwriting being the only choice. Because brain-based learning indicates that each brain is unique, the best way for students to illustrate their knowledge may be differ by individual. Multiple means of engagement is designed primarily for the learning aspects; however, it could mean providing students with different ways of interacting with an assessment like a field test (e.g. the classification of tree by walking around with a dichotomous key) or allowing students to work in partners to collaboratively interact with the assessment. This idea of engaging students in many ways is supported within brain-based learning because students remember information best when it is embedded within more natural contexts, like alternate assessments often provide.

According to Goswami (2008) when children are taught information in multiple modalities, like brain based research and universal design suggest, the brain connections and neural pathways are stronger. If applied to assessments, the more potential neural pathways they have to access the information, the better the chance is for retrieval and expression of the knowledge and skills. Within a traditional classroom environment, it can be a struggle for teachers to provide concurrent varieties of activities for students to acquire content, interact with it, and illustrate their understanding in terms of both supplies and monitoring capabilities. As a result, researchers are investigating the integration of universal design for assessment considerations through technology, because it affords a vast array of options for teachers and the assessment itself (Burgstahler, 2007).

Technology as a Tool for Multiple Methodologies

Technology provides a wider range of possibilities for integrating multiple sensory experiences within instruction and assessment. According to Tucker (2009), technology can help teachers gain a better perspective on student learning and enrich their classroom instruction through more complex assessments. In addition, technology-enabled assessments provide teachers the capability of simultaneously broadening and deepening evaluations through evaluating additional skills and understandings by tracking student actions through multiple step problems. Furthermore, he explains an educational strength for this form of assessment is that students can perform investigations in order to solve a problem. Computer programs can record student’s trajectories through a recorded log of each individual action in solving multistep problems, which can help identify gaps in understanding rather than a single data point that only informs whether the answer was correct or incorrect (Lewis & Sewell, 2007; Lowry, 2005; Thissen-Roe, Hunt, & Minstrell, 2004). By providing teachers with student pathways and answer sequences, technology-enabled assessments can help teachers determine student misconceptions and more easily provide individualized feedback to guide students toward conceptual understanding and application (Thissen-Roe, Hunt, & Minstrell, 2004). Unlike paper-and-pencil testing, technology-enabled assessments provide students with the opportunity to display their flexibility in determining their solutions, which is a necessary skill in the rapidly changing workplace (Code, Clarke-Midura, Mayrath, & Dede, 2011).

Immersive Virtual Environments

Immersive virtual environment based assessments are the newest form of technology-enabled assessments purporting to more authentically assess skills (Clarke-Midura, Code, Zap, & Dede, 2011) and have the capability of integrating more accessibility via Universal Design and brain-based principles. Typically immersive virtual environments, like *River City*, *Quest Atlantis*, and *Whyville*, have been used solely for learning and instruction within an education context (Barab, Sadler, Heiselt, Hickey,& Zuiker, 2007; Kafai, 2010; Nelson, Ketelhut, Clarke, Bowman, & Dede, 2005). Through interactive multimedia, they allow for a more brain-based experience via engaging contextualization and inquiry processes, especially in curriculum areas like science where not all topics can logistically be covered safely within a classroom, e.g. adding random chemicals together (Donnelly, McGarr, & O'Reilly, 2011). However, virtual environments are now being used as technology- enabled assessments that transport students into the problems they must solve, which provides a rich data source of student inquiry skills and knowledge understanding pathways. Because the literature reports primarily on two immersive virtual environment-based assessment research projects, each will be explained in detail.

Virtual Performance Assessment

Virtual Performance Assessment is a scientific inquiry, immersive virtual environment, assessment project out of Harvard University. In this immersive virtual environment, students use an avatar to explore the world and gather evidence to solve a specific problem such as declining aquatic life in *Save the Kelp!* or genetic mutations in *There's a New Frog in Town*, (Mayrath, Clarke-Midura, Dede, & Code, 2011). The students can observe, use tools to gather data, or use inquiry skills to generate a conclusion that solves the in-world problem. The researchers designed these virtual assessments using Mislevy's Evidence Centered Design approach, which provides a structured perspective of the knowledge, skills and abilities specific to each content area and general inquiry skills (Clarke-Midura, Code, Mayrath, & Dede, 2011). The hope of this project is to provide a supplement to paper-and-pencil exams, which have not historically been able to measure inquiry skills authentically (Clarke, 2009). In accordance with brain-based learning, this project provides students the ability to create their essay responses by incorporating pictures of key concepts by dragging and dropping, in case they do not know the terminology (Mayrath et al., 2011). This supplies students with different options for expressing their content understanding. Since this program is still in its infancy, no data has been presented on its effectiveness to assess inquiry skills in comparison to other assessment types. Preliminary results indicate that students reported that the virtual environment added to their prior knowledge, provided feedback and flexibility, and enabled learner control (Code, Clarke-Midura, Mayrath, &

Dede, 2011); however, this does not speak to the overall effectiveness of the assessment just that students had a positive learning experience.

SAVE Science

Situated Assessment using Virtual Environments for Science Content and Inquiry (SAVE Science) is a project that is designing and examining a series of science content and inquiry assessments based in a virtual environment called *Scientopolis* (Shelton & Ketelhut, 2012). Each module assessment is designed to evaluate both inquiry skills and a specific science content topic for middle school science students. Specifically, the modules consist of *Sheep Trouble*, which focuses on adaptation and speciation; *Weather*, which centers around the topics of fronts, air masses, and atmospheric conditions; and *Basketball*, which features the properties of gases as its content area. In each module, students solve a problem based on externally constructed benchmark questions that fit within the purview of state standards (Nelson, Ketelhut, & Schifter, 2010). Furthermore, it is hoped that this innovative form of assessment will bridge the distance between scientific inquiry classroom experiences and how they are assessed on conventional assessments, which often evaluates scientific inquiry as definitions. Designed using the theory of situated cognition, which states that learning is embedded within a context and thus should occur in a context that closely resembles what will be used, SAVE Science positions the learner and the content inside an immersive virtual environment that represents an authentic context (Ketelhut, Nelson, Schifter, & Kim, 2010). Because the content and assessment are contextualized and embedded within a virtual environment, students can perform observations and generate hypotheses to solve a problem that they are immersed in rather than one they are reading about. This is representative of the orchestrated immersion that Caine and Caine (1990) stress is a key principle in brain-based learning. Moreover, in order to reduce perceived threat and cognitive load for students, there is visual signaling above key digital objects (Nelson et al., 2010) making the test more intuitive and accessible, satisfying elements of both universal design and brain-based learning.

SAVE Science assessments are pushing students to move beyond selecting a response from a multiple choice list towards using more complex thinking in order to express their science content knowledge and inquiry skills. While complete data analysis has not yet been reported, during a pilot study of *Sheep Trouble*, there was a difference in success on the contextualized immersive virtual environment-based assessment in comparison to decontextualized multiple-choice questions. A small group of middle school students who took the situated assessment were much more likely to answer the contextualized speciation multiple choice question correctly than the control group when the question was presented with no context via a paper-and-pencil test. This finding occurred despite a much larger time delay between learning the content and completing the assessment for virtual test takers (Ketelhut, Nelson, Schifter, & Kim, 2010). Researchers in this project are only beginning to analyze raw data as well as to determine how to use student paths of discovery and investigation in the world, via automatically recorded click data, to develop an overall picture of each student's science content knowledge and inquiry abilities.

Implications for Technology-enabled Assessments

Both VPA and SAVE Science show promise for creating an assessment that will illustrate the extent of student understanding accurately. These projects use technology as a vehicle to orchestrate student immersion within a virtual world and view student thought processes and application of content matter. By embedding the assessment within a virtual environment, these projects are incorporating the principles of brain-based learning in several ways. Solving problems within virtual environments is more realistic to student lives than solving well-defined problems on paper-and-pencil assessments. Within real life, people do not stumble across problems that are well defined with a definitive answer very often. Rather, solving problems in life is an iterative process that does not necessarily have one correct path or answer. In solving the problems of virtual characters, students are able to remember facts and inquiry processes since it is embedded within a natural spatial context. Furthermore, another brain-based learning tenet focuses on emotions being a critical factor because challenging, non-threatening activities increase learning and proficiency. Many students have test anxiety associated with paper-and-pencil assessments. Virtual environment assessments provide students with a challenging task that requires them to apply content knowledge, but the vehicle is perceived as less threatening than a high-pressure paper-based test. Moreover, because assessments based in virtual environments allow for multiple pathways towards success, students can use their unique brains to solve the problem in the pathway they choose. Through the contextualization of science assessments within virtual environments, these research projects focus on attempting to evaluate student content understanding rather than test taking skills, like many paper-based multiple choice questions do. However, these projects have not investigated whether contextualization within virtual environments eliminates the measuring of reading skills, which is commonly evaluated as an unintended construct causing validity issues.

Validity and Construct Irrelevance Variance

According to Messick (1989), validity is the degree to which evidence and theory support the interpretations of actions and answers derived from assessments, in other words, how accurately an assessment correctly evaluates one's knowledge of a specific construct. According to the standards for Educational and Psychological Testing outlined by AERA, APA, and NCME (1999), a construct is a "concept an assessment is intended to measure." Therefore, validity is a measure of how accurately an evaluation achieves what it purports to accomplish (Duran, 2008). Construct-Irrelevant Variance, described as "excess reliable variance associated with other constructs" (Messick, 1989, p.742), is one source that can erode the validity of an assessment instrument. Variance due to construct irrelevance can come in the form of additional difficulty or easiness, but this dissertation will focus on the former. When knowledge, skills, or abilities that are ancillary to the tested construct become a focal point or a gateway to demonstrating knowledge on an assessment, construct irrelevant difficulty is established (Haertel & Linn, 1996). For example, on assessments where reading comprehension is necessary to understand the questions in order to provide an answer, this question becomes unnecessarily more difficult for poor readers, students with reading difficulties, and English Language Learners. Therefore, if someone is only looking to assess scientific knowledge, not reading skills, the reading construct should not be part of the required skill set.

Oakland and Lane (2004) explained that language is often the medium of communication between the assessment and the student, which can be helpful or prohibitive, depending on student skill set. Further, unless reading is being tested specifically, designers should try to minimize the reliance on this skill in order to reduce the impact of language related construct irrelevance (Camara, 2009; Oakland & Lane, 2004). AERA et al. (1999) stressed that any test that uses language is also a test of language skills, which can introduce components of construct irrelevance variance, in turn, inaccurately representing student knowledge and understanding.

Abedi (2002) found that English language learners perform lower than students who know English as their first language on reading, science, and math, with the largest gap being in reading and the smallest in math. Therefore, this suggests that as subject matter testing decreases in its dependency on reading skills, the difference in scores between English language learners and those without language barriers decreases similarly. In other words, as the construct irrelevance variance disappears, so does the difference in achievement, making a case for the removal of unintended constructs from assessments. In order to ameliorate the issues of construct irrelevant variance for students with learning disabilities or language barriers, schools are administering assessments with accommodations for students.

Accommodations

According to Tindal and Fuchs (2000), accommodations are alterations to the appearance, time allotment, location, or response for the test administration that do not change the targeted constructs. It is important to the validity of assessments that accommodations not alter the construct being measured (Hollenbeck, 2002; Sireci, 2004). Further, Sireci (2004) explains that in order to be a good accommodation in eliminating construct irrelevant variance, the modification should improve the measurement of understanding and skills, maintain comparability with the non-accommodated assessment and not provide an unfair advantage. If the measured construct is changed, it becomes a modification to the assessment, which diminishes comparability between scores.

Accommodations come in a variety of forms from extended time, to alternate presentation methods, to students being read each question aloud, depending on what is required by a student's Individualized Education Plan (IEP). Accommodations are designed to minimize the impact construct irrelevant variance causes on assessment answers (Camara, 2009; Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Sireci, 2004). Moreover, they provide the opportunity for more accurate score interpretations since only the intended construct is being measured. While there is contradictory evidence indicating that some accommodations can introduce additional construct irrelevant variance (Cahalan, Mandinach, & Camara, 2002; Sireci, Scarpati, & Li, 2005), other studies provide data that suggests that a read-aloud accommodation does not introduce this on a non-reading assessments (Cook, Eignor, Steinberg, Sawaki, & Cline, 2009; Huynh, Meyer, & Gallant, 2004).

Although a misconception exists that reading only involves the visual system of processing, neuroscience research suggests that it is intertwined and dependent on all sensorimotor systems (Block & Parris, 2008). For example, reading is linked to auditory processing because children learn to understand language and communicate through this system, long before learning how to read. Frequently, poor readers are more proficient in understanding language orally than by visually processing text, which are grounds for providing a reading aloud accommodation on assessments evaluating constructs other than reading.

Read Aloud Accommodation

One of the most commonly used accommodations is the RAA, which entails reading aloud some portion or the entire assessment to students who struggle with reading difficulties (Cormier, Altman, Shygan, & Thurlow, 2010; Johnstone, Altman, Thurlow, & Thompson, 2006; Clapper, Morse, Lazarus, Thompson, & Thurlow, 2005; Thompson, Blount, & Thurlow, 2002; Dolan & Hall, 2001). According to Burns (1998), a standard method for implementing the reading aloud accommodation does not seem to exist. Differences lie in how many times the reader verbalizes the information, how much of the information is read, and how many students the information is read to at a time. While some research suggests that this variation can cause issues in reliability both within and across studies (Bolt & Ysseldyke, 2006), other research found little evidence to show that the variance in implementation causes a difference in effectiveness (Calhoun, Fuchs, & Hamlett, 2000). Moreover, Clapper et al. (2005) found that all states allow the reading aloud accommodations with provisions, usually the preclusion of the accommodation on reading assessments, whereas Massachusetts, Missouri, and Vermont have no limitations on its use. Because of the frequency of its use, this accommodation has been researched with a variety of methods, participants, and content matter.

On an eighth grade standardized reading assessment, McKevitt and Elliott (2003) examined student perceptions of read aloud accommodations. Students with learning disabilities preferred taking the test with the RAA and reported feeling it was easier to illustrate their understanding, whereas those without disabilities reported the conditions were similar despite the accommodation, but that they thought they performed better in the accommodation condition. In a study examining how college students perceived the utility of testing accommodations during their high school and collegiate experiences, Bolt, Decker, Lloyd, and Morlock (2011), provided data that showed college students self-reported on a survey that all forms of reading aloud accommodations range from somewhat helpful to helpful on average.

According to a review by Bolt and Thurlow (2007), three categories encapsulate the majority of RAA research: differential improvement, construct measurement equivalence, and question level studies and these categories will be used in this section. Most commonly in differential improvement studies, researchers seek to determine if students with learning disabilities benefit more from a RAA, which indicates that the accommodation is performing as it should by providing students with learning disabilities a differential boost and eliminating the construct irrelevant variance (Phillips, 1994). Typically, this is accomplished using a group of students with learning disabilities and a group without, where both groups take an assessment with and without accommodations and a repeated measures analysis determines the differences in scores across implementations. In studies of the construct measurement equivalence, researchers investigate if the accommodation alters the measured construct in a significant way (Huynh & Barton, 2006). Investigators compare the factor structures or differential item functioning analysis of both accommodated and non-accommodated tests to conclude whether the conditions alter the measured construct. Question level studies are designed to compare the performance of students per item on each evaluation to determine different levels in performance by student characteristics (Helwig, Rozek-Tedesco, Tindal, Heath & Almond, 1999). In general, researchers hypothesize that RAA should make a greater difference on items of high reading difficulty. Alternately, to the other two methods, this one can illuminate differences between the accommodation not working and students lacking in content knowledge or associated skills. While all three methods have provided mixed results, findings from each method will be presented in the following sections.

