8-1-2013

Evaluation of the effect of a digital mathematics game on academic achievement

Christine M. Wale
University of Northern Colorado

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

EVALUATION OF THE EFFECT OF A DIGITAL MATHEMATICS GAME ON ACADEMIC ACHIEVEMENT

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Christine M. Wale

College of Educational and Behavioral Sciences
School of Psychological Sciences
Educational Psychology

August 2013

This Dissertation by: Christine M. Wale

Entitled: *Evaluation of the Effect of a Digital Mathematics Game on Academic Achievement*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Education and Behavioral Sciences in School of Psychological Sciences, Program of Educational Psychology

Accepted by the Doctoral Committee

_____
Marilyn C. Welsh, Ph.D., Research Advisor

_____
Kathryn F. Cochran, Ph.D., Committee Member

_____
John B. Cooney, Ph.D., Committee Member

_____
Jenni L. Harding-DeKam, Ed.D., Faculty Representative

Date of Dissertation Defense _____

Accepted by the Graduate School

_____
Linda L. Black, Ed.D., LPC
Acting Dean of the Graduate School and International Admissions

ABSTRACT

Wale, Christine M. *Evaluation of the Effect of a Digital Mathematics Game on Academic Achievement*. Published Doctor of Philosophy dissertation, University of Northern Colorado, August 2013.

Digital games are widely popular and interest has increased for their use in education. Digital games are thought to be powerful instructional tools because they promote active learning and feedback, provide meaningful contexts to situate knowledge, create engagement and intrinsic motivation, and have the ability individualize instruction. However, claims about the potential benefits of digital games in education have outpaced quality empirical research on their effectiveness in K-12 settings.

The purpose of this study was to investigate the effectiveness of a mathematics digital game, Ko's Journey, on seventh grade students' mathematics achievement as defined by a researcher-constructed test aligned with the Common Core Mathematics Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010) and measured on a Rasch (1960) unidimensional equal-interval scale. This research was conducted using secondary data from a pretest-posttest control group design study with a total of 371 seventh grade students from 10 classrooms. Classroom teachers randomly assigned their classroom sections to play the mathematics digital game or served as a wait-listed control group and continued using the typical mathematics curriculum. Hierarchical linear modeling (HLM) and Rasch differential item functioning were used to determine the effect of the intervention on student's mathematics achievement.

Hierarchical linear modeling analyses, using person ability logit estimates derived from the Rasch scaling, concluded that the Ko's Journey intervention did not have a significant effect on posttest scores. The HLM analyses revealed a significant positive relationship between the students' individual pretest and posttest scores and the classroom average pretest and posttest scores. Using the Rasch differential item functioning, six assessment items were significantly less difficult for the experimental group compared to the control group; this suggested that the intervention was successful in teaching the mathematics targeted by the items. Technological problems experienced in the classrooms and differential implementation of the game among teachers confounded an accurate estimate of the efficacy of the digital game to improve academic achievement.

ACKNOWLEDGEMENTS

Completing this dissertation required immense perseverance, dedication, and effort; however, this dissertation would not have come to fruition without the support of many wonderful individuals. It is with pleasure and gratitude that I acknowledge their invaluable support and assistance.

First, I would like to thank Imagine Education for providing access to the data for this project.

I would like to express my deepest gratitude to my research advisor, Dr. Marilyn Welsh, for her unfailing support and advocacy, speedy review of my research, and setting me on a path to complete my degree. Thank you for seeing my potential and helping me to realize it.

Without the guidance and persistent support of Dr. John Cooney, this dissertation would not have been possible. I am indebted to him for patiently teaching me Rasch modeling, HLM, and providing me with this amazing research opportunity. His excellence as a teacher and scholar has made me a better researcher and I will forever strive to follow his example.

I am grateful to Dr. Kathryn Cochran and Dr. Jenni Harding-DeKam for being on my committee and for their encouraging words and thoughtful criticism that made the research better. I would be remiss not to acknowledge the impact of Dr. Steven Pulos on

initial interest in measurement and my realization that “it depends” answers relatively any academic-related question.

I owe significant thanks to Deb Myers for her help in pulling together a comprehensive literature review far from campus. I also appreciate the friendship of Keyleigh Gurney and Roberta Oschsner and all their assistance and cheerleading throughout the years. Connie Beard deserves special thanks for keeping me sane by helping with the particularly onerous formatting. Thanks to all my friends who let me talk about school, asked about my dissertation, or helped me to forget about the dissertation.

Last, but definitely not least, I am indebted and overwhelmed by the support of my family throughout this journey. My parents, Michael and Mary Pirraglia, helped me value education and provided me endless support and encouragement to make this dream a reality. My mother, Mary Pirraglia, was my main source of inspiration and her unwavering support and joyful care of my children provided me the opportunity to complete my dream of a doctorate.

Most importantly, I would like to thank my husband, Ian Wale, and children, Lily and Lola, for their support and love. Thank you for loving me through the difficult moments and giving me the confidence and time to finish “the paper.”

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND BACKGROUND

Digital computer and video games are played by 97% of teens (Lenhart et al., 2008) with 60% percent of teens playing more than an hour each day (Rideout, Foehr, & Roberts, 2010). Consequently, the video game industry is rapidly growing with a reported $25 billion in revenue in 2011 (Entertainment Software Association, 2013). Given the massive appeal and commercial success of these games, interest has increased in the use of instructional digital games (digital games designed specifically for training or educational purposes) for education, military training, and healthcare (Cannon Bowers, Bowers, & Procci, 2011; Chatham, 2011; Federation of American Scientists, 2002; Prensky, 2001; Ricci, Salas, & Cannon-Bowers, 1996).

A number of factors underscore the belief that instructional games are beneficial learning tools. Due to technological advancement, digital games can be played on simple platforms such as mobile devices, computers, and game consoles (i.e., Sony PlayStation 2, Microsoft Xbox and the Nintendo GameCube), making instructional games more accessible to individuals who do not have personal computers (Mitchell & Savill-Smith, 2004; Rideout et al., 2010). In addition, instructional digital games might better correspond with the learning needs of "digital natives" or today's K-12 students who have grown up immersed in and accustomed to interactive and fast-paced media presentations (Prensky, 2001). Most importantly, instructional digital games are thought

to be good instructional tools because they (a) promote active learning and feedback, (b) provide meaningful contexts to situate knowledge, (c) create engagement and intrinsic motivation, and (d) have the ability individualize instruction (Csikszentmihalyi, 1990; Gee, 2006; Kyriacou, 1992; Prensky, 2001).

Given the many hypothesized benefits of instructional games, educators have become increasingly interested in ways they could be used to improve learning. A National Summit on Educational Games convened by the Federation of American Scientists (2006) concluded that complex digital games develop many higher-order thinking skills needed in the 21st century workforce and provide practical skills training opportunities. Using digital games might also be an innovative teaching strategy that could particularly support underrepresented student populations, such as females and minority students, in their awareness and educational preparation for science, technology, engineering, and mathematics (STEM) careers (Hacker & Kiggens, 2011).

Digital games could be particularly effective tools for the improvement of mathematics achievement, especially given the needs of upper elementary and middle school students. National and international comparisons of mathematics achievement have shown that between the fourth grade and eighth grades, U.S. students start falling rapidly behind desired levels of proficiency in mathematics, consequently making them ill-prepared to succeed in college preparatory mathematics course in high school (Balfanz & Byrnes, 2006; Balfanz, Ruby, & MacIver, 2002; Beaton et al., 1996). Mathematics proficiency not only increases the probability of college and career success (U.S. Department of Education, 1997; Vogel, 2008) but is also considered to influence the nation's long-term economic future (National Mathematics Advisory Panel, 2008).

However, this issue is particularly critical for high poverty and minority students because the middle grades are when the achievement gaps further widen (Balfanz & Byrnes, 2006). On the 2011 National Assessment of Educational Progress (NAEP; 2011) mathematics exam, 40% of fourth grade students scored in the proficient or advanced categories with wide variations between students of different ethnicities (52% White, 17% Black, 24% Hispanic, 22% American Indian) and economic status as indicated by qualifying for free or reduced lunch (57% not eligible due to higher family income, 35% reduced lunch, and 23% free lunch; U.S. Department of Education, 2011). National Assessment of Educational Progress mathematics scores have shown improvement over time; however a wide achievement gap persists between White students and students of other ethnicities. Fewer students are proficient in mathematics by the eighth grade. In the 2011 NAEP test in eighth grade mathematics, 44% of White students scored in the proficient and advanced categories while only 13% of Black students, 20% of Hispanic students, and 17% of American Indian students rose above the basic level of mathematical concepts and skills. In addition, students eligible for free and reduced lunches had strikingly lower proficiency rates (17% and 28% proficient or advanced, respectively) compared to students who were not eligible (47%) due to higher family income levels (NAEP, 2011).

Despite the renewed interest in using digital games in education, previous attempts to introduce digital games in education have not been widely implemented in the classroom (Van Eck, 2006). Irrespective of the existence of games designed with educational objectives in mathematics (*DimensionM*--Kebritchi [2008]; *Zombie Division* --Habgood, Ainsworth, and Benford [2005]) and in science (*Quest Atlantis*--Barab,

Thomas, Dodge, Carteaux, and Tuzun [2005]; *Supercharged!*--Squire, Barnett, Grant, and Higginbotham [2004]), the use of digital games in the classroom remains rare (Kirriemuir & McFarlane, 2004). Teachers and administrators often have difficulty determining how a particular game is aligned with the required curriculum, especially in the case of commercial off-the-shelf (COTS) games or games available commercially (Van Eck, 2006), and lack of time to properly familiarize themselves with the game to properly implement the game and support its use (McFarlane, Sparrowhawk, & Heald, 2002). Other significant barriers were the lack of up-to-date computers and infrastructure and the lack of technical support (Van Eck, 2006). Also given the emphasis from No Child Left Behind (NCLB; 2002) legislation and the emphasis on use of rigorous "scientifically based" research interventions and standards testing, teachers must evaluate the efficacy of specific games on achievement and determine if the content of the game adequately addressed tested content (U.S. Department of Education, 2002).

The problem was a lack of empirical evidence to support firm conclusions about the effect of digital games in K-12 educational settings and particularly in the area of mathematics (Girard, Ecalle, & Magnan, 2012; Mitchell & Savill-Smith, 2004; Vogel et al., 2006). While published research on games has increased since 2006 (Hwang, Wu, & Chen, 2012), a significant portion of the literature reflects enthusiastic descriptions of the affordances of games without data (Randel, Morris, Wetzel, & Whitehill, 1992; Tobias, Fletcher, Dai, & Wind, 2011; Vogel, et al., 2006). The empirical work is plagued by serious methodological flaws and lacks theoretical foundations (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Tobias et al., 2011; Wu, Hsiao, Wu, Lin, & Huang, 2012).

The few articles that tested the impact of mathematics games on academic achievement empirically were not always positive and findings were not replicated on new samples. A literature review of mathematics digital games for middle school students revealed few games had been tested in multiple school settings, thus impeding more specific questions about for whom and under what conditions mathematics games were best suited (Dede, 2011). In the few upper elementary and middle school mathematics digital games empirically studied in multiple classroom settings (ASTRA EAGLE, Dimension M), there were inconsistent findings in terms of improvement in achievement and motivation (Bai, Pan, Hirumi, & Kebritchi, 2012; Ke, 2008a, 2008b; Ke & Grabowski, 2007; Kebritchi, Hirumi, & Bai, 2010; Ritzhaupt, Higgins, & Allred, 2011). However, it was unclear if the inconsistencies were due to unique attributes of the sample (i.e., low socioeconomic status), teacher effects, differential implementation of the digital game into the classroom curriculum, or a combination of factors.

The most frequent methodological flaw in the literature was lack of a control group, thus impairing the ability to accurately determine the effect of digital games (Girard et al., 2012; Vogel et al., 2006). Vogel et al. (2006) noted that many studies also did not include important demographic details or did not adequately describe the programs or interventions in sufficient depth to be categorized in a meta-analysis or used to generalize the effectiveness of digital games. Further complicating the clear analysis of the effect of digital games was evidence of the trend that studies with small sample sizes tended to have larger effect sizes on average then did studies with larger sample sizes, potentially leading to misleading meta-analysis findings (Slavin & Smith, 2009). Due to the rapid advancement of technology, the time lag between reseach and

publication in peer reviewed journals and inability for academic researchers to get access to proprietory digital games, games described as “current” in journal articles, might be significantly behind the latest market trends (Kirriemuir & McFarlane, 2004).

The digital games literature was further complicated by an abundance of unique, but largely interchangeable, terms and little consensus on the defining features of games and simulations. The term “serious game” was used frequently to refer to digital games that were not designed for commercial purposes; instead, they were used to teach a specific skill in training or education. The more specific term “serious educational game” referred to games that targeted K-20 content knowledge (Annetta, 2010). Similarly, Prensky (2001) and others used the term “digital game-based learning” to refer to learning from digital games that mixed educational content and entertainment. “Edutainment” was also used in the literature to refer to computer or video games that mix education and entertainment; however this term carries negative connotations of early attempts at educational software that were not successful in making games educational or entertaining (Egenfeldt-Nielsen, 2007). The current paper uses the term “instructional digital games” to refer to both video and computer games designed specifically for training or educational purposes.

There are no standardized or widely used definitions of digital games or simulations. Habgood and Ainsworth (2011) asserted the key defining characteristic of a game that separates it from films and toys is that is an “interactive challenge” (p. 171). Salen and Zimmerman (2004) synthesized other definitions of games and concluded that a “game is a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome” (p. 80). Similarly, Hays (2005) defined a game as “an

artificially constructed, competitive activity with a specific goal, a set of rules and constraints that is located in a specific context" (p. 15). As evidenced by similarities in the definitions of Hays (2005) and Salen and Zimmerman (2004), two frequent elements of a game included rules and goals and outcomes. Kirriemuir and McFarlane (2004) further defined a digital game as one that (a) "provides some visual digital information or substance to one or more players;" (b) "takes some input from the players," (c) processes the input according to a set of programmed game rules," (d) "alters the digital information provided to the players," and (e) is played on video game consoles, computers, or mobile devices (p. 6).

Furthermore, blurry distinctions between simulations and games hinder clear analysis of the impact on education. Salen and Zimmerman (2004) defined a simulation as a "procedural representation of aspects of reality" (p. 457); according to their definition, all games are a type of simulation. Gredler (1996) clarified that one difference between the two terms was that the goal for games was winning while the goal of a simulation was discovering causal relationships. Therefore, once a goal was achieved in a game, players advanced to working toward new goals in a linear fashion. Whereas in a simulation, once a goal was achieved, the player could make modifications to the variables and examine their effect on outcomes multiple times, thus characterizing a nonlinear goal structure (Gredler, 1996).

Theoretical Framework

The framework of this paper draws upon previous research from theorists in psychology and education, mathematics learning, and current mathematical principles and standards. The Learning Principle of the National Council of Teachers of

Mathematics (NCTM) standards and principles (National Council of Teachers of Mathematics, 2000) stated that students must learn mathematics with understanding by actively building new knowledge from experience and prior knowledge. As evidenced in the learning principle and throughout the standards, current mathematical theory and practice is grounded in the belief that students actively build meaning and do not passively absorb experiences (National Council of Teachers of Mathematics, 2000; Piaget & Inhelder, 1969; Sfard, 2003). Consequently, rote memorization of facts or procedures without understanding is unlikely to result in stable and useful knowledge (Schoenfeld, 1988; Stylianides & Stylianides, 2007).

The NCTM Learning Principle also stressed the need for alignment of factual knowledge (knowledge of facts) and procedural proficiency (sequence of actions) with conceptual knowledge (understanding of relationships) for students to be effective mathematical learners (Hiebert & Carpenter, 1992; National Council of Teachers of Mathematics, 2000). In contrast to previous theoretical disagreement on the absolute importance of procedural and conceptual knowledge, recent research has focused on the interrelation between factual and procedural competence and learning with understanding (Hiebert & Carpenter, 1992). Evidence suggested that conceptual understanding of written mathematical symbols and rules was important to establish before procedures became automatic due to consolidation of declarative knowledge into set procedures and the flexibility of problem solving and strategy use was decreased (Anderson, 1983; Hiebert & Carpenter, 1992). Connected and conceptually grounded ideas enable students to better remember information and promote transfer--defined as the ability to use information learned in new and unfamiliar problems (Hiebert & Carpenter, 1992).

Advocates of the situated learning theory and anchored instruction proposed that use of complex authentic mathematical problems encouraged meaningful learning and making connections between conceptual and procedural knowledge (Brown, Collins, & Duguid, 1989; Cognition and Technology Group at Vanderbilt, 1990).

The NCTM Teaching Principle states that "effective mathematics teaching requires understanding what students know and need to learn and then challenges and supports them to learn it well (National Council of Teachers of Mathematics, 2000). Teacher instruction impacts not only students' understanding of mathematics, ability to use it to solve problems, but also their confidence in using mathematics (National Council of Teachers of Mathematics, 2000). According to this principle, teachers must not only be knowledgeable about their students but also deeply understand mathematics to flexibly use a variety of pedagogical and assessment strategies to increase student knowledge. This principle draws from the work of Vygotsky (1987) and the concept of the zone of proximal development (ZPD):

> It is the difference between the actual developmental level as determined by the independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers. (p. 86)

To move students to more complex levels of understanding, teachers must use strategies to assist and guide understanding that are commonly referred to as scaffolding (Wood, Bruner, & Ross, 1976). De Jong and van Joolingen (1998) suggested that cognitive scaffolds that structure a task, take over components of a task, or offer hints or support are imperative in digital-based simulations and games to support learning. Building on the work of cognitive load theory (Sweller, 1999) and cognitive theory of

multimedia learning (Adams, Mayer, MacNamara, Koenig, & Wainess, 2012; Mayer, 2009), Lemmkuil and de Jong (2011) stated:

> Some form of guidance is needed in rich, problem-based experiential learning environments to prevent learners from missing essential information, incorrectly performing learning processes, or experiencing a cognitive overload that makes them unable to construct adequate mental representations. (p. 355)

In addition, the NCTM (2000) Assessment Principle suggested the importance of assessment to inform and guide teachers' instructional decisions as well as provide students with feedback to promote goal setting, assuming responsibility for their own learning, and becoming independent learners. Formative feedback, defined by Shute (2008) as "information communicated to the learner that is intended to modify his or her thinking or behavior to improve learning" (p. 154), generally enhances learning of low achieving students when it is simple, specific, and immediate.

Purpose of Study

The purpose of this study was to investigate the effectiveness of a mathematics digital game, Ko's Journey (Imagine Education, 2011), on seventh grade students' mathematics achievement of the Common Core State Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers [NGA], 2010). Mathematics achievement was defined by a researcher-constructed test aligned with the Common Core Mathematics Standards (NGA, 2010) and measured on a unidimensional equal interval scale (Rasch, 1960). Additionally, the Rasch (1960) measurement theory was used to identify the differential impact of Ko's Journey on the assessment items, such as what items were learned, and to evaluate need for further refinement of the current assessment.

The relatively few empirical studies of digital games, the contradictory results when used in mathematics education, and methodological flaws in empirical studies indicated a clear need for further rigorous empirical investigation of digital games to better understand if the promise of the use of digital games in education was warranted and how to best implement them in the classroom. This study addressed this need and empirically evaluated the effect of a digital mathematics game, Ko's Journey, on seventh grade students' mathematical achievement. Ko's Journey is a digital online computer game that follows Ko, a young girl in an ancient wilderness, who must make her way back to her kin. Students' progress through the game by using the guidebook and story-based math modules targeted critical areas of the seventh grade Common Core State Mathematics Standards. The mathematics topics encountered were anchored to the game and not superfluous to the overall story, e.g., helping Ko set a compass to the proper degree or mix medicine into ratios for a sick wolf pup (Imagine Education, 2012).

Research Questions

The following questions guided this study:

Q1 To what extent are the test items of the researcher-constructed test used to measure mathematics achievement aligned with the Common Core State Standards for seventh grade mathematics?

Q2 To what extent does the item level data of the mathematics assessment conform to the requirements of the Rasch (1960) model to produce a unidimensional equal-interval scale of measurement?

Q3 What is the effectiveness of Ko's Journey on students' mathematics achievement as measured by the researcher-constructed assessment of the seventh grade Common Core Mathematics Standards relative to students who do not play Ko's Journey?

Q4 Do the items of the assessment function differently for students using Ko's Journey as a supplement to normal instruction than for students who do not play the game?

Research Design

This research was conducted through the use of secondary data analysis. Although the current author was involved in the construction and evaluation of the instrument used to measure the impact of the Ko's Journey intervention, selection of participants and procedures used in the study was managed by Imagine Education program directors. The study was a pretest-posttest control group design (Gall, Gall, & Borg, 2003) of 371 seventh grade students at three middle school schools with high levels of poverty and ethnic minority students. Further information on research design, procedures, and prior work developing the instrument is included in Chapter III.

Limitations

Many of the limitations of this study were due to the use of secondary data. Secondary data were used in this research study due to the affordances of evaluating an innovative digital game using a national sample that would not be feasible otherwise. However, due to the use of secondary data, desired information about implementation, participants, and schools was often unavailable; therefore, the research questions were restricted to available data. Although individual demographic information was not gathered in the study by program directors, school demographic and achievement information was available and collected to be able to infer to whom the results of this study might be generalizable.

CHAPTER II

REVIEW OF LITERATURE

This chapter reviews literature on the promise of using digital games in education and specifically the use of digital games to learn mathematics. The first section details the theoretical rationale behind using digital games in education. The second section provides an empirical review of the effects of digital games on student achievement in mathematics starting with the broad topic of computer-aided instruction, further refining to computer-aided instruction and mathematics, and finally digital mathematics games. The third section reviews specific characteristics of current digital mathematics games and identifies unique characteristics pertaining to Ko's Journey including Common Core State Standards (NGA, 2010), narrative, and intrinsic integration.

Promise of Digital Games in Education

Given the pervasive use of video games and the dedication of the players, it is not surprising educators and policy makers are interested in the effects of the games on children and adolescents and how some of the motivating aspects of video games might be harnessed to facilitate learning. There is general agreement with Clark (1983) that what makes digital games good for learning is not only that they are games but that they incorporate significant learning principles and instructional design (Gee, 2007; O'Neil, Wainess, & Baker, 2005; Van Eck, 2006). Proponents of using digital games to teach educational content point to the potential of games to better individualize instruction

through a learner centered approach (Prensky, 2001). In reviewing literature on the promise of using digital games for instruction, several key components theorized to lead to the success of digital games included active learning and adaptivity, situated in meaningful context, and the motivational theories of flow and cognitive evaluation theory.

Active Learning and Adaptivity

Advocacy for the use of active learning strategies in the teaching of mathematics is not a new phenomenon (Bonwell & Eison, 1991; Kyriacou, 1992). The term "active learning" has been applied to a wide variety of learning techniques including computer-assisted learning, small group discussions, and collaborative problem-solving; it is contrasted with passive learning methods such as lecture (Kyriacou, 1992). Active learning is described by Kyriacou (1992) as the use of learning activities where students are given a degree of ownership and control in the learning activity and learning experiences are typically open-ended where students can actively participate and shape the learning experiences (p. 309). In addition to considering the nature of the learning experience, active learning can also be examined in terms of the students' mental processing. In this way, active learning is characterized by developing meaningful understanding through cognitive restructuring of the information, which is contrasted with passive mental processes such as rote learning (Kyriacou, 1992).

Digital games and simulations are particularly suited to engage students in active learning (Prensky, 2001). Some digital games allow individual students to direct their learning by making personal decisions in the game including the development of plans and goals for which they receive instantaneous feedback (Prensky, 2001). Additionally,

digital games can provide opportunities for practice of academic skills and feedback, typically referred to as drill and practice (Prensky, 2001). One of the most powerful tools of digital games is adaptivity. Adaptivity refers to the changing of the game in response to the player's success and progress and is intended to maintain players in an optimal level of difficulty and challenge (Prensky, 2011). Through artificial intelligence and continuous recording and monitoring of player performance, a complex digital game can adapt. For example, a digital game can provide players more challenge by adding difficulty or removing scaffolding, or making the game easier by adding resources, or reducing the difficulty or number of challenges. Facilitating an optimal level of challenge for students was a key component in Vygotsky's (1987) zone of proximal development and Csikszentmihalyi's (1990) theory of flow.