Differential Improvement

Usually, when a student receives an RAA, scores improve, regardless of the subject matter or learning disability status; however, since the purpose of the accommodation is to reduce the construct irrelevant variance and more accurately assess each student on only the tested construct, many researchers look for differential improvement for students with learning disabilities over those in the general education population. Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) found the RAA to be helpful for all students, however, larger gains appeared in students with learning disabilities than general education students on a fourth grade mathematics assessment. In a related study examining math tests in elementary students, Weston (2003) also found that both groups scored higher when assessed using a RAA, but the effect sizes were twice as large for learning disabled students. In addition, both groups of students overwhelmingly said they understood the questions better, believed they performed better, and preferred being tested with the accommodation in comparison to standard administration. Meloy, Deville, and Frisbie (2000) investigated math, science, and reading assessments. Data illustrated that the RAA increased all students’ scores in both math and science on the Iowa Tests of Basic Skills, but differentially increased the scores of students with learning disabilities. These mean differences in scores were much higher in science than in math, indicating that the accommodation is more useful in this subject area and does reduce construct irrelevant variance. Moreover, they found the RAA raised scores of all students on the reading portion of the assessment, making the evaluation easier for all, rather than providing differential access to those who need it. Since the assessment is measuring reading, this change in the measured construct is an unnecessary and counterproductive one. This assertion is supported by some research studies (Bielinski, Thurlow, Ysseldyke, Friedebach, & Friedebach, 2001; Burns, 1998; McKevitt & Elliott, 2003; Phillips, 1994) while others refuted this by suggesting that students with learning disabilities benefit on reading tests to a much greater extent than those without them (Huynh & Barton, 2006; Laitusis, 2010).

RAA also seems to help children with low reading skills but who do not have identified learning disability. Elbaum (2007) found that high school mathematics students without learning disabilities gained statistically significantly more when using the RAA, in comparison to students with learning disabilities. Thus, the RAA eliminated a difficulty shared by many students, since data indicated an overall low proficiency in reading among students in both groups. Rather than a differential increase, results indicated that the accommodation provided the greatest benefit to the students with better math skills, which tended to be in the general population, not those with learning disabilities. She further elaborates that if a more accurate measurement of math or science skills is the goal, the RAA should be offered since it alleviates the construct irrelevant variance for all students, provided reading is not the construct being measured.

Randall and Englehard (2010b) studied two accommodations, RAA and a resource guide, on a reading competency assessment for approximately 750 seventh grade students in Georgia. Through differential item functioning analysis, they found construct invariance among students regardless of disability status, but that RAA made only one question out of twelve easier for the students with learning disabilities who received the accommodation in comparison to those who received other accommodations or nothing. In a related study, Randall and Englehard (2010a) studied both RAA and resource guide accommodations with both fourth and seventh grade students in a pre-post design. Students took the original high stakes reading assessment in grades 3 and 6, and then different schools re-administered the test in fourth and seventh grade with the same students; some randomly selected schools provided one of the two accommodations. While the resource guide did not prove to be effective in either grade, the RAA provided a differential boost for students with learning disabilities in fourth grade. Differences in the effectiveness of the RAA by grade could be attributed to a few different factors, such as a greater perceived importance of high stakes testing in lower grades and an ever-increasing gap in ability levels due to reading difficulties. Because the results are a repeated measure using a high stakes test, the second administration, which tested the treatment versus the control, students were aware that the test only mattered to researchers and did not affect them personally. Older students may have been more cognizant of the lower stakes and not exerted as much effort affecting the results. Another source of the difference could be that the differences in reading skill levels increase over time and it becomes increasingly difficult to overcome the differences with accommodations.

Other research tends to mimic this finding and shows that often the effectiveness of the RAA wanes as students progress through their educational career (Elbaum, 2007; Laitusis, 2010). Stanovich (1986) describes the Matthew effect as the accumulated advantage for good readers or disadvantage for poor readers as time progresses. Furthermore, he explains that because reading acts as an entryway to develop other necessary cognitive skills and acquire content matter in other subjects, it is a looping mechanism that continues to benefit good readers and disenfranchise poor readers over time because deficits grow not only in reading, but also in all other areas. Because reading is frustrating and not enjoyable to poor readers, they experience less practice and reinforcement to increase their skill growth, while good readers continue their ascent through the acquisition of skills and content matter.

The rate of acquisition of content matter and related skills because of reading difficulties also affects some English language learners on assessments, and thus construct irrelevant variance can invalidate the testing instrument. In order to determine the effectiveness of RAA for English language learners, Wolf, Kim, Kao, and Rivera (2009) compared the test performances of English language learners receiving the RAA and those who have the language barrier but did not receive an accommodation in two different anonymous states in the United States. Researchers found mixed results in that the comparison in state one revealed no significant differences in performance but in state two, higher scores were obtained for those receiving the treatment. The students from state two were more familiar with the RAA in the classroom and the implementation was more standardized as proctors received a script. Through further analysis, it was discovered that the RAA made a greater difference among students who had acquired more content knowledge. Just as with the Bolt and Thurlow (2007) study, reading aloud questions is not going to provide students with any more knowledge than they already have. In order for this accommodation to be successful, students must know how to complete math problems, as it is meant only to act as a reduction to a barrier of access to enable the accurate measurement of student knowledge and valid comparability among students regardless of inhibiting characteristics (Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999).

Comparability studies in the Reading Aloud Accommodation

To determine whether the tested constructs maintain their integrity when accommodations are introduced, researchers compare the factor structures of questions or differential item function between accommodated and non-accommodated tests to check for congruency. Research studies use factor analyses, differential item function analysis, and item level analysis to investigate whether the accommodation alters test items or constructs. Factor analyses are used to determine if factor loadings are still similar when students experience accommodations, while differential item function analysis examines the differences in scores between groups, assuming minimized differences if accommodations are proficient. While factor analysis and differential item function analysis both consider groups and questions at a larger level, item level analysis specifically investigates how effective the affordance is for different types of questions. Each method is important in considering if the RAA eliminates construct irrelevant variance, without changing the measurement.

Factor analyses

Factor analyses are used in these studies to establish factor loadings and indicate whether the test items measure similar information or the construct is changed by the accommodation. Koretz (1997) found that Kentucky students with learning disabilities who had a combination of accommodations on science assessments outperformed students who only received a RAA and students in the general population without accommodations. Students who had a combination of the reading aloud, dictation, and paraphrasing accommodations, scored .2 to .5 standard deviations above students without disabilities on a science assessment. Moreover, students with mild retardation that received multiple accommodations scored similarly to general education students without accommodations and higher than students with learning disabilities that only received the RAA treatment. Since students with mental retardation typically score below average, these results indicate that combinations of accommodations are providing students with learning disabilities an unfair advantage, in math and science. Data suggest that combinations of accommodations are changing the measured construct for students with learning disabilities. Although the sample size was too small to demonstrate true statistical significance, data in this study showed RAA provided the smallest overall change to the construct. Because separating the students into groups of mutually exclusive accommodations made comparison groups too small, results were inconclusive as to which accommodations made the biggest difference alone or in combination. Using exploratory and confirmatory factor analyses to compare students with learning disabilities who received the RAA accommodation with those that did not Middleton (2007) found dissimilarities between the factor structures of accommodated and non-accommodated tests. The constructs were more similar without accommodations, indicating that the RAA changed the measured reading comprehension construct to be easier for accommodated students.

In contrast, upon examining the factorial structure of a fourth-grade mathematics assessment, Pomplun and Omar (2000) acquired data indicating an invariance in measured construct among their three populations: student without learning disabilities or accommodation, students with learning disabilities but no accommodation, and students with disabilities accommodated by a read aloud affordance. Because the factorial structure was maintained throughout the tested groups, the results offer support for the comparability among testing conditions, in contrast to previously cited studies. In a later and analogous study on an eighth-grade mathematics assessment, Huynh, Meyer, and Gallant (2004) found no variance in factor structure among students with learning disabilities receiving the accommodation and students without disabilities who did not have a RAA. Moreover, Cook, Eignor, Sawaki, Steinberg, and Cline (2010) found congruence in factor structures across students with disabilities receiving RAA and a control group of students without disabilities receiving no treatment on fourth grade language arts tests.

Overall, research studies using factor analysis are inconclusive; however, when comparability studies are broken down by subject matter, RAA is not usually found to be appropriate for assessments of reading comprehension. However, mathematics and language arts comparability studies suggest that the RAA eliminates the construct irrelevant variance for students with disabilities and maintains the integrity of the measured construct in elementary and middle school students. The small number of studies indicates a need for more research in order to generalize findings, especially in science to consider the RAA individually.

Differential item function analysis

Comparability studies in differential item function analysis have similar results. According to Elbaum (2007), if the constructs measure the same information and skills and have no construct irrelevant variance, the differential item function analysis difference, or variance between scores of two different groups, should be the low between students. Differential item functioning analysis consists of researchers examining the scores between two groups with an expectation that their scores are the same. It is expected that in using the RAA with only students with learning disabilities, they should score approximately the same as students without disabilities who do not receive accommodations, if the RAA is removing the construct irrelevant variance. However, if there is a large differential item functioning in how students perform between groups, there is an indication that the accommodation is not functioning as intended and the measured construct is being changed. Further, if there is construct irrelevant variance, like reading being measured on a science test and no accommodation is provided, a large difference in item functioning is expected between students with and without reading disabilities. An accommodation specifically designed to alleviate the construct irrelevant variance should improve the amount of differential item functioning between groups of students, if it is removing the unintended tested construct. Fuchs, Fuchs, Eaton, Hamlett, Binkley, and Crouch (2000) found that the differential item functioning decreased and the measurement improved greatly on a math construct for half of the questions for students with learning disabilities that received the RAA in comparison to student with no disabilities or accommodations. This finding suggests that the affordance was indeed eliminating the construct irrelevant variance due to reading.

Bolt and Ysseldyke (2006) performed a differential item functioning analysis between three groups of students, those with learning disabilities receiving an RAA, those with disabilities not receiving an accommodation, and regular education students receiving no accommodations on both math and reading/language arts assessments. Results provided illustrate a larger differential item functioning when students experienced the RAA in comparison to those not receiving a treatment. Although the researchers found differential item functioning for both math and reading/language arts, the differential item functioning was much higher for the latter, suggesting that the accommodation alters the reading construct more than the math. Therefore, as with the majority of previously cited studies, the RAA tends to be more appropriate for usage in content areas other than reading (e.g., math) because differences are minimized.

Bielinski, Thurlow, Ysseldyke, Freidebach, and Freidebach (2001) found similar results in a differential item functioning analysis study. They performed differential item functioning analysis to establish the frequency of difficult items for students with disabilities that were receiving the treatment versus those who were not. Data from a reading test suggested that the RAA was ineffective for students with learning disabilities because there was a large differential item functioning between students who received the accommodation and those who did not, with the difficulty of questions actually increasing when using this affordance. Within this study, the math results also mirrored the Bolt and Ysseldyke research results. In the math section, overall, the differential item functioning was the same whether students received the RAA or not. However, when students received the RAA, six questions became easier for students with learning disabilities. Because there was no item analysis performed, it is inconclusive whether it was the higher level in reading difficulty of the word problems or a different reason that caused the differential item functioning.

Item level analysis

Examining the read aloud accommodation (RAA) per question is important in establishing how effective the affordance is for different types of questions. It can help researchers, teachers, and policy makers understand which types of questions students with learning disabilities need accommodations for most and possibly determine why. Studies at this level indicate positive results for the RAA for students with learning disabilities, but not always with questions of high reading difficulty. Keterlin-Geller, Yovanoff, and Tindal (2007) found that third grade students with low reading skills had a differential boost in math scores compared to students without disabilities when both received the RAA, but this increase was detected only on problems with both high math and reading complexity. Using a video-delivered RAA and a repeated measures design, Helwig, Rozek-Tedesco, Tindal, Heath, and Almond (1999) found a similar result. Students with low reading abilities and high mathematical skills scored statistically higher on problems associated with a high reading difficulty (large word count, higher number of verbs, less familiar words) when they experienced the video delivered RAA in comparison to their control scores. There was no difference in scores for students with high reading abilities, respective to the treatment or control condition. These two similar findings indicate that some students have high mathematical capabilities but are unable to show it when the reading level is also difficult or that as problems become more mathematically complex, differences in reading abilities make a larger difference. Further, the authors indicate that students with low reading fluency but high mathematical skills and English language learners have similar issues in that navigating the test and comprehension are the barriers that preclude them from illustrating their mathematical knowledge. Bolt and Thurlow (2007) also found a differential improvement in scores of students with learning disabilities when using the RAA. However, they found the positive impact on the performance only occurred on questions that were high in reading difficulty but low in mathematical difficulty. In contrast to the Keterlin-Geller et al. and Helwig et al. studies, this result suggests that students with learning disabilities do have construct irrelevant variance on math tests when the reading level is difficult, but their mathematical abilities may not be equal to those without disabilities. There is a possibility that the Matthew effect could potentially explain the difference in results since Bolt and Thurlow consider older students, which could add the variable of lower math skill growth concurrent with the increasing gap in reading skills over time. Hence, item level research indicates that construct irrelevant variance due to measuring the unintended reading construct can be eliminated and in essence, even the playing field, though its success varies by mathematics skills. Research into content areas like science, which also has word problems and difficult reading passages, is an area in need of additional research.

RAA conclusion

RAA research studies primarily focus on students with learning disabilities and English language learners. Some research on English language learners suggests that students with the most content acquisition benefit the most from the RAA, while other research suggests that it is students with the lowest proficiency in English (Enriquez, 2008). While there is still debate whether RAA alters the construct of reading comprehension for students with learning disabilities or English language learners, among the methods of verifying construct equivalence, results are more conclusive and positive for other subject areas like mathematics. Research into the content areas of Social Studies and Science tends to be minimal, most likely due to the lack of or recent appearance on high stakes testing. Because accountability determines the success of students, teachers, and schools, the validity and fairness of tests for all students in high stakes assessments appear to take precedence over classroom measures, generally in research. As more states include science in their high stakes assessment batteries, increasing evidence is needed to ensure the accurate measurement of science understanding and skills. Since reading appears to provide construct irrelevant variance among students with learning and/or language difficulties in similar constructs, like math, accommodations are necessary for some students in order to acquire an accurate and comparable measurement of all students.

Read aloud Accommodation in Immersive Virtual Environments

While research into the reading aloud accommodation is limited in science, it is not existent in immersive virtual environment-based assessments. Embedded accommodations like pre-recorded audio of text has the potential to reduce the chances of introducing additional construct irrelevant variance, by providing identical conditions. The use of embedded accommodations, designed through universal design for assessment, provide a standardization of delivery of accommodations (e.g. RAA provided via a pre-recorded audio narration for dialogue) (Majerich, Schifter, Shelton, & Ketelhut, 2011), not possible in teacher-provided accommodations due to variation in human readers (Dolan & Hall, 2001; Ketterlin-Geller, 2005). According to Smedley and Higgins (2005), when used for learning, immersive virtual environments and simulations provide students with accommodations such as text-to-speech, which is somewhat analogous to a RAA, and the opportunity to experience content at their own pace. Since brain-based research suggests that all brains are unique (Caine & Caine 1989), allowing students to have capabilities to move at an individual rate is preferred. With immersive virtual environment-based assessments, accommodations must be available through computer software and yet still maintain construct invariance, which can be made easier by providing a multitude of media and options for response so that the format is less important (Rose & Dolan, 2000). According to Almond et al. (2010), rather than being detrimental, the computer-based RAA may be more beneficial to students than the standard teacher RAA because it provides the student with individual, unlimited access and control, combating the issue of students not asking for a re-read when being read to in groups as previously outlined by Dolan and Hall (2001).

According to Dalton and Rose (2008), providing accompanying audio to allow students to listen to digital text in order to alleviate learning disability issues in reading comprehension is a way to incorporate universal design and potentially improve engagement. Because brain based research indicates that brains process both parts and wholes to learn, universal design researchers like Proctor, Dalton, and Grisham (2007) are integrating multiple methodologies to scaffold recognition and expression of knowledge to facilitate a more positive and effective involvement for students within digital learning environments focused on reading comprehension for English language learners. Furthermore, within learning environments, Dalton and Proctor (2007) who similarly integrated digital scaffolds to help with reading comprehension, explain that these scaffolds should vary depending on the difficulty of the task in order to help students develop metacognitive skills about reading; however, this is in reference to a learning environment, not one that acts as a summative assessment. Since the brain discards any sensory information it cannot recognize (Sprenger, 2010), if students cannot read the text within the virtual assessments, their brains may not attend to or be able to process the information. Thus, without an RAA within a virtual assessment, some students may not be able to understand what questions they are being asked, despite visual text and contextual clues.