Situated Cognition and Anchored Instruction

Digital games are well-suited for providing students with a meaningful and relevant context for learning mathematics. According to proponents of situated learning, knowledge is contextually situated and is significantly influenced by the activity, context, and culture in which it is learned and used (Brown et al., 1989; Lave & Wenger, 1991). Thus, digital games and other technology have the potential to expand meaningful learning experiences in schools through authentic activities and social interaction (Brown et al., 1989). By situating educational content in a digital game or simulation, it is possible to engage students to develop goals, have legitimate roles in managing learning, and develop deep understanding of content (Barab, Gresalfi, & Ingram-Goble, 2010; Gee, 2007). Students are able to take the role of a scientist, mathematician, or engineer and develop "situated understanding in the context of activity and experience grounded in

perception" (Gee, 2006, p. 4). Teaching is guided by a cognitive apprenticeship model that uses authentic activities and social interaction to enable students to begin to think meaningfully and purposefully (Brown et al., 1989). Gee (2006) uses the example of the work of diSessa (2000) to illustrate how students use a computer programming language called *Boxer* (diSessa & Lay, 1986) to understand the algebra behind Galileo's principles of motion. Because students are able to manipulate and elaborate the programming, they develop a deeper understanding of the algebraic equations because they make connections to the material in a situated and embodied way rather than simply seeing the algebra as a set of symbols to be passively regurgitated on a test (Gee, 2006).

Anchored instruction is a practical application of the situated learning framework, utilizing technology in authentic tasks and apprenticeship models. Anchored instruction, based on the framework of situated cognition (Brown et al., 1989), is an attempt to avoid "inert knowledge" (Whitehead, 1929) or knowledge that can be recalled but not effectively used. Anchored instruction incorporates an authentic problem (also termed an anchor), visual technology (i.e., videodisc, computer), and apprenticeship learning to make information more transferable and usable (Cognition and Technology Group at Vanderbilt, 1992b). One successful example of anchored instruction is the Jasper Woodbury Series created by the Cognition and Technology Group at Vanderbilt (CTGV; 1992a) in which students engage in critical thinking and mathematical problem formation and problem solving in 12 videodisc adventures. The series is based on the following seven theory-based design principles: video based format, narrative with realistic problems (rather than a lecture on video), generative format, embedded data design, problem complexity, pairs of related adventures, and links across the curriculum

(Cognition and Technology Group at Vanderbilt, 1992a). A large evaluation of the Jasper Woodbury Series indicated that students using the Jasper activities significantly outperformed control classrooms in mathematic word problems and planning and had less anxiety and better attitudes toward mathematics (Cognition and Technology Group at Vanderbilt, 1992b).

Digital games might also be a tool to help disadvantaged students overcome inequalities in content and technological knowledge. Gee (2003) suggested that socioeconomically disadvantaged students have fewer opportunities inside and outside the classroom for authentic and embodied experiences. These immersive experiences allow students to situate abstract principles in context and allow students to fully appreciate what it is like to think like a professional in a field (geologist, chemist, or mathematician, etc.). Although families and schools of higher socioeconomic students could be expected to provide these immersion experiences, such experiences are typically not available for lower socioeconomic students and result in inequalities in knowledge and experience (Dai & Wind, 2011). Digital games in school or after school programs could provide opportunities to reduce or close gaps of socioeconomic experiences and prior knowledge by exposing players to virtual environments likely to be more meaningful, less stressful, and less anxiety-producing than typical classroom learning (Squire, 2006). Following similar reasoning, digital games are also being considered as an avenue to increase minority and female students' interest in science, technology, engineering, and mathematics (STEM) areas--two populations that are typically underrepresented in STEM courses and careers (DiSalvo, Crowley, & Norwood, 2008; Hacker & Kiggens, 2011; Meluso, Zheng, Spires, & Lester, 2012) .

Intrinsic Motivation: Flow and Cognitive Evaluation Theory

A key rationale for the use of games in education is the desire to harness the motivational power of games to make learning fun and engaging. Csikszentmihalyi's (1990) theory of flow is often used to explain the affective and motivational power of digital games. After interviewing artists, chess players, rock climbers, and others, Csikszentmihalyi developed flow theory to describe the "state where people are so involved in an activity that nothing else seems to matter; the experience itself is so enjoyable that people will do it even at great cost, for the sheer sake of doing it" (p. 4). According to flow theory, an activity is seen as rewarding in relation to an individual's assessment of attractiveness and challenge and whether they believe they have the skills needed to accomplish the task. According to Csikszentmihalyi, state of flow, also called optimal experience, has one or more of the following characteristics: (a) a challenging, but accomplishable task; (b) ability to concentrate on the task; (c) clear goals; (d) immediate feedback; (e) deep but effortless involvement (losing awareness of everyday worry and frustration of everyday activity); (f) ability to exercise control over actions; (g) concern for self disappears during flow, but sense of self is stronger after flow activity; and (h) sense of duration of time is altered (p. 49).

Although the theory was not developed using computer game players, many of the characteristics and criteria for flow could be useful for describing individuals learning with computers (Ghani & Deshpande, 1994) and the potential for learning effectiveness of games (Prensky, 2001). Kirriemuir and McFarlane (2004) suggested that game designers should look to conditions of flow to create game environments that best support learning. However, Prensky (2001) articulated that the biggest challenge was keeping

someone in the flow state in the game and learning simultaneously—something to which he believed game designers and good digital games were well adapted. Because flow is interrupted for players when the game is too easy or difficult, good digital games are easy to learn, hard to master, and highly adaptable to adjust to the needs of a variety of skill levels to provide an adequate challenge to keep players in flow states (Prensky, 2001; Salen & Zimmerman, 2004). The flow of gameplay could also be interrupted by educational content if the core mechanics and content were not intrinsically integrated (Habgood et al., 2005). Therefore, educational content must be seamlessly connected to gameplay for the flow experience to enhance learning (Habgood et al., 2005).

The motivational appeal of playing digital games could also be explained by self-determination theory and cognitive evaluation theory (Ryan & Deci, 2000). Self-determination theory is a theory of human motivation primarily concerned with how social context provides exeriences that satisfy universal human needs (Przybylski, Rigby, & Ryan, 2010; Ryan & Deci, 2000). Cognitive evaluation theory, a subtheory of self-determination theory applied to sports, educationm and leisure activities, suggests activities that satisfy the fundatmental human needs of competence (sense of efficacy), autonomy (volition and personal agency), and relatedness (social connectedness) are more likely to be intrinsically motivating (Przybylski et al., 2010). Factors that enhance extrinsic motivation such as rewards, punishments, and evaluations typically decrease intrinsic motivation (Deci, Koestner, & Ryan, 1999). Although games do have virtual rewards such as points and promotion to advanced levels, few players receive external extrinsic rewards to play; they typically play because the games are fun (Przybylski et al., 2010).

Gender, Mathematics, and Digital Games

Gender is an important factor in this study given gender differences reported in both mathematics achievement and the use of digital games. Despite general agreement of the existence of a gap between males and females in mathematics, the precise size of the gap varies from test to test and to some degree from research method (Ellison & Swanson, 2010). Results from Program for International Student Assessment and National Assessment of Educational Progress assessments suggest that gender gaps in mathematics are inconsistent and sufficiently small to be of no practical importance (Ellison & Swanson, 2010; Freeman, 2004; Lindberg, Hyde, Petersen, & Linn, 2010). However, research using the Early Childhood Longitudinal Study data (ECLS-K) found that although males and females had equivalent mathematical skills in kindergarten, females lost ground in elementary school and were still significantly behind boys in mathematics in eighth grade (Robinson & Lubienski, 2011).

Consistently males of all ages are found to be more avid players of digital games and report significantly more daily playtime (Bourgonjon, Valcke, Soetaert, & Schellens, 2010; Greenberg, Sherry, Lachlan, Lucas, & Holmstrom, 2010; Lowrie & Jorgensen, 2011; Rideout, Foehr, & Roberts, 2010). Although both genders play digital games, a large Kaiser Family Foundation study found that American boys reported playing digital console games nearly an hour a day while girls reported playing under 15 minutes (Rideout et al., 2010). Digital game usage peaks for both genders around the age range of 11-14 years old (Greenberg et al., 2010; Rideout et al., 2010). Accordingly, Bourgonjon et al. (2010) suggested wide diversity in experience with game technology

and that less experienced students, including many girls, might need additional support and instruction on game play to receive equivalent benefits of instructional digital games.

In addition to differences in the amount of time playing digital games, girls and boys tend to prefer different game genres. Boys are more likely to be intensive gamers—playing for an extended period of time and also playing a wider range of game genres (Lenhart et al., 2008). Girls and boys are equally likely to play racing, rhythm, simulation, and virtual world games; while girls are significantly more likely to play puzzle games than boys (Lenhart et al., 2008). In all other genres including action, sports, adventure and first-person shooters, boys are significantly more likely to play (Lenhart et al., 2008). In terms of educational games, girls prefer games that require problem solving, quantitative computations, and interpretation of graphs, while boys prefer adventure games that have a journey-based storyline and require visual and spatial reasoning skills (Lowrie & Jorgensen, 2011). Nonetheless, the types of games that are more applicable for learning such as strategy, adventure, and role-playing are popular for both girls and boys (Steiner, Kickmeier-Rust, & Albert, 2009).

One of the unique characteristics of Ko's Journey is that the main character is a Native American female. In commercial video games, a lack of primary female characters, gender stereotypes, and overly sexualized female characters have been well documented and is potentially a reason traditional games have not had strong appeal for girls (Steiner et al., 2009). It has been found boys and girls strongly prefer avatars of the same gender and could be a strength of the Ko's Journey to interest girls (Inal, Sancar, & Cagiltay, 2006). Despite this preference, a study of the digital game Phoenix Quest found that the presence of a female avatar did not discourage boys from playing or

enjoying the game. De Jean, Upitis, Koch, and Young (1999) found that the game appealed to girls because the main character was their age and gender and included problem solving; however, boys were also engaged in the game.

Barriers to the Implementation of Digital Games

Although teachers and parents recognize that digital games can support valued skills, numerous barriers exist for teachers and administrators to successfully integrate them in the classroom. Teachers reported difficulty with quickly identifying how a specific game would be relevant to the required curriculum as well as a lack of time to familiarize themselves with a game to properly implement and support the game's use (Kirriemuir & McFarlane, 2004). Teachers and school leaders also had difficulties persuading stakeholders of the educational benefits of using digital games in the classroom, making funding difficult (Kirriemuir & McFarlane, 2004). However, the most critical obstacle was the lack of correspondence between the skills and knowledge in a game and those recognized and tested in educational settings (McFarlane et al., 2002). Without clear information about the content of the game, correspondence with current educational standards, and trustworthy information on empirical effectiveness, teachers do not have sufficient information or evidence to incorporate games in the classroom—a clear gap in the literature this study addressed.

Empirical Review of the Impact of Games on Mathematics Achievement

Despite an increase in published literature on digital games in education, empirical evidence regarding the effectiveness to impact educational achievement is still building. Due to a scarcity of quantitative research on mathematics games, this literature

review starts with the broad topic of computer-aided instruction, further refines this to computer-aided instruction and mathematics, and finally to digital mathematics games.

Computer Aided Instruction

Computer aided instruction and computer assisted instruction (CAI) are broad terms that refer to programs that use technology to enhance instruction. Generally, the terms encompass a wide variety of applications including games and simulations, intelligent tutoring systems (Cognitive Tutor), and drill-and-practice software (Li & Ma, 2010; Slavin, Lake, & Groff, 2009). Research on educational technology has been prevalent. At least 60 meta-analyses on technology in education have been published in the literature since 1980--each focused on specific aspects such as subject matter, grade level, type of technology, and each employed unique article inclusion criteria (Tamim, Bernard, Borokhovski, Abrami, & Schmid, 2011). Despite the multitude of reviews, findings were inconsistent due to different procedures for article inclusion and the issue of different computer technology interacting with student and environmental characteristics (Li & Ma, 2010). In an attempt to gain more insight on the overall impact of technology in education, second-order meta-analyses were used to synthesize the findings of the growing body of meta-analyses on the topic. Tamin et al. (2011) conducted a second-order meta-analysis of 25 meta-analyses on computer technology integration published after 1985. The meta-analyses were selected because of minimal overlap in primary literature and together totaled 1,055 primary studies (Tamim et al., 2011). The average effect size of the 25 meta-analyses was +0.33, indicating that the average student in a classroom using technology performed 12 percentile points higher than the average student in a traditional setting that did not use technology to enhance

achievement (Tamim et al., 2011). Technology used as direct instruction (i.e., computer assisted instruction) was found to have a greater impact when used to support instruction (i.e., word processing or simulations) and when used with K-12 populations compared to postsecondary use (Tamim et al., 2011). The average effect size of the four included meta-analyses on mathematics instruction was +0.32 (Tamim et al., 2011).

Computer-Aided Instruction in Mathematics

Numerous meta-analyses exist that examine the effect of learning with computer technology (Kulik, 2003; Kulik & Kulik, 1991; Kulik, Kulik, & Bangert-Drowns, 1985; Liao, 1998, 2007; Lou, Abrami, & d'Apollonia, 2001), with most addressing multiple school subjects and ages. However until recently, most meta-analyses did not focus specifically on the impact of computer technology on mathematics achievement (Li & Ma, 2010). Similar to Tamin et al. (2011), a secondary review of 20 meta-analyses on the effectiveness of educational technology applications for enhancing mathematics achievement in K-12 (Cheung & Slavin, 2011) concluded there were positive effects of educational technology on mathematics achievement with an overall study-weighted effect size of +0.31; however, effect sizes ranged from +.10 to +.62. One major factor in the wide range of average effect sizes in the studies was attributable to the different procedures used for study inclusion and analysis and the different types of technology included (Cheung & Slavin, 2011).

Given the rapidly changing technology, recent meta-analyses on technology applications specifically designed for mathematics instruction provided the most information on the impact of current computer assisted instruction (CAI) in mathematical applications. A meta-analytic review by Slavin and Lake (2008) on elementary

mathematics CAI found an effect size of +0.19 for 12-weeks or more elementary technology interventions. A similar study by Slavin et al. (2009) on secondary students found a smaller effect size (+.10) of CAI for middle and high school students' mathematics achievement. Similarly, a national study by the Department of Education (Campuzano, Dynarski, Agodini, & Rall, 2009), using a large national sample of over 132 schools and 428 teachers and experimental design with random assignment of teachers to treatment or control groups, found no significant effects of technology applications (e.g., Cognitive Tutor, PLATO, Larson Pre-Algebr*a*) on sixth grade or Algebra 1 students' mathematics standardized test scores during the two year study.

A meta-analysis by Li and Ma (2010) found greater effects of computer technology on secondary student's mathematics learning (+.28) but employed more lenient criteria for inclusion in the meta-analysis than did Slavin, Lake, Chambers, Cheung, and Davis (2009) and Slavin et al. (2008). In contrast to the best-evidence synthesis by Slavin et al. (2009) and the Department of Education study by Campuzano et al. (2009), the Li and Ma meta-analysis included short -term interventions and pre-post test studies without a control group—both moderator variables were shown to increase effect size (Liao, 2007; Slavin & Lake, 2008). Similar to the Slavin et al. studies (Cheung & Slavin, 2011; Slavin & Lake, 2008; Slavin, Lake, & Groff, 2009), Li and Ma found a greater impact of technology on elementary students than with secondary students. Additionally, smaller effect sizes were found for the most recent CAI applications compared to earlier technology (Li & Ma, 2010; Liao, 2007)--studies using rigorous scientific methods such as randomized experiments, randomized quasi-experiments (Liao, 2007; Slavin & Lake, 2008), and shorter CAI treatments (Kulik &

Kulik, 1991; Li & Ma, 2010). Possible methodological issues such as using non-standardized tests (vs. standardized; Li & Ma, 2010) and using different teachers for treatment and control groups (vs. the same teacher; Kulik & Kulik, 1991) were also significant moderator variables that increased the effect size of the use of technology applications on academic achievement.

Methodologists examining trends in mathematics computer assisted instruction (CAI) have noted that studies with small sample sizes tended to have larger effect sizes than did studies with larger sample sizes (Slavin & Smith, 2009). This phenomenon has been described in medicine (Finckh & Tramèr, 2010; Richy, Ethgen, Bruyere, Deceulaer, & Reginster, 2003) and education, specifically elementary and secondary mathematics (Slavin & Lake, 2008; Slavin, Lake, Chambers, et al., 2009; Slavin & Smith, 2009). A meta-analysis (Slavin & Smith, 2009) of 185 studies of elementary and secondary mathematics program found that studies with sample sizes below 250 had significantly larger effect sizes (+0.27) as compared to those studies with larger samples (+0.13). Furthermore, differences in sample size appeared to have a greater effect on effect size than differences in methodology including random or matched assignment to treatments (Slavin & Smith, 2009).

There are various explanations hypothesized for small study effects including publication bias, teacher effects, and superrealiztion. The main reason suggested, also called the file drawer effect (Rosenthal, 1979), was that small studies with large significant effects were more likely to be published than small “underpowered” studies with nonsignificant effects (Slavin & Smith, 2009; Wouters & van Oostendorp, 2013). Studies with larger sample sizes were more likely to have adequate power to detect

significant effects and funding from outside sources. Therefore, studies with large samples were more likely to publish at least a technical report that could later be included in future reviews even in the absence of significant findings. In small randomized studies, teacher effects could also significantly impact effect sizes; however the issue of publication bias suggested that only those finding large positive effects would end up published. Therefore, the small sample effect could potentially artificially inflate the true effectiveness of educational technology when the sample size was not considered a moderator variable in meta-analtyic analyses.

Another consequence of small study effects was superrealization, a term coined by Cronbach et al. (1980). Superrealization refers to small studies where the implementation or treatments by experimenters created unrealistic conditions, such as intense one-on-one tutoring for all students, that could never be replicated in other environments or scaled-up (Slavin & Smith, 2009). Evidence of small study effects and superrealization might partially explain the null result of the large national Department of Education study on computer technology (Campuzano et al., 2009) compared to the larger effects found from the various meta-analyses on the same topic that were potentially biased.

In summary, although specific mathematics CAI applications varied widely in their effectiveness, systematic reviews of CAI for mathematics indicated consistent small positive effects of this technology on improving academic achievement. However, caution must be used in interpreting and generalizing these findings because of the influence of small study effects and superrealization.

Digital Educational Mathematics Games

Research more specifically on educational computer games and interactive simulations, a subset of computer aided instruction, is a relatively new focus; therefore, evidence of the impact of gaming and interactive simulations is still tentative and building. Published research on games has increased dramatically since 2006 (Hwang et al., 2012); however, a significant portion of the literature reflected enthusiastic descriptions of the affordances of games (Randel et al., 1992; Tobias, et al., 2011; J. J. Vogel, et al., 2006) and the empirical work is plagued by serious methodological flaws and the lack of theoretical foundations (Connolly, et al., 2012; Tobias et al., 2011; Wu et al., 2012). Another difficulty was that given the rapid advancement in computer technology and games and the substantial lag between research and publication, games described as "current" in journal articles might be significantly behind the latest market trends (Kirriemuir & McFarlane, 2004).

Due to the scarcity of empirical research and the heterogeneity of the existing empirical studies, narrative reviews were more common than meta-analytic studies in the area of digital games and simulations. A review article about games (Randel et al., 1992) covering the years 1984 to 1991 reported that of the 67 articles included, 38 found no differences between digital games and traditional teaching methods, 22 favored games, five favored games but had questionable control groups, and three favored conventional instruction. Randel et al. (1992) reported math was the most promising domain for the use of digital games with seven out of eight studies showing significant gains in mathematics compared to traditional instruction.

Similar to Randel et al. (1992), other researchers (Connolly et al., 2012; Girard et al., 2012; Young et al., 2012) with intentions to synthesize research on the effectiveness of games and simulations opted for a narrative review given the lack of empirical research needed for a meta-analysis. Connolly et al. (2012) found only four studies with sufficient empirical evidence about effects of digital games on mathematics from an initial search result of 7,392 articles . In contrast to Randel et al. (1992), Young et al. (2012) found evidence for effects of games on language learning, history, and physical education but only mixed evidence in the areas of math and science. Girard et al. (2012) included only one empirical study of a mathematics computer game (*DimensionM*) out of the 11 games and it was one of three games that had a positive effect on learning. One of the few empirical reviews of games and simulations (Vogel et al., 2006) found significantly higher cognitive gains (+0.13) for participants using simulations or games compared to traditional instruction but the low number of included articles (32) prevented further investigations of moderator variables.

Instructional support, defined as support for cognitive processing such as providing feedback, scaffolding, or giving advice, that is built into games can influence their effectiveness. A meta-analysis on game-based learning and the benefit of instructional support (Wouters & van Oostendorp, 2013) found that instructional support improved learning (+0.34), especially for mathematics games (+0.40). Instructional support that helped learners select relevant information was more beneficial to learning than instructional support facilitating the organization or integration of information (Wouters & van Oostendorp, 2013). Use of a story line to help students organize educational information (i.e., narrative structure) had the lowest impact of all types of

instructional support. Wouters and van Oostendorp (2013) discovered a publication bias in effect sizes for articles found in peer-reviewed journals (+0.44) compared to conference proceedings (+.08) and unpublished studies (+0.14), and of posttest only (+.056) compared to pretest-posttest (+0.16).

In the research literature, some researchers asserted some types of games, such as drill and practice games, were not as effective in improving learning or skills as other game genres such as simulations (Ke, 2008a; Wenglinsky, 1998). However, the support for this claim was not strong based on empirical studies. Reviews of computer games and technology (Li & Ma, 2010; Randel et al., 1992; Vogel et al., 2006) found that all types of technology applications had the same effects on the mathematics achievement of students. In the area of computer assisted instruction (CAI), Slavin, Lake, and Groff (2009) found that supplemental CAI applications had greatest impact on academic achievement as compared to core CAI applications and computer-managed learning systems.

Echoing findings from the previous mentioned reviews on games, modern 3D mathematics computer games are currently being used and researched for upper elementary, middle school, and secondary students. The following research review on mathematics games targeted for older children and adolescents provides a foundation to discuss the unique properties of Ko's Journey.

Freeware

Ke (2013) investigated the impact of using multiple freeware mathematics games on middle school students' mathematics achievement, attitudes toward mathematics, and mathematics self-efficacy. Computer games included Detention, Factor Dazzle, Fantasy

Stock Exchange, Sim Lemonade Stand, Ker-Splash, Late Delivery, Square Off, Bathroom Tiles, Turtle Pond and Lure of the Labyrinth. Participants from a rural Native American school (N = 15) and an urban Hispanic school (N = 51) played the computer games with a trained tutor for 10 hours over five weeks. The study found significant improvement in the Pueblo school students' state test performance but no significant difference for the urban school (Ke, 2013); however, caution must be used interpreting this research due to the lack of control group and small number of participants. Also, given the number of different games and short time span playing each game, it was impossible to determine specific effects of individual games.

Zombie Division

Zombie Division (Habgood & Ainsworth, 2011) is a 3D adventure game designed to teach 7-11-year-olds division. The game was designed to empirically test the impact of intrinsic integration or integration of gameplay and/or game mechanics with the learning objectives. In the main intrinsic version, players defeat enemy zombie skeletons wearing numbers on their chests by attacking them with a divisor that divides the zombie skeleton's number into whole numbers. In the extrinsic version, students battled skeletons using symbolic representations of possible attacks (i.e., swords, shields, gauntlets) instead of the mathematical content and then took an end-of-level quiz. Students who played the intrinsic version had significantly higher learning gains at both the immediate and delayed posttests of mathematics achievement. There were no significant differences between girls and boys and no interaction within game conditions (Habgood & Ainsworth, 2011). When given a choice between playing an intrinsic or extrinsic version of Zombie Division during a computer club, students played the

intrinsic version seven times longer than the extrinsic version of the same game (Habgood & Ainsworth, 2011).

DimensionM

DimensionM is a modern 3D digital game with multiplayer options teaching pre-algebra and algebra. DimensionM was designed to be similar to a first person shooter game in which players experienced the game through a first-person perspective of the protagonist (21-6 Productions, 2013). Players are immersed in a fast-paced 3D environment and players must solve mathematical problems quickly to continue missions (Kebritchi et al., 2010). A study of 193 algebra and pre-algebra students in 10 high school classrooms found that the treatment group using DimensionM 30 minutes a week for 18 weeks significantly increased mathematics achievement on district tests; however, there was no significant difference in motivation. Prior mathematics achievement, computer skills and English skills were not found to be indicators of the students' posttest motivation or mathematics achievement (Kebritchi, 2008).