Pre-recorded voices, that narrate the entire test as many times as students wish to hear it, have been provided in the only published study currently published regarding immersive virtual environment-based assessment and an auditory accommodation. In this study on an assessment on weather curriculum, 31 middle school students randomly received an RAA, while 41 students received no accommodation (Majerich, Schifter, Shelton, & Ketelhut, 2010). Although only borderline significance was found ($p=.07$), mean rubric scores of students who received the RAA were higher than those from the control group. This study did not control for students with learning disabilities or English language learners or determine differential effects by individual characteristics. Providing students with an RAA within a virtual environment allows for multiple modalities to be used fulfilling brain based learning principles while alleviating construct irrelevance for students with reading difficulties.

Conclusion

Based on the review of the read aloud accommodation findings, it is evident that more research is needed to make definitive conclusions on the effectiveness of this accommodation in removing construct irrelevant variance associated with reading difficulties on science assessments, especially within virtual environments. The majority of research pertains to high stakes testing in math and reading, most likely, because that has been the major focus of accountability in the last decade. While results from research studies using a differential improvement model, comparability studies of accommodated and non-accommodated tests, and item level analyses primarily suggested that there are mixed results in reading and language arts, studies show less variance in subjects not typically associated with language, like mathematics. The purpose of this accommodation is to allow a test to evaluate its intended construct validly that may otherwise be inhibited due to unrelated learning difficulties. Generally, research indicates that the effectiveness of the RAA is the same regardless of the delivery, a teacher or a computer. Computer-based RAA can provide students with an individualized experience and unlimited repeats, a disadvantage of group read aloud implementations.

Additionally, providing an RAA via a computer can provide students with a choice in using the accommodation and not quarantine students with learning disabilities or language barriers to separate testing locations. In order to create the most accessible test, designers should use principles of brain based learning and universal design from the planning through implementation stage. Universal design for assessment integrates as many test features as feasible to allow for the customization and choice of accommodations at the individual student level. By making assessments more accessible, more students can benefit via flexible features while fewer resources will be required in the form of teachers or aides. By using brain-based principles, more students can benefit by being assessed more accurately and technology is one vehicle to accomplish this.

Immersive virtual environments, previously used in learning, are now being designed to test scientific knowledge and inquiry skills as well as providing students with a more orchestrated experience. While assessments situated within virtual environments provide visual clues, contextualization, and less of an emphasis on reading, some reading is still required during the final assessment and thus it is important to determine if a reading aloud accommodation is necessary to obtain more valid measurements of student knowledge. Because research into using immersive virtual environment-based assessments is still in its infancy, there is very little existing literature specifically on removing reading construct irrelevant variance from this type of science assessment. This dissertation hopes to begin that dialogue and determine if it is necessary and/or possible.

CHAPTER THREE

RESEARCH DESIGN AND METHODOLOGY

Overview

This chapter describes the research design and methodologies utilized in this study. Additionally, the rationale for the selected research design, unit of analysis and sample, the specific data sources, collection and analyses, and research issues comprise the other important sections of the chapter.

Focusing on the assessment of science knowledge of middle school students from two large school systems in the Mid-Atlantic region, this study seeks to determine how a Read Aloud Accommodation (RAA) in an immersive virtual environment -based assessment affects student performance. Students attending these schools have inquiry-based science curricula in place aligned with State and National Science Standards; however, existing assessments evaluate the students' levels of content learned and inquiry skills developed from exposure to these curricula through long passages to provide situational context or separately in decontextualized questions. Tests comprised mostly of multiple-choice questions are pervasive in these two school systems. With many multiple-choice questions, students are expected to recall only content (i.e., facts, concepts, law theories, laws, principles) and terms about the scientific method (i.e., observation, inference, conclusion, and hypothesis) instead of being assessed on their science content and inquiry skills used to solve science-related problems.

In the immersive virtual environment-based assessments utilized in this study, students used their science content and science process skills to make and describe observations, used tools to extend their observations, make inferences and generate hypotheses when charged with solving a problem, as outlined by National Science Teachers Association [NSTA], (2004) and the American Association for the Advancement of Science [AAAS], (1993). To ensure that the text the students are asked to read is grade level appropriate, the narrative was subjected to the SMOG readability formula and found to be on the eighth grade reading level. The modules were implemented in sixth, seventh, and eighth grade classrooms, which could be a problem for lower level reading students. In order to combat the issue of construct irrelevant variance due to reading, approximately half of students in each class experienced verbal interactions with characters in the virtual environment. Instead of relying on reading skills as a gateway to understanding the questions and available data, students in the treatment group experienced an RAA. Because research into virtual environments is still in its infancy, this has yet to be researched in depth. I proposed that students with the RAA will create better hypotheses and perform more careful observations and measurements because they will better understand the problem via the treatment. In theory, this accommodation will allow students who need help determining the words and those who are auditory learners. Applying findings from previous studies, students with learning disabilities and those who are English language learners, might differentially benefit more by this affordance.

Research Questions

This proposed study will attempt to answer the following research questions:

1. While engaged in an immersive virtual environment assessment, to what extent and how does the expression of science content knowledge and in-world actions differ between students who experience a reading aloud accommodation (RAA) when compared to students who experience text only as indicated by their “sci-tools” usage, interactions with characters and artifacts, and their answers provided at the end of the module?

2. How do significant student actions in question one vary by gender, ethnicity, English language learner status and disability status, if at all?

3. To what extent and how do students perceive the RAA as helpful or distracting?

Rationale for Controlled Experiment

Because students entered either the treatment group or the control group randomized at the student level and this was chosen by a computer, not a person, this study falls into the category of true experimental design (Millsap, 2009). The database assigned participants to either the control or treatment in an every other pattern, as students access the worlds. Each student in every class had an equal chance of belonging to either group, despite any demographic variables. By using true randomized experimental design, threats to internal validity are minimized (Fraenkel & Wallen, 2009). In order to determine if the RAA affected student performance, it was important to control for other differences and only have one manipulated variable, as is common in true experimental design (Campbell & Stanley, 1963; Cohen, Manion, & Morrison, 2003). Because this study is positioned within another as a means of determining if immersive virtual environments are universally designed enough to accommodate all students by including contextualization of test questions, true experimental design using a control group is possible (Creswell, 2009).

Unit of Analysis and Sample

The unit of analysis for this study is the individual middle school student and his or her expression of science content knowledge and processing skills. Each student's answers, in-world interactions, and survey responses were analyzed using quantitative measures mentioned in the analysis section found later in this chapter. Participants in this study were students from the middle grades, consisting of grades 6-8, in several different school districts. These school systems are situated in the Mid-Atlantic region of the United States; one is a large urban school district, while the other is an intermediate unit comprised mainly of near urban districts. Within each class, the students from all nine teachers were randomly assigned to either the treatment or the control group. Although there was no way to alleviate the Hawthorne effect (participants modifying behavior because they know they are being observed) since students knew they are part of a research project, those in the treatment group did not know they were receiving one until after the implementation. All students wore disposable headphones, but some students only heard environmental sounds like animal noises or the sound of rain. Blind students were unable to play this game because the auditory affordance was not enough to serve as an accommodation for them because of the intrinsic visual format. Data were collected over two consecutive school years and implementations, with a total of 791 participants, 285 of them urban students and the remaining classified as near urban (more urban than suburban).

In order to minimize the teacher variability, the SAVE Science project provides professional development for teachers to ensure teaching through inquiry. In order to make certain that students are acquainted and comfortable with maneuvering and participating in a immersive virtual environment before their scientific understanding is assessed, students completed an introductory module before beginning the assessment modules tied to the local and state curriculum.

Participants

In order to obtain data to answer the research questions, 791 subjects from urban and near urban school districts participated in this study across two school years. Although a pilot study was conducted in the 2009-2010 school year, these data have been presented elsewhere and were previously referenced within Chapter Two so they will not be included within this dataset. In the 2010-2011 school year, a total of 445 students from nine different teachers participated, while there were 346 students from five teachers in the 2011-2012 school year, see Table 1. Though there was an additional teacher whose students participated in all modules with a total of 156 students in year one and 172 students for the second year of analyzed data, the results of her students are not included, as she did not implement the RAA treatment with fidelity. Upon observing an implementation in the second year, it was discovered that she did not distribute headphones properly to all students and thus her data were corrupted since there was no way to determine which students actually received the RAA accommodation with headphones. Because of the loss of these data from an urban school district, the number of participants per location is skewed towards near urban participants in the first year and the number of weather participants is small, see Table 2. The districts and schools had varying racial compositions and the values represented in Table 2 are averages across the locations, not necessarily indicative of any one school in particular. As a result, ethnicity may be conflated across the district types.

Table 1.

Participants per module, school year, and school location

Location	Sheep	Basketball	Weather	Total
	Year 1			
Urban	103	0	7 * (54)	110
Near urban	335	0	0	335
	Year 2			
Urban	62	113	0	175
Near urban	93	78	0	171
Total	593	191	7*	791

Note: *54 students completed the weather module, but 47 of these participants also completed the Sheep module and do not count towards the overall total of participants.

Due to constraints, there are missing data on all demographic variables. In year one, the near urban schools did not provide information about English Language Learner status or students with learning disabilities. Moreover, the large urban school district did not provide the same information for sixty students. Further, many students from both school districts did not volunteer their ethnicities; however, from the acquired data, the most common ethnicity is white, see Table 2. The amount of available and missing data varies by demographic variables, see Tables 2, 3, and 4. The grade level with the largest number of participating students was seventh grade, which has data from both school year implementations (see Tables 1 and 3). Within this sample the gender distribution is close to equal, with slightly more females (see Table 3). The proportion of English Language Learners and students with learning disabilities is typical of classroom populations, even though it appears low in comparison within this sample (see Table 4). Analyses presented later in this chapter were performed with only the available data being considered.

Table 2.

Distribution of Ethnicities in Sample

	Frequency	Percent
White	384	48.5
African American	94	11.9
Latino/a	78	9.9
Asian (including Indian)	66	8.3
Mixed	18	2.3
American Indian	8	1.0
European	4	.5
Other (Not specified)	5	.6
Missing	134	16.9
Total	791	100

Table 3.

Participant's gender by grade

Gender	Grade			
	Sixth	*Seventh*	*Eighth*	*Total*
Male	50	232	76	358
Female	44	265	90	398
Missing	0	9	25	35
Total	94	506	191	791

Table 4.

Participant's status as English Language Learners or Learning Disabled

College	English Language Learner	Learning Disability
Yes	58	32
No	612	250
Missing	121	509
Total	791	791

Data Sources

In order to determine the extent the auditory affordance affects students when in the SAVE Science assessment, I gathered several types of data and compared the results of the control and treatment groups. Data were gathered via surveys as well as students' actions and answers from the virtual environment-based assessment.

Surveys

Pre-survey.

This initial survey obtained demographic information, such as gender and ethnicity. In addition, the survey focuses on other constructs unrelated to this study, such as, science anxiety through a modified version of the Attitudes towards science inventory (mATSI; Goglin & Swartz, 1992), and self-efficacy in scientific inquiry via the Self Efficacy in Technology in Science (SETS) survey (Ketelhut, 2007).

Post-survey.

With identical questions, this survey is similar for students in the control group with the exception of the omission of the request for demographic information. Students who received the RAA received an additional seven questions randomized across constructs that are specifically designed to elicit their feelings about receiving this treatment. Because research in the area of auditory affordances in immersive virtual environment assessments is so new, there are no validated surveys currently. These questions were split among two separate Likert scales: a five point scale from "Strongly disagree" to "strongly agree" and a four point scale from "Not at all" to "To a great extent", respectively, the survey questions appear in Table 5

Table 5.

RAA Survey Questions by Likert Scale

"Strongly disagree" to "Strongly agree"	**"Not at all" to "To a great extent"**
1. I wish I did not have to listen to the characters speak.	1. I listened to the characters talk to me.
2. Hearing the assessment questions at the end helped me understand what I was being asked.	2. It was helpful to me when I heard the character explain the problem to me.
3. The characters speaking to me was a distraction.	3. I gathered information from the characters by reading the text.
	4. I gathered information from the characters by listening to them speak.

In-world Documentation

Data were gathered from all three assessment modules: Sheep Trouble, Basketball, and Weather. Each assessment module is geared towards a specific content area. Sheep Trouble focuses on adaptation and speciation, Basketball assesses understanding of gas laws, and Weather evaluates fronts and air masses. Each module has different "sci-tools," a set of virtual scientific measuring devices, to gather data and determine whether they find the information important enough to record into their virtual "notes", see figure 1. Below is a more in-depth explanation of each assessment module.

Figure 1.

Sci-tools from the Basketball Module

The Sheep Module.

This module examines student understanding of adaptation and speciation. Students are transported back to medieval times to help Farmer Brown, a farmer who imported new sheep to increase the size of his flock. After their arrival, the new sheep started to lose weight and the townspeople are convinced that these animals are cursed and need to be killed so the bad magic does not spread. Students are tasked with determining if there is a scientific explanation for why the sheep that were once thriving at a farm on Flatlands Island are dying or if the sheep really are cursed. Within the world, students can interact with two characters and 16 sheep. They have access to tools that measure weight loss; the length of ears, legs, and bodies; age; and gender. Further, there are four other tools to aid in the investigation: a notepad to scribe observations, a clipboard to save important measurements or conversations, a help button that functions as a glossary, and a graphing tool that will visually display a comparison of selected measurements. Once students complete their investigation, Farmer Brown asks them one open-ended question, two multiple-choice questions, and four variable change questions. The answers to these questions along with their measurements and interactions are used to determine if the reading aloud accommodation treatment affects students expression of science understanding.

Basketball module.

In this module, students enter a cartoon virtual environment where a basketball tournament is being held. One location of the tournament is inside a gymnasium, while the other is an outdoor park. It is an unseasonably cold day as there is snow on the ground, but according to characters when this tournament was held the year before, it was warm. When students interact with the lead character, Julius, they discover that the basketballs located at the outdoor court are not bouncing very high, even though they were functioning correctly when they originally left the indoor gym. Students are tasked with determining why the basketballs are no longer functioning properly. In the virtual assessment, students can interact with eight characters and perform up to five measurements (e.g. temperature, bounce height, pressure) on six different basketballs and/or balloons, see Figure 1 (Chapter 3). Because the students can transport balloons and basketballs between locations, they could actually double their measurements in an attempt to eliminate possible variables. To aid in the synthesis of data, students have access to the same four non-measuring science tools as the sheep module. Once students finish gathering and analyzing their data, Julius asks them two multiple-choice questions, three variable choice questions, and one open ended question. The answers to these questions along with their measurements and interactions are used to determine if the reading aloud accommodation treatment affects students expression of science understanding or amount of in- world involvement.

The Weather Module.

This module is designed to complement the sheep module, in that they are set in the same time period and have some of the same characters. As mentioned earlier in this chapter, only 54 students completed this module and 87 percent of these students completed the sheep module as well. The seven students who only completed the Weather module either were absent on the day of the sheep implementation or did not have consent at that time, as the two teachers who implemented weather had previously provided the sheep module to their students. Thus, for most students who completed this module, it was their second time in the virtual town of Scientopolis. Within this module, students are informed that Scientopolis is experiencing a drought. There has not been rain for over 30 days and the crops and animals are starting to expire. The townspeople are again afraid that Scientopolis is cursed and that they must move away to escape this. Farmer Brown has lived in Scientopolis his entire life and does not want to move. He asks students to investigate and determine if there is a scientific reason behind the drought or if the town really is doomed. Students have access to several artifacts like newspapers, paintings, and maps to gather data. They can also use their sci-tools to perform three different measurements: wind direction, temperature, and pressure. Students have access to five non-measuring tools to help them acquire and synthesize their data. The tool that is additional to that of sheep and basketball is a teleporter that provides them access to three towns within the weather module. This tool allows them to perform measurements in separate locations in order to determine where fronts lie using a map. Within this module there are seven characters including Farmer Brown for students to interact with and gather information. After they have developed a hypothesis to the cause of the drought and whether it will ever end, students engage in a sequence of questions with Farmer Brown. He asks students six questions: one open-ended, two multiple-choice, and three variable choice. It is important to note that a typo in one of the multiple-choice questions nullified the correct answer, so it was removed from the subsequent analyses.

From these three modules, every action students performed while in each of the immersive virtual environment assessments was recorded into a database. In each world, students' interactions with characters and measurements performed using their sci-tools are recorded per student. By comparing character interactions and measurements by treatment group, statistical tests can illustrate if the reading aloud accommodation makes students more or less active during the assessment. Because the sci-tools are different in each assessment and more measurements allow students to eliminate possibilities and determine an answer, the frequency of overall usage was calculated.

While students' information gathering is important, how they use it to illustrate their science content knowledge by answering the questions was also considered. In each immersive virtual environment assessment, a lead non-player character (NPC) provides a series of questions for the students to answer at the end of each module. The questions are a mix of multiple-choice questions, variable changes, and open-ended constructed responses, see Table 6 The multiple-choice questions and variable change questions have specific, objectively scored responses, while the constructed response were scored via a 5-point rubric, see Table 7.

Table 6.

Example Questions from Modules

Question Type	Example from Sheep Module	Example from Weather Module	Example from Basketball Module
Multiple Choice	Which is LEAST LIKELY to be an adaptation for sheep survival in Farmer Brown's hilly farm? 1. Short legs 2. Bright Coloration 3. Large Gripping Hooves 4. Excellent Vision	Warm, wet weather covers eastern Pennsylvania from the south in April. What type of air mass is the most likely cause? 1.Maritime Polar 2. Maritime Tropical 3.Continential Tropical 4. Continental Polar	A sample of oxygen is being stored in a closed container at a constant temperature. What will happen to the gas if it is transferred to a container with a smaller volume? 1. Its weight will increase 2. Its weight will decrease 3. Its pressure will increase 4. The size of its particles will decrease
Variable Change	What variable could you change on the NEW sheep to help them? <u>Make the gender</u> 1. Male 2. Female 3. It doesn't make a difference <u>Make the legs</u> 1. Longer 2. Shorter 3. It doesn't make a difference <u>Make the age</u> 1. < 3 years 2. > 3 years 3. It doesn't make a difference	How would you change these variable to make it rain in Scientopolis? <u>Pressure</u> 1. Increase 2. Decrease 3. It does not make a difference <u>Wind direction</u> 1. Make it blow from the North 2. Make it blow from the South 3.Make it blow from the East 4.Make it blow from the West	What variable would you change to correct this basketball problem? <u>Temperature</u> 1. Make it 75^{o}F 2. Make it 55 ^{o}F 3. Make it 35 ^{o}F <u>Court Type</u> 1. Concrete only 2. Wood only 3. Court Type makes little to no difference <u>Basketball used</u> 1. Replace one Wade Park ball with one Jordan Gym Ball 2. Purchase a new set of balls for Wade Park 3. New basketballs will not help this problem
Open-Ended	"Can you tell me why the new sheep are dying? What are 3 sources of evidence I can give the judge to support this answer?"	"Why do you think it has not been raining? What evidence do you have to support this idea?"	"What is wrong with the basketballs? Please give me three sources of evidence to support this"

Table 7.

Open ended scoring rubric with example responses from the Sheep Module

Score	Criteria	Student responses for the question, “Can you tell me why the new sheep are dying?”
0	Provides no hypothesis	-----
1	Provides a hypothesis	"one ate a tape worm then passed it on to there offspring” <u>Evidence:</u> “they are losing weight rapidly”
2	Provides a somewhat correct answer	“well i noticed that the sheep that are uphill have lost few pounds while the ones down here are loosing 80 pounds and over. i guess that the food they need is up in the hill. try taking them up the hil” <u>Evidence:</u> “the sheeps down here are loosing weight because they dont have the food they need”
3	Provides a correct hypothesis with only folk or incorrect evidence	“They are not adapted to this land.” <u>Evidence:</u> “They are used to flat lands.[…]They don't eat this type of grass.”
4	Provides a correct hypothesis with supporting data gathered from within the world	“ Th sheeparenot getting enough food. The sheep on top of the hill have adapted to climbing up the hill to where the food is while the others on the ground have no food and are losing wieght and dying.” <u>Evidence:</u> “Most of the sheep on have the right size of legs to supprot their bodies when they walk up the hill.[…] All the sheep on the ground have hada severe drop of weight.[…]The sheep on the hill have lost little wieght.”

Procedures

Before interacting with even the introductory module, students assigned unique ids and completed the pre-survey in order to obtain demographic information. After the introductory survey, students participated in an introductory assessment that is used to acclimate them to the virtual environment. Next, students completed the contextualized science assessment by solving the embedded problem. Following the completion of each assessment module, students answered individual wrap-around questions and a post-survey. Subsequently, the researcher scored student's open-ended responses with a rubric and calculated the number of student's interactions with characters and artifacts as well as measurements with sci-tools. Open-ended responses, short answers, and multiple-choice questions were analyzed as an overall percentage score per module to determine any differences in performance between the control and treatment groups, as outlined further in the data analysis section.

Quantitative Data Analysis

The instruments previously mentioned helped the researcher obtain a variety of sources of evidence to support claims. Findings from the statistical analysis of objective, numerical data are the basis for claims. The first research question was answered in a quantitative manner and outlined specifically below.

Research Question 1: While engaged in an immersive virtual environment assessment, to what extent and how does the expression of science content knowledge and in world actions differ between students who experience a reading aloud accommodation (RAA) when compared to students who experience text only as indicated by their "sci-tools" usage, interactions with characters and artifacts, and their answers provided at the end of the module?

As mentioned previously student tool usage and interactions with characters and artifacts were automatically recorded into the project-created database. Each of these were considered as separate dependent variables, as was an overall score percentage per module. In order to create an overall score percentage for displayed science content knowledge, scores from multiple-choice questions, short answer questions, and the constructed response were combined to form one variable average per student. Then a one way MANOVA was run on the three dependent variables: overall score percentage, number of interactions, and number of measurements for each module separately with treatment as the independent variable. Determining if any significant differences are stronger among subgroups or only existent among these more specific groups is important and thus the purpose of the second research question.

Research Question 2: How do student actions in question one vary by gender, ethnicity, English language learner status and disability status, if at all?

The second research question was answered using significant dependent variables from the first research question. Using a multiple regression, the researcher looked for significance on overall score percentage as the criterion variable and gender, ethnicity, English language learner and disability statuses, treatment and their respective interactions with the treatment as predictor variables. Because ethnicity is categorical and not dichotomous, this was dummy coded in order to be used in the analyses.

Research Question 3: *To what extent and how do students perceive the RAA as helpful or distracting?*

To gain an overall view of student feelings towards the affordance, those in the treatment group took a researcher-created, seven question survey. In order to provide illustrative details of the findings, the in-depth responses acquired during paper-based reflections of five students from a random stratified sample of varying performance and demographics from the treatment group are included (e.g. answers from open-ended reflection question, “Did you hear the characters speak to you? If yes, did this affect your information gathering and question answering? Please explain. If not, would you have liked to have heard the characters speak to you? Why or why not?”).

Together the answers to the three research questions allowed the investigator to determine if the RAA treatment has any effect on virtual science assessment outcomes and to what extent.

Research Issues

This study is designed to look at the implementation of an audio component to the SAVE Science project. The aforementioned project is a science assessment contextualized in an immersive virtual environment. The audio piece was researched to determine if it helped students understand the questions being asked of them. It was hypothesized that because the reading barrier is diminished, this environment could potentially increase engagement and participation and thus scores. Currently, reading is being measured as an unintended, irrelevant construct on many high and low stakes tests, while the focus should only be content specific to skills and knowledge. In addition, this puts students with learning disabilities, especially reading ones, and English language learners at an even greater disadvantage on assessments. The only ethical issues that could be raised is if all the students from the RAA treatment group performed better and we did not allow all students access after discovering this. Other than that, the only issues that arose are my personal biases.

Another bias comes from my experience as a teacher. During my tenure as a high school science teacher, I had quite a few students with learning disabilities and students who were English language learners. These students often did poorly on exams. When I worked with them individually or another teacher read the test to the students, they almost always did better. When I tutored them one-on-one, I would read test questions that they answered incorrectly. The majority of the time, the students could immediately tell me what the right answer was and would remark how easy the question seemed. For many of my students, the issue seemed to be in processing information from paper to their mind. While my experience has been with high school students and the SAVE Science project focuses on middle school students, I think they may have similar issues.

My expectations were that most students would perform better on assessments that have the auditory support. Since the dialogues were recorded and students could listen to them as many times as they wanted, they are getting to access two different modalities. I hypothesized that this would eliminate the accidental measuring of the construct, reading. As long as students can hear, they have access to the problems and all the textual data in the world. Even if students could not read, they could still use science skills to investigate the problem, potentially solve it, and express their content knowledge and inquiry skills. I had to be objective in my interpretation of their hypotheses.

Reliability and Validity

Because research is so new in the realm of educational assessments via immersive virtual environments, there are some issues of reliability and validity of this assessment. Few studies exist on the accuracy of immersive virtual environment-based assessments as an accurate and reliable measure in comparison to paper-and-pencil testing. However, it has been argued that the two types of assessments are not congruent and are measuring different things. There is a chance that students will not take this test with full fidelity in that it is up to the teacher's discretion if any stakes, like a grade, are attached to the immersive virtual environment assessment. Furthermore, because it looks like a video game, students may take it less seriously because it seems fun.

Limitations

Within one of the participating school districts, schools that were in danger of being taken over by the state for not meeting Adequate Yearly Progress (AYP) were not eligible to participate, even though they might be the schools with students that would benefit the most. Therefore, it limits the sample to the students in schools that are already at least performing at the state minimum, not schools with the lowest performance rates on high stakes testing.

In addition, the research team is comprised of five members to gather data during implementations, which means that some schools were missed or only auditory recordings of implementations were available. Further, participating schools are all from one Mid-Atlantic state, with only urban and near urban demographics being represented. The lack of representation among students from other regions, states, and the rural environment could potentially limit the generalizability of findings. Moreover, in order for students to be eligible for participation their teachers must volunteer to be part of the research project SAVE Science, which requires a time commitment above required school district duties. Thus, students in the study may be exposed to different types of teachers (e.g., motivated teachers). By using a random sampling within class, it is hoped that some of these limitations are minimized.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

Introduction

This study was conducted to assess the effects of providing a Reading Aloud Accommodation (RAA) as a treatment on science assessments situated within virtual environments. Research indicates that many paper based science assessments evaluate reading in addition to scientific understanding due to construct irrelevant variance. Within classrooms experiencing traditional assessments, students with learning disabilities or those not yet proficient in English often are provided with a reading aloud accommodation to alleviate this construct irrelevance. While designers of virtual assessments posit that their tests evaluate student content understanding more validly by reducing the length of test questions through contextualizing them within a virtual world, it is important to understand if that is indeed enough. In order to determine if the RAA is useful within virtual assessments, whether the effectiveness is influenced by demographic variables, or student's perceptions about this accommodation, participants were randomly assigned to the control or treatment group and each completed a virtual science assessment. The remainder of this chapter provides details about the sample of participants, any explanations as to why some data were missed or removed, as well as the presentation of data and respective analyses that answers the research questions.

Data Analysis

In order to answer the three research questions posed in Chapter Three, a combination of statistical tests were performed. Within this section, each research question is restated followed by a presentation of the chosen analyses and results. The interpretation of the presented results will be explained in Chapter 5.

Research Question 1.

> While engaged in an immersive virtual environment assessment, to what extent and how does the expression of science content knowledge and in world actions differ between students who experience a reading aloud accommodation (RAA) when compared to students who experience text only as indicated by their “sci-tools” usage, interactions with characters and artifacts, and their answers provided at the end of the module?

Since all modules assess different content, have varying amounts of in-world objects to interact with and measure, as well as evaluate students in different grades, each module was considered separately to answer this question. Further, since there are only results for basketball and weather in one year each, results are presented by module rather than implementation year.

The Sheep Module.

In order to ensure that the two years of sheep implementations were not statistically different and could be compared as one sample, an independent samples t-test was performed. The t-test was conducted to compare the percentage of questions correct in year one and two. There was no significant difference between the scores for year one ($M=43.91$, $SD=19.05$) and year two ($M=46.16$, $SD=17.66$); $t(590) = -1.287$, $p = .199$, see Tables 8 and 9. These results suggest that the year of implementation does not have an effect on student percentage and they can be analyzed as one sample.

Table 8.

Group statistics for Overall Percentages by School year

School year	N	Mean	Std. Deviation
1 (2010-2011)	438	43.9083	19.04955
2 (2011-2012)	155	46.1629	17.65525

Table 9.

Values for Independent t-test for Overall Percentages by School year

	t-test for Equality of Means		
	t	Df	Sig. (2 tailed)
Equal variances assumed	-1.287	590	.199

To determine if and how the expression of science content knowledge and in world actions differed by treatment within the sheep module a multivariate analysis of variance (MANOVA) was run on these three variables. While it would have been possible to run three separate ANOVAs, the results would have a greater chance for a type I error and the correlation between variables would not be considered. The multivariate result was significant for the treatment within the sheep module, Wilks' λ = .983, *F* (3, 583) = 3.321, p=.002, partial eta squared = .017, while the power to detect the effect was .756, see Table 10. Given the significance of the overall MANOVA, the univariate main effects were examined. Significant univariate main effects were only present for the overall score percentage, F(1)=9.744, p=.002, partial eta squared=.016, power =.876, indicating a difference in score on the sheep questions between students in the treatment and control groups. The results for measurements and collisions (e.g. using sci-tools or interacting with characters) were not significant among treatment groups, see Table 11. Furthermore, in comparing mean differences, the treatment group had a higher mean by 4.87, see Table 12. This finding indicates that on average those receiving the RAA outscored those in the control on the overall sheep module percentage by close to five percentage points, while it had no effect on measurements or collisions for the Sheep module.

Table 10.

MANOVA Results for Treatment and Percentage Correct in Sheep Module

	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Observed Power
Wilks Lambda	.983	3.32	3.000	583.0	.020*	.017	.756

Note: Indicates p<.05*

Table 11.

Between Subject Effects for Treatment by Percentage Correct in Sheep Module, Measurements, and Collisions

	Depend. Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Part Eta^2	Obsd Power
Treat.	sheep_%	3369.89	1	3369.9	9.74	.002*	.016	.876
	Measure	395.086	1	395.09	.07	.791	.000	.058
	Collision	187.950	1	187.95	.16	.686	.000	.069
Error	sheep_%	202321.7	585	345.85				
	measures	3277633.1	585	5602.8				
	Collision	672513.6	585	1149.6				

Table 12.

Means and Standard Deviations for Percentages and In-world actions by Treatment group.

	Control (n=346)		Treatment (n=242)	
Dependent Variable	Mean	SD	Mean	SD
Sheep Percentage Score	42.564	18.61	47.435*	18.785
Measurements	57.85	72.062	59.52	78.619
Collisions	38.41	36.967	37.26	28.963

Notes: $p<.05$, Within the database, 5 students' treatment statuses were unknown*

The Basketball Module.

Similar to the analysis of the sheep module records, a MANOVA was used to analyze the student in-world actions. During this implementation, there were 87 and 104 students in the control and treatment groups, respectively. According to the multivariate test, there was no overall significance between treatment and control groups, among percentage of correct answers, measurements or collisions, Wilks' λ = .988, *F* (3, 187) = .740, p=.529, see Table 13. In comparison of the dependent variables, the treatment group had a higher mean in all three categories; however, none of these values accounted for any significant part of the variance between the two groups.

Table 13.