A second study (Ritzhaupt et al., 2011) investigating DimensionM with 225 low-SES middle school students found no impact of the game on academic achievement but did find positive changes in students' attitudes toward mathematics and mathematics self-efficacy. Similar to Kellbritchi, Hirumi and Bai (2010), a study (Bai, et al., 2012) on the impact of DimensionM with 437 eighth grader found significantly larger academic gains in the treatment group, but no significant increases in motivation. Gain scores were used because the treatment group had significantly lower academic achievement and motivation scores at the pretest compared to the control group. Research on DimensionM

illustrates the situated nature of computer games and the lack of solid evidence for the effectiveness of specific games across different populations.

ASTRA EAGLE

ASTRA EAGLE is a set of web-based games developed by the Center for Advanced Technologies (2002). Games include "Treasure Hunt"--students locate X and Y coordinates on a map in order find treasures, "Cashier"--students must play a cashier and do math calculations of money, "Tic Tac Toe"--students win by correctly answering mathematics questions, and "Up, Up and Away" students solve math problems to continue to fly a hot air balloon (Ke, 2008a). Eight ASTRA EAGLE mathematics games for upper elementary students (fourth and fifth grades) were used to facilitate mathematics achievement and positive attitudes toward math learning. A study with 125 fifth graders found the game-playing conditions increased academic achievement regardless of gender and socioeconomic status. Cooperative gameplay significantly increased low socioeconomic students' positive attitudes compared to competitive gameplay or paper-based drills (Ke & Grabowski, 2007). Follow-up studies with 15 fourth and fifth grade students (Ke, 2008a) and 487 fifth grade students (Ke, 2008b) failed to find game-playing differences in academic achievement or meta-cognition. However, there was evidence the games increased positive attitudes toward math learning. It was unclear what factors contributed to the varying results of the game on academic achievement but differences in student and school characteristics (i.e., low SES), implementation differences, and teacher effects were possible considerations. Qualitative observations indicated students did less random guessing in games in which mathematics learning and game objectives were intrinsically integrated ("Treasure Hunt"

and “Cashier”) than in other games where the mathematics content was extrinsic and more of a necessary chore to complete to continue to play the game (“Tic, Tac, Toe” and “Up, Up and Away”; Ke, 2008a).

In conclusion, few modern digital games have been systematically researched or used for middle school mathematics. While some studies (Bai et al., 2012; Kebritchi et al., 2010) found DimensionM improved academic achievement but had no impact on motivation, another study (Ritzhaupt et al., 2011) found the algebra game did not significantly improve achievement but did increase motivation. Similarly, ASTRA EAGLE was found to positively increase mathematics achievement in one study (Ke & Grabowski, 2007) but only increased motivation in follow-up studies (Ke, 2008a, 2008b). Given the lack of empirical research on games, especially using the same game in multiple contexts, it was difficult to clearly understand why the games were successful in improving academic achievement in one school but not in another. Given the mixed results using the same games, extreme caution must be used to avoid universal claims that a game is effective and instead concentrate on “what works, when, for whom,” and conditions needed for success (Dede, 2011, pp. 237-238). Research on *Zombie Division* and ASTRA EAGLE also provided evidence of the importance of intrinsic integration of mathematics content to increase learning and student interest (Habgood & Ainsworth, 2011; Ke, 2008a).

Characteristics of Ko’s Journey

The digital game, Ko’s Journey, was the focus of this study; it is an online computer mathematics game created by Imagine Education (2012) to improve middle school students’ mathematics achievement by increasing interest and understanding

through intrinsic integration of Common Core State Mathematics Standards and a narrative story. The concept of Ko's Journey was created by a middle school mathematics teacher for his rural low-income and traditionally low performing middle school students in the southwestern United States (Imagine Education, 2012). Originally developed as a large board game, the game was later developed into a web-based model.

The game follows Ko, a young girl in an ancient wilderness, who must make her way back to her kin after her village was destroyed by fire. Students progress through the game by using a guidebook and story-based math modules targeting critical areas of the seventh grade Common Core State Math Standards. The mathematics topics encountered are intrinsically integrated to the game mechanics and overall story, e.g., helping Ko set a compass to the proper degree or mix medicine into ratios for a sick wolf pup (Imagine Education, 2012). The student playing the role of Ko receives scaffolding and support in problem solving from interaction with her Spirit Grandfather and the guidebook (Dickey, 2006). Imagine Education suggests the program will improve mathematics achievement through a functional and repetitive approach targeting the Common Core State Standards as well as make the mathematics authentically interesting without arbitrary rewards such as points and awards. Ko's Journey is not intended to function as a stand-alone mathematics program and the authors have developed supplemental materials to assist teachers with integrating the game into to their traditional curriculum. The mathematics content of Ko's Journey is detailed in an educator guide (Imagine Education, 2013) with suggestions for additional classroom activities. Table 1 summarizes the mathematical content and scenarios.

Table 1

Ko's Journey Mathematical Content and Scenarios

Lesson	Mathematics Content	Task
Compass and Travel	Simple multiplication Degrees of a Circle Adding Fractions with non-common denominators Reading a Graph Determining Distance Applying to Scale	Using a compass, learn about degrees of a circle. Using the guidebook, determine speed of travel and then calculate distance and apply the number to scale.
Arrow Balance and Travel	Number Relationship Division Units of Measure	Match different arrow lengths with weighted points and fletch lengths.
Medicine Poultice and Travel	Estimation Percentage of a Number Complex Ratios Determining a Variable	Use the guidebook to make a medicinal poultice with the correct ratios to save a wounded wolf pup.
Luna and Travel	Cartesian Coordinates Line Equations Percentage of a Number Rounding	Calculate a change in "basic velocity" and use Cartesian coordinates and line equations to determine the location of the North Star.
Bolsa and Travel	Large number multiplication Volume Computation Multi-step equations Determining a Variable	Determine how many droplets of water the travelers need for the crossing and cutting the height of the bolsa to the proper length.
Crystal Oasis	Estimating angles Subtraction Fractions with non-common denominators	Students must take a percentage of a number every time they travel and round it to the nearest tenth. Students choose a crystal and record angle of exiting and incoming light. Weight of crystals (1/3 stone) is then calculated.

(table continues)

Table 1 Continued

Lesson	Mathematics Content	Task
The Secret Circle	Diameter, radius and circumference Determining an unknown variable Relationship of parts of a circle	Students must determine the radius of three circles from the circumference.
The Great Mountain Climb	Determining slope Division Rounding Order of Operations Using a basic algorithm	Determine the slope of a mountain climb and then enter it into the guidebook to determine velocity.
The Crystal Cave	Supplementary Angles Division Rounding Determining variables Problem organization Working with decimals and scale Order of Operations Using a basic algorithm All basic functions	Place a crystal upon a staff of the correct height (matching the radius of the secret circle), which is placed on the correct etching on the cave floor for the light to come in at the correct angle to send a 180 degree beam into the cave opening.

Although a comprehensive literature review of design and pedagogical principles inherent in Ko's Journey was out of the scope of this paper, a review of unique aspects of Ko's Journey is useful in understanding key foundational tenets underpinning the objective for increased academic achievement.

Common Core State Standards Initiative. One unique aspect of Ko's Journey was that the game was created to teach the critical areas of the Common Core State Mathematics Standards. The Common Core State Standards Initiative (CCSSI) is a current state-led effort by the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO) to develop a set of common standards in the areas of English language arts and mathematics NGA, 2010).

The standards were developed in collaboration with teachers, school administrators, and subject matter experts to prepare students for college and careers and have currently been adopted by 45 states and three territories. According to the Common Core Standards website, the standards

> (1) are aligned with college and work expectations; (2) Are clear, understandable and consistent; (3) Include rigorous content and application of knowledge through high-order skills; (4) Build upon strengths and lessons of current state standards;(5) Are informed by other top performing countries, so that all students are prepared to succeed in our global economy and society; and (6) Are evidence-based. (NGA, 2010)

The mathematics standards define what students should know and what they should be able to do with that information. For teachers to be able to assess whether or not a student has that mathematical understanding, students must be able to justify why a mathematical rule is correct (NGA, 2010)). The standards emphasize challenging multistep and authentic questions that would be encountered in everyday life. In the seventh grade, the key instructional components are:

> (1) developing understanding of and applying proportional relationships; (2) developing understanding of operations with rational numbers and working with expressions and linear equations; (3) solving problems involving scale drawings and informal geometric constructions, and working with two- and three-dimensional shapes to solve problems involving area, surface area, and volume; and (4) drawing inferences about populations based on samples. (NGA, 2010)

The mathematics scenarios in Ko's Journey (Imagine Education, 2013) incorporate multi-step problems that are emphasized in the Common Core State Mathematics Standards (NGA, 2010). A comparison of the mathematical content of Ko's Journey summarized in Table 1 and the seventh grade Common Core State Standards mathematics domains (ratios and proportional relationships, number system, expressions and equations, geometry, and statistics and probability) suggested that while the game

incorporated mathematics content in the first four domains, content from the statistics and probability domain was absent.

Narrative. According to a various number of researchers, there are numerous motivational and cognitive benefits of integrating a narrative or story-line within a digital game. Use of narrative within a digital game is hypothesized to increase the engagement of learners in game activity by helping them emotionally connect and identify with characters (Dickey, 2006; Lee, Park, & Jin, 2006; Yelland & Masters, 2007), use effortful and meaningful learning strategies (Rieber, 1996; Salomon, Perkins, & Globerson, 1991), and scaffold and support problem solving (Dickey, 2006; Gee, 2007; Vygotsky, 1987; Wood et al., 1976; Yelland & Masters, 2007). Through identifying with the virtual character of Ko, persistence and interest increase because players become invested in the goals of the game and resolving challenges presented (Gee, 2007; Hefner, Klimmt, & Vorderer, 2007). Players take on the role of Ko and become active participants in a complex system that encourages the use mathematics to solve authentic problems (Gee, 2007). Use of narrative in games also assists presentation of new content in meaningful contexts instead of as random sets of facts and procedures, thereby increasing the likelihood that knowledge will be learned in meaningful and useful ways (Cognition and Technology Group at Vanderbilt, 1992b; Rieber, 1996). Within games, the narrative or story-line can provide players with information about the boundaries and what is plausible in game play to guide problem solving (Dickey, 2006).

Narratives can support student learning by providing numerous types of scaffolding for learning and problem solving. Use of a familiar story structure helps students better organize information and consequently makes it easier to remember

(Graesser, Hauft-Smith, Cohen, & Pyles, 1980). Additionally, cognitive instructional supports can be built into the narrative to facilitate and guide student learning (Yelland & Masters, 2007). In Ko's Journey, cognitive scaffolding is provided by advice and guidance from her Spirit Grandfather and interactive guidebook as well as through diagrams and illustrations.

Conversely, some theories suggest that narratives are not directly related to instructional objectives and can distract the learner from focusing on key information. The cognitive load theory (Sweller, 1999) and cognitive theory of multimedia learning (Mayer, 2009) propose that learners are only able to process a limited amount of information in working memory at one time. Therefore, if learners are using processing capabilities to understand the story narrative or figure out how to maneuver in the game, capabilities for mental representation and processing of key instructional materials might be diminished (Adams et al., 2012). Narrative can also distract the learners from focusing on key instructional materials by introducing interesting but irrelevant narrative that draws attention away from academic content (Mayer, Griffith, Jurkowitz, & Rothman, 2008). Therefore, Adams et al. (2012) suggested the importance of close integration of the instructional content and story narrative.

Intrinsic integration. Intrinsic motivation, wherein people are motivated to learn in the absence of obvious external rewards or punishments (Deci, 1971), is fundamental to user engagement created by digital games (Garris, Ahlers, & Driskell, 2002). However, educational software has typically used an extrinsic "chocolate-covered broccoli" approach (Bruckman, 1999) by using the gaming portion as enticement to complete the educational content (Habgood & Ainsworth, 2011). In a review of digital

games' motivational aspects, Lepper and Malone (1987; Malone, 1981) concluded content needs to be intrinsically related to the fantasy or storyline of the game to produce the best learning and flow. This concept, originally called intrinsic fantasy, refers to games where the skill being used is closely related to the fantasy (such as playing darts by hitting balloons on a number line); whereas in an extrinsic fantasy, the skill is only weakly related to the content (such as in the game of Hangman) and could therefore be used for different subject matter (Malone, 1981). Habgood and Ainsworth (2011) used the game Trash Zapper from the classic Math Blaster series (Davidson, 1983) to provide a clear distinction between the intrinsic and extrinsic fantasy (also known as endogenous versus exogenous). In the game, players answer simple arithmetic sum problems by shooting moving trash particles with the correct answer on it; however, the educational content could easily be changed to spelling content (Habgood & Ainsworth, 2011). Therefore, this game is considered an extrinsic game. Rieber (1996) added that the benefit of an endogenous fantasy was that if the player was interested in the fantasy, then the player would be interested in the game and more likely to be intrinsically motivated.

Habgood et al. (2005; Habgood & Ainsworth, 2011), however, claimed that the emphasis on the intrinsic nature of fantasy was misplaced. The fantasy context was relatively arbitrary and could be switched for another as long as the underlying rule systems of the game and player interaction were intact. For example, the fantasy of Ko's Journey could be changed from a quest through an ancient wilderness to a space odyssey as long as the basic mechanics of the game were unchanged. Habgood et al. proposed the term intrinsic integration to emphasize the importance of aligning core mechanics, rather than fantasy, with educational content. According to Salen and Zimmerman (2004), the

core mechanic was the "essential nugget of game activity, the mechanism through which players make meaningful choices and arrive at meaningful play experience" (p. 317). Developed by Habgood et al., the definition for intrinsic integration has two components:

> 1. Intrinsically integrated games deliver learning material through the parts of the game that are the most fun to play, riding on the back of the flow experience produced by the game and not interrupting or diminishing its impact.
> 2. Intrinsically integrated games embody the learning material within the structure of the gaming world and the player's interactions with it, providing an external representation of the learning content that is explored through the core mechanics of the gameplay. (p. 494)

Although many researchers continue to use intrinsic vs. extrinsic labels (or endogenous vs. exogenous), many incorporate aspects of game mechanics when discussing the importance of integration of the game and educational content (Gunter, Kenny, & Vick, 2007). As mentioned previously, findings from research on Zombie Division and ASTRA EAGLE offered evidence supporting the value of intrinsic integration in educational games (Habgood & Ainsworth, 2011; Ke, 2008a).

In conclusion, Ko's Journey's use of narrative, intrinsic integration, and, most importantly Common Core State Standards, made it distinct from current mathematics games described in the literature. Use of narrative and intrinsic integration in digital games could potentially enhance learning by sustaining motivation and provide needed instructional support. Additionally, the literature review described many advantages of the use of digital games in the classroom including instructional supports, active learning, adaptivity, motivation, and meaningful learning context that might be particularly relevant for lower students with low socioeconomic status who are lacking in knowledge or experience. Games might be an avenue to increase interest and preparation of females and minority students for STEM careers. However, the few quality empirical studies of

digital games, the contradictory results of games in mathematics education, and methodological flaws in empirical studies indicated a clear need for further rigorous empirical investigation of digital games. Without clear information about the content of the game, correspondence with current educational standards, and trustworthy information on empirical effectiveness, teachers do not have sufficient information or evidence to incorporate games in the classroom. Given the unique properties of Ko's Journey including the use of narrative, intrinsic integration, and most importantly Common Core State Standards, there was a need to investigate the efficacy of the game to improve mathematical achievement.

This study addressed this need and empirically evaluated the effect of a digital mathematics game, Ko's Journey, on seventh grade students' mathematical achievement using secondary data analysis. The effects of Ko's Journey were measured by a researcher-constructed test of the Common Core Mathematics Standards (NGA, 2010). In addition, the Rasch (1960) measurement theory was used to determine the differential impact of Ko's Journey on the assessment items, such as what items were learned, and evaluated the need for further refinement of the current assessment in order to inform future research.

CHAPTER III

METHODOLOGY

The purpose of this study was to evaluate the effects of Ko's Journey on students' achievement of the conceptual and procedural knowledge specified by the Common Core State Standards (CCSS; NGA, 2010) for seventh grade mathematics using a pretest-posttest control group design (Gall et al., 2003). This evaluation study was the final phase of a three-phase project. Phase I focused on the development of an instrument to measure student achievement aligned with the CCSS for seventh grade mathematics due to the fact that a commercially developed instrument was not available at the time data were collected. Phase II focused on the development of a measurement model for the assessment. The goal of Phase II was to establish a unidimensional, true equal interval scale of measurement for use in the evaluation. Both the classical test theory and Rasch (1960) model theory were used to evaluate the instrument. Phases I and II are briefly summarized to provide a foundation for the current methodology utilized in this paper. The methodology is organized by research question.

Phase I: Instrument Development and Validation

Q1 To what extent are the test items of the researcher-constructed test used to measure mathematics achievement aligned with the Common Core State Standards for seventh grade mathematics?

Item Development

Items were developed using the Common Core State Standards for seventh grade mathematics (NGA, 2010) and the seventh grade Prentice Hall (2010) *Mathematics Course 2: All-In-One Student Workbook--Version A.* The author worked closely with a former middle school mathematics teacher with 29 years of classroom experience to develop items corresponding with the critical content of the CCSS. The CCSS for seventh grade mathematics include the following high-level domains: ratios and proportional relationships, the number system and operations, expressions and equations, geometry, and statistics and probability. A total pool of 43 items was constructed using a table of specifications from the revised Bloom's Taxonomy (Krathwohl, 2002).

Unfortunately, the length of the assessment detracted from teacher participation in the project. Thus, the Imagine Education project directors selected 20 items from the 43-item pool to be used as the final assessment. Selection was based on items that were most closely aligned with the content of Ko's Journey without omitting any area of the standards. The final 20-item assessment (two multiple-choice items and 18 constructed-response items) included six items on ratios and proportional relationships, three items on the number system, three items on expressions and equations, five items on geometry, and three items on statistics and probability. The instrument is located in Appendix A and the table of specification linking items to CCSS standards is located in Appendix B.

Teacher Validation

Participants. Five veteran middle school math teachers were recruited to review the full 43-item pool to evaluate content validity of the items with the CCSS. The five teachers reviewing the exam had an average of 16 years teaching experience (SD = 11)

and four teachers had advanced degrees. Teachers agreeing to participate received an honorarium at the conclusion of reviewing the assessment. Teachers reviewed the full 43-item pool for possible future revisions of the assessment.

Procedure. Teachers rated each item on how well it corresponded with the targeted standard using the following format (Lawshe, 1975): "Is the skill (or knowledge) measured by this item: (a) Essential; (b) Useful but not essential; or (c) Not necessary to the mastery of the standard.?"

Data analysis and results. Lawshe (1975) argued that items where more than half of the subject matter experts agree that an item is essential to the measurement of a standard have viable content validity. An interrater reliability analysis using the Fleiss Kappa statistic (Fleiss, 1971; King, 2004) was performed to determine consistency of agreement among the five teachers on the 20-item test. As shown in Table 2, the proportion agreement was 56%; however, kappa was not above zero when adjusted for chance agreement. Using Lawshe's content validity ratio (CVR), the agreement (.44) was below the corrected critical value of (.736) for $\alpha = .05$ (Wilson, Pan, & Schumsky, 2012). As shown in Table 3, agreement for 16 items was greater than 50%; five of those items had 100% agreement. When teacher responses were aggregated across the first two categories (i.e., Essential and Useful but not essential), the CVR for 11 items was 1.0 and greater than .5 for eight items with the average CVR =. 80. Thus, agreement among teachers was above chance under the more lenient criteria.

Table 2

Phase I: Fleiss Kappa Agreement for Teachers

Basic Information

Number of Items	Number of Categories	Number of Raters	Proportion of Rater Agreement
20	3	5	0.56

Empirical Confidence Limits: Overall Fleiss Kappa

Kappa	Standard Error	z	*p*	Lower 95% CI	Upper 95% CI
-0.00182	0.05463	-0.03334	0.5133	-0.1089	0.10525

Note: CI = confidence interval

Table 3

Phase I: Content Validity Ratio and Percentage Agreement by Item

Item Number	Item Code	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Content Validity Ratio (CVR)	Lenient CVR	% Agree
1	Q1_RP	1	3	1	3	2	-0.2	0.2	40%
2	Q2_RP	3	2	3	3	3	0.6	1	80%
3	Q3_RP	3	3	1	3	3	0.6	0.6	80%
4	Q4_RP	1	3	3	2	3	0.2	0.6	60%
5	Q5_RP	2	3	2	3	1	-0.2	0.6	40%
6	Q6_RP	3	2	3	2	3	0.2	1	60%
7	Q7_NS	3	2	3	3	3	0.6	1	80%
8	Q8_NS	3	3	3	3	3	1	1	100%
9	Q9_NS	3	2	3	3	3	0.6	1	80%
10	Q10_NS	3	3	3	3	3	1	1	100%
11	Q11_EE	1	2	2	3	3	-0.2	0.6	40%
12	Q12_EE	2	3	3	1	3	0.2	0.6	60%
13	Q13_G	3	3	3	3	3	1	1	100%
14	Q14_G	3	3	3	3	3	1	1	100%
15	Q15_G	2	2	3	2	3	-0.2	1	40%
16	Q16_G	3	3	3	3	3	1	1	100%
17	Q17_G	1	2	3	3	3	0.2	0.6	60%
18	Q18_SP	2	3	3	3	3	0.6	1	80%
19	Q19_SP	1	3	2	3	3	0.2	0.6	60%
20	Q20_SP	1	3	3	3	3	0.6	0.6	80%
Average							0.44	0.8	72%

Note. Lenient CVR combines ratings of 3 (Essential) and 2 (Useful, but not essential).

Summary of phase I. The first phase of research developed a mathematics instrument aligned with the CCSS for seventh grade mathematics. Twenty questions (two multiple-choice and 18 constructed-response) were selected to evaluate Ko's Journey. Ratings from five middle school mathematics teachers were used to assess the alignment of the test to the CCSS. Using Lawshe's (1975) CVR and condensing the top categories, the assessment showed an acceptable level of content validity for use in this

evaluation research. However, the degree of agreement was not sufficient for use in high stakes testing.

Phase II: Development of a Measurement Model

Q2 To what extent does the item level data of the mathematics assessment conform to the requirements of the Rasch (1960) model to produce a unidimensional, equal-interval scale of measurement?

Participants

Twenty-four teachers from nine schools with large concentrations of low-income minority students were recruited by Imagine Education to use Ko's Journey as a supplement to their normal curriculum. Program directors used a method of non-probability, purposive sampling (Gall et al., 2003; Patton, 2002) and schools were selected based on meeting the criterion of having a high percentage of students receiving free or reduced-school lunch (FRSL)--a proxy measure for the concentration of low-income students within a school (National Center for Educational Statistics [NCES], 2012). High-poverty schools are defined by NCES (2012) as public schools where 76% or more of the students are eligible for FRSL. The schools selected had an average of 85% (*SD* = 14.8) of students receiving FRSL. A total of 1,148 students were included in the study. Although demographic information for individual students was unavailable, school demographic information and seventh grade achievement data were retrieved for seven of the schools (see Table 4). Information was not available for an online virtual academy and a new school. Gender information was obtained in the final sample in Phase III discussed below.

Table 4

Phase II: School Demographics

	State	Lunch %	Lunch % State	Math %	Math % State	Ethnicity	
School 1:	AZ	63%	51%	46%	61%	64% 29%	Hispanic White
School 2:	CA	92%	52%	28%	50%	98% <1%	Hispanic Black
School 3:	NY	74%	44%	68%	65%	53% 25%	White Hispanic
School 4:	NM	99%	62%	43%	38%	78% 16%	American Indian White
School 5:	NM	95%	65%	30%	38%	95% 5%	Hispanic White
School 6:	SC	72%	52%	70%	73%	93% 3%	Black White
School 7:	NM	99%	62%	38%	38%	100%	American Indian
	Mean	84.86%	55.43%	46.14%	51.86%		
	SD	14.80	7.66	16.90	14.62		

Note. Lunch % = Students eligible for free or reduced-price lunch program
Lunch % State = State average for students eligible for free or reduced-price lunch program
Math % = Percentage of students at or above proficient on state standardized mathematics
Math % State = State average of students at or above proficient on state standardized mathematics
Adapted from Great Schools (2013).

Procedures

All participants received and gave informed consent and parental consent forms (gathered by Imagine Education, 2013) and completed the 20-item mathematics pretest developed in the previous phase (see Appendix A) and via web browser at the beginning of their seventh grade year. Students took the 20-item assessment on computers in their

regular classroom or on a home computer in the case of virtual education. Order of the test items was randomized for each student and the assessment was not timed. The accuracy of the 18 open-ended items was reviewed by two judges and any disagreements were resolved. Permission to use the data was granted by the University of Northern Colorado Institutional Review and Imagine Education program directors (see Appendix C).

Results

Classical test theory (CTT).

Mean scores. The 1,148 students had a mean score of 5.82 (*SD* = 3.87) on the 20-item test with a minimum score of 0 and maximum score of 20. Two scores of 0 and one perfect score of 20 were eliminated from the data set to mirror settings used in Rasch (1960) modeling. The final data set of 1,145 students had a mean score of 5.82 (SD = 3.85) and a minimum score of 1 and maximum score of 19.

Reliability. The 20-item pretest had a high reliability--Cronbach's $\alpha = .808$, a value suitable for tests of academic ability (Kline, 1999).

Item difficulty and item discrimination. Item difficulty and item discrimination estimates for the entire pretest are presented in Table 5. The items had a mean difficulty of .29 (29% of students answered correctly); question 17 was the most difficult question with a pass rate of .02 and question 2 was the easiest item with a pass rate of .56. Item discrimination calculated by point-biserial correlations of item and total scores had a mean value of .46 (*SD* = .10). Positive correlations indicated that the item was good at discriminating between high and low ability test takers. Question 17 (.15) and question 3 (.34) had the lowest point-biserial correlations of the test items.

Table 5

Phase II: Item Difficulty and Discrimination

Half-Test	Item Number	Item Difficulty	Item Discrimination*
B	1	0.16	0.54
B	2	0.43	0.50
A	3	0.56	0.34
B	4	0.48	0.49
A	5	0.23	0.39
A	6	0.11	0.53
A	7	0.55	0.47
B	8	0.20	0.54
B	9	0.56	0.49
A	10	0.14	0.55
B	11	0.34	0.49
A	12	0.14	0.55
A	13	0.51	0.56
B	14	0.24	0.37
B	15	0.05	0.35
B	16	0.27	0.48
A	17	0.02	0.15
B	18	0.08	0.44
A	19	0.48	0.46
A	20	0.27	0.57
	Mean	0.29	0.46
	SD	0.184	0.101

**Note*: Point-biserial with scores of 0 and 20 eliminated

Exploratory factor analysis. Because the assessment instrument consisted of items from different domains of mathematical knowledge, it is possible the instrument was comprised of two or more dimensions. In CTT, this question is typically investigated via exploratory factor analysis. A principal component analysis using polychoric correlations was conducted using the Factor (Lorenzo-Seva & Ferrando, 2006) computer program on the 20 items using both orthogonal rotation (varimax) and oblique rotation (promax; see Tables 6 and 7). Use of polychoric correlations is advised when the univariate distributions of ordinal items are asymmetric or with excess kurtosis

(Muthén & Kaplan, 1992). The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = .86 (Hutcheson & Sofroniou, 1999). Bartlett's test of sphericity, $x^2(190) = 10598.1,\ p < .001$, indicated that correlations between items were sufficiently large for principal component analysis. An initial analysis was run to obtain eigenvalues for each component in the data. Four components had eigenvalues over Kaiser's criterion of 1, but the first factor had an eigenvalue of 7.8 and accounted for 38.9% of the variance. The scree plot and parallel analysis (Horn, 1965; Timmerman & Lorenzo-Seva, 2011), a technique using randomly generated eigenvalues, both suggested the presence of only one dimension. Additional dimensions were not interpretable or of sufficient magnitude. Tables 6 and 7 show the factor loadings after rotation.

Rasch measurement model. Rasch measurement theory (RMT) is a family of unidimensional item response models (Bond & Fox, 2007; Rasch, 1960) based on principles of fundamental and conjoint measurement that is used to assess the quality of instruments and to construct true interval scale measures from raw scores (Luce & Tukey, 1964; Sick, 2008). Bond and Fox (2007) stated,

> Basic Rasch assumptions are that (a) each person is characterized by an ability and (b) each item by a difficulty which (c) can be expressed by numbers along one line. Finally, (d) from the difference between the numbers (and nothing else) the probability of observing any particular scored response can be computed. (p. 27)

Table 6

Phase II: Exploratory Factor Analysis with Four Factors: Varimax

	Varimax Rotated Loadings							
Test items	1	2	3	4	Factor	Difficulty	Domain	Item stems
1	0.49	0.05	0.34	0.42	1.00	0.12	RP	If a person walks 2/5 mile in each 20 minutes, what is the rate per hour?
2	0.44	0.20	0.03	0.39	1.00	0.43	RP	Three seventh grade classes were asked if they wanted pizza or hamburgers for their special Friday lunch.
4	0.49	0.12	0.18	0.20	1.00	0.48	RP	According to your smoothie recipe, you need 2 cups of ice and 3 cups fruit for 5 smoothie servings.
7	0.46	0.04	0.06	0.33	1.00	0.56	NS	Find the value of n. Write your answer as a decimal.
9	0.54	0.01	0.31	0.10	1.00	0.56	NS	Recorded temperatures at the South Pole Station in Antarctica have ranged from a high of -24°F to a low of -115°F in the month of June.
13	0.58	-0.15	0.31	0.28	1.00	0.51	G	Keisha drew a scale drawing of the local swimming pool. In real life, the pool is 164 feet long. It is 4 inches on the drawing.
14	0.29	0.19	0.27	0.12	1.00	0.24	G	A pizza has a diameter of 12 inches. Find the circumference of the pizza.
16	0.47	0.02	0.24	0.21	1.00	0.27	G	You want to paint the outside of this cube. What is the surface area?
19	0.63	0.12	0.00	0.13	1.00	0.48	SP	Jose's parents kept records of the number of text messages he sent per day of the week. What is the median number of text messages for the week?
20	0.56	0.21	0.27	0.26	1.00	0.31	SP	You spin a spinner numbered 1 through six. Each number is equally likely. Find the probability of landing on an even number
17	0.15	0.98	0.18	0.12	2.00	0.02	G	An aquarium is built in the shape of a triangular prism. What is the volume of the aquarium?
5	0.04	0.16	0.73	0.17	3.00	0.24	RP	To make curtains in your home, you purchase 7 ½ yards of fabric at $13 per yard. If there is a 7% sales tax, what is the total cost of the fabric?
6	0.42	0.03	0.67	0.29	3.00	0.11	RP	The waiter arrives with the bill after dinner. The total is $25.75. You tip at the rate of 18%. What is the total bill?
10	0.44	-0.01	0.56	0.34	3.00	0.14	NS	Three gallons of paint cost $33.12. How much would a pint cost?
18	0.22	0.13	0.55	0.48	3.00	0.09	SP	The Mars Company website states that each bag of original milk chocolate M&M's contains 1.69 ounces and has an average of 55 M&Ms.
3	0.09	0.14	0.13	0.40	4.00	0.56	RP	$10 = 5 items. If you purchase 5 items for $10, how would you calculate unit price?
8	0.30	-0.14	0.24	0.74	4.00	0.20	NS	Lily has 2 ½ cups of Neapolitan ice cream. A serving size is 1/3 of a cup. How many friends can she serve full servings?
11	0.30	0.05	0.21	0.48	4.00	0.34	EE	Solve for 11 - 3n = 5
12	0.40	0.21	0.41	0.48	4.00	0.14	EE	Oliver runs 8 miles per hour in cross country. He has already run 4 miles. His goal is to run a total of 64 miles per week.
15	0.29	0.13	0.31	0.46	4.00	0.05	G	In the diagram below <1 = 4x and <2 = 2x+10. Solve for x to find <3

Table 7

Phase II: Exploratory Factor Analysis: Promax

	Varimax Rotated Loadings							
Test items	1	2	3	4	Factor	Difficulty	Domain	Item stems
1	0.41	-0.03	0.16	0.29	1.00	0.12	RP	If a person walks 2/5 mile in each 20 minutes, what is the rate per hour?
2	0.38	0.16	-0.19	0.31	1.00	0.43	RP	Three seventh grade classes were asked if they wanted pizza or hamburgers for their special Friday lunch.
4	0.51	0.07	0.02	0.03	1.00	0.48	RP	According to your smoothie recipe, you need 2 cups of ice and 3 cups fruit for 5 smoothie servings.
7	0.45	-0.01	-0.13	0.25	1.00	0.56	NS	Find the value of n. Write your answer as a decimal.
9	0.62	-0.06	0.20	-0.13	1.00	0.56	NS	Recorded temperatures at the South Pole Station in Antarctica have ranged from a high of -24°F to a low of -115°F in the month of June.
13	0.61	-0.24	0.16	0.11	1.00	0.51	G	Keisha drew a scale drawing of the local swimming pool. In real life, the pool is 164 feet long. It is 4 inches on the drawing.
14	0.27	0.16	0.20	-0.04	1.00	0.24	G	A pizza has a diameter of 12 inches. Find the circumference of the pizza.
16	0.48	-0.04	0.11	0.05	1.00	0.27	G	You want to paint the outside of this cube. What is the surface area?
19	0.74	0.08	0.19	0.08	1.00	0.48	SP	Jose's parents kept records of the number of text messages he sent per day of the week. What is the median number of text messages for the week?
20	0.56	0.16	0.09	0.04	1.00	0.31	SP	You spin a spinner numbered 1 through six. Each number is equally likely. Find the probability of landing on an even number
17	0.00	1.03	0.05	-0.12	2.00	0.02	G	An aquarium is built in the shape of a triangular prism. What is the volume of the aquarium?
5	0.12	0.11	0.77	0.01	3.00	0.24	RP	To make curtains in your home, you purchase 7 ½ yards of fabric at $13 per yard. If there is a 7% sales tax, what is the total cost of the fabric?
6	0.34	-0.06	0.60	0.07	3.00	0.11	RP	The waiter arrives with the bill after dinner. The total is $25.75. You tip at the rate of 18%. What is the total bill?
10	0.37	-0.10	0.46	0.17	3.00	0.14	NS	Three gallons of paint cost $33.12. How much would a pint cost?
18	0.01	0.06	0.45	0.40	3.00	0.09	SP	The Mars Company website states that each bag of original milk chocolate M&M's contains 1.69 ounces and has an average of 55 M&Ms.
3	-0.09	0.12	0.02	0.44	4.00	0.56	RP	$10 = 5 items. If you purchase 5 items for $10, how would you calculate unit price?
8	0.07	-0.21	0.02	0.84	4.00	0.20	NS	Lily has 2 ½ cups of Neapolitan ice cream. A serving size is 1/3 of a cup. How many friends can she serve full servings?
11	0.15	0.00	0.05	0.47	4.00	0.34	EE	Solve for 11 - 3n = 5
12	0.25	0.15	0.23	0.36	4.00	0.14	EE	Oliver runs 8 miles per hour in cross country. He has already run 4 miles. His goal is to run a total of 64 miles per week.
15	0.13	0.08	0.16	0.41	4.00	0.05	G	In the diagram below <1 = 4x and <2 = 2x+10. Solve for x to find <3

These basic Rasch assumptions translate into the assumptions of unidimensionality, local independence, no error due to guessing, and equal discrimination. The assumption of unidimensionality requires that the items function in unison and all non-random variance in the data can be accounted for by person ability and item difficulty. Local independence asserts the probability of an individual responding correctly to a particular item is not dependent on previous responses or the responses given by other individuals to the same item. The Rasch model requires that the raw scores can be explained through only person ability and item difficulty and guessing is counted toward misfit. The Rasch model supposes equal slope or discrimination of the items so that persons and items can be ordered in terms of ability and difficulty. Although these criteria are often referred to assumptions of the Rasch model, they are seen more as requirements of fundamental measurement to Rasch measurement theorists (Sick, 2010). The extent to which Rasch model assumptions were violated can be tested through fit statistics.

In the dichotomous one-parameter model developed by Rash (1960), person ability (B_n) and item difficulty (D_i) are first created by calculating percentage correct for each person or item and then converting the raw score percentages into odds of success (Bond & Fox, 2007). The natural log of these odds becomes the person ability and item difficulty estimates. For example, a raw score of 40% correct (p = .40) would be divided by the proportion incorrect (1-p = .60) to obtain the ratio 40/60. The natural log of these odds (-0.4) becomes the person ability (B_n) estimate. The logarithmic transformation of the odds of success transforms the scale from a simple ordinal scale to a more useful interval scale and avoids compression at the ends of the raw scale scores (Bond & Fox,

2007). The person ability (B_n) and item difficulty (D_i) estimates are located on a common scale of log odd ratios or logits. The average logit is arbitrarily set at 0; higher positive logits indicate higher probabilities of success and lower logits indicate lower probabilities of success.

Rasch model fit estimates. To use the Rasch model to estimate mathematics achievement on an equal interval scale, it was necessary to first determine if the fit of the data to the model was acceptable. The dichotomous model constructs true interval measures based on a probabilistic relation between only the item difficulty and person ability. If other factors systematically affect the response probability, the requirements for the model are not met and it would not be advisable to use the Rasch model.

In Rasch measurement, residual based fit statistics provide information for the fit of the model as well as individuals and items. The fit statistics are calculated by comparing each pair of observed and model-expected responses, squaring the differences, summing over all pairs, averaging to create a mean square (MNSQ) chi-square statistic variate with an expected value of 1 for data that fit the model (Wright & Panchapakesean, 1969). The cube root transformation (Wilson & Hilferty, 1931) of the MNSQ produces a *t*-statistic that approximates a normal (z) distribution.

Wright and Master (1982) proposed two different fit statistics for persons/items: weighted and unweighted. The weighted MNSQ (Infit) weighs the square residual by the variance of the item while the unweighted MNSQ (Outfit) gives the residual the same weight (i.e., 1). When item difficulty is similar to a person's ability, the variance is greater than when item difficulty and person ability are distant (too easy/difficult). Therefore, the Infit statistic gives more weight to persons whose ability is closer to the

item value and less to persons for whom the item was off-target (too easy or hard). The Outfit statistic is not weighted and therefore is influenced more significantly by outlying scores and extreme responses (Bond & Fox, 2007).

The MNSQ residual summary statistics have an expectation of one and *t*-statistics (ZSTD) have values centered at zero. Mean square values greater than one indicate underfit (the data are less predictable than the model predicts) and values less than one indicate overfit (the data are more predictable than the model predicts; Wright, Linacre, Gustafson, & Martin-Lof, 1994). If there are only a few misfitting items or persons, they can be removed from the analysis; however, numerous misfitting items indicate violations of the requirements for fundamental measurement. High MNSQ values or underfit might indicate unpredicted responses due to poor item construction (due to ambiguous wording or concepts, etc.) or indicate that the item is different from the other items and might be measuring another construct (Linacre, 2012). High person MNSQ values might indicate the person filled in responses randomly, had untypical gaps in their knowledge or responded differently to the item than other students (Linacre, 2012).

Item and person fit corrections. Parameter estimates and standard errors are considered unbiased when the program control parameter STBIAS = Y (Wang & Chen, 2005). Therefore, item difficulty estimates, standard errors, and MNSQ Infit and Outfit estimates were unbiased: however, standard deviations (*SD*) of the MNSQ estimates were not (Wang & Chen, 2005). The *SD*s were inversely related to sample size, the Infit *SD*s were substantially smaller than the Outfit SD, and MNSQ estimates were not symmetrical around the expected value of 1. Therefore, the recommendation (Wright et al., 1994) to use symmetrical values for identifying item misfit was not appropriate.

Thus, critical ranges for the Infit and Outfit MNSQ were adjusted by the sample size. Using the Wang and Chen (2005) table for a 20-item test and a sample size of 1,000, the effective range for Infit MNSQ was 1.1 - .91, whereas the effective range for Outfit MNSQ was 1.31 - .83.

Additionally, Wang and Chen (2005) reported the standardized *t*-statistic (ZSTD) did not conform to the assumption of a standard normal distribution. The mean value of the ZSTD was slightly less than the expected value of 0 and the *SD* was slightly less than the expected value of 1 under all test length-sample size combinations. Infit *SD*s were smaller than Outfit *SD*s; however, both *SD*s approached unity for items of moderate difficulty (i.e., 0 logits). Thus, using the criteria of ± 2.0 *SD*s (.05 level) to screen misfitting items would only be appropriate for items in the middle of the difficulty distribution. Wang and Chen proposed correction factors for both Infit and Outfit ZSTDs that yielded unbiased critical values (CV) for screening items:

$$Infit\ CV\ =\ ZSTD\ \times\ (1 - |\bar{b} - d| / 4)$$

$$Outfit\ CV\ =\ ZSTD\ \times\ (1 - |\bar{b} - d| / 8)$$

General test characteristics. As shown in Table 8, overall summary statistics are presented for persons and items. Person reliability index, an estimate of the stability of person ability scores if the sample was given another set of items measuring the same construct, was .76 (Bond & Fox, 2007). The person reliability index was analogous to Cronbach's α and was between 0 and 1. Person separation (1.69), an estimate of the spread of people on the measured variable in standard error units, indicated a small range of person ability measures but was sufficient to separate the sample into low and high achievers (Linacre, 2012). Item reliability, the estimate of the stability of item difficulty

if the same items were given to a sample of comparable ability, was equal to 1. However, perfect reliability was unlikely and likely inflated due to large sample size and overfitting items. The item separation (14.21), an estimate of the spread of items on the measured variable in standard error units, indicated a wide spread of items along the continuum and ability to divide items in multiple levels if desired (Linacre, 2012). The Wright Map and the Pathway Map, visual descriptions of the data, can be found in Figures 1 and 2.

Summary of Phase II. As expected, the pretest which measures seventh grade proficiency on the CCSS mathematics standards before they have been taught the content was difficult. Reliability from CTT analyses (Cronbach's $\alpha = .808$) suggest that the assessment has adequate precision to be used to evaluate academic achievement (Kline, 1999). The 20 item pretest conforms to the requirements of the Rasch model and the majority of the items and persons fit the model well. The fit statistics do not indicate that any of the requirements of fundamental measurement have been violated; therefore, use of Rasch measurement to construct a true interval scale is appropriate. The multiple choice format questions (Q3, Q5) are potentially misfitting items, as well as, Q17 which appears to misfit due to confusing labeling of dimensions. Therefore, preliminary analyses of the pretest test using CTT and Rasch modeling suggest the test can be used for the current evaluation purposes. CTT and Rasch model analyses conducted in this phase will be repeated in the next phase with the pretest and posttest data to ensure that the requirements for constructing a true interval scale hold.

Table 8

Phase II: Summary Statistics for Persons and Items

Summary of 1,145 Measured Persons

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	5.8	20.0	-1.45	.65	1.00	.1	1.00	.2
S.D.	3.8	.0	1.37	.15	.22	.8	.97	.8
MAX.	19.0	20.0	3.92	1.12	1.84	2.9	9.90	4.2
MIN.	1.0	20.0	-3.65	.53	.45	-2.7	.11	-1.7

Real RMSE	.70	True SD	1.18	Separation	1.69	Person Reliability	.74
Model RMSE	.67	True SD	1.19	Separation	1.77	Person Reliability	.76

S.E. of Person Mean =.04
Person Raw Score-to-Measure Correlation = .99
Cronbach's Alpha (KR-20) Person Raw Score "Test" Reliability=.80

Summary of 20 Measured Items

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	333.3	1145.0	.00	.09	.99	-.1	1.02	.0
S.D.	204.0	.0	1.38	.04	.10	2.5	.27	2.5
MAX.	646.0	1145.0	3.35	.22	1.21	7.0	1.59	5.8
MIN.	22.0	1145.0	-1.76	.06	.87	-4.7	.64	-4

Real RMSE	0.10	True SD	1.38	Separation	14.21	Item Reliability	1
Model RMSE	0.09	True SD	1.38	Separation	14.77	Item Reliability	1

S.E. of Item Mean =.32

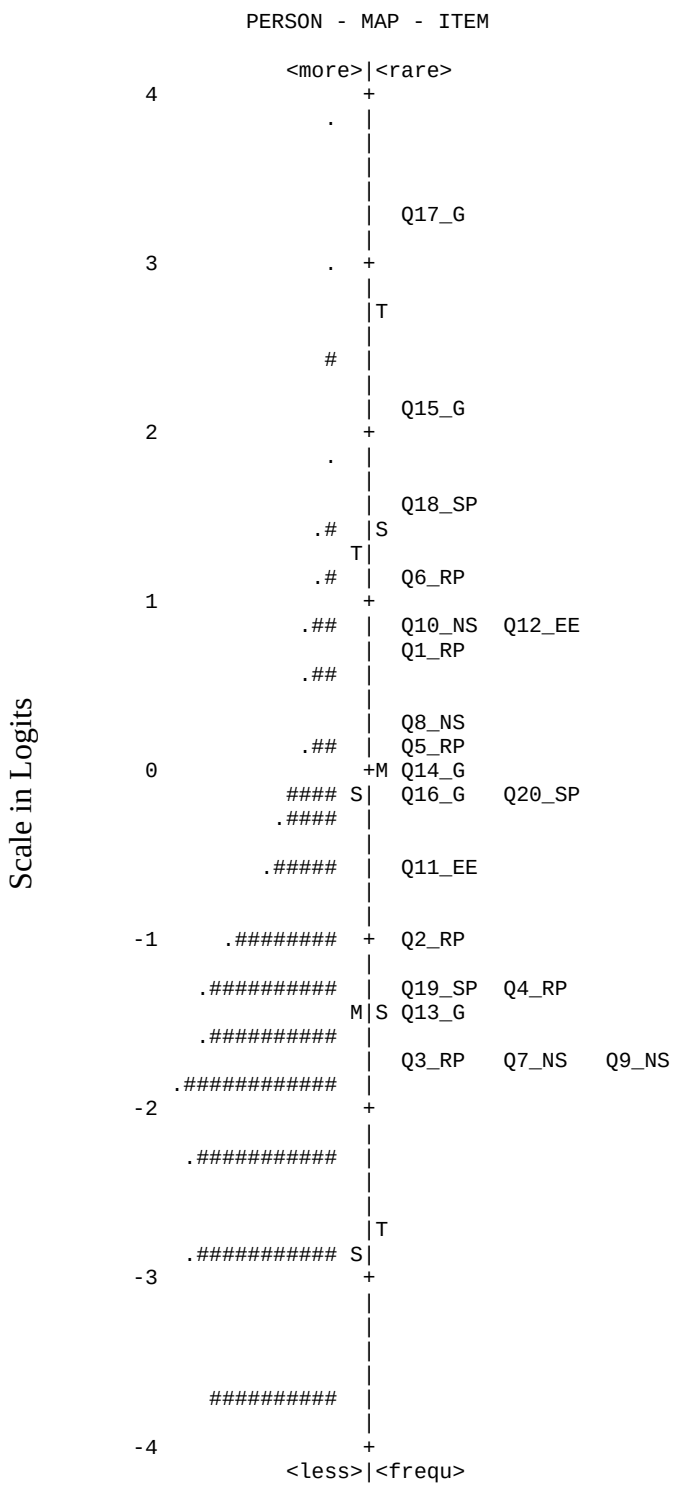

Figure 1. Phase II: Wright map.