MANOVA Results for Treatment in the Basketball Module

	Value	F	Hypothesis Df	Error df	Sig.	Partial Eta Squared	Observed Power
Wilks Lambda	.988	.740	3.000	187.000	.529	.012	.206

The Weather Module.

Like both the Sheep and Basketball modules, a MANOVA was used to determine if the treatment affected the expression of science content knowledge or their interactions with characters or tools within the assessment. Thirty-three students were randomly assigned to the control group, while 21 were provided with the reading aloud treatment. Analogous to the basketball module results, in a pairwise comparison, students in the treatment group had a higher mean average of overall score percentage, number of measurements, and number of character interactions, but the results were not significant, Wilks' $\lambda = .505$, $F\ (3, 50) = .505$, p=.680 , refer to Table 14.

Table 14.
MANOVA Results for Treatment in the Weather Module

	Value	F	Hypothesis Df	Error df	Sig.	Partial Eta Squared	Observed Power
Wilks' Lambda	.971	.505	3.000	50.00	.680	.029	.146

Overall Results.

The only finding that presented statistical significance was the overall percentage score in the sheep module. Students in the treatment group earned, on average, 4.87 percentage points higher in comparison to those in the control group. All other results from performed MANOVAs indicated no significant difference between the treatment and control groups, though the results did follow a similar pattern to those of the sheep module.

Research Question 2.

> How do student actions in question one vary by gender, ethnicity, English language learner status and disability status, if at all?

Only students who completed the Sheep module and Basketball module are included in this regression because the weather module would allow some students to be in the participant sample more than once. Therefore, a multiple regression was run with all demographic variables as predictors.

Multiple Regression

In order to investigate if and how demographic variables affect the overall percentage performance on the sheep and basketball assessments differentially by treatment, a multiple regression was conducted with the module questions percentage as the outcome variable and the following predictor variables: gender, English Language Learner (ELL) status, Learning disability (LD) status, and treatment as well as the interactions between the variables. The interaction variables were created by subtracting each predictor from its mean value, in order to alleviate multicollinearity problems. Table 15 summarizes the descriptive statistics and analysis results and the sample size does not include all participants as missing values were excluded listwise.

As can be seen both English language learner and learning disability status are both positively and significantly correlated with the criterion, indicating that those with higher values for these variables tend to have a higher overall percentage of correct answers within the module. Because students with these barriers were coded with lower numbers, English Language Learners and learning disabled students are more likely to have a lower percentage score. Furthermore, the interactions between both gender and learning disability and treatment is significantly and positively correlated. Because

females were coded as the higher number as was the treatment group, females who received the treatment were more likely to score higher than their counterparts. While the interaction between learning disability status and treatment was positively statistically correlated, learning disability alone had a higher p value. The multiple regression model with all fifteen predictors produced $R^2 = .130$, $F(15, 247) = 2.456$, $p = .002$.
Thus, this model accounts for 13 percent of the variance in scores. Shown in Table15, English language learner and learning disability status had significant positive regression weights, indicating students without learning barriers were expected to have higher percentage scores on module questions, after controlling for the other variables. Gender, treatment, ethnicities, and other interactions among variables did not contribute to the multiple regression model.

It is important to note that the regression models run to answer this research question used a listwise deletion, which drastically reduced the sample size due to missing data. The status of Learning disability was missing for almost 400 students which precluded them from this analysis. See the Appendix A and B, for alternate representations of this analysis including pairwise deletion and the exclusion of Learning Disability status.

Table 15.*Correlations between Percentage Scores per module and Predictors (n=263)*

Variable	Mean	SD	Pearson Correlation	Sig. (2 tailed)	MR Weights b	β
Gender	1.55	.499	.035	.284	.478	.011
ELL	1.87	.336	.195**	.001	14.878	.224**
LD	1.88	.323	.222***	.000	12.629	.183*
Treatment	1.48	.500	.063	.153	3.525	.079
African American	.1445	.35225	-.058	.17	-4.206	-.066
Latino/a	.1445	.35225	-.050	.210	-.980	-.015
Asian	.1673	.37396	-.003	.481	3.377	.057
Other	.0646	.24636	.020	.373	3.103	.034
Gender*Treatment	.0033	.25125	.122*	.024	8.552	.096
ELL* treatment	.0012	.16914	-.023	.353	-4.733	-.036
LD * treatment	-.0091	.16620	.148**	.008	14.553	.109
Af. Am * Treat.	.0116	.18113	-.006	.463	-.644	-.005
Latino*Treatment	-.0134	.17140	-.012	.426	-4.889	-.038
Asian* Treatment	-.0027	.18888	.024	.348	-2.975	-.025
Other* Treatment	.0006	.12406	-.08	.097	-14.44	-.080

Notes: * p < .05 ** p < .01 ***p<.001

In order to simplify the model, a second multiple regression was run by discarding some of the predictors that did not contribute to the model. The reduced multiple regression included overall percentage per module as the criterion and only English language learner status, learning disability status, treatment, and their respective interactions with the treatment variable as predictor variables. Eliminating the ethnicity variables increased the sample size, due to missing values. While the Pearson correlations for the above module remained significant, treatment also correlated significantly with percentage score in the second analysis, see Table 16. Because students in the treatment group were coded as the higher number, the positive correlation indicates that students in the treatment group were more likely to have a higher percentage score on the module questions. However, in the reduced multiple regression, no additional predictor variables were found to significantly predict for this model. The multiple regression model with only five predictors produced $R^2 = .111$, $F(5, 318) = 7.848$, $p = .000$. While this model accounts for slightly less of the variance, 11.1 percent, it has ten less predictors and a more significant p value.

Table 16.

Correlations between Percentage Scores per module and Predictors (n=324)

					MR Weights		
Variable	Mean	SD	Pearson Correlation	Sig. (2 tailed)	b	β	Sig.(2 tailed)
Treatment	1.48	.500	.147***	.000	2.975	.069	.196
ELL	1.89	.311	.106**	.004	14.390	.208***	.000
LD	1.87	.336	.242***	.000	14.426	.225***	.000
ELL*Treatment	.0006	.15629	.001	.973	-1.255	-.01	.865
LD*Treatment	-.010	.17332	.127*	.022	11.515	.093	.092

Overall Results.

Though treatment, the interaction between treatment and learning disability, as well as the statuses of learning disability and English language all correlate significantly with the overall answer percentage, only the two barriers to learning and assessment significantly predict the variance in scores within the multiple regression. No other demographic variables or interactions between treatment and these variables significantly predicted for overall percentage score in the Sheep or Basketball modules.

Research Question 3.

To what extent and how do students perceive the RAA as helpful or distracting?

In order to answer the third research question, only students in the treatment group are considered, since the control group did not experience the accommodation. Further, the survey and written debriefing questions were presented to the students in year two only. The resulting sample includes 183 students, 88 from an urban school district and 95 from two near urban middle schools. Seventy of these students did not complete the survey, 51 from urban schools and 19 from near urban schools. Thus, the sample for survey questions is only a total of 113 students. Descriptive statistics of students' answers to the survey questions are provided in addition to four randomly selected responses that students supplied to illustrate in-depth feelings about the treatment.

Survey Questions.

The first three questions of the survey had students rate statements on a five-point likert scale ranging from strongly agree (1) to strongly disagree (5), while the remaining four used four-point likert scale ranging from to a "not at all" (1) to "to great extent" (4), refer back to Table 5. All survey responses fall within the acceptable skewness of -2.0 to + 2.0, see Appendix C.

Survey Question 1. The first survey question asked students to respond to the statement, "I wish I did not have to listen to the characters speak." As shown is Appendix C, scores averaged 3.39 with a mode of 3, indicating that overall students reported neutral feelings about listening to the character speak to them within the virtual environment. Close to half of students disagreed or strongly disagreed with wishing they did not experience the treatment, while less than 20 percent agreed to some extent, see Table 17.

Table 17.

Frequency and Percentage of students not wanting to listen to the characters.

	Frequency	Percent
Strongly Agree	6	5.3
Agree	16	14.2
Neutral	41	36.3
Disagree	28	24.8
Strongly Disagree	22	19.5
Total	113	100.0

Survey Question 2. In order to determine if students found hearing the questions helped them with comprehension, students responded to the following prompt, "Hearing the assessment questions at the end helped me understand what I was being asked." The most frequent response as denoted by the mode was agree (2) and the mean average of 2.44 is closest also to that response, see Appendix C. Fifty-seven percent of students who answered the survey agreed that the affordance helped their comprehension, while a quarter reported neutral feelings, see Table 18. Only 18.6 percent of students said that hearing these questions did not help them comprehend the questions.

Table 18.

Students reports of if the RAA helped them understand the test questions

	Frequency	Percent
Strongly Agree	21	18.6
Agree	44	38.9
Neutral	29	25.7
Disagree	15	13.3
Strongly Disagree	4	3.5
Total	113	100.0

Survey Question 3. In the interest of ascertaining if students felt that the treatment diverted their attention in an unhelpful way during the virtual assessment, students responded to the statement, "The characters speaking to me was a distraction." Participants reported being neutral or disagreeing with the characters talking created a distraction during the assessment, (M=3.55, SD=1.12), see Appendix C. Approximately 55 percent of students disagreed to some extent that the characters speaking was a distraction, see Table 19.

Table 19.

Students reports as to whether the characters speaking was a distraction

	Frequency	Percent
Strongly Agree	6	5.3
Agree	13	11.5
Neutral	31	27.4
Disagree	38	33.6
Strongly Disagree	25	22.1
Total	113	100.0

Survey Question 4. This survey question was the first of the survey items to provide a different likert scale. The four point likert scale did not include a neutral option with "not at all," "very little," "somewhat," and "to a great extent" as the possible responses coded as 1-4, respectively. In order to ensure that students used the audio affordance and did not just remove the earphones or ignore the characters, they were asked to respond to the sentence, “I listened to the characters talk to me.” Seventy-six percent of students chose listened to the characters *somewhat* or *to a great extent*, see Table 20. This is further evidenced with a mean of 2.95 and a mode of 3, refer to Appendix C.

Table 20.

The extent that students listened to the characters speak.

	Frequency	Percent
Not at all	15	13.3
Very Little	11	9.7
Somewhat	52	45
To a great extent	35	31.0
Total	113	100.0

Survey Question 5. Similar to the second survey question, which asks specifically about the RAA helping with comprehension of the test questions, this question investigates whether students found it helpful to hear the characters describe the in-world problem. Students responded to the prompt, “It was helpful to me when I heard the character explain the problem to me.” Possible responses were from the previously defined 4-point likert scale. Sixty-nine percent of students expressed that hearing the character explain the problem was at least somewhat helpful to them during the virtual assessment, see Table 21. The most common response was “somewhat” and the calculated average is very close to its associated coded value (M=2.82, SD=1.06), see Appendix C.

Table 21.

The extent students said hearing the characters explain the problem was helpful.

	Frequency	Percent
Not at all	20	17.7
Very Little	15	13.3
Somewhat	43	38.1
To a great extent	35	31.0
Total	113	100.0

Survey Questions 6 and 7. In order to determine how students gathered information during character interactions, two similar questions were asked. Question 6 had students rate the extent to which they gathered information by reading text, while the last survey item had rank the degree to which they did so by listening to the characters. Both questions used the four point likert scale. These survey items were meant to determine if students acquired more information through the RAA than text alone; however, the two questions are not necessarily dichotomous or exclusive.

"I gathered information from characters by reading the text" is what students responded to in question 6. Overwhelmingly the majority of students specified that they gathered information via reading the text. Over 40 percent of students responded that they acquired information this way to a great extent, which was the mode result, see Table 22. Further, by adding students who responded very little and somewhat, the percentage becomes 93.7 percent of all students, refer to Table 22. This is furthered supported by the average score for all students who completed this question (M=3.13, SD=.905), see Appendix C.

Table 22.

The extent that students gathered information by reading character dialogue.

	Frequency	Percent
Not at all	7	6.3
Very Little	18	16.1
Somewhat	40	35.7
To a great extent	47	42.0
Total	112	100.0

*Note: * One student did not answer this question, thus the mean is calculated with n=112*

The last survey question, "I gathered information from the characters by listening to them speak" had an identical mode and similar increasing trend of students per group as the reported extent of gathering information also increased; however, the average was not as high as question six. The mean was slightly lower, 2.84, indicating less reliance on audio and a slightly larger standard deviation (SD=1.05) indicating a wider variance. Further, only 33.9 percent of students agreed that they gathered information to a great extent by listening to the characters speak, while 13.8 percent reported that they did not obtain knowledge through the reading aloud accommodation, see Table 23. Therefore, overall more students reported gathering details to a great extent about the in-world problem by reading than listening, while more students reported that they did not gather information by listening at all.

Table 23.

The extent that students gathered information by listening to character dialogue.

	Frequency	Percent
Not at all	15	13.8
Very Little	24	22
Somewhat	33	30.3
To a great extent	37	33.9
Total	109	100.0

*Note: * Four students did not answer this question, thus the mean is calculated with n=109*

Individual Student Responses.

In order to provide more in-depth perceptions, four seventh grade students who completed the Sheep trouble module and one eighth grade student who completed Basketball were chosen randomly from the treatment sample and then provided pseudonyms for confidentially. One selected student, Greg, has a learning disability, one student, Augusta, is an English Language Learner, while the three others, Ferris, Vivienne, and Michael, do not have these learning barriers. Ferris, Michael, and Greg all listed their ethnicity as white, while Vivienne is Asian and Augusta is Latina. All five students were asked to answer the following question in a written debrief with respect to experiencing the RAA, “did this affect your information gathering and question answering? Please explain.” If this followed the trend of previously displayed data from the first research question, the overall average answer would be “no” to the gathering aspect and “yes” to the question answering part.

Vivienne. Vivienne, a female student, earned an 81.8 overall percentage in the sheep module. Of the five cases, she was the only student who was overtly negative about experiencing the RAA. Vivienne stated, “it was a bit distracting in my opinion, because I often read faster than the character talks." This explanation indicates that listening at a different speed than she reads unnecessarily diverted some of her attention when interacting with the characters, though she does not specifically address whether this affected her information gathering or how she answered the questions.

Michael. Michael, a male student, answered 81.8 percent of the sheep questions correctly, just as Vivienne did. However, unlike Vivienne, Michael did not report disapproving feelings with respect to the RAA. He indicated neutral feelings towards the RAA in explaining, "No it just let me hear it instead of read it." This response suggests that the treatment allowed for audio support, but this did not help or hinder his in-world processes.

Greg. Greg had a response similar to Michael, yet he had the lowest percentage of correct answers in the sheep module, 36.4 percent. He explained that hearing the characters talk to him did not affect his information gathering or question answering. In response to the open-ended question, he stated, "No it's just easier then reading it." Therefore, according to him, it did not affect his performance but did make the assessment easier. This is somewhat unexpected in that he has a learning disability and this is a possible associated accommodation for paper and pencil assessments. Further, his score is very low, although it could have been lower without the accommodation.

Ferris. Ferris is a male student who answered 45.6 percent of the sheep questions correctly. He is the only student of these five who does not have a learning barrier and who expressed that the RAA was helpful and affected his in-world actions. Ferris qualified the effect by responding, "Yes, because I could hear them and I didn't misread." In this statement, he implies that hearing the characters enabled him to acquire correct information. Ferris does not explain whether he has trouble decoding words when reading or goes through the process too quickly and misunderstand the meaning of the passages. He just provided the information that the RAA ensured that he correctly obtained what the characters were conveying during the interaction and that this helped his information gathering and question answering.

Augusta. Augusta is a female student whose first language is not English. Within her school district, she is still classified as an English Language Learner, indicating that she does not have a proficient command and usage of English. Within the module, she answered 44.44% of the questions correctly. It is possible that she might not have done as well, without the RAA. When responding to the question as to if and how the RAA affected her experience, she said, “it does because it give information.” She explains that by hearing the characters, she obtains information, perhaps indicating that textual content is not easily accessible to her.

Overall Results.