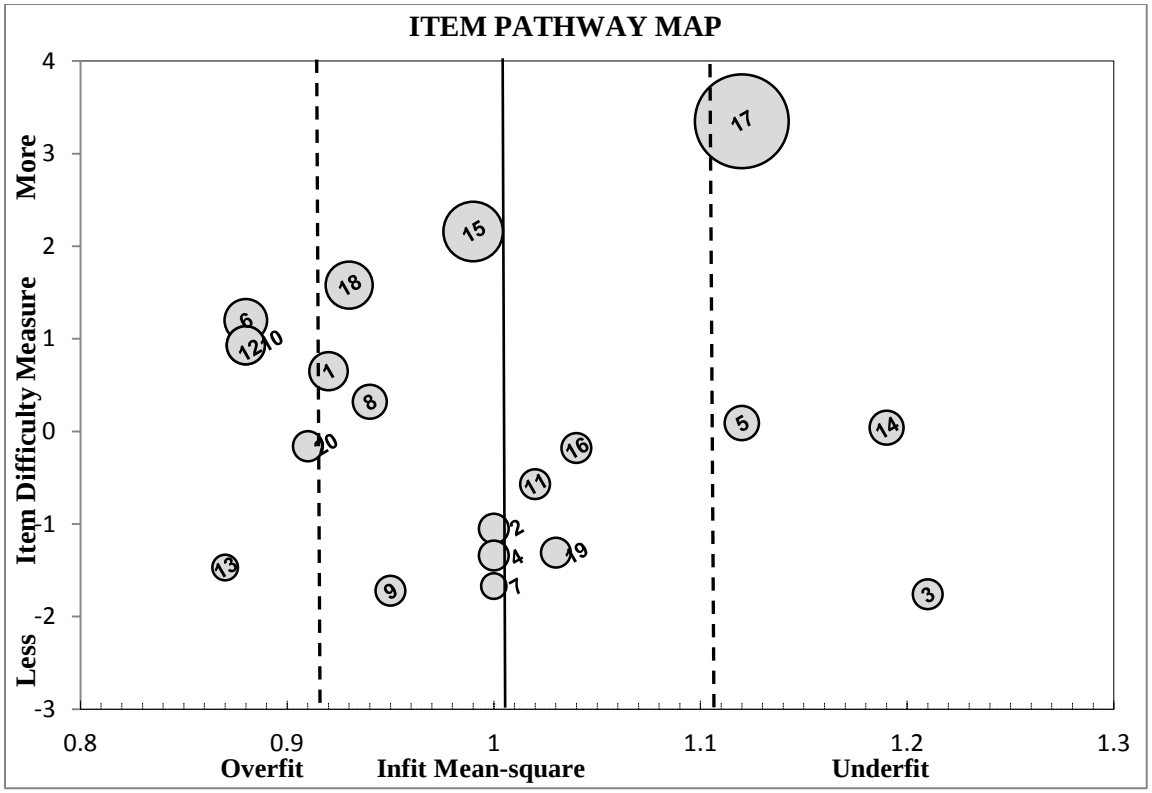

Figure 2. Phase II: Pathway map (bubble plot) of 20 items.

Phase III: Effect of Ko's Journey on Mathematics Achievement

Q3 What is the effectiveness of Ko's Journey on students' mathematics achievement as measured by the researcher-constructed assessment of the seventh grade Common Core Mathematics Standards relative to students who do not play Ko's Journey?

Participants

Five of the original 24 participating teachers from three schools administered both the pretest and posttest to students in their classrooms. Mathematics teachers were instructed to assign their classes using a random lottery to receive the Ko's Journey digital game or serve as a wait-listed control group and continue using the typical mathematics curriculum. A total of 196 seventh grade students (89 females and 107

males) were assigned to experimental classrooms and 175 students (91 females and 84 males) were assigned to control classrooms for a total of 371 students. Although more detailed demographic information for individual students was unavailable, school demographic information and seventh grade achievement data were retrieved for the three schools from Great Schools.org (see Table 9). The two schools located in New Mexico had a large average proportion of students qualifying for FRSL (99%), an indication of a high level of poverty, with a majority (89%) being American Indian students. The school in South Carolina had 72% of students qualifying for FRSL and the majority of students were Black (93%). The two schools in New Mexico had low averages of students scoring at proficient or advanced levels on the state mathematics assessment (45% and 37%, respectively). The school in South Carolina had 62% of students scoring at proficient or advanced levels on the state mathematics assessment; however, this was below the state average of 72%. Due to the differences in state assessments and corresponding proficiency levels, proficiency on mathematics should not be directly compared for schools in different states.

Table 9

Phase III: School Demographics for Three Schools Completing Pretest and Posttest

School	State	Lunch %	Lunch % State	Math %	Math % State	Ethnicity	
A	NM	99%	62%	45%	42%	78% 16%	American Indian White
B	SC	72%	52%	62%	72%	93% 3%	Black White
C	NM	99%	62%	37%	42%	100%	American Indian

Note. Lunch % = Students eligible for free or reduced-price lunch program
Lunch % State = State average for students eligible for free or reduced-price lunch program
Math % = Percentage of students at or above proficient on state standardized mathematics assessments
Math % State = State average of students at or above proficient on state standardized mathematics assessments
Adapted from Great Schools .org (2013)

Procedures

After completion of the web-based pretest assessment (described in the first phase and second phase of research), students in the experimental group were given access to the Ko's Journey digital game as a supplement to their normal mathematics curriculum. The control group continued with the normal seventh grade curriculum. Students in the Ko's Journey experimental group played the digital game on classroom computers via the web. The eight lessons in Ko's Journey were designed to be completed in approximately ten 50-minute sessions; however, teachers were given flexibility in implementation of Ko's Journey and accompanying supplemental material due to time and school constraints. Additionally, some classrooms experienced technical difficulties that interfered with the smooth implementation of the digital game including slow download of the game itself, slow internet connections, and compatibility of the digital game with

web browsers. Additionally, technical problems with the application also prevented tracking the amount of time students spent playing Ko's Journey.

Data Analysis

Rasch measurement. Data analysis started by employing Rasch measurement theory (Rasch, 1960) using Winsteps (Linacre & Wright, 2004) to construct a true equal interval scale of measurement for items and the subset of persons that completed the pretest and posttest. Similar to procedures used in Phase II, classical test theory and Rasch modeling were used to assess the model. The basic Rasch assumptions of unidimensionality, local independence, no error due to guessing, and equal discrimination were assessed through analyzing fit statistics, a Rasch principal component analysis of residuals. Additional assumptions were examined using residual based fit statistics, MNSQ and ZSTD Infit and Outfit statistics, described in the Rasch component of Phase II. Using the fit statistics corrections proposed by Wang and Chen (2005) for a 20-item test and a sample size of 400, the effective range for Infit MNSQ was 1.19-0.82; whereas the effective range for Outfit MNSQ was 1.63-0.69.

Hierarchical Linear Modeling

Given the hierarchical structure of the data, students nested in classrooms, and the varying implementation of Ko's Journey, use of hierarchical linear modeling (HLM) was appropriate and necessary for answering this research question. A short rationale of the use of hierarchical modeling, assumptions, and sample size are given to support the use of HLM for this analysis. Next a section on model building details the a priori model proposed to be analyzed in HLM7 (computer software; Bryk, Raudenbush, & Cogdon, 1996).

Rationale for use of hierarchical linear modeling. Hierarchical levels of grouped data are a common occurring phenomenon in social, developmental, and educational research (Raudenbush & Bryk, 2002). In education, data are often organized at student, classroom, school, school district, and state levels. Similarly, in repeated measures research, data collected over time are nested within each study participant. A hierarchical dataset can be structured in many forms and all that is required is that some level-1 units of some type (e.g., students or measurements) be nested inside level-2 units (e.g., schools, classrooms, or students). Although two-level structure is common, multilevel models are not restricted to only two levels but must have at least two levels (Roberts, 2004).

Data that have hierarchical structures in which units of analyses (e.g., students) are nested in higher units (e.g., classrooms or teachers) are problematic because an important assumption of many statistical analyses is independence of observations among level-1 units (Raudenbush & Bryk, 2002). The magnitude of the dependence among individuals within a level-2 unit is measured via an intraclass correlation. The hierarchical linear model (HLM) is a complex form of ordinary least squares regression that is used to analyze variance in outcome variables when the predictor variables are at varying hierarchical levels (Bickel, 2007). The HLM technique is also referred to as a multilevel linear model, mixed-effects model, random-effects model, random coefficient regression model, or covariance components model in different domains of research (Raudenbush & Bryk, 2002). Advances in algorithms used to estimate covariance components of unbalanced data (Dempster, Laird, & Rubin, 1977; Dempster, Rubin, & Tsutakawa, 1981) allowed for widespread application of HLM for hierarchal data

analysis. Further advancement of computer programs such as HLM (Bryk et al., 1996), MlwiN (Rasbash, Charlton, Browne, Healy, & Cameron, 2005) and PROC MIXED, a routine of the SAS statistical package (Singer, 1998) have also increased the use of HLM procedures.

Prior to HLM, hierarchical data were analyzed using two fixed parameter, simple linear regression procedures--disaggregation and aggregation--that were insufficient because of their failure to deal with shared variance. One such approach, disaggregation, essentially ignored the hierarchical structure and disaggregated all the higher level variables (teacher, class, school characteristics) to the individual level. However, students in the same class shared values of the class variable and this violated assumptions of independence of observations, an important premise of traditional linear model analysis (Raudenbush & Bryk, 2002). In addition, ignoring the structure of the data resulted in underestimated standard errors (no between-unit variation) and thus increased type I errors. Conversely, aggregation dealt with hierarchical data by ignoring the lower level individual differences and level-1 individual variables were aggregated to higher levels (i.e., classes). The problem with this method was that a large percentage of the total variation (possibly 80% or 90%) attributable to within-group differences was thrown away before the start of the analysis (Raudenbush & Bryk, 2002). Therefore, aggregating and disaggregating were both unsatisfactory methods and demonstrated the need for HLM that simultaneously investigated relations within and between hierarchical levels of group data.

In addition to the ability to assess cross-level relationships and accurately parse the effects of between- and within-group variance, it was also the best method for nested

data because it required fewer assumptions to be met than other statistical methods (Raudenbush & Bryk, 2002). Woltman, Feldstain, MacKay and Rocchi (2012) noted that HLM could "accommodate non-independence of observations, a lack of sphericity, missing data, small and/or discrepant group sample sizes and heterogeneity of variance across repeated measures" (p. 56).

Sample size in hierarchical linear modeling. One disadvantage of HLM is the need for a substantial sample size. Maximum-likelihood estimation methods commonly used in multilevel are asymptotic and therefore have assumptions of a large sample size (Maas & Hox, 2005). Although it is generally recognized that group-level sample size is generally more important than total sample size, some simulations suggested that large individual sample sizes partially compensated for a small number of groups (Maas & Hox, 2005). A general rule of 30/30 (30 groups/30 observations per group; Kreft, 1996; Kreft & de Leeuw, 1998) is a commonly cited estimate; however, guidelines vary widely depending on the complexity of the model and aims of the study (Raudenbush & Bryk, 2002). Snijders and Bosker (1996) suggested that multilevel modeling becomes attractive when the number of second-level groups is larger than 10. Even in the case of a small group sample, a simulation by Maas and Hox (2005) provided evidence that parameter estimates, variance components, and standard error of the coefficients were estimated accurately and only second-level variances were biased (i.e., underestimated). A simulation study by Shih (2008) confirmed the results of Maaz and Hox and additionally suggested that higher ICC values could compensate for smaller level-1 or level-2 values.

Hierarchical linear model development. Hierarchical modeling often follows a "model building" approach in which the final or an *a priori* model is compared to a baseline and more basic models (Roberts, 2004, p. 31). In this particular study, HLM7 (Bryk et al., 1996) with maximum likelihood estimation was used to build and test the proposed model. A null or baseline model was used to compare future models and predictors were added individually to see the unique contribution to the total model (Roberts, 2004). Chi-square tests versus degrees of freedom tests were then used to determine if the models differed significantly and helped to select the most parsimonious model (Raudenbush & Bryk, 2002). Additionally, questions of statistical significance were addressed by examining Akaike information criterion (AIC) and/or the Bayesian information criterion (BIC), statistics that calculate goodness of fit of a model based on previous model estimates and number of parameters estimated (Roberts, 2004).

The most basic hierarchical linear model is the null or baseline model, which is equivalent to a one-way ANOVA with random effects. It is also called the unconditional model and has no level-1 or level-2 predictor variables (Raudenbush & Bryk, 2002). The null model is often used as a preliminary step in hierarchical data analysis to calculate point estimates and confidence intervals for the grand mean. It provides information about the outcome at the within-group level (i.e., individuals) and between-groups (i.e., classes). In the null model, $\beta_{1\mathrm{J}}$ is set to zero for all j or level-2 units. The null model is given by

Level-1: $Y_{ij} = \beta_{0j} + r_{ij}$

Level-2: $\beta_{0j} = \gamma_{00} + u_{0j}$

(Model 1): Combined: $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$

where,

Y_{ij}= Posttest score for student *i* in classroom *j*;

β_{0j}= school mean posttest measure for the *j*th classroom;

r_{ij}= random error associated with variability within classrooms;

γ_{00}= grand mean of posttest measure across classrooms;

u_{0j}= random error associated with variance between classrooms.

With the following assumption that r_{ij} and u_{0j} are independent and normally distributed,

$r_{ij} \sim$ iid $\mathrm{N}(0, \sigma^2)$;

$u_{0j} \sim \mathrm{N}(0, \tau_{00})$.

The intraclass correlation coefficient (ICC) is an important parameter calculated from the results of the null model that measures the proportion of variance in the outcome that is due to level-2 groups (i.e., classes; Raudenbush & Bryk, 2002). Using notation consistent with Raudenbush and Bryk (2002), the unconditional intraclass correlation coefficient is given by the following formula:

$$\rho = \tau_{00}/(\tau_{00} + \sigma^2)$$

Where,

τ_{00}=between-group variability (Var γ_{00}),

σ^2= within-groups variability (Var r_{ij}).

The null model was used to formally test whether the estimated value of τ_{00} was significantly greater than zero, formally stated as (H_0: $\tau_{00} = 0$). If significant variation existed among schools in their mathematics achievement, further predictor variables could be considered.

The next model investigated the impact of two level-1 (person model) predictors--gender and students' pretest score (centered on classroom mean)--on the mean posttest achievement scores using a random coefficients model. In a random coefficient model, each classroom was allowed to have its own regression equation where gender and pretest were used as explanatory variables in the level-1 (person model) for each classroom. The separate regression equations for each classroom provided information about the variation of classroom means, magnitude and variation of gender differences across classrooms, and magnitude and variation of the relationship (i.e., slope) between pretest and posttest scores. Raudenbush and Bryk (2002) suggested that group-mean centering is more effective than grand-mean centering in diminishing correlations among random components and minimizing bias in estimating variances of random components. Thus, the level-1 student is specified as

$$Y_{ij} = \beta_{0j} + \beta_{1j} Gender_{ij} + \beta_{2j}\left(X_{ij} - \bar{X}_{.j}\right) + r_{ij}$$

Where,

$Gender_{ij}$ is a design variable coded 0 for male students and 1 for female students;

$\left(X_{ij} - \bar{X}_{.j}\right)$ is a deviation of student-level pretest scores from their classroom average;

β_{0j} is the mean posttest achievement of classroom *j* for males;

β_{1j} is the mean difference between males and females in classroom *j*;

β_{2j} is the slope of the pretest and posttest scores of classroom j;

r_{ij} is the residual level-1 error after controlling for students' gender and pretest score.

The level-2 (classroom) model describes the level-1 parameters as varying across classrooms as a function of the grand mean and random error:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

where,

γ_{00} = average of the classroom means on the posttest for male students;

γ_{10} = effect of being female on the posttest;

γ_{20} = average pretest-posttest slope across classrooms;

u_{0j} = unique effect of school *j* on the mean posttest score;

u_{1j} = unique effect of school *j* on the female-male achievement differences;

u_{2j} = unique effect of school *j* on the pretest-posttest relationship.

Using substitution, the combined model 2 became

$$Y_{ij} = \gamma_{00} + \gamma_{10}(Gender_{ij}) + \gamma_{20}\left(X_{ij} - \bar{X}_{.j}\right) + u_{0j} + u_{1j}\left(Gender_{ij}\right) + u_{2j}\left(X_{ij} - \bar{X}_{.j}\right) + r_{ij}$$

Only level-1 predictors found to be statistically significant were included in subsequent models. Additionally, level-2 classroom models included the treatment effect (TRT_j), a design variable indicating if the classroom was assigned to the control group ($T_j = 0$) or the Ko's Journey experimental group ($T_j = 1$), and deviation of the classroom average pretest scores from their grand mean $\left(\bar{X}_{.j} - \bar{X}_{..}\right)$ to control for selection bias. This model was referred to as an intercepts-and-slopes-as-outcomes model, in which level-1 slopes and intercepts were modeled by the level-2 grouping variable (classroom

mean pretest score) and the treatment variable. Only those terms in the random coefficients model found to be statistically significant were used to construct the model to evaluate the effects of Ko's Journey on the outcome measure.

Assumptions of hierarchical linear modeling. To ensure the validity of inferences based on the results of hierarchical linear models, the following assumptions were carefully tested. The HLM model assumed the following (Raudenbush & Bryk, 2002):

1. Conditional on the student variables, the within school errors (r_{ij}) are normally distributed and independent with a mean of 0 in each school and equal variance across schools (i.e. $r_{ij} \sim$ iid $\mathrm{N}(0, \sigma^2)$).
2. Any student level predictors of math achievement that are excluded from the model and thereby relegated to the error term (r_{ij}) are independent of the level-1 variables that are included in the model (covariance equal 0).
3. The residual school effects, u_{0j} and u_{1j} are assumed bivariate normal with variances τ_{00} and τ_{11}, respectively, and covariance τ_{01}.
4. The effects of whatever school predictors are excluded from the model for the intercept are independent of other school level variables.
5. The error at the student level, r_{ij}, is independent of the residual school effects, u_{0j} and u_{1j}.
6. Any student level predictors that are excluded from the level-1 model and as a result relegated to the error term, r_{ij}, are independent of the school level predictors in the model (covariance equal 0). In addition, any school level predictors that are excluded from the model and as a result relegated to the level-2 random effects, u_{qj}, are uncorrelated with the level-1 predictors (covariance equal 0). (p. 255)

Raudenbush and Bryk (2002) explained that assumptions 2, 4, and 6 focused on the relationships of the variables in the structural portion of the model (level-1 and level-2 predictor variables) and factors relegated to the error terms, r_{ij} and u_{qj}. They pertained to the adequacy of model specification; misspecification of the model could bias estimating level-1 and level-2 fixed effects. Assumptions 1, 3, and 5 were related to the random part of the model, r_{ij} and u_{qj}. Their tenability affected the consistency of the

estimates of standard errors of level-2 fixed effects, accuracy of level-1 random effects, the variances for level-1 and level-2, and the accuracy of hypothesis tests and confidence intervals.

As recommended by Raudenbush and Bryk (2002), data analysis began with examination of the univariate frequency distribution of each variable to provide a check of the quality of the data and identify outliers. Next, plots of the bivariate relationships were used to identify possible nonlinear relationships. Assumptions of homogeneity of the level-1 variance were tested using the chi-square test statistic provided in HLM7 (Bryk et al., 1996). A visual analysis of the residuals using scatter plots, histograms and normal Q-Q plots at each level was used to assess normality and homoscedasticity. Normality of the fixed effects were indirectly measured by examining level-2 residuals using a Q-Q plot of the Mahalanobis distance (Raudenbush & Bryk, 2002). Selection of the most parsimonious model was guided by chi-square tests versus degrees of freedom tests and Akaike information criterion and/or the Bayesian information criterion tests described previously.

Q4 Do the items of the assessment function differently for students using Ko's Journey as a supplement to normal instruction than students who do not play the game?

Rasch measurement theory (Rasch, 1960) using Winsteps (Linacre & Wright, 2004) was used to explore if the intervention of Ko's Journey changed the way the assessment functioned from pretest to posttest. When using Rasch measurement theory to study change over time, Wright (2003) proposed two methods of structuring the data: stacking and racking. Stacking the data showed how the students had changed as a result of the intervention and racking the data showed how the items had changed.

Cunningham and Bradley (2010) and Herrmann-Abell, Flanagan, and Roseman (2012) used these techniques to evaluate a science teacher training program and a science curriculum unit for eighth grade students.

Stacking the data was used to study the effect of the intervention on students' understanding of the targeted mathematics on the constructed assessment. Stacking was done by preparing a data file that contained two rows of data per student (see Figure 3). One row contained the student's responses during the pretest and the second row contained posttest responses. In essence, this put the pretest and posttest on the same ruler so changes in achievement could be measured. This analysis resulted in two ability measures per student: pretest and posttest ability; the difference between these measures represented the change in understanding or ability as a result of the intervention. If the Ko's Journey intervention was effective in increasing mathematics achievement on the posttest, students' ability measures would be expected to increase from pretest to posttest. The stacking technique was used in the analysis of research question 3.

Racking the data was used to see the differential effects of the intervention on the items' difficulty level. The racked data contained one row per student and two columns per item with one column containing the students pretest responses and one with the posttest responses (see Figure 3). In this type of analysis, it was assumed that items became less difficult from pretest to posttest as a result of the intervention and that the students remained unchanged. Racking resulted in two difficulty measures per item for the experimental and control groups--one from the pretest and one from the posttest. Differential changes in item difficulty between the experimental and control groups from pretest to posttest represented the degree the intervention successfully targeted those

items. If the intervention was successful in improving mathematics achievement targeted on the items, it was expected that the items' difficulty would decrease from pretest to posttest.

Using a racked data structure in Winsteps (Linacre & Wright, 2004), visual item maps and item difficulty were examined to determine the differential effects of the intervention on the items. Standard errors of the item difficulty estimates were used to detect differences greater than chance. Typically, differences of .5 of a logit are considered noteworthy (Linacre, 2012).

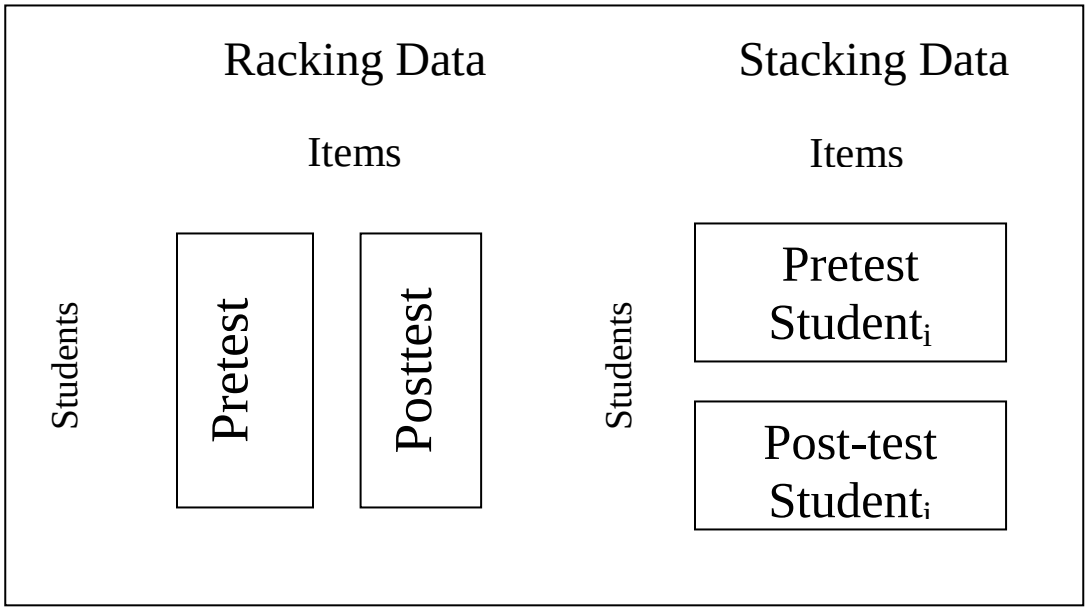

Figure 3. Illustration of stacking and racking data for Rasch modeling.

CHAPTER IV

ANALYSIS

Introduction

This chapter presents the results of Rasch (1960) and hierarchical linear model (HLM) analyses to answer the research questions guiding this phase of the evaluation research:

Q2 To what extent does the item level data of the mathematics assessment conform to the requirements of the Rasch (1960) model to produce a unidimensional equal-interval scale of measurement?

Q3 What is the effectiveness of Ko's Journey on students' mathematics achievement as measured by the researcher-constructed assessment of the seventh grade Common Core Mathematics Standards relative to students who do not play Ko's Journey?

Q4 Do the items of the assessment function differently for students using Ko's Journey as a supplement to normal instruction than students who do not play the game?

To answer these research questions, secondary data were used from a pretest-posttest control group design study (Gall et al., 2003) described in detail in Chapter III. Five teachers from three schools administered the pretest and posttest to students in their classrooms. Mathematics teachers were instructed to assign their classes using a random lottery to play the Ko's Journey digital game or continue to use the typical mathematics curriculum. A total of 371 seventh grade students participated in the study with 196 (89 females and 107 males) assigned to experimental classrooms and 175 (91 females and 84 males) assigned to control classrooms. As shown in Table 9, the participating schools

had high levels of poverty as indicated by the percentage of students qualifying for free or reduced-school lunch (FRSL) and American Indian and Black students.