In general, results indicated favorable feelings towards the RAA treatment and its helpfulness within the virtual assessment. Students were neutral on average of whether or not they wished they did not have to receive the treatment. However, according to mean values, students reported listening when the characters spoke and that this helped them comprehend the problem and assessment questions. Further, the majority of students reported that listening to the characters speak during their experience was not a distraction. The examples illustrate that students have mixed feelings about the accommodation. One student felt it was distracting during the assessment, while two students were neutral in general, and the other two students felt it helped with comprehension and reading accuracy, yet not all of these students performed to a passing standard. Between the survey responses and individual case responses, no consensus is evident.

Conclusion

Data indicate that with respect to the first research question, the RAA treatment influenced student performance on only one module, sheep. There was no significant effect within the basketball or weather module. Furthermore, the RAA did not affect student's in-world data gathering via character interactions or tool usage. With respect to the second research question, Learning Disability and English Learner statuses are the only significant predictors for student assessment performance. Moreover, the effect of the RAA observed in the first research question disappears when demographic variables are considered. The examination and explanation of why the significance of the RAA disappears when considering these variables will be presented within the next chapter. Student's perceptions of the RAA appear to vary. Overall, survey means indicate that students have positive feelings about experiencing a reading aloud accommodation, although consensus is not overwhelming. Within the randomly selected case studies, only two students indicated that it helped them perform better during the assessment. These inconsistencies will be explored during Chapter Five.

CHAPTER FIVE

DISCUSSION

Introduction

This study was conducted in order to determine if a reading aloud accommodation (RAA) is an effective treatment in removing construct irrelevance and evaluating student knowledge in accordance with the brain-based theory of learning within virtual science assessments,. Research indicates that many paper based science assessments unintentionally evaluate unnecessary language skills (e.g. reading) and only provide students with one way to access the information (Oakland & Lane, 2004). Within classrooms experiencing traditional assessments, students with learning disabilities or those not yet proficient in English often are provided with a reading aloud accommodation to alleviate construct irrelevance (Duran, 2008). Thus, to determine the utility of an RAA to better assess students, whether this is affected by demographic variables, or how students perceived this accommodation, participants completed a virtual science assessment as either part of the control or RAA treatment group. This chapter provides interpretations of the data presented within Chapter four.

Data Analysis Interpretation

The analyzed data presented within Chapter four illustrate mixed results for the effectiveness of the Reading aloud accommodation as a treatment.

Research question 1.

> While engaged in an immersive virtual environment assessment, to what extent and how does the expression of science content knowledge and inquiry skills differ between students who experience a reading aloud accommodation (RAA) when compared to students who experience text only as indicated by their "sci-tools" usage, interactions with characters and artifacts, and their answers provided at the end of the module?

In comparison of the measurements and collisions performed within all modules, there were no significant differences between the control and the treatment group for any of the assessments. On average, the overall percentages were higher for students in the treatment group than in the control; however, significant differences were only found within the sheep module. When not considering any other factors, the mean treatment percentage scores in the sheep module (47.44) are significantly higher ($p=.002$) than the mean control scores (42.56). These values indicate that the treatment helps students outperform others who do not receive it by on average 4.87 percent, see Table 12. This finding is concurrent with what would be expected out of brain-based learning research in that the brain processes information multiple ways, and thus by providing audio support to text-based questions, the brain has multiple ways to access the information (Caine & Caine, 1991). Further, as referenced earlier, Block and Parris (2008) provided neuroscience research to indicate that reading involves audio processing as this system is how people learn to understand language before reading instruction begins. Moreover, in accordance with research into the reading aloud accommodation, many studies produce results indicating that students perform better when experiencing this treatment, although primarily in the context of paper-based assessments (Bielinski, Thurlow, Ysseldyke, Friedebach, & Friedebach, 2001; Huynh & Barton, 2006; Laitusis, 2010; McKevitt & Elliott, 2003; Meloy, Deville, & Frisbie, 2000; and Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998).

Within the two other modules, Basketball and Weather, treatment scores were higher but the difference between the two groups was small, as were the sample sizes. The lack of significant differences in overall percentage scores on basketball and weather could be attributed to the small sample size for both. Each module was only implemented with a small subset of the population for one year each. This caused each student's score to weigh in more on the overall average and potentially skew it. Moreover, both of these samples were different from the Sheep sample in that many participants had completed other assessment modules prior. Many students who completed basketball also completed either Sheep or Weather in year one and the majority of students who completed weather also completed Sheep. There is a possibility that students did not take this module seriously, because teachers had never graded the sheep assessment and therefore they knew it did not count towards their class grade. Researchers discovered through personal communication that in the 2010-2011 school year no teacher score the sheep module and returned it to students.

In addition, both the weather and basketball modules include a much larger number of non player characters. There are some characters that provide important information in both worlds; however, both assessment modules also include several characters that provide erroneous information that is not important to understanding and solving the problem. In contrast, the sheep module only has two characters who are obviously at odds with one another. This difference in design could have accounted for some of the lack of difference among students between the treatment and control groups. In modules like Basketball and Weather, where the majority of character interactions provide folk or incorrect evidence, if students rely more heavily on their character interactions instead of measurements, this could negatively affect their score. While this would need a more controlled study to determine specific results, this is an important aspect for game designers to consider. Furthermore, former research studies have found that the RAA makes less impact on older students (Elbaum, 2007; Laitusis, 2010). While students who complete the weather module are slightly older than when completing the sheep module, this would be more realistically applied to students completing the basketball module. Within this study, the students who completed basketball were at least one grade ahead of those who did Sheep. Therefore, the accommodation making less of a difference for Basketball than Sheep fits into the research.

In accordance with previous research, it would be expected that students who received the treatment would have a statistically higher overall percentage in comparison to the control because they experienced a more brain-based approach (Ali et al. 2010; Duman, 2010; Goswami, 2008). This was only found in the Sheep module, but for the Basketball and Weather modules, sample size, experience levels of students, and design features are the possible factors contributing to the lack of statistical variance between students experiencing the treatment and those in the control groups.

Research Question 2.

> How do student actions in question one vary by gender, ethnicity, English language learner status and disability status, if at all?

In order to consider student averages only once, the percentage scores for the Sheep and Basketball modules were considered for this question. All demographic variables were evaluated using a multiple regression. This analysis revealed that English language learner and disability statuses were the only two statistically significant predictors ($p<.001$). These variables predicted that scores for students without a learning disability or language barrier would be higher on average. Because students with a learning disability were coded as zero and those without were coded as one and the standardized beta value was positive (Beta=.225, t=4.115, p=.000), students without learning disabilities are more likely to earn a higher percentage score on module questions. This could mean that the accommodation is not providing students with learning disabilities a differential increase in comparison to students without disabilities or that it is not helping them enough. According to the first research question, the student average in the treatment group was statistically higher in the presence of the RAA on the sheep module. In considering expected results in differential improvement studies and the first two research questions together, the students in the learning disability status should have had a greater increase in their performance. However, this accommodation alone did not diminish that gap because learning disability is a significant predictor for overall percentage score and there is no interaction with treatment. On the contrary, it could indicate that this accommodation provided the greatest benefit to the students with better science skills, which tended to be in the general population, not those with learning disabilities, similar to a previously cited study on mathematics (Elbaum, 2007). This accommodation only provides greater access and students must already know the content knowledge and be able to apply it (Bolt &Thurlow, 2007).

Moreover, it is important to note that this analysis considers only about half of the participant sample scores, because of unavailable data. The original sample size was 592, but once students with missing learning disability status information was removed, only 324 participants were considered, resulting in a higher mean in the percentage for the control group and a lower mean for the treatment group. Furthermore, this translated to a much smaller overall mean difference in percentage score between the treatment and control groups before the loss of participants in the sample, observed in the first research question. It is unclear if these missing pieces would have changed the outcome of the analysis; however, the change in the sample clearly affected the findings.

An analogous relationship occurs between students who are not English language learners being more likely to attain a higher percentage score (Beta=.208, t=3.90, p=.000). While the sample sizes were similar between the control and treatment groups for students that English was not their first language, 23 and 22 respectively, there was a larger number of non-ELL students in the control group. Arguably, this sample size difference factors into the analysis in that ELL scores affect the average mean in the treatment group more than the control group. However, this does not negate the finding that there is a significant difference between students who have English language proficiency and those who do not while there is no significant interaction between treatment and this predictor variable. Within paper-based assessments, Enriquez (2008) found that students with lower levels of English language proficiency benefited the most from the read-aloud accommodation. Because the data within this study were only supplied dichotomously as students who are English Language learners or not, it is impossible to determine if these results parallel this finding; however, this could be an area of future research to determine if this treatment helps some English Language Learners more than others. The results of students being compared by English Language Learner and learning disability status were also considered in tandem with gender and ethnicity in a multiple regression analysis.

Gender, treatment, ethnicity, and their interactions with the treatment were not significant as predictors in the multiple regression. Ethnicity not being a significant predictor for success is positive because this study hoped to eradicate some of the scoring differences consistently found between white students and minority subgroups, like the No Child Left Behind act was designed to do (Packer, 2007). According to Brown and Hunter (2006), tailoring assessments to a variety of learning styles is a way to diminish the differences in scores among ethnic subgroups. In this study, students were provided with visual, auditory, and kinesthetic portions of the assessment and ethnicity was not a statistical predictor. By integrating brain-based learning and universal design for assessment principles (i.e. providing multiple ways of processing information), the elimination in ethnicity being a predictor was erased, as is common in standardized testing. However, this could be a result of more than just the accommodation because the modules are designed to incorporate many different learning styles. Without a control classroom taking paper and pencil assessments with half receiving an audio treatment, it is impossible to identify if the accommodation, the virtual world, or the combination caused the change to typically observed gaps in ethnicity.

Thus, in considering all variables, the treatment only made a significant difference in the percentage of correct answers in the sheep module, while it had no effect on the other two: basketball and weather. When controlling for other variables, such as English Language learner and learning disability status, the previously discovered effect disappeared. In essence, the treatment did affect scores of students; however, other variables predicted for a greater amount of the variance between the subjects. English language learners, students with learning disabilities, and minorities typically score lower than those without barriers to learning and white students, respectively. The disparities between ethnicities disappeared during this analysis. While this treatment did help students within the treatment group in comparison to the control, it was not significant enough to overcome all the obstacles that some students face. The treatment provided no statistical differential benefit for students with learning disabilities or limited English proficiency; nevertheless, generally, students receiving the RAA had a higher mean average score. The overall increase in scores for those in the treatment group may help to explain the general positive perception towards the reading aloud accommodation.

Research Question 3.

To what extent and how do students perceive the RAA as helpful or distracting?

Students in the treatment group answered both a seven-question survey and an open-ended question concerning their perceptions of the reading aloud accommodation. While only five students' open-ended responses were selected to be displayed in Chapter Four, frequencies and averages for all survey results from the treatment group were calculated and presented.

Overall, the majority of students, 80.1 percent, self-reported being neutral (36.3) towards or in disagreement (44.3) with the statement, "I wish I did not have to listen to the characters speak," indicating that it either did not affect their experience or did not affect it negatively. Of the students who agreed that they did not wish to hear the characters speak, only 5.3 percent strongly agreed with this statement, refer to Table 17. Thus, overall students were neutral or positive in this answer.

Similarly, most students were positive or neutral with respect to the second survey question. More than half of students, 57.5 percent, agreed or strongly agreed that hearing the assessment questions helped them understand what they were being asked. The open-ended response from Ferris helps to clarify why, "yes, because I could hear them and I didn't misread." The RAA helped some students ensure that they understood what they were being asked and possibly helped them process the information. Michael's response, "no, it just let me hear it instead of read it," is indicative of the 25.7 percent of participants who were neutral towards the helpfulness of the RAA . Only 15.8 percent of students disagreed with the survey question that asked about the helpfulness of the RAA during the assessment. Generally, though, the responses were positive that the RAA helped students understand the assessment questions.

Another survey question investigated how helpful students found the RAA. Survey question five asked students to respond to the statement, "It was helpful to me when I heard the character explain the problem to me." A total of 69.1 percent of students responded somewhat helpful or two a great extent, see Table 21. This suggests that students found it beneficial to hear the characters explain the problem verbally. Greg's open-ended response helps explain that the RAA was only somewhat helpful, because "it's just easier then reading it." Only 17.1 percent of students responded, "not at all." Like question two, it is unclear if responses of "not at all" indicated that students had indifference like Greg or if it was detrimental; however, question three does investigate if students found the RAA interfered with their virtual experience.

Specifically, the third survey question asked students to react to the statement, "the characters speaking to me was a distraction." More than half of participants either disagreed or strongly disagreed with this statement, 33.6 and 22.1 percent respectively. In including the 27.4 percent of students who reported neutral feelings, 83.2 percent of the 113 students who answered this question did not find the RAA to be a distraction. Only 16.8 percent of students indicated that the RAA was distracting during their virtual environment experience. One theory as to why students found this accommodation to be a distraction can be gleaned from the case study Vivienne, “it was a bit distracting in my opinion, because I often read faster than the character talks.” The characters speak at a relatively slow rate to ensure that students have time to process the information and try to read along. Students who read very quickly could easily get distracted or frustrated as the sounds are different from the words they are reading. This idea could also explain variances in the fourth survey question.

In order to determine if students used the RAA, students were asked to respond to the prompt, “I listened to the characters talk to me.” Students like Vivienne most likely disagreed or strongly disagreed because the voices were a nuisance due to their difference in speed in comparison to reading rates. Approximately 13 percent of students responded to the question, “not at all.” It is unclear if this suggests that students removed their headphones because the characters speaking was a distraction or if they just ignored the voices. The remaining 86.7 percent of students indicated listening to the characters, even if it was very little (9.7 %).   Thus, the vast majority of students reported using the affordance, regardless of if they found it helpful or distracting.

While the fifth question investigated if students physically attended to the RAA, the last two questions were designed to discover if students gathered information more by reading or listening. These two questions asked students to report the extent that they gathered information during their experience by reading the text or listening to the characters speak. Because the two questions were separate, there is a possible overlap from students who felt they used both methods “to a great extent” in obtaining information. A larger majority of students responded that they gathered information "somewhat" or "to a great extent" by reading (77.7 percent) than by listening (64.2 percent). Augusta, the selected English Language Learner, explained that hearing the characters speak affected her information gathering and question answering since, "because it give information." Since only one English Language Learner was chosen, it is unclear if it helps the majority or only some, but it did affect her experience. Generally though, more students indicated that reading the text was important in the obtaining data to solve the in-world problem. However, in both cases the majority of students indicated gathering information through both mediums, which is consistent with brain-based learning principles in that students can process the information in parallel and more naturally (Caine & Caine, 1991; Jensen, 2008).

Whereas there are some studies indicating positive student perceptions of both virtual environment assessment (Clarke-Midura, Code, Zap, & Dede, 2011; Code, Clarke-Midura, Mayrath, & Dede, 2011) and of reading aloud accommodations on assessments (Bolt, Decker, Lloyd, & Morlock, 2011; McKevitt & Elliott, 2003), there have been no studies investigating student's feelings about a RAA within a virtual science assessment. Therefore, the results from the survey questions are generally in line with what current research in virtual environment assessments and reading aloud accommodations separately. Because this is a new application of reading aloud accommodation into a new type of assessment, it is important to uncover student perceptions of its utility and helpfulness as a treatment situated within a different context. Furthermore, because the reading aloud accommodation is typically only provided to students with diagnosed learning disabilities or English language learners, the overall general population could have had a different perspective than students who already use the accommodation on assessments.

An interesting side finding was that when the survey results were calculated only sampling students with learning disabilities, the mean averages changed for survey question responses. Students were more neutral towards wishing they did not have to hear the characters and they found the characters speaking to be less of a distraction. Moreover, these students reported listening to the characters more, and they had higher mean averages for finding the characters talking to them being helpful in understanding the problem and assessment questions. Furthermore, when recounting their experience, they had a lower mean average for gathering information by reading and a higher one for gathering by listening in comparison to general education students. This is congruent with a finding from McKevvitt and Elliot (2003) in that they found that students with learning disabilities reported the reading aloud accommodation more helpful than a standard condition in reading assessments.