After completion of the web-based pretest assessment (see Appendix A), students in the experimental group were given access to the digital game via the web as a supplement to their normal mathematics curriculum. The control group continued with their normal mathematics curriculum. The nine mathematics lessons (see Table 1) in Ko's Journey were designed to be completed in approximately ten 50-minute sessions. Supplementary classroom activities were also given to teachers in an educator's guide (Imagine Education, 2013).

Research Question Analysis

Q2 To what extent does the item level data of the mathematics assessment conform to the requirements of the Rasch (1960) model to produce a unidimensional equal-interval scale of measurement?

Rasch (1960) measurement theory using Winsteps (Linacre & Wright, 2004) was used to construct a true interval scale of measurement for items and persons. The pretest was previously analyzed in Phase II using classical test theory and Rasch modeling with a sample of 1148 students that included the current participant's data. The current analysis examined the properties of the pretest, posttest, and stacked analysis of the 371 students using classical test theory and Rasch modeling to answer Research Questions 2 and 3.

Classical Test Theory

Mean scores. As displayed in Table 10, the control group students had a mean pretest score of 5.82 (*SD* = 3.31) and posttest score of 6.93 (*SD* = 3.63) and the experimental group had a mean pretest score of 5.29 (*SD* = 3.03) and posttest score of

6.51 (*SD* = 4.08). The pretest and posttest both had a minimum score of 1 and the maximum score on the posttest was 18.

Table 10

Descriptive Statistics for Students Who Played Ko's Journey and Students in the Control Classrooms by Teacher

		Control Group		Experimental Group		
		Pre-Test	Post-test		Pre-Test	Post-test
Teacher	*N*	Mean (*SD*)	Mean (*SD*)	*N*	Mean (*SD*)	Mean (*SD*)
A	22	7.18 (3.26)	10.68 (4.28)	23	5.48 (2.66)	9.00 (5.48)
B	26	4.04 (2.03)	4.96 (2.99)	37	4.24 (2.03)	5.03 (2.62)
C	52	5.33 (2.44)	6.81 (3.10)	69	5.86 (3.51)	6.43 (4.23)
D	42	4.12 (1.92)	5.07 (2.19)	26	3.00 (1.83)	4.12 (2.07)
E	33	9.24 (3.87)	8.52 (3.35)	41	6.63 (2.70)	8.10 (3.69)
Total	175	5.82 (3.31)	6.93 (3.63)	196	5.29 (3.03)	6.51 (4.08)

Reliability. The 20-item pretest had a reliability of Crohbach's α=.719; whereas, the posttest had a slightly higher reliability, Crohbach's α=.803. Both values were suitable for tests of academic ability (Kline, 1999).

Item difficulty and item discrimination. The item difficulty and item discrimination estimates for the pretest and posttest are presented in Table 11. The pretest had a mean difficulty of .28 (28% of the students answered correctly); whereas, the mean difficulty decreased slightly to .32 on the posttest. The most difficult items on the pretest were Item Q6 and Item Q17 with a pass rate of .02. Item Q7 was the least

difficult question with a pass rate of .61. On the posttest, Item Q17 was the most difficult item with a pass rate of .01, down from .02 on the pretest. Item Q13 was the least difficult question on the posttest with a pass rate of .66.

Table 11

Item Difficulty and Item Discrimination for Pretest and Posttest Items

	Pretest		Posttest	
Item Number	Item Difficulty	Item Discrimination	Item Difficulty	Item Discrimination
1	.12	0.54	0.16	0.64
2	.41	0.47	0.51	0.49
3	.58	0.32	0.56	0.45
4	.58	0.41	0.54	0.52
5	.17	0.18	0.12	0.23
6	.02	0.21	0.18	0.59
7	.61	0.40	0.61	0.22
8	.20	0.51	0.22	0.62
9	.53	0.41	0.60	0.46
10	.09	0.45	0.18	0.62
11	.33	0.48	0.37	0.46
12	.09	0.46	0.18	0.63
13	.51	0.52	0.66	0.55
14	.27	0.33	0.43	0.48
15	.04	0.31	0.11	0.46
16	.23	0.55	0.33	0.45
17	.02	0.07	0.01	0.09
18	.05	0.34	0.03	0.31
19	.48	0.43	0.63	0.51
20	.20	0.54	0.25	0.30
Mean (*SD*)	0.28 (.21)	0.40 (.13)	0.34 (.22)	0.45 (.15)

Item discrimination was calculated by point-biserial correlations of item and total scores; positive correlations indicated that an item was good at discriminating between high and low ability test takers. The pretest had mean value of .40 (*SD* = .13) and the posttest had a mean value of .45 (*SD* = .15). Item Q7 had the lowest point-biserial correlation of the test items on both the pretest (.07) and the posttest (.09).

Rasch Model Analysis

General test characteristics of the pretest and posttest. As shown in Table 12, summary pretest statistics are presented for persons and items. The person reliability index was .67. The person reliability index was analogous to Cronbach's α and was an estimate of the stability of person ability scores if the sample was given items measuring the same construct (Bond & Fox, 2007). The person separation value of 1.41 indicated a limited range of person ability measures but it was sufficient to separate the sample into low and high achievers (Linacre, 2012). The item separation (8.19) indicated a wide spread of the items and an ability to divide into multiple categories if desired. The pretest had an item reliability of .99; this suggested that item difficulty estimates would be stable if the items were given to a sample of comparable ability (Bond & Fox, 2007).

Table 12

Person and Item Summary Statistics for the Pretest

Summary of 371 Measured Persons on the Pretest

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	5.5	20	-1.63	0.65	1	0	1	0.2
SD	3.2	0	1.2	0.13	0.26	0.9	1.08	0.8
MAX.	17	20	2.57	1.06	1.98	2.8	9.9	3.5
MIN.	1	20	-3.85	0.56	0.43	-2.6	0.3	-1.7

Real RMSE	0.69	True *SD*	0.98	Separation 1.41		Person Reliability	0.67
Model RMSE	0.66	True *SD*	1	Separation 1.50		Person Reliability	0.69

S.E. of Person Mean = .06

Person Raw Score-to-Measure Correlation = .99
Cronbach's Alpha (KR-20) Person Raw Score "Test" Reliability= .72

Summary of 20 Measured Pretest Items

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	102.8	371	0	0.17	0.99	0	1.01	0.1
SD	75.7	0	1.58	0.08	0.11	1.3	0.34	1.9
MAX.	226	371	3.11	0.41	1.25	3.1	1.9	4.3
MIN.	6	371	-2.11	0.11	0.84	-2.1	0.54	-2.4

Real RMSE	0.19	True *SD*	1.57	Separation	8.19	Item Reliability	0.99
Model RMSE	0.19	True *SD*	1.57	Separation	8.39	Item Reliability	0.99

S.E. of Item Mean =.32

As reported in Table 13, the posttest had increased person reliability (.77) and person separation (1.8) compared to the pretest. The posttest item reliability remained stable (.99) and the item separation was 8.45.

Table 13

Person and Item Summary Statistics for the Posttest

Summary of 371 Measured Persons on the Posttest

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	6.7	20	-1.26	.64	1	.0	1.06	.2
SD	3.9	0	1.41	.13	.26	.9	1.23	.8
MAX.	18	20	3.34	1.05	2.17	-3.5	9.90	.78
MIN.	1.0	20	-3.87	.55	.37	-2.5	.11	-1.0

Real RMSE	.68	True *SD*	1.24	Separation	1.82	Person Reliability	.77
Model RMSE	.65	True *SD*	1.25	Separation	1.93	Person Reliability	.79

S.E. of Person Mean = .07

Person Raw Score-to-Measure Correlation = .99
Cronbach's Alpha (KR-20) Person Raw Score "Test" Reliability= .80

Summary of 20 Measured Posttest Items

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	124.4	371	0	.16	.99	-.1	1.16	.3
SD	78.8	0	1.71	.10	.16	1.9	.63	2.2
MAX.	246	371	4.52	.57	1.33	3.7	2.87	4.6
MIN.	3	371	-2.11	.11	.78	-3.3	.51	-2.6

Real RMSE	.20	True *SD*	1.69	Separation	8.45	Item Reliability	.99
Model RMSE	.19	True *SD*	1.70	Separation	8.78	Item Reliability	.99

S.E. of Item Mean = .39

Wright map. The Wright map is a visual depiction of the data in the Rasch (1960) analysis. The Wright Map provides an overall picture of the assessment by placing the difficulty of exam items on the same measurement scale as the ability of the participants. The Wright Map for the pretest is shown in Figure 4 and is organized into two vertical histograms; the left side displays the participants and the right side displays the items. The participants are distributed according to mathematics ability in logits with the most ability at the top and least ability at the bottom. The items on the right side are distributed from the most difficult items at the top to the least difficult items at the

bottom. Each “X” represents four participants and each “.” equals one to three participants. Theoretically, when the candidates and items are opposite each other on the map, the difficulty of the item and the ability of the person are comparable; thus, the candidate has approximately 50% of answering the item correctly (Bond & Fox, 2007).

On the left side, the Wright Map in Figure 4 shows the mean student ability level (*M* = -1.63 logits), standard deviation (*SD* = 1.20), and two standard deviations (T) for measured candidate ability. The map shows that the mean person ability (M) is one standard deviation (S) lower than the mean (M) item difficulty. It is understandable that the participants found the current assessment difficult as the purpose of the current assessment was a pretest of seventh grade proficiency on the CCSS mathematics standards before they had been taught the content.

The Wright Map of the posttest is displayed in Figure 5. Mean student ability increased from the pretest (*M*=-1.26, *SD* = 1.41); however, the assessment remained difficult for the majority of students as illustrated on the Wright Map.

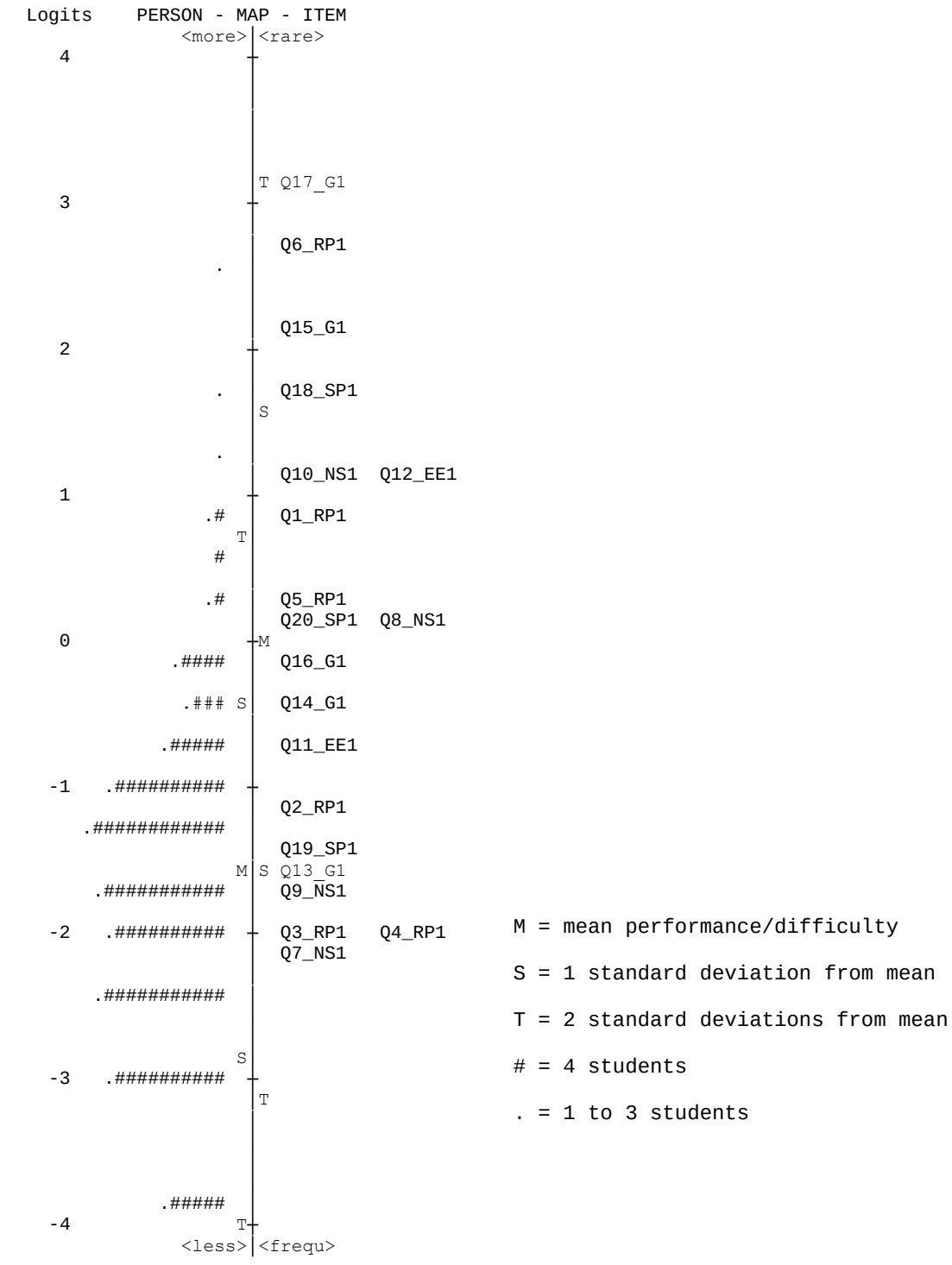

Figure 4. Wright map for pretest.

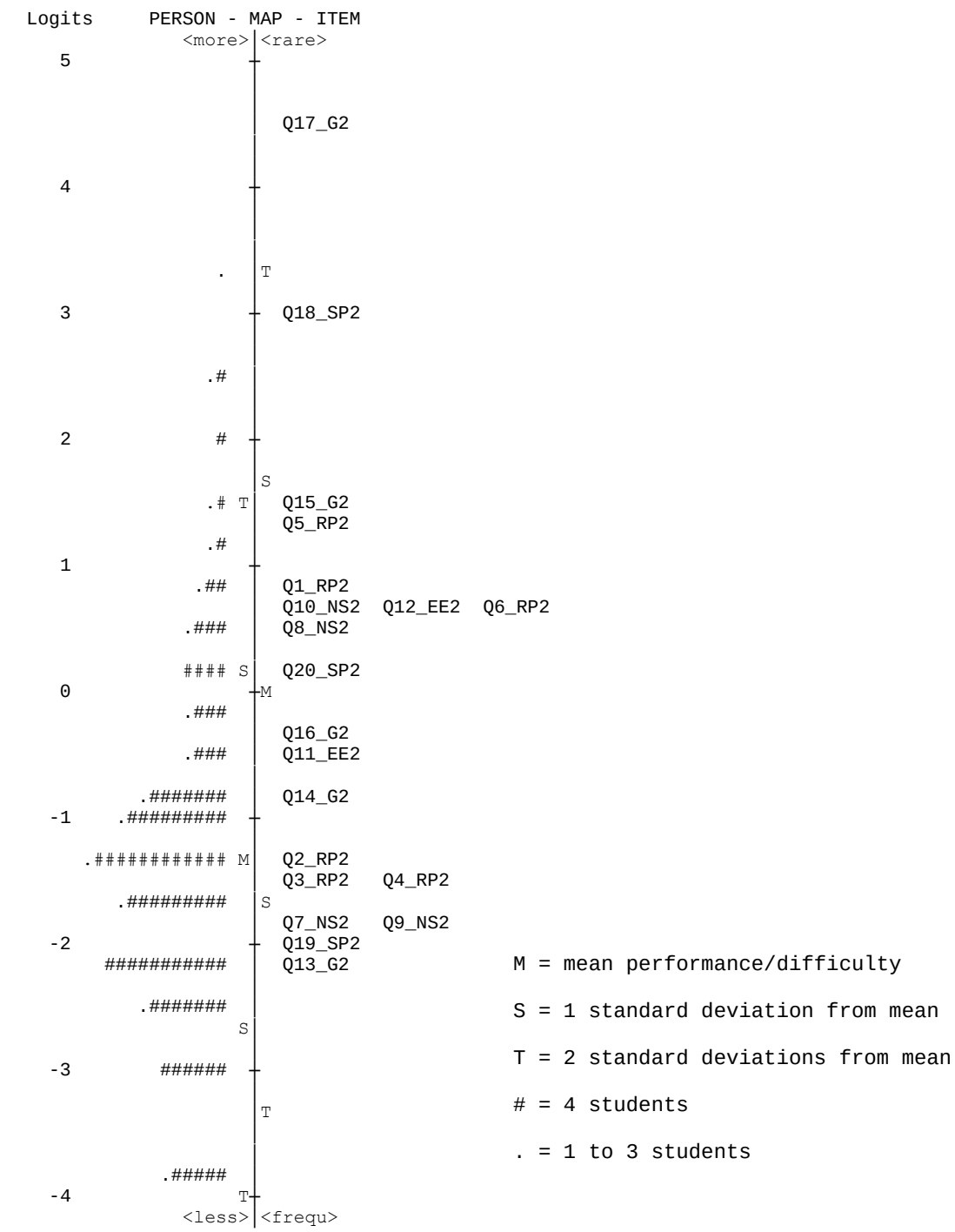

Figure 5. Wright map for posttest.

Rasch stacked analysis. To measure changes in achievement from pretest to posttest, assessments must be put on the same "ruler" (Wright, 2003). This was done by stacking the data so the data file contained two rows per student--the first for the pretest responses and the second for the posttest responses. The pretest and posttest data were given equal weight in the analysis. The person and item summary statistics are shown in Table 14. The stacked analysis had a person reliability estimate of .73 and person separation of 1.64. Item reliability was equal to the pretest and posttest reliability of .99; item separation was 11.89. Similar to the pretest and posttest, person reliability and separation estimates were lower than item estimates due to the limited number of items and limited range of abilities.

Table 14

Person and Item Summary Statistics for the Stacked Analysis

Summary of 742 Measured Persons on Stacked Analysis

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	6.1	20	-1.41	0.64	1	0	1.02	0.2
SD	3.6	0	1.31	0.13	0.25	0.9	1.08	0.8
MAX.	18	20	3.16	1.05	2.14	3.4	9.9	5.1
MIN.	1	20	-3.81	0.55	0.42	-2.6	0.17	-1.6

Real RMSE .68 True *SD* 1.12 Separation 1.64 Person Reliability .73
Model RMSE .65 True *SD* 1.13 Separation 1.74 Person Reliability .75
S.E. of Person Mean = .05

Person Raw Score-to-Measure Correlation = .05
Cronbach's Alpha (KR-20) Person Raw Score "Test" Reliability = .77

Summary of 20 Measured Items Stacked

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	227.2	742	0	0.11	0.99	-0.1	1.06	0.1
SD	152.5	0	1.56	0.06	0.12	2.1	0.46	2.5
MAX.	454	742	3.79	0.33	1.32	3.9	2.35	6.5
MIN.	9	742	-1.94	0.08	0.8	-4.2	0.53	-3.5

Real RMSE .13 True *SD* 1.56 Separation 11.89 Item Reliability 0.99
Model RMSE .13 True *SD* 1.56 Separation 12.29 Item Reliability 0.99
S.E. of Item Mean = 0.36

Item fit statistics. The stacked analysis was used to examine the basic Rasch (1960) assumptions of unidimensionality, local independence, no error due to guessing, equal discrimination through analyzing fit statistics, and a Rasch principal component analysis (PCA) of residuals.

Item fit statistics displayed in Table 15 were examined using Infit and Outfit MNSQ and corrected Infit and Outfit ZSTD values. Using the Wang and Chen (2005) table for a 20-item test and a sample size of 800, the effective range for Infit MNSQ is 1.12 - .86; whereas the effective range for Outfit MNSQ is 1.45 - .79. Two items (Item

Q5 and Item Q17) had Infit and Outfit MNSQ values over the effective ranges (Infit MNSQ 1.45-.86; Outfit MNSQ 1.45 - .79) and Item Q14 had only a high Infit value. These items are highlighted in Table 15. Items over the MNSQ Infit and Outfit values are considered to underfit and indicate noise or unexpected response patterns in the items, which degrade measurement. Item Q5, for example, had a MNSQ Infit value of 1.32. Thus, there was 32% more variation in the response pattern than what was expected in the Rasch model (Bond & Fox, 2007).

Table 15

Item Fit Statistics for the Stacked Model

MODEL			INFIT			OUTFIT		
ITEM	MEASURE	S.E.	MNSQ	ZSTD	Corr. ZSTD	MNSQ	ZSTD	Corr. ZSTD
Q1_RP1	0.87	0.11	0.8	-2.6	-1.6	0.53	-3.5	-2.8
Q2_RP1	-1.19	0.08	0.97	-0.8	-0.5	1.12	1.9	1.5
Q3_RP1	-1.72	0.08	1.12	3.5	2.2	1.21	2.7	2.2
Q4_RP1	-1.7	0.08	1.01	0.3	0.2	0.97	-0.3	-0.2
Q5_RP1	0.81	0.11	1.32	3.9	2.4	2.35	6.5	5.3
Q6_RP1	1.32	0.13	0.89	-1.1	-0.7	0.67	-1.7	-1.4
Q7_NS1	-1.94	0.08	1.1	2.8	1.8	1.36	4.1	3.3
Q8_NS1	0.3	0.1	0.88	-2.1	-1.3	0.71	-2.8	-2.3
Q9_NS1	-1.72	0.08	1.01	0.3	0.2	1.09	1.3	1.1
Q10_NS	0.9	0.11	0.84	-2.1	-1.3	0.65	-2.4	-1.9
Q11_EE	-0.6	0.08	1.02	0.5	0.3	1.05	0.7	0.6
Q12_EE	0.91	0.12	0.81	-2.5	-1.5	0.6	-2.8	-2.3
Q13_G1	-1.8	0.08	0.86	-4.2	-2.6	0.79	-3	-2.4
Q14_G1	-0.63	0.08	1.08	2	1.3	1.1	1.4	1.1
Q15_G1	1.73	0.15	0.92	-0.7	-0.4	0.94	-0.1	-0.1
Q16_G1	-0.22	0.09	1	0	0.0	0.9	-1.1	-0.9
Q17_G1	3.79	0.33	1.14	0.5	0.3	2.21	2.2	1.7
Q18_SP	2.39	0.19	1	0	0.0	0.95	0	0.0
Q19_SP	-1.67	0.08	0.97	-1	-0.6	0.89	-1.6	-1.3
Q20_SP	0.16	0.1	1.1	1.7	1.1	1.16	1.4	1.1

Overfit

Underfit

Item Q5 was one of two multiple choice items on the test; therefore, it was not unexpected that students responded differently to this item compared to the rest of the test. Item Q17 was the most difficult question, requiring student to calculate volume of a

triangular prism. Analyses from CTT (reliability if item deleted, discrimination) and the Rasch model (MNSQ Infit and Outfit) suggested problems with Item Q17. Item Q17 required students to use three out of four labeled values to find the volume of a prism. Confusion over the labeling of the diagram (i.e., what was length, height, and width) might have contributed to performance of this item.

Items with corrected Infit MNSQ values less than .86 (Items Q1, Q10, and Q12) and Outfit MNSQ values less than .79 (Items Q1, Q6, Q10, and Q12) were considered to overfit (Wang & Chen, 2005). Overfit indicated less variation in the response pattern than predicted by the Rasch (1960) model. For example, a perfect Guttman response string (11111100000) was highly unlikely, had much less variation than would be predicted by the probabilistic Rasch model, and consequently the MNSQ would be less than 1. Item Q1, for example, had an Infit MNSQ value of .53. Thus, there was 47% (1 - .53 = .47) less variation in the observed response pattern than was modeled. Although overfit might mislead one to think the measure was better than it was, there were no practical consequences of overfit (Bond & Fox, 2007).

As recommended by Linacre (2012), person measures were recalculated after eliminating the misfitting items. Eliminating Item Q5, Item Q17 did not significantly improve person fit reliability or separation; therefore, it was decided to leave in all 20 questions for further analysis.

Rasch principal component analysis. The requirement for unidimensionality was tested by a Rasch (1960) principal component analysis (PCA) of residuals. Unlike in CTT factor analysis where factor loadings are interpreted as correlations with latent traits, the Rasch PCA of residuals aimed to falsify the hypothesis that the residuals were

random noise by finding components that explained the largest possible variance in the residuals—this was the first contrast in PCA (Linacre, 2012). Both the size of the factor and the response style or item characteristics were used to determine if there was systematic variation in the residuals not explained by the Rasch model or if they were simply random noise (Bond & Fox, 2007). Rasch item difficulties and person abilities explained 38.2 % of the raw variance in the observations. Person abilities accounted for 12% of the raw score variance, while items accounted for 26.2%. The first contrast had an eigenvalue of 1.5 (a strength of less than two items) and accounted for 4.5% of the raw variance not explained by the measures. Eigenvalues are considered to be at noise level when they are two items or less; thus, the hypothesis that residuals were random noise was not falsified (Linacre, 2012).

Summary of research question 2. Research Question 2, which asked if the item level data of the mathematics assessment conformed to the requirements of the Rasch (1960) model to produce a unidimensional equal-interval scale of measurement, was analyzed using CTT techniques and Rasch modeling. The pretest and posttest were stacked to put both of the tests on the same scale and assumptions of Rasch modeling were checked through residual based fit statistics and a Rasch PCA. The 20-item pre-assessment generally conformed to the requirements of the Rasch model and the majority of the items and persons fit the model well. The fit statistics did not indicate that any of the requirements of fundamental measurement (unidimensionality, equal item discrimination, and error due to guessing) had been violated; therefore, use of Rasch measurement to construct a true interval scale was appropriate. On the stacked analysis, items Q5 and Q17 were misfitting items. However, eliminating them from the

assessment did not significantly alter item or person measures. Another check of unidimensionality was confirmed through the use of the Rasch PCA.

Q3 What is the effectiveness of Ko's Journey on students' mathematics achievement as measured by the researcher-constructed assessment of the seventh grade Common Core Mathematics Standards relative to students who do not play Ko's Journey?

As a preliminary estimate of the effectiveness of Ko's Journey on mathematics achievement, parametric gain effect sizes were computed for each of the teachers' classrooms and overall for the experimental and control classrooms using the following formula (Rudner, Glass Gene, Evartt, & Emery, 2002):

$$d = \left[\frac{\bar{X}_{ePost} - \bar{X}_{ePre}}{S_{ePooled}}\right] - \left[\frac{\bar{X}_{cPost} - \bar{X}_{cPre}}{S_{cPooled}}\right],$$

with pooled standard deviation for each group computed by the following formula recommended by Dunlap, Cortina, Vaslow, and Burke (1996):

$$S_{pooled} = \sqrt{\frac{(n_{Post}-1)S^2_{Post} + (n_{Pre}-1)S^2_{Pre}}{(n_{Post}-1)+(n_{Pre}-1)}}.$$

Effect sizes were standardized, scale-free measures of the relative size of the effect of an intervention and helpful for comparing the impact of the effects across studies (Coe, 2002). The effect size d, or the standardized difference of gain scores for the experimental and control classrooms for each teacher, are presented in Table 16. The effect sizes ranged in size from +.652 (indicating a large achievement gain of the experimental class as compared to the control class) to -.381(indicating academic gain of the control class compared to the experimental group), with a small overall mean effect size of 0.02, 95% CI [-0.18, 0.224]. An effect size of .6, as in the case of Teacher A, could be interpreted as 73% of the control class would be behind the average person in the experimental class (Coe, 2002). However, due to the likelihood of strong

dependencies among students within classrooms, effect sizes must be interpreted with caution.

Table 16

Standardized Difference in Gain Scores Between Experimental and Control Conditions and Fidelity of Implementation

Teacher	Parametric Gain Effect Size	95% Confidence Intervals		Std Error of Effect Size	Technical Difficulty	Discontinued use of Ko's	Majority completed Ko's	Primary mode of instruction	Used supplemental activities
		Lower	Upper						
A*	-0.087	-0.671	0.498	0.298	y	y	n	n	n
B*	-0.018	-0.520	0.483	0.256	y	u	y	n	n
C*	-0.381	-0.744	-0.017	0.185	y	n	y	u	n
D*	0.117	-0.372	0.607	0.250	y	n	y	y	n
E*	0.652	0.181	1.122	0.240	y	n	y	y	y
Total	0.020	-0.184	0.224	0.104					

Note. * Effect size corrected for small sample size. Fidelity of implementation data is summarized by yes (y), no (n) or unknown (u).

Fidelity of implementation data collected by Imagine Education and displayed in Table 16 indicated that all classrooms had technical difficulties and Teacher A discontinued use of Ko's Journey. The two classrooms that used Ko's Journey as the primary mode of instruction had the greatest gains.

The differential effect of the treatment on the experimental group can be seen in Figure 6. These results were consistent with an aptitude treatment interaction (ATI) (Cronbach & Webb, 1975; Snow, 1989) where control group students who scored low on the pretest had higher posttest scores than did experimental students also scoring low on the pretest, while experimental students scoring high on the pretest had higher posttest scores than control group students with similarly high pretest scores. This interaction

suggested there might be a threshold of mathematical ability in order to benefit from the Ko's intervention.

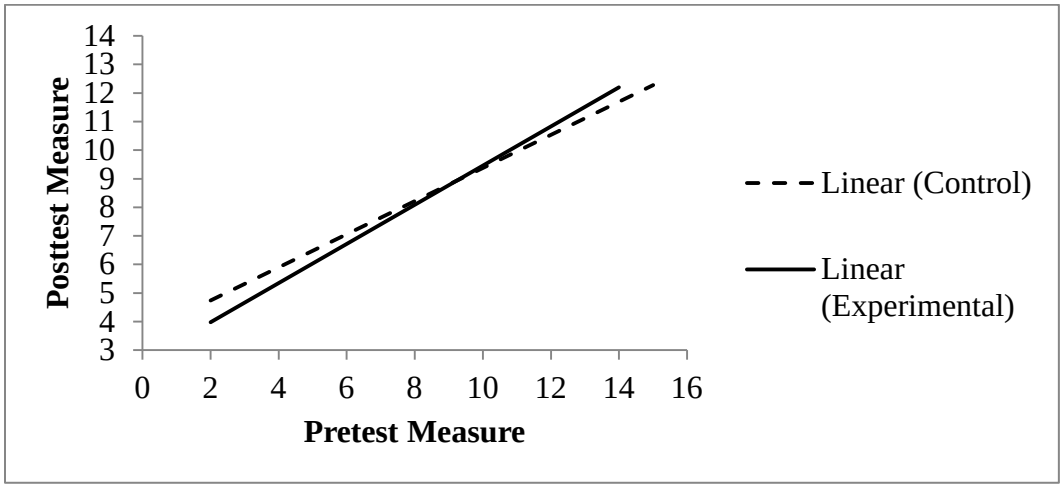

Figure 6. Differential relationship of pretest and posttest for experimental and control group students.

Hierarchical Linear Modeling

Due to the hierarchical structure of the data--students nested in classrooms and the varying implementation of Ko's Journey in classrooms, hierarchical linear modeling was used to decompose variance within and between classrooms in order to get an accurate estimate of the effect of Ko's Journey. Person ability logit scores produced in the Rasch analysis of the pretest and posttest were used and rescaled from 0 to 20 to aid interpretability (Bond & Fox, 2007). Fixed effects estimates and variance-covariance estimates for all models are presented in Table 17.

Table 17

Fixed Effects Estimates (Top) and Variance-Covariance Estimates for Models of the Predictors of Mathematics Achievement

Parameter	Model 1	Model 1.1	Model 2	Model 3	Model 4
	Unconditional	Unconditional	Random Coefficient	Intercepts and Slopes as Outcomes	Intercepts and Slopes as Outcomes
			Fixed effects		
Intercept	7.35 (.43)	7.29 (.44)	7.30 (0.41)	7.63 (0.64)	7.34 (0.39)
Level 1					
Gender			-0.02 (0.24)		
Pretest			0.56 * (0.05)	0.54* (0.04)	.54* (0.05)
Level 2					
TRT				-0.66 (0.90)	-0.12 (0.56)
AvgPre					0.92* (0.24)
			Random parameters		
Intercept, u_0	1.80* (1.34)	1.80* (1.34)	1.54* (1.24)	1.96* (1.97)	0.64* (0.80)
Gender slope, u_1			0.23 (0.48)		
Pretest slope, u_2			0.01 (0.09)		

Note. Standard errors are in parentheses. Model 1.1 has outliers (18) removed.

Model 1: Null model. The null model, equivalent to a one-way ANOVA with random effects, was used as a first step in model building to calculate a point estimate for the grand mean and calculate the intraclass correlation coefficient (ICC) or the amount of dependency among students within classroom (Raudenbush & Bryk, 2002). In the null model, $\beta_{1\mathrm{J}}$ was set to zero for all j or level-2 units. The null model was given by

Level-1: $Y_{ij} = \beta_{0j} + r_{ij}$

Level-2: $\beta_{0j} = \gamma_{00} + u_{0j}$

(Model 1): Combined: $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$

where,

Y_{ij}= Posttest score for student *i* in classroom *j*;

β_{0j}= mean posttest measure for the *j*th classroom;

r_{ij}= random error associated with variability within classrooms;

γ_{00}= grand mean of posttest measure across classrooms;

u_{0j}= random error associated with variance between classrooms.

Due to the small number of classrooms, restricted maximum likelihood (MLR) was used to estimate the model parameters. Although full maximum likelihood (MLF) and MLR would have had similar results with a large number of level-2 units, analyses with small number of groups had biased variance and covariance using MLF (Raudenbush & Bryk, 2002). The grand mean, γ_{00} of the posttest measure was 7.35 with a standard error of .40. Accordingly, we could expect that 95% of the means for similar samples would fall within 6.95 and 7.75. The estimate of the grand mean had high reliability (.906), indicating the sample means were reliable indicators of the true classroom level means. The chi-square test, $x^2(9) = 83.76$, $p <$ .001, indicated the variance in the posttest by the level-2 grouping (class rooms) was statistically significant. Thus, there was variance among the classroom means that could be explained by the treatment variable. The ICC for the null model was .225. This result suggested that approximately 23% of the variance in the posttest was at the classroom level and 77% was at the individual level (Raudenbush & Bryk, 2002). With an ICC of this size, standard errors would be significantly underestimated in traditional analyses that did not consider this dependency and led to a much greater likelihood of Type 1 errors than the a priori established alpha values indicated (Thomas, Heck, & Bauer, 2005).

The test of homogeneity of level-1 variance was statistically significant, $x^2(9) = 31.00$, $p = 0.00$. Therefore, in this model, the assumption of homogeneity of variance was violated. The residuals from this model were examined to search for outliers.

Outliers of two or more standard deviations were eliminated (18 students), resulting in a total of 353 students for subsequent analysis. Model 1.1: Null model was rerun with this modified dataset. Similar to the previous results, the estimate of the grand mean, γ_{00}, of the posttest measure was 7.29 with a standard error of .44 and reliability of .93. The chi-square test, $x^2(9) = 115.48$, $p < .001$, indicated significant variability in classrooms and the ICC for this model was .30. Therefore, over 30% of the variance of the posttest measure was at the classroom level. In contrast to the null model with the full dataset, the revised model had a test of the homogeneity of level-1 variance $x^2(9) = 7.29$, $p > .500$, signifying that this assumption had been met. Because of the significant amount of variation among school means, further predictor variables were considered in a random coefficients model.

Model 2: Random coefficients model. The random coefficients model investigated the impact of the two level-1 (person model) predictors, gender and students' pretest score (centered on classroom mean), on the mean posttest achievement scores. Thus, the level-1 student was specified as

$$Y_{ij} = \beta_{0j} + \beta_{1j} Gender_{ij} + \beta_{2j}\left(X_{ij} - \bar{X}_{.j}\right) + r_{ij}$$

Where,

$Gender_{ij}$ is a design variable coded 0 for male students and 1 for female students;

$\left(X_{ij} - \bar{X}_{.j}\right)$ is a deviation of student-level pretest scores from their classroom average;

β_{0j} is the mean posttest achievement of classroom *j* for males;

β_{1j} is the mean difference between males and females in classroom *j*;

β_{2j} is the slope of the pretest and posttest scores of classroom j;

r_{ij} is the residual level-1 error after controlling for students' gender and pretest score.

The level-2 (classroom) model described the level-1 parameters as varying across classrooms as a function of the grand mean and random error:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

where,

γ_{00} = average of the classroom means on the posttest for male students;

γ_{10} = effect of being female on the posttest;

γ_{20} = average pretest-posttest slope across classrooms;

u_{0j} = unique effect of school *j* on the mean posttest score;

u_{1j} = unique effect of school *j* on the female-male achievement differences;

u_{2j} = unique effect of school *j* on the pretest-posttest relationship.

Using substitution, the combined Model 2 became

$$Y_{ij} = \gamma_{00} + \gamma_{10}(Gender_{ij}) + \gamma_{20}\left(X_{ij} - \bar{X}_{.j}\right) + u_{0j} + u_{1j}\left(Gender_{ij}\right) + u_{2j}\left(X_{ij} - \bar{X}_{.j}\right) + r_{ij}$$

As presented in Table 17, the estimate of the grand mean, γ_{00}, of the posttest measure was 7.25 (.23) and reliability was .876. The regression coefficient relating gender to posttest measure achievement was negative (-0.02) but it was not significant, $t(9) = -.070$, $p = .946$. Therefore, because gender was not a significant predictor of posttest achievement, it was eliminated from subsequent models.

The regression coefficient, 0.56 (0.5), relating the pretest to posttest measure, was significant, $t(9) = 11.04$, $p < .001$. Thus, the pretest explained a sizeable proportion of variance in posttest scores. In terms of variance components, u_{0j}, the unique effect of school j on the posttest measure was the only random effect that was significant in this model. As expected from the results of the fixed effects, the variability in gender differences (u_{1j}) across classrooms was not statistically significant. Furthermore, the magnitude of the relationship between the pretest and posttest scores was consistent among schools. Consequently, u_{2j}, was set to 0 in subsequent models.

Model 3: Intercepts-and-slopes-as outcomes model. The intercepts-and-slopes-as outcome model included level-1 predictors that were found to be statistically significant (student pretest scores centered by class mean) and added the classroom level-2 variable, treatment (TRT_j). At level-1, the model was specified as

$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{ij} - \bar{X}_{.j}\right) + r_{ij}$$

where,

β_{0j} is the intercept or grand mean;

β_{1j} is the slope of the pretest and posttest scores of classroom j;

$\left(X_{ij} - \bar{X}_{.j}\right)$ is a deviation of student-level pretest scores from their classroom average;

r_{ij} is the residual level-1 error after controlling for students' pretest score.

The level-2 (classroom) model was specified as

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\left(TRT_j\right) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

with the combined model,

$$Y_{ij} = \gamma_{00} + \gamma_{01}(TRT_j) + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + u_{0j} + r_{ij}.$$

The treatment effect was entered into the model first as a single level-2 predictor. The treatment did not significantly affect the posttest mean, $t(8) = -.730$, $p = .486$, thus signifying that Ko's Journey did not have an significant effect on achievement.

Model 4: Intercepts-and-slopes-as outcome model controlling for selection bias. Due to the possible bias in the selection of classrooms to receive Ko's Journey (i.e. selection of the lower ability classes instead of random assignment), average pretest score (grand centered) were added to the level-2 model. At level-1, the model was specified as

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij}$$

where,

β_{0j} is the intercept or grand mean;

β_{1j} is the slope of the pretest and posttest scores of classroom j;

$(X_{.j} - \bar{X}_{..})$ or $(AVGPRE_j)$ is a deviation of student-level pretest scores from their classroom average;

r_{ij} is the residual level-1 error after controlling for students' pretest score.

The level-2 (classroom) model was specified as

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(TRT_j) + \gamma_{01}(X_{.j} - \bar{X}_{..}) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

with the combined model,

$$Y_{ij} = \gamma_{00} + \gamma_{01}(TRT_j) + \gamma_{02}(X_{.j} - \bar{X}_{..}) + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + u_{0j} + r_{ij}.$$

As presented in Table 17, the grand mean of the posttest measure for the full model was 7.33 (.39) with a high reliability estimate of .87. The regression coefficients of treatment, -0.12 (.56), $t(8)= -0.206$, $p = .843$, indicated that students who received Ko's Journey had lower posttest scores; however, this effect was not statistically significant. The classroom average pretest was a significant predictor of posttest scores, 0.92 (.24), $p = .006$. Thus, even when differences in average classroom achievement on the pretest were controlled (i.e., held constant), treatment effects were not evident.

Checking assumptions of hierarchical linear modeling. In addition to the assumptions addressed in the models (i.e., homogeneity of variance), the residuals at each level were examined for normality and homoscedasticity using visual inspection of histograms and Q-Q plots of the residuals. Given that most of the assumptions required examining the residual at each level, checking the assumptions was done after fitting the models. Residuals from the intercepts-and-slopes-as-outcome model in the HLM7 (Bryk et al., 1996) program were imported into SPSS Version 17 (SPSS Inc., 2007). Visual inspection of the Q-Q plots (see Figure 7) of the total level-1 residuals, pretest, and posttest measures did not show any major departures from normality; thus, the assumption was considered tenable and the level-1 model was appropriately specified. At level-2, a Q-Q plot of the Mahalanobis distance for classrooms indicated a lack of normality. Lack of normality of level-2 residuals did not bias the estimation of fixed effects in the model; however, it could lead to biases of standard errors at all levels and thus affect the validity of statistical tests and confidence intervals (Raudenbush & Bryk, 2002). The striking heterogeneity of the residuals suggested that the level-2 model was not adequately specified. In other words, an important classroom characteristic, other

than the treatment variable and classroom average pretest achievement, was omitted from the model.

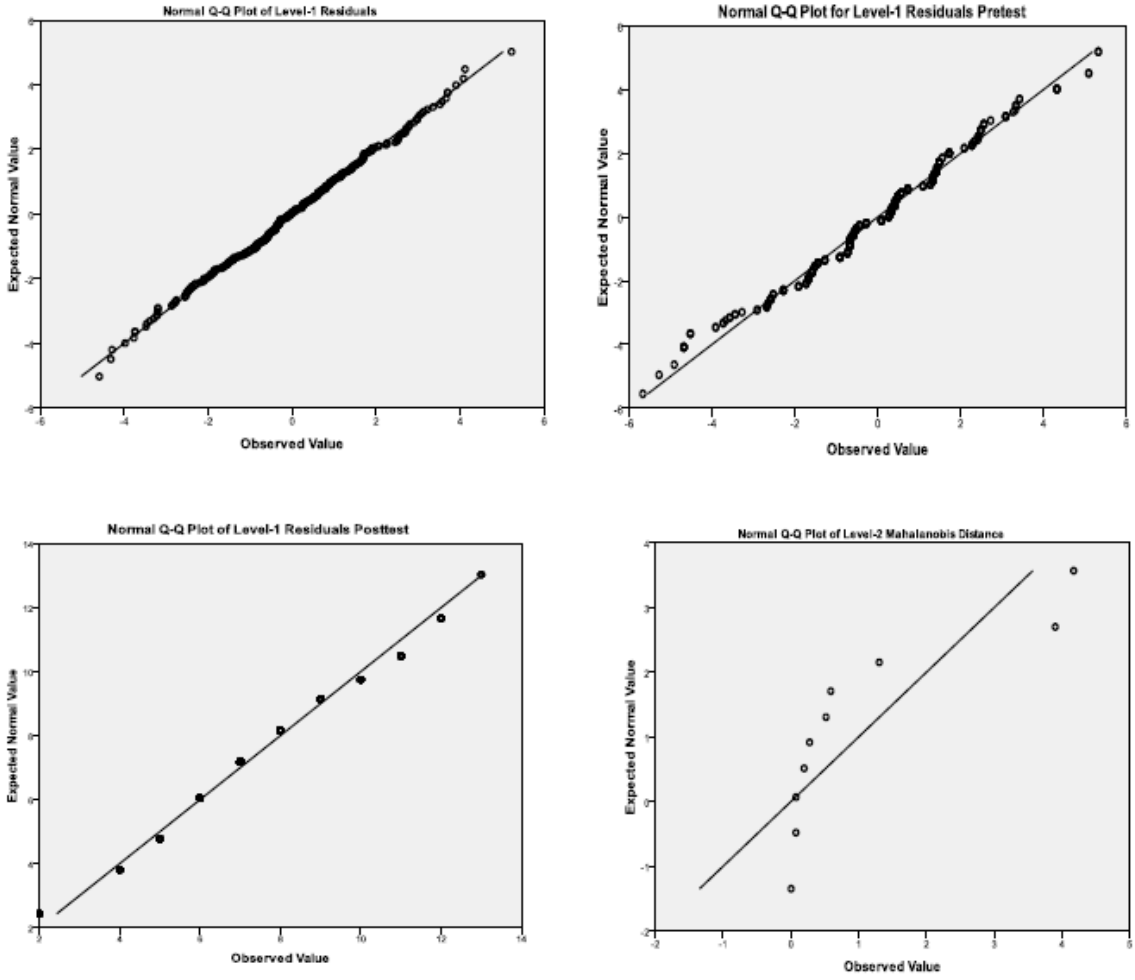

Figure 7. Q-Q plots of level-1 and level-2 residuals.

Summary of research question 3. Research Question 3, which asked what the effectiveness of Ko's Journey was on students' mathematics achievement relative to students who did not play Ko's Journey, was analyzed using Rasch (1960) modeled scores in hierarchical linear modeling (HLM). Hierarchical linear modeling was used

because of the hierarchical structure of the data and the high level of dependency of students' posttest scores attributable to their classroom. The HLM analysis indicated that the posttest score was significantly influenced by the pretest score but not significantly by the Ko's Journey treatment.

A simple graph of the relationship between pretest and posttest without taking into account the hierarchical groups revealed an aptitude treatment interaction (ATI). However, the interaction between treatment and pretest score did not emerge from the HLM analysis. As discussed in Cronbach and Webb (1975), the ATI is an artifact of the data that did not exist within each classroom when between and within class variance was accounted for on the outcome measure. Similar to this analysis, Cronbach and Webb reanalyzed an achievement study showing an ATI; when class membership and between and within class variance were controlled, the ATI artifact disappeared.

Q4 Do the items of the assessment function differently for students using Ko's Journey as a supplement to normal instruction than for students who do not play the game?

To better understand what type of mathematical content students learned from playing the Ko's Journey digital game, Rasch (1960) modeling using Winsteps (Linacre & Wright, 2004) was used to see how the intervention changed the functioning of the items relative to those students who did not play the game. The pretest and posttest data were racked with one row per student and two columns per item (pretest response and posttest response) in the data set. This resulted in two difficulty measures per item (pretest and posttest) for the experimental and control groups. Differential changes in item difficulty between the experimental and control groups were assumed to represent the degree the intervention successfully targeted the item (Linacre, 2012). If Ko's

Journey was successful in improving mathematics achievement targeted on the assessment items, it would be expected that item difficulty would decrease from pretest to posttest relative to the control group (Wright, 2003) . Typically, differences of .5 logits are considered significant enough to not to have happened by chance (Linacre, 2012).

The group that received the Ko's Journey intervention had item difficulty decrease for 16 of the 20 items (Q1, Q2, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13, Q14, Q16, Q17, Q18,and Q20) and increase for four items (Q3, Q4, Q15, and Q19). To visualize the change in item difficulty estimates, experimental and control group pretest item difficulty estimates were plotted against posttest item difficulty estimates (see Figure 8). By plotting a line (x=y) indicating no changes in item difficulty, items over the line were more difficult at the posttest assessment while items under the line were less difficult.

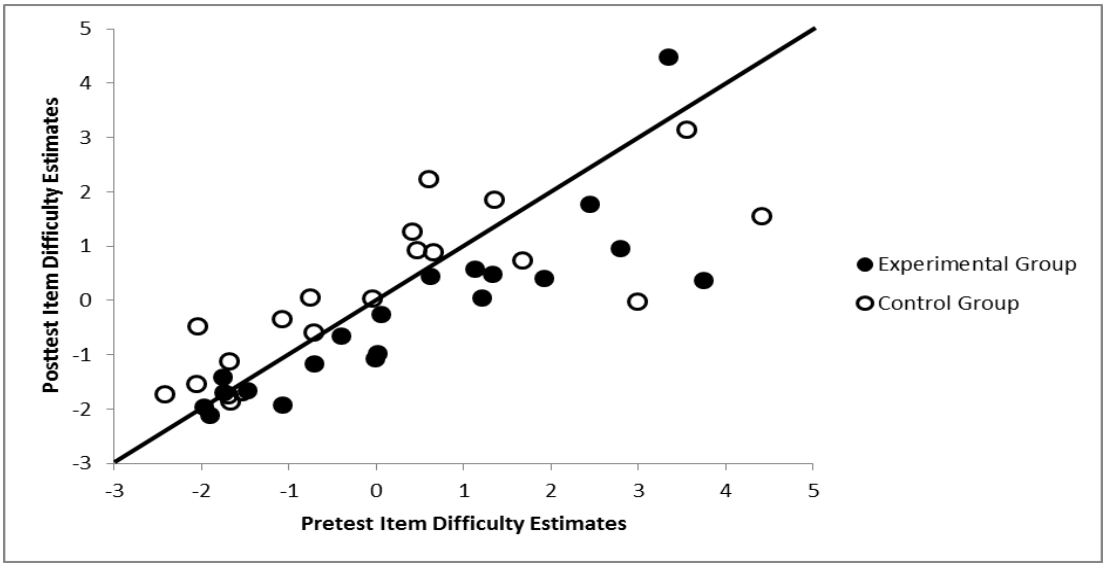

Figure 8. Scatterplot of experimental group and control group item difficulty estimates from racked analyses. Line (y=x) indicates no change in item difficulty.