Results Conclusion

Based on these findings it appears that a reading aloud accommodation helps students earn higher percentages on virtual science assessments, but not in all instances. This could indicate that the accommodation is relieving some construct irrelevant variance and more validly assessing student science content knowledge; however, the reading aloud accommodation does not provide a differential boost for students with learning disabilities or English Language learners. Rather, when demographic variables like the aforementioned barriers to learning are considered, the significant differences between the treatment group and the control group disappear, although a dramatic data loss potentially skews these results. Therefore, even though students receiving the RAA score higher than the control groups, English language learner and learning disability status account for a much larger portion of the variance between the groups. This could indicate that the reading construct irrelevance was not a major issue or that something else within these demographic variables is interfering. Either way, it suggests that the RAA treatment does not remove systemic barriers, as it was hypothesized to do. However, in general, students had positive perceptions of the accommodation being integrated within their tests. Many students identified that this accommodation helped them understand the in-world problem and the questions they were asked at the end of the assessment. Nevertheless, though students have positive perceptions and scores generally increased in the presence of this treatment, it alone does not account for a significant variance in the score when demographic variables are taken into consideration and thus cannot be used as a sole means of decreasing the achievement gap or more validly assessing student content knowledge.

Practical Applications

Although the findings in this study did not indicate that the reading aloud accommodation implemented within a virtual world was the panacea to shrinking the achievement gap and completely eliminating construct irrelevance variance, there are some practical applications to the field of science assessment, specifically those based in virtual environments. This study initiates the dialogue of investigating specific accommodations in virtual environment assessments to make them more aligned with brain-based learning, particularly that the RAA treatment is not enough to reduce structural inequities significantly. Within a different virtual environment assessment project, *Virtual performance assessment,* picture dragging is incorporated as a way to accommodate students in answering science questions (Mayrath et al., 2011). This accommodation is being piloted to determine if vocabulary acquisition is a barrier in illustrating conceptual knowledge by allowing those in the treatment group to drag and drop pictures of difficult words within their answer responses. Therefore, these two studies are investigating two separate barriers to determine if either will more validly assess student content understanding. It is possible that a better solution is a combination of the two barriers, which is discussed in the future research section within this chapter.

While it was not the main purpose of this study, these findings add to the body of research investigating the incorporation of multiple methodologies in technology-enabled assessments. Unlike the findings of many studies involving technology-enabled assessments, students' actions (e.g. measurements and collisions) did not provide additional insights into student understanding (Lewis & Sewell, 2007; Lowry, 2005; Thissen-Roe, Hunt, & Minstrell, 2004). The treatment did not statistically affect how students interacted with tools or characters within the modules and individual trajectories would have been impossible to analyze with almost 800 students.

Because the structural inequities were still evident despite the treatment, this suggests that virtual environment assessments may also be perpetuating the unfairness of assessments. These assessments are designed to assess student scientific content knowledge more accurately by embedding the questions within a more natural context and having students synthesize data rather than memorize it; however, the results are similar to standardized assessment results when demographic variables are considered. Though this was not the purpose of this study, the results indicate that the breakdown of student performance is similar to typical results. In order to determine the accuracy of this truly, a more controlled study separate from this would need to be completed, yet the presented findings suggest the necessity for this.

Limitations

The sample constitutes the largest portion of the limitations within this study. Because teachers must voluntarily participate in the SAVE Science project, it is hard to obtain a large dataset that is equal in terms of participants and demographic variables. Over sixty percent of the entire sample was from a near urban school, which limits the generalizability of these findings. Beyond that, discovering that one teacher did not implement correctly devastated the equality of the sample because she had a large amount of participants from an urban school with a variety of demographic variables. In order to maintain the integrity of this study, there was a loss of 328 students among all three modules for both years. While it was damaging to the sample to remove this teacher, it was necessary to ensure the accuracy of the results. Furthermore, within the sample of analyzed data, not all students or schools provided the necessary demographic variables, which in some analyses, reduced the available sample to half broadening the limit to the generalizabilty and possibly validity of the results. Due to institutional review board regulations, participants may omit answers on the survey, so it was necessary to run the statistical analyses with only available information for each variable.

Technological and scheduling difficulties also presented limitations to this study. In the first implementation of the basketball module, a glitch within the database prevented the random assignment of the treatment. Although half of students were supposed to be assigned to the treatment group, all students received the control only.

This was not discovered until both implementing teachers had completed the process. Thus, the sample for the basketball module consisted of only one year's data. Similarly, the weather data are only available for year one, since teachers will not implement the weather module again, until after this study is complete. Teachers must schedule the implementation of the modules close to when they teach the associated content material in order to assess student's understanding within an applicable period of time. The lack of multiple years of data or low sample sizes could be why no statistical differences were found in these modules.

Future Research

During the analysis and interpretation of data, unanswered questions arose, which lead to areas of future research. It became apparent through the surveys and open ended student responses, that not all students liked the reading aloud accommodation. One case study explained that the voices were a distraction because they spoke at a speed slower than she read. It would be interesting to determine how students perceptions change if participants have control over the speed of the voices. This could be investigated by allowing students to choose the speed of the voices of the characters or by allowing students to mouse over words in order to hear them. Either scenario would allow students to have more control over their reading aloud accommodation. Moreover, repeating this study but adding in qualitative methods could help illuminate what would help students more. Using focus groups, think alouds, and semi-structured interviews would provide researchers with greater access into students' thought processes and understanding their perceptions. Further, investigating student trajectories through the module in tandem with interviews could help clarify how much students really know and what barriers exist for each student. This would have been impossible with a large sample size, but in choosing three or four case studies to investigate in depth, could elucidate how the accommodation helps and hinders.

In order to determine if treatment is helping reduce the difference in scores among ethnicities, a differently designed follow up study would need to be completed. Adding a control class of mixed ethnicities where all students completed similar questions on paper-and-pencil test with half of participants receiving an individualized reading aloud accommodation could help investigate this further. It would allow the comparison of the percentage of success for students across ethnicities both in and out of the virtual world while also considering the affordance. This would provide researchers the opportunity to understand if the reading aloud accommodation is solely responsible for diminishing the difference in performance, in tandem with the virtual world, or if it is the virtual world experience. Following these studies, additional accommodations could be considered and implemented. Through the presented data, it is impossible to determine if the RAA helped students but they still could not express their conceptual understanding. Two additional accommodations can be implemented, speech to text and picture dragging. The speech to text option would allow students to tell their answers rather than typing it, while the picture dragging, provides students with pictures of words they may not know how to say or spell. It is possible that in order to remove the inequities commonly observed in standardized testing, that all construct irrelevant variance needs to be removed. In the inception of the design of this study, the focus was on the irrelevance of reading; however, it is necessary to consider and investigate all potential forms of construct irrelevant variance.

REFERENCES

Abedi, J. (2002). Assessing and accommodations of English language learners: Issues, concerns and recommendations. Journal of School Improvement, 3(1), 83-89.

Ali, R., Hukamdad, Ghazi, S., Shahzad, S., & Khan, H. (2010). The impact of brain based learning on students academic achievement. *Interdisciplinary Journal of Contemporary Research in Business. 2*(2), 542-556.

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke, J., Torres, C., Haertel, G., Dolan, B., Beddow, P., & Lazarus, S. (2010). Technology enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. Dover, NH: Measured Progress and Menlo Park, CA: SRI International

American Association for the Advancement of Science (AAAS). (1993). Project 2061: Benchmarks for Science Literacy. Washington: Oxford University Press.

American Educational Research Association (AREA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Atkinson, R.C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K.W. Spence (Ed.), *The psychology of learning and motivation: Advances in research and theory* (2) p. 89-105. San Diego, CA: Academic Press.

Barab, S., Sadler, T., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating narrative, inquiry, and inscriptions: Supporting consequential play. *Journal of Science*

Education and Technology, 16 (1), 59-82.

Bellah, K., Robinson, J., Kaufman, E., Akers, C., Haase-Wittler, P., & Martindale, L. (2008) Brain-based Learning: A Synthesis of Research. *North American Colleges and Teachers of Agriculture Journal, 52*(2), 15-22.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodation: Effects on multiple-choice reading & math items (Technical Report 31).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Block, C. C. & Parris, S. R. (2008). Using neuroscience to inform reading comprehension instruction. In C. C. Block (Ed.), *Comprehension instruction: research-based best practices* (2nd ed.). New York: Guilford Press. p. 114-126.

Bolt, S.E., Decker, D. M., Lloyd, M., & Morlock, L. (2001). Students' Perceptions of Accommodations in High School and College. *Career Development for Exceptional Individuals*, 34(3), 165-175.

Bolt, S. E., & Thurlow, M. L. (2007). Item-level effects of the read-aloud accommodation for students with reading disabilities. *Assessment for Effective Intervention*, *33*, 15-28.

Bolt, S. E., & Ysselddyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. *Applied Measurement in Education*, *19*(4), 329-355.

Bolt, S. E., & Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment*, *26*, 121-138.

Bourne, L.E., Dominowski, R. L., & Loftus, E.F. (1979). *Cognitive Processes* Englewood Cliffs, NJ: Prentice-Hall.

Brown, F., & Hunter, R. C. (2006). *No child left behind and other federal programs for urban school districts*. Amsterdam: JAI Press.

Burgstahler, S. (2007). *Equal access: Universal design of instruction*. Seattle: DO-IT, University of Washington. Retrieved November 12, 2008, from *http://www.washington.edu/doit/Brochures/Academics/equal_access_udi.html*

Burns, E. (1998) *Test accommodations for students with disabilities.* Springfield: Charles C. Thomas, Publisher, LTD.

Bush, G. H. W. (1990, July 18). Decade of the Brain: Presidential Proclamation 6158 (Library of Congress). *Library of Congress Home.* Retrieved December 5, 2011, from http://www.loc.gov/loc/brain/proclaim.

Cahalan, C., Mandinach, E. B., & Camara, W. J. (2002). Predictive validity of SAT I: Reasoning test for examinees with learning disabilities and extended time accommodations. The College Board Research Report No. 2002-5. New York: College Entrance Examination Board.

Caine, R.N., & Caine, G. (1991). *Making connections: teaching and the human brain.* Alexandria, VA: Association for Supervision and Curriculum Development.

Caine, R.N., & Caine, G. (1990). Understanding a brain-based approach to learning and teaching. *Educational Leadership*, *48*(2), 66-70.

Calhoon, M.B., Fuchs, L.S., & Hamlett, C.L. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly*, *23*, 271-282.

Camara, W. J. (2009). College admissions testing: Myths and realities in an age of admissions hype. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 147-180). Washington, DC: American Psychological Association.

Center for Applied Special Technology (CAST), (2011). *Universal Design for Learning Guidelines version 2.0.* Wakefield, MA: Author.

Chapman, D.W., & Snyder, C.W. (2000). Can high stakes national testing improve instruction: reexamining conventional wisdom? *International Journal of Educational Development. 20*(1), 457-474.

Clarke, J. (2009). Studying the potential of virtual performance assessments for measuring student achievement in science. Paper presented at the American Educational Research Association (AERA), San Diego: April 13-17.

Clarke-Midura, J., Code, J., Mayrath, M. & Dede, C. (2011). *Using evidence centered design to develop immersive virtual assessments.* Paper presented at the AERA 2011 Annual Meeting, New Orleans, LA.

Clarke-Midura, J., Code, J., Zap, N. & Dede, C. (2011). Student Perceptions of Immersive Virtual Environments for the Meaningful Assessment of Learning. In T. Bastiaens & M. Ebner (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2011* (pp. 358-367). Chesapeake, VA: AACE

Clapper, A. T., Morse, A. B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. L. (2005). *2003 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 56). Minneapolis, MN: University of

Minnesota, NCEO.

Code, J., Clarke-Midura, J., Mayrath, M. & Dede, C. (2011). Student perceptions of the assessment utility of immersive virtual assessments. Paper presented at the AERA 2011 Annual Meeting New Orleans, LA, April 11, 2011.

Cohen, L., Manion, L., & Morrison, K. (2003). *Research methods in education* (5th ed.). London: Routledge Falmer.

Connell, B. R., Jones, M., Mace, R., Mueller, J., Mullick, A., Ostroff, E., et al. (1997). *Principles of Universal Design*. Retrieved August 29, 2011, from h http://www.design.ncsu.edu:8120/cud/univ_design/princ_overview.htm

Cook, L., Eignor, D., Sawaki, Y., Steinberg, J., & Cline, F. (2010). Using factor analysis to investigate accommodations used by students with disabilities on an English-language arts assessment. *Applied Measurement in Education, 23*(2), 187-208.

Cormier, D. C., Altman, J. R., Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007-2008* (Technical Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Crawford, L. (2007). State testing accommodations: A look at their value and validity. New York: National Center for Learning Disabilities.

Creswell, J.W. (2009). *Research design: qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Dalton, B. & Rose, D. (2008). Scaffolding Digital Comprehension. In C. C. Block (Ed.), *Comprehension instruction: research-based best practices* (2nd ed.). New York: Guilford Press. p. 114-126.

Dalton, B., & Proctor, C. P. (2007). Reading as thinking: Integrating strategy instruction in a universally designed digital literacy environment. In D. S. McNamara (Ed.), Reading comprehension strategies: Theories, interventions, and technologies (pp. 423-442). Mahwah, NJ: Erlbaum.

Dolan, R. P. & Hall, T.E. (2001). Universal design for learning: Implications for large scale assessment. *IDA Perspectives*, *27*(4), 22-25.

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying Principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment, 3*(7). Available from http://www.jtla.org

Donnelly, D., McGarr, O. & O'Reilly, J. (2011). The promotion of scientific inquiry in Irish post-primary schools through the use of a virtual chemistry laboratory: Implication for teacher education. In *Proceedings of Society for Information Technology & Teacher Education International Conference 2011* (pp. 3638-3645). Chesapeake, VA: AACE.

Duman, B. (2010). The effects of brain-based learning on the academic achievement of students with different learning styles. *Educational Sciences: Theory & Practice 10*(4), 2077-2103.

Duran R. P. (2008). Assessing English-language learners' achievement. Review of Research in Education, (32), 292–327.

Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *The Journal of Special Education*, *40*, 218-229.

Elliot, S. N., T. R. Kratochwill, & A. G. Schulte. (1999). Assessment accommodations checklist. Monterey, CA: CTB/McGraw Hill.

Enriquez, M. (2008). Examining the effects of linguistic accommodations on the Colorado student assessment program—mathematics. *Dissertation Abstracts International.*

Flavell, J.H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10) 906-911.

Flowers, C., Do-Hong, K., Lewis, P., & Davis, V. (2011). A Comparison of computer-based testing and paper-and-pencil testing for students with a read-aloud accommodation. *Journal of Special Education Technology, 26*(1), 1-12.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). Boston: McGraw Hill Higher Education.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. Exceptional Children, 67, 67-81.

Fuchs, L.S., Fuchs, D., Eaton, S.B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments about test accommodations with objective data sources. *School Psychology Review, 29 (1)*, 65-85.

Gabriel, A. E. (1999). Brain-based learning: The scent of a trail. *The Clearing House, 72(5)*,288-290.

Galas, C. & Ketelhut, D. J. (2006). River City, the MUVE. *Learning and Leading with Technology, 33*(7), 31-32.

Goglin, L., & Swartz, F. (1992) A quantitative and qualitative inquiry into the attitudes towards science of non-science college students. *Journal of Research in Science Teaching*, 29, 487-504.

Goswami, U. (2008). Principles of learning, implications for teaching: A cognitive neuroscience perspective. *Journal of Philosophy of Education, 42*(3-4) 381-399.

Haertel, E. H. & Linn, R. L. (1996). Comparability. In Phillips, G. W. (Ed.), *Technical issues in large-scale performance assessment* (pp. 59-78). Washington: National Center for Education Statistics, U. S. Department of Education, Office of Educational Research and Improvement.

Harris, L. W. (2009). Comparison of student performance between teacher read and CD-ROM delivered modes of test administration of English language arts tests. *Dissertation Abstracts International Section A, 69.*

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *Journal of Educational Research, 93*(2), 113–25.

Higbee, J.L. (2001). Implications of universal instructional design for developmental education. *Research and Teaching in Developmental Education, 17*(2), 67-70.

Hollenbeck, K. (2002). Determining when test alterations are valid accommodations or modifications for large-scale assessment. In G. Tindal & T. Haladyna (Eds.), Large scale assessment programs for all students (pp. 109–148). Mahwah, NJ: Erlbaum.

Huynh, H., & Barton, K. (2006). Performance of students with disabilities under regular and oral administration of a high stakes reading examination. *Applied Measurement in Education, 19,* 21–39.

Huynh, H., Meyer, J. P., & Gallant, D. (2004). Comparability of student performance between regular and oral administrations for a high-stakes mathematics test. Applied Measurement in Education, 17, 39-57.

Jensen, E. (2008). *Brain-based learning: The new paradigm of teaching* (2nd ed.). Thousand Oaks, CA: Corwin Press .

Johnstone, C. J., Altman, J., Thurlow, M. L., and Thompson, S. J. (2006). A summary of research on the effects of test accommodations: 2002 through 2004 (Technical Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 2011, from http://education.umn.edu/NCEO/OnlinePubs/Tech45/.

Johnstone, C.J., Thompson, S.J., Moen, R.E., Bolt, S., Kato, K., & National Center on Educational Outcomes, M.N. (2005). Analyzing results of large-scale assessments to ensure universal design. Technical Report 41. *National Center on Educational Outcomes.*

Kafai, Y. (2010). World of Whyville: An Introduction to Tween Virtual Life *Games and Culture. 5*(1), *3-22, doi:10.1177/1555412009351264*

Ketelhut, D. (2007). The Impact of Student Self-efficacy on Scientific Inquiry Skills: An Exploratory Investigation in River City, a Multi-user Virtual Environment. *Journal of Science Education & Technology, 16*(1), 99-111. doi:10.1007/s10956-006-9038-y.

Ketelhut, D.J., Nelson, B., Schifter, C., and Kim, Y. (2010). Using immersive virtual environments to assess science content understanding: the impact of context. In Kinshuk, D. G. Sampson, J. M. Spector, P. Isaías, D. Ifenthaler and R. Vasiu (Eds.), Proceedings of the IADIS international conference on cognition and exploratory learning in the digital age (CELDA 2010). p 227-230.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment, 4*(2). Available from http://www.jtla.org.

Ketterlin-Geller, L. R. (2008).Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*(3), 3-16.

Ketterlin-Geller, L. R., Yovanoff, P., & Tindal, G. (2007). Developing a new paradigm for conducting research on accommodations in mathematics testing. *Exceptional Children, 73,* 331-347.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky (CSE Technical Report No. 431).* Los Angeles, CA: Center for Research on Standards and Student Testing.

Laitusis, C. (2010). Examining the impact of audio presentation on tests of reading comprehension. *Applied Measurement in Education, 23*(2), 153-167.

Lehr, C., & Thurlow, M. (2003). *Putting it all together: Including students with disabilities in assessment and accountability systems* (Policy Directions No. 16). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 20, 2011 from http://education.umn.edu/NCEO/OnlinePubs/Policy16.htm.

Lewis, D.J.A., and Sewell, R.D.E. (2007). Instructional design and assessment: Providing formative feedback from a summative computer-aided assessment. *American Journal of Pharmaceutical Education, 71* (2), article 33.

Lowry, R. (2005). Computer aided self-assessment – an effective tool. *Chemistry Education Research and Practice*, 6(4), 198-203.

Mace, R.L. (1998). Universal design in housing. *Assistive Technology. 10*(1), 21-28.

Mace, R., Hardie, G. & Place, J. (1991), Toward universal design. In Preiser, W. Vischer, J. & White. E. (Eds.), *Design intervention: Toward a more humane architecture.* New York: Van Nostrand Reinold. p: 155–175.

Majerich, D. M., Schifter, C. S., Shelton, A., Ketelhut, D. J. (2011, April). Effects of reading-while-listening affordance on students' scientific hypotheses and supporting evidence developed in an inquiry-based, virtual environment assessment. Paper presented at the American Educational Research Association Conference, New Orleans, LA, April 9.

Mayrath, M., Clarke-Midura, J., Dede, C. & Code, J. (2011). A Framework for Designing Assessment Activities for Virtual Worlds. Paper presented at the American Educational Research Association Conference, New Orleans, LA, April 9.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities (NCES 2000-473). Washington, DC: U.S. Department of Education, National Center for Education Statistic.

McGuire, J. M., Scott, S. S., & Shaw, S. F. (2006). Universal design and its applications in educational environments. *Remedial & Special Education, 27*(3), 166-175.

McKevitt, B. C., & Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review*, *32*(4), 583-600.

Meloy, L.L., Deville, C., & Frisbie, D.A. (2002). The Effect of a read aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial & Special Education, 23*(4), 248.

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational measurement, (3rd ed.). Washington, D.C.: American Council on Education, 13-103.

Middleton, K. V. (2007). The effect of a read-aloud accommodation on items on a reading comprehension test for students with reading-based learning disabilities. *Dissertations International,* University of Iowa.

Miles, M.B., Huberman, A.M. (1994) *Qualitative Data Analysis: An expanded sourcebook* (2nd ed.), Sage: London & Thousand Oaks, California.

Millsap, R.E. (2009). *The SAGE handbook of quantitative methods in psychology.* Los Angeles: SAGE. National Joint Committee on Learning Disabilities (NJCLD). 1998. Learning disabilities: Pre-service preparation of general and special education teachers. *Learning Disability Quarterly 21*:182–186.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Nelson, B., Ketelhut, D. J., Clarke, J., Bowman, C., and Dede, C. (2005). Design-based research strategies for developing a scientific inquiry curriculum in a multi-user virtual environment. Educational Technology, *45* (1), 21–27.

Nelson, B., Ketelhut, D. J. & Schifter, C. (2010). Exploring Cognitive Load in Immersive Educational Games: The SAVE Science Project. *International Journal for Gaming and Computer Mediated Simulations 2* (1), 31-39.

Oakland, T. & Lane, H.B. (2004). 'Language, reading, and readability formulas:

Implications for developing and adapting tests', *International Journal of Testing, (4)*3, 239-252.

Packer, J. (2007). The NEA supports substantial overhaul, not repeal, of NCLB. *Phi Delta Kappan*, *89*, (4), 265- 269.

Phillips, S.E. (1994). High stakes testing accommodations: Validity vs. disabled rights. *Applied Measurement in Education, 7* (2), 93-120.

Pomplun, M. & Omar, H. M. (2000). Score comparability on a state mathematics assessment across students with and without reading accommodations. *Journal of Applied Psychology*, 85(1), 21–9.

Preiser, W.F. & Ostroff, E. (Eds.) (2001). *Universal design handbook.* New York: McGraw Hill.

Proctor, C. P., Dalton, B., & Grisham, D. L. (2007). Scaffolding English language learners and struggling readers in a universal literacy environment with embedded strategy instruction and vocabulary support. *Journal of Literacy Research, 39* , 71-93.

Randall, J., & Engelhard, G. (2010a) Performance of students with and without disabilities under modified conditions: Using resource guides and read-aloud test modifications on a high-stakes reading test. *The Journal of Special Education 44*(2), 79-93.

Randall, J., & Engelhard, G. (2010b). Using confirmatory factor analysis and Rasch Measurement Theory to assess measurement invariance in a high stakes reading assessment . Applied Measurement in Education, 23, 286-306.

Rose, D. & Dolan, B. (2000). Universal design for learning: associate editor's column. *Journal of Special Education Technology*, *15*(4), 47-51.

Rose, D.H. & Gravel, J.W. (2010). Universal design for learning. In E. Baker, P. Peterson, & B. McGaw (Eds.) *International Encyclopedia of Education*, *3rd Ed.* Oxford: Elsevier.

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Rose, D., Meyer, A., Strangman, N., & Rappolt, G. (2002). Teaching every student in the digital age: Universal design for learning. Alexandria, VA: ASCD.

Rose, D., & Strangman, N. (2007). Cognition and learning: Meeting the challenge of individual differences. *Universal Access in the Information Society*, *5*(4), 381-391.

Roussos, M., Johnson, A., Mober, T., Leigh, J., Vasilakis, C., & Barnes, C. 1999. Learning and building together in an immersive virtual world. *Presence. 8*(3), 247-263.

Shelton, A. & Ketelhut, D. J. (2012, March). Comparing Student Performances, Anxieties, and Preferences between Situated, Virtual Environment Assessments and Multiple-Choice Assessments. Paper presented at the annual meeting of the National Association for Research in Science Teaching (NARST), Indianapolis, IN, March 28.

Shih-Wei , C., & Chien-Hung, L. (2005). Learning effectiveness in a Web-based virtual learning environment: a learner control perspective. *Journal of Computer Assisted Learning*, *21*(1), 65-76. doi:10.1111/j.1365-2729.2005.00114.x.

Sireci, S. G. (2004). *Validity issues in accommodating NAEP reading tests.* Center for

Educational Assessment (Research Report No. 515), Amherst, MA: School of Education, University of Massachusetts Amherst.

Sireci, S. G., S. E. Scarpati, and S. Li. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research,* 75 (4): 457-490.

Sylwester, R. (2010). *A child's brain: The need for nurture.* Thousand Oaks, CA: Corwin Press.

Smedley, T.M., & Higgins, K. (2005). Virtual technology: Bringing the world into the special education classroom. *Intervention in School & Clinic, 41*(2), 114-119.

Sprenger, M. (2010). *Brain-based teaching :)in the digital age.* Alexandria, VA: Association for Supervision and Curriculum Development.

Stanovich,K. (1986). Matthew effect in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 306-407.

Stevens, J., & Goldberg, D. (2001). *For the learner's sake: brain-based instruction for the 21st century.* Tucson, AZ: Zephyr Press.

Thissen-Roe, A., Hunt, E, and Minstrell, J. (2004). The DIAGNOSER project: Combining assessment and learning. *Behavior Research Methods, Instruments, & Computers 36 (2), 234-240.*

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (NCEO Synthesis Report 44). Minneapolis, MN:

University of Minnesota, National Center on Educational Outcomes. Retrieved August 20, 2011, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Tindal, G., & Fuchs, L. (2000). *A Summary of Research on Test Accommodations: An Empirical Basis for Defining Test Accommodations*. Lexington, KY: Mid-South Regional Resource Center. (ERIC Document Reproduction Service No. ED 442 245).

Tindal, G. Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study of student response and test administration demands. *Exceptional Children, 64* (4), 439-450.

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

Tucker, B. (2009). Beyond the bubble: Technology and the future of educational assessment. Washington, DC: Education Sector. UDL (Universal Design for Learning) (2009). UDL guidelines - Version 1.0 – Research evidence. Wakefield, MA: National Center on Universal Design for Learning.

United States Department of Education (USDOE). (2004). What is the purpose of the No Child Left Behind Act? -- Teacher Update. *U.S. Department of Education*. Retrieved May 26, 2011, from http://www2.ed.gov/teachers/how/tools/initiative/updates/040513.html

Weiss, R.P. (2000). Brain-based learning. *Training & Development, 54*(7), 20-24.

Welch, P. (Ed.). (1995). *Strategies for teaching universal design: An interpretation of the*

ADA. NewYork: Van Nostrand Reinhold.

Weston, T. J. (2003). *NAEP Validity Studies: The validity of oral accommodation testing.* Washington, DC: National Center for Education Statistics.

Wolf, M. K., Kim, J., Kao, J. C., & Rivera, N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a math assessment* (CRESST Report 766). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Appendix A.

Results from 15 predictor Multiple Regression pairwise deletion

Variable	N	Mean	SD	Pearson Corr.	Sig. (2 tailed)	MR Weights b	β
African American	657	.1431	.35042	-.119	.001	-7.600	-.124*
Latino	657	.1187	.32371	-.112	.002	-6.919	-.104
Asian	657	.1005	.30084	.019	.310	-1.022	-.014
Other	657	.0533	.22475	.001	.487	-4.151	-.043
AA*Treatment	652	.0008	.17164	.012	.384	.282	.002
Latino*Treatment	652	-.009	.15695	.017	.336	1.221	.009
Asian*Treatment	652	.0001	.14926	-.024	.268	-6.725	-.047
Other*Treatment	652	.0018	.11237	-.051	.098	-11.14	-.058
Treatment	786	1.44	.497	.147	.000	6.535	.151
ELL	717	1.92	.275	.106	.002	7.720	.099
LD	329	1.86	.344	.242	.000	14.656	.235**
Gender	766	1.53	.499	.021	.282	.273	.006
Gender*Treatment	761	-.005	.24792	.015	.338	-2.121	-.024
ELL*Treatment	712	-.002	.13744	-.005	.445	-1.841	-.012
LD* Treatment	329	-.005	.17344	.120	.015	13.056	.105

Notes: * $p < .05$ ** $p < .01$ ***$p<.001$

The multiple regression model with all fifteen predictors and pairwise deletion produced $R^2 = .132$, $F(15, 252) = 2.564$, $p = .001$. In comparison to the listwise deletion, five additional students are included in the analysis and significance of the model is .001 higher. The difference between the performance of African American students and Caucasian students is slightly significant now ($p=.044$). Further, ELL is no long a statistically significant predictor of performance.

Appendix B.

Results from 13 predictor Multiple Regression pairwise deletion (LD and interaction with treatment removed)

Variable	N	Mean	SD	Pearson Corr.	Sig. (2 tailed)	MR Weights b	MR Weights β
African American	653	.1431	.35042	-.119	.001	-9.04	-.15***
Latino	653	.1187	.32371	-.112	.002	-6.83	-.103
Asian	653	.1005	.30084	.019	.310	.743	.010
Other	653	.0533	.22475	.001	.487	-1.97	-.021
AA*Treatment	648	.0008	.17164	.012	.384	.309	.002
Latino*Treatment	648	-.009	.15695	.017	.336	-.640	-.005
Asian*Treatment	648	.0001	.14926	-.024	.268	-4.09	-.028
Other*Treatment	648	.0018	.11237	-.051	.098	-8.95	-.047
Treatment	782	1.44	.497	.147	.000	6.259	.145***
ELL	713	1.92	.275	.106	.002	7.382	.095
Gender	762	1.53	.499	.021	.282	1.785	.042
Gender*Treatment	757	-.005	.24792	.015	.338	1.348	.016
ELL*Treatment	653	-.002	.13744	-.005	.445	-3.25	-.021

Notes: * $p < .05$ ** $p < .01$ ***$p<.001$

The multiple regression model with thirteen predictors (eliminating learning disability) and pairwise deletion produced $R^2 = .065$, $F(13,634) = 3.381$, $p = .000$. This analysis accounts for a smaller amount of the variance than either multiple regression that was run with all fifteen predictors, but it is more statistically significant and has over 400 participants in the sample. In this model though, only treatment and whether a student is African American or White are significant predictors for the multiple regression model. While it is important to include Learning Disability within the model, it is necessary to show what the regression results would look like without the missing data.

Appendix C.

Frequencies for averages and variance for Audio Survey Questions

	S1_ Iwish	S2_ Hearing	S3_The characters	S4_ listen	S5_Itwas helpful	S6_gath_ reading	S7_gath_ listening
Valid N	113	113	113	113	113	112	109
Missing N	0	0	0	0	0	1	4
Mean	3.39	2.44	3.55	2.95	2.82	3.13	2.84
Median	3.00	2.00	4.00	3.00	3.00	3.00	3.00
Mode	3	2	4	3	3	4	4
Std. Deviation	1.11	1.052	1.118	.971	1.063	.905	1.047
Variance	1.24	1.106	1.249	.944	1.129	.820	1.096
Skewness	-.19	.483	-.498	-.785	-.547	-.789	-.421
Std. Error of Skew	.227	.227	.227	.227	.227	.228	.231
Range	4	4	4	3	3	3	3