The control group had item difficulties decrease for 13 of the 20 items (Q1, Q2, Q6, Q8, Q9, Q10, Q11, Q12, Q13, Q14, Q15, Q16, and Q19) and increase for seven items (Q3, Q4, Q5, Q7, Q17, Q18, and Q20). Figure 9 is a Wright map of the racked data and change in item difficulty for the entire sample.

In terms of differential rate of change of the item difficulty between the experimental group and the control group, items Q5_RP, Q6_RP, Q12_EE, Q15_G, Q18_SP, and Q20_SP decreased in difficulty by more than .5 logits for the experimental group as compared to the change in the control group (see Table 18). This might suggest that the intervention was successful in targeting the content for these items.

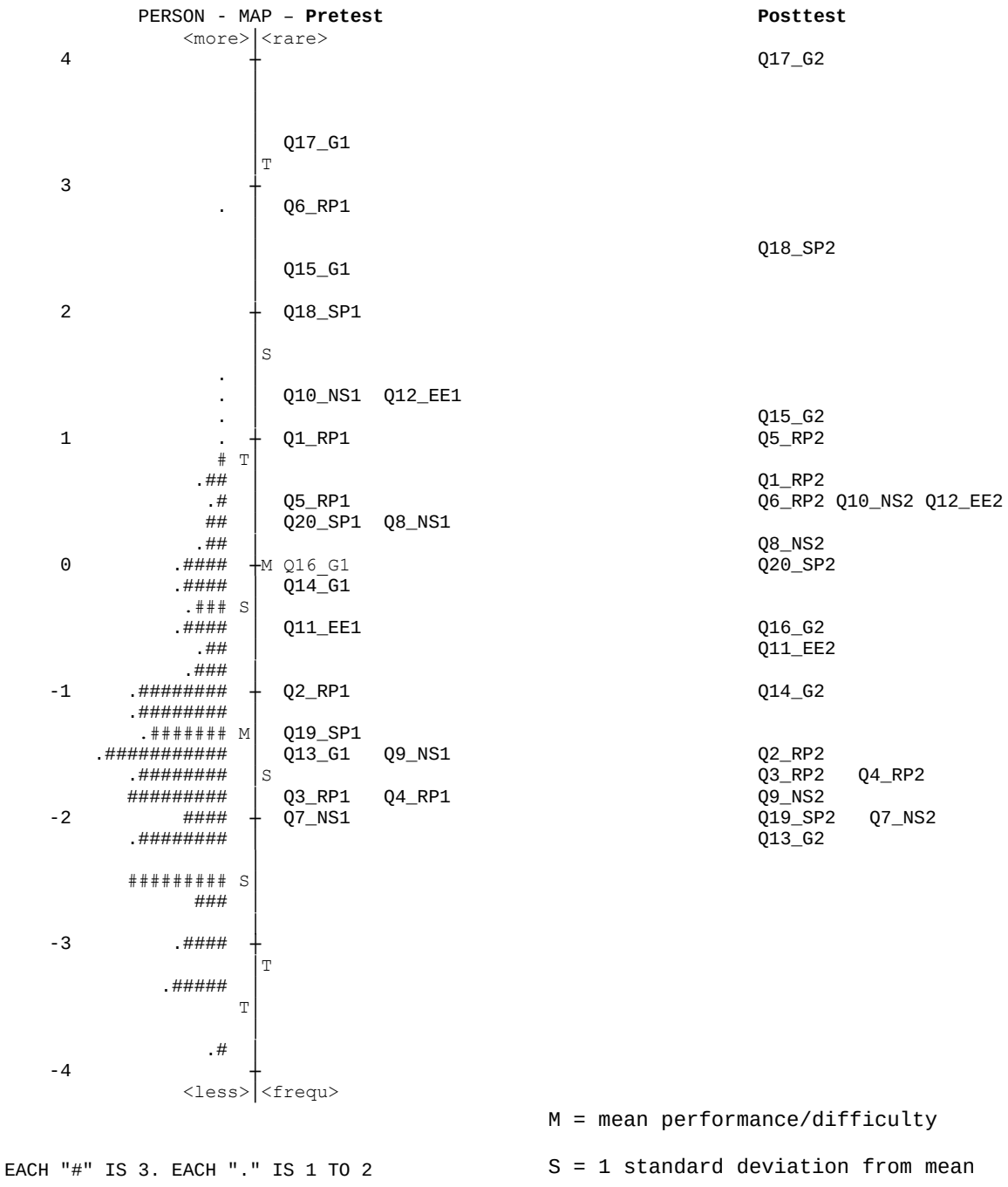

Figure 9. Wright map of racked data and item difficulty change.

Table 18

Item Difficulty Changes for Racked Pretest and Posttest

	Item Difficulty						DIF
	Control		Experimental		Item Change		
ITEM	Pre	Post	Pre	Post	Control	Exp.	
Q1_RP	0.89	0.64	1.13	0.57	-0.25	-0.56	-0.31
Q2_RP	-1.12	-1.69	-0.71	-1.17	-0.57	-0.46	0.11
Q3_RP	-1.68	-1.54	-1.75	-1.7	0.14	0.05	-0.09
Q4_RP	-1.74	-1.71	-1.76	-1.42	0.03	0.34	0.31
Q5_RP	-0.02	2.98	1.21	0.05	3	-1.16	-4.16**
Q6_RP	2.24	0.59	3.75	0.36	-1.65	-3.39	-1.74**
Q7_NS	-1.85	-1.68	-1.91	-2.12	0.17	-0.21	-0.38
Q8_NS	0.05	-0.05	0.62	0.45	-0.1	-0.17	-0.07
Q9_NS	-1.52	-2.06	-1.48	-1.67	-0.54	-0.19	0.35
Q10_NS	1.27	0.41	1.33	0.49	-0.86	-0.84	0.02
Q11_EE	-0.58	-0.73	-0.4	-0.66	-0.15	-0.26	-0.11
Q12_EE	0.94	0.46	1.92	0.4	-0.48	-1.52	-1.04**
Q13_G	-1.72	-2.43	-1.07	-1.92	-0.71	-0.85	-0.14
Q14_G	-0.34	-1.09	0.01	-0.97	-0.75	-0.98	-0.23
Q15_G	1.86	1.34	2.8	0.96	-0.52	-1.84	-1.32**
Q16_G	0.06	-0.76	0.06	-0.25	-0.82	-0.31	0.51
Q17_G	3.15	3.54	3.34	4.48	0.39	1.14	0.75
Q18_SP	1.57	4.4	2.44	1.77	2.83	-0.67	-3.5**
Q19_SP	-0.47	-2.05	-1.97	-1.96	-1.58	0.01	1.59
Q20_SP	0.74	1.67	-0.01	-1.08	0.93	-1.07	-2.00**

Note. DIF logit values > | .5 | for experimental group are denoted by **

A comparison of the content of the assessment items (see Appendix A) decreasing in difficulty for the experimental group and the mathematical content of the Ko's Journey intervention (see Figure 1) showed partial overlap. Item Q5 and Q6 were both multi-step ratio and proportion questions and both required students to calculate sales tax. Although Ko's Journey did not specifically address calculating sales tax, the Medicine and Poultice, Luna, and Crystal Oasis lessons all required working with ratios and proportions. The Medicine and Poultice lesson specifically required students to make a medicinal poultice with the correct ratios to save a wounded wolf pup.

Item Q12 was an expressions and equations item requiring students to solve a real-life mathematical problem using numerical and algebraic equations. The problem read:

> 12. Emma runs 8 miles per hour in cross country. She has already run 4 miles. Her goal is to run a total of 64 miles per week. How much more time, in hours, does she have to run to reach her goal? Write your answer as a decimal.

The lessons with the most overlap to this item included The Great Mountain Climb (determining slope, order of operations, using a basic algorithm), Bolsa and Travel (multi-step equations, determining a variable), and The Crystal Cave (determining variables, decimals, order of operations, using a basic algorithm).

Item Q15 was a multi-step geometry item that required students to use facts about supplementary angles to solve a simple equation for an unknown angle in a figure. The Crystal Oasis lesson required students to estimate angles and the last lesson of Ko's Journey specifically taught about supplementary angles. Given the match of content and recency, it was not surprising that this question became significantly less difficult for the experimental group on the posttest.

Item Q18 was a statistics and probability question that addressed the concept of random sampling and item Q20 asked students to find the probability of a spinner landing on an even number. Although these items became notably less difficult for the experimental group compared to the control group, there was no apparently comparable statistics and probability mathematics content in Ko's Journey as detailed in the educator guide (Imagine Education, 2013). One possibility was that students used alternate techniques such as use of ratios and proportions to solve Q18 or knowledge about percentages and fractions to solve Q20. Nonetheless, the omission of the statistics and probability content in Ko's Journey caused a gap in the full coverage of the seventh grade Common Core State Standard domains (NGA, 2010).

Summary of research question 4. Research Question 4, which asked if items of the assessment functioned differently for students using Ko's Journey compared to students who did not play the game, was analyzed using Rasch (1960) modeling racking techniques using Winsteps (Linacre & Wright, 2004). Pretest and posttest item difficulties were calculated for the experimental and control groups and decreased in item difficulty for the experimental group relative to the changes for the control group. As recommended by Linacre (2012), items with differences of .5 logits were considered notable and possible evidence of content that the Ko's Journey intervention was successful in targeting. Four of the six items (Q5, Q6, Q12, and Q15) had clear connections to Ko's Journey content. However, it is unknown why the experimental students found the two statistics and probability items (Q18 and Q20) less difficult given the lack of comparable content in Ko's Journey.

CHAPTER V

DISCUSSION

This final chapter provides a summary and discussion of research findings regarding the effects of Ko's Journey on seventh grade students' mathematical academic achievement. Implications of the research findings and limitations of the study are discussed. The chapter concludes with recommendations for future studies.

Digital games are widely popular and interest has increased for their use in education. A number of factors underscore the belief that instructional games are beneficial learning tools including they (a) better correspond with the learning needs of today's students who are accustomed to interactive and fast-paced media presentations (Prensky, 2001), (b) are a potential avenue to develop higher order thinking skills needed in the 21st century workforce (Federation of American Scientists, 2006), and (c) have the potential to support underrepresented student populations in their awareness and educational preparation for science, technology, engineering, and mathematics (STEM) careers (Hacker & Kiggens, 2011). Most importantly, instructional digital games are thought to be powerful instructional strategies because they (a) promote active learning and feedback, (b) provide meaningful contexts to situate knowledge, (c) create engagement and intrinsic motivation, and (d) have the ability individualize instruction (Csikszentmihalyi, 1990; Gee, 2006; Kyriacou, 1992; Prensky, 2001).

There are few empirical studies on the use of digital games in middle school mathematics and many include serious methodological flaws (Connolly et al., 2012; Tobias et al., 2011). This indicated a clear need for further empirical work to investigate if the promise of the use of digital games in education was warranted and how to best implement them in the classroom.

Summary and Discussion of Research Findings

The purpose of this study was to investigate the effectiveness of a mathematics digital game, Ko's Journey (Imagine Education, 2011) on seventh grade students' mathematics achievement of Common Core State Mathematics Standards (CCSS; NGA, 2010). This research was conducted using secondary data from a pretest-posttest control group design study (Gall et al., 2003) with a total of 371 seventh grade students. Five mathematics teachers from three schools administered both the pretest and posttest to students in their classrooms. They were instructed to randomly assign their classes to receive the Ko's Journey digital game or serve as a wait-listed control group and continue using the typical mathematics curriculum. A total of 196 students were assigned to experimental classrooms and 175 students were assigned to control classrooms.

This evaluation study was the final phase in a three-phase project. Phase I involved the development and validation of an instrument to measure student achievement because a commercially developed instrument aligned to the CCSS was not available at the time data were collected. Phase II used Rasch (1960) modeling to establish a unidimensional, equal interval scale of measurement for use in the evaluation.

In the current phase of the evaluation, mathematics achievement was defined by a 20-item, researcher-constructed test aligned with the seventh grade CCSS and measured

on unidimensional equal interval scale (Rasch, 1960). Additionally, Rasch (1960) measurement theory was used to identify the differential impact of Ko's Journey on the assessment items including what items were learned as a result of the intervention and to evaluate the need for further refinement of the assessment.

Findings from the current data analysis as well as questions that guided the three phase evaluation study are addressed.

Research Question 1

The first phase of research developed and validated a mathematics instrument aligned with the CCSS for seventh grade mathematics due to the fact that a commercially available instrument was not available at the time of data collection. The final assessment consisted of 20 items (two multiple-choice, and 18 constructed-response) selected by Imagine Education (2011) project directors from a 43-item pool. A validation study using veteran middle school mathematics teachers as subject matter experts found that the majority of the assessment items had an acceptable level of content validity for use in the evaluation research.

Given the comprehensive nature of the CCSS for seventh grade mathematics, it is unrealistic to assert that the final 20-item instrument had complete coverage of the critical domains. The final assessment did have questions in each of the critical domains that were adequate for this research; however, more items would have to be added for sufficient content coverage for high stakes testing. The decision to limit the number of items was made by program directors to gain teacher participation in the project. It was assumed that Imagine Education project directors (2011) selected items that were most closely aligned with the content of Ko's Journey without omitting any domain in the

standards. A comparison of Ko's Journey mathematical content and items, however, did not show bias of the selection given that items more closely aligned were not selected from the item pool.

If the instrument is used in future research, the issue of item format might need to be considered. Constructed-response format was initially selected for items because reliability is typically higher compared to multiple choice questions because guessing is minimized and students are not able to derive the correct solution by a process of elimination (Kastner & Stangla, 2011). However, due to the open-ended response format, it was necessary to review every student's responses for each question. For the initial pretest given to 1,148 students in Phase II, this equaled 22,960 responses including 4,834 unique responses. Although most judgments were straightforward, moving to a multiple choice format would significantly simplify data collection by allowing for automated scoring. Also, having given the items previously as open response, common incorrect answers could be given as plausible distracters to minimize the success of chance guessing (Kastner & Stangla, 2011).

Research Question 2

Analysis of the item level data of the mathematics assessment generally conformed to the requirements of the Rasch model (1960) to produce a unidimensional, equal-interval scale of measurement. Although pilot testing the assessment prior to adoption would have been preferable, using Rasch modeling to scale the assessment and analyze the questions for fit provided evidence that the instrument could be used to measure mathematical proficiency.

Reanalysis of the pretest and posttest with the sample of 371 students was generally consistent with results from the larger sample analyzed in Phase II of the research. The fit statistics did not indicate that any of the requirements of fundamental measurement (unidimensionality, equal item discrimination, and error due to guessing) were violated; therefore, use of Rasch (1960) measurement to construct a true interval scale was appropriate. Items Q5 and Q17 had high Infit values but eliminating them from the assessment did not significantly alter item or person measures. Therefore, it was decided to leave them in the analysis. Although the assessment was composed of different mathematial domains, analysis from classical test theory (CTT) and Rasch modeling supported considering the assessment as a composite test of seventh grade mathematics. The assessment remained difficult for the majority of students even at the time of the posttest. This lack of range in student scores limited the precision of person ability scores and separation into distinct ability levels (Bond & Fox, 2007).

The major advantage of using Rasch (1960) measurement theory was the ability to create a true equal interval scale of measurement for persons and items. Although Rasch theory has been used for development and revision of various educational assessments including developing an early mathematics assessment (Clements, Sarama, & Liu, 2008) to exploring differential item functioning (DIF) in subpopulations in the Trends in International Mathematics and Science Study (TIMMS; Klieme & Baumert, 2001), traditional CTT analyses still dominate educational mathematics journals (Callingham & Bond, 2006). However, Rasch measurement techniques bridge the gap between traditional techniques by using rigorous measurement and also allowing qualitative analysis of individual students. The use of visual displays of persons and

items helps to communicate the information in a way that is informative and understandable by novices. The ability to investigate an individual student's response pattern is a huge advantage over traditional CTT techniques and makes it ideal for use in educational research and in the classroom.

Research Question 3

The main purpose of the research was to evaluate the effects of Ko's Journey, an instructional digital game, on student achievement in the domains of the Common Core State Standards (NGA, 2010) for seventh grade mathematics. Hierarchical linear modeling (HLM) analyses, using person ability logit estimates derived from the Rasch scaling, concluded that the Ko's Journey intervention did not have a significant effect on posttest scores. Gender of the students did not have an impact on pretest score and did not have an interaction with the treatment. The HLM analyses revealed a strong positive relationship between students' pretest and posttest scores within classrooms. Although the differences between the experimental and control classrooms were not statistically significant except for one teacher's classroom, control classrooms exhibited higher pretest scores than did the experimental classrooms. Thus, the deviation of the classroom average pretest scores from the grand mean were included in the model to control for possible selection bias. The addition of classroom average pretest scores (grand-mean centered) to the model explained a significant proportion of variance on posttest scores but did not alter conclusions about the treatment effect.

Analysis of the residuals also suggested that additional level-2 variables other than treatment and average pretest score might be missing from the model. Additional information about teachers such as their comfort and support of technology in education

could be important factors in the successful implementation of Ko's Journey (Butler & Sellbom, 2002).

The finding that Ko's Journey did not significantly improve mathematics achievement corresponded with other findings by Ritzhaupt et al. (2011) and Ke (2008a, 2008b) on other middle school mathematics digital games. The wide variation in terms of effect sizes across teachers in this study was reminiscent of the lack of consistent results of the effects of games such as *DimensionM* and *ASTRAEAGLE* on improving mathematics achievement across multiple contexts (Bai et al., 2012; Ke, 2008a, 2008b; Ke & Grabowski, 2007; Kebritchi et al., 2010; Ritzhaupt et al., 2011). Given the lack of empirical research on digital games, it is difficult to determine why the games were successful in one school but not another. The lack of gender effects was consistent with research on other mathematics games such as *Zombie Division* (Habgood & Ainsworth, 2011).

Fidelity of implementation information gathered by Imagine Education (20120 did shed some light on possible factors that interfered with an ideal estimate of the impact of Ko's Journey on academic achievement. All classrooms reported technological difficulties and this was the reported reason one teacher discontinued use of Ko's Journey. Ko's Journey was not intended to function as a stand-alone mathematics program; it was accompanied by supplementary materials to assist teachers with additional classroom extension activities and how to integrate the game with their traditional curriculum. Interestingly, only one of the teachers (Teacher E) reported using the supplementary activities and she was the only teacher to have significant improvement in achievement for students using the game (ES=+.652). Nonetheless, in

order to make definite conclusions about the efficacy of the game, future studies would need more control over the fidelity of implementation and technological difficulties would need to be resolved.

Research Question 4

To better understand what type of mathematical content students learned from playing the Ko's Journey digital game, Rasch (1960) modeling with racked data was used to see how the intervention changed the functioning of the items relative to those students who did not play the game. Items Q5_RP, Q6_RP, Q12_EE, Q15_G, Q18_SP, and Q20_SP decreased in difficulty by more than .5 logits for the experimental group as compared to the change in the control group. Four of the six items (Q5, Q6, Q12, and Q15) had clear connections to Ko's Journey content in the ratio and proportions, expressions and equations, and geometry domains of the seventh grade CCSS NGA, 2010) .

However, it is unknown why the experimental students found the two statistics and probability items (Q18 and Q20) less difficult given the lack of comparable content in Ko's Journey (Imagine Education, 2013). One possibility could be that students used alternate techniques such as use of ratios and proportions to solve Q18 or knowledge about percentages and fractions to solve Q20. Nonetheless, omission of the statistics and probability content was a notable gap in Ko's Journey's full coverage of the seventh grade Common Core State Standard domains (NGA, 2010).

Rasch (1960) modeling stacking and racking methods could be helpful methods to determine the impact of intervention in the absence of significant overall results (Cunningham & Bradley, 2010; Herrmann-Abell et al., 2012). Such techniques could be

used successfully to inform revisions to the curriculum, assessment instrument, or gauge the efficacy of the intervention to target certain topics.

Implications of Research Findings

Unfortunately, the results of the study did not provide conclusive evidence about the efficacy of Ko's Journey to improve academic achievement. The lack of effectiveness in four out of five classrooms could arise from a myriad of issues including differential implementation of the digital game across classrooms, the lack of integration with the teacher's guide, and the significant technical difficulties encountered by the classroom teachers. The literature review and present results seemed to concur with Clark (1983), Dede (2011), and Sivin-Kachala and Bialo's (2000) position that technology and specifically digital games are not a "silver bullet" in educational reform. Sivin-Kachala and Bialo (2000) succinctly stated:

> Technology can improve teaching and learning, but just having technology doesn't automatically translate to better instructional outcomes. Whether a given school experiences the potential benefits of technology depends on the software it chooses, what students actually do with the software and computer hardware, how educators structure and support technology-based learning and whether there is sufficient access to the technology. (p. 7)

Therefore, future research that considers more of an aligned intervention that incorporates professional development, curriculum, and digital game technology might be a more complex, but fruitful approach, than the technology/no-technology design (Dede, 2011; Roschelle et al., 2010).

However, the existence of technological difficulties in this study was not unique and was one of the main barriers to widespread implementation of computer assisted instruction (CAI) or digital games. Wood, Mueller, Willoughby, Specht, and Deyoung (2005) noted that teachers experienced a high level of problems in terms of computer

hardware and software compatibility; this was a huge barrier to planning and integration of technology. In this particular study, the technical problems included slow download of the game, slow internet connections, and compatibility of the digital game with web browsers. In addition, technical problems with the application also prevented tracking the amount of time students spent playing Ko's Journey.

Another important implication of this study was the importance of using data analytic tools such as hierarchical linear modeling (HLM), which accounted for the hierarchical nature of classroom level data and possible dependencies. While a simple plot of the relationship between pretest and posttest revealed an aptitude treatment interaction, the HLM analysis revealed that this was an artifact of the data when between and within class variance was accounted for on the outcome measure (Cronbach & Webb, 1975). Using analyses that ignore the structure of the data could produce artifacts and result in significant underestimation of standard errors (no between-unit variation) and increased Type I errors (Raudenbush & Bryk, 2002). Notably, only one article encountered during the literature review search used HLM to analyze the effectiveness of technology (Roschelle et al., 2010).

Limitations of the Study

Several limitations existed in this study. Many of the limitations of this study were due to the use of secondary data. Secondary data were used due to the affordances of evaluating an innovative digital game using a large sample that would not be feasible otherwise. However, due to the use of secondary data, desired information about implementation, participants, and schools was often unavailable; therefore, the research questions were restricted to available data.

Inability to control the selection of items in the final assessment and ensure equal representation of all the domains on the test decreased the intended correspondance with the CCSS. Items with high agreement in the teacher content review could have replaced items with low agreement and potentially confusing questions could have been eliminated if the pretest was able to be revised. Given the lack of more detailed demographic information about the students, school data were used to provide information about for whom this study might be generalizable.

In addition, the technological problems that were encountered in the classrooms confounded finding an accurate estimate of the efficacy of Ko's Journey to impact academic achievement. It is unclear if under more ideal classroom conditions, results would have been different. Information about the teacher's level of computer skills and comfort of technology might also be an important factor in the successful implementation of the digital game.

Recommendations for Future Studies

There are numerous ways future studies of Ko's Journey could be improved. Standardizing the implementation of the digital game would be essential to being able to make accurate conclusions about the effect of treatment (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005). Although use of the educators' guide and supplementary classroom activities were encouraged, only one teacher reported using them as designed. It is unclear if the significant improvement in this classroom was a result of the ideal implementation of the program or a result of chance. It would be critical to understand why teachers were reluctant to use the supplementary activities in order to provide improved support for teachers to integrate the game in the classroom. Future studies

must develop a plan for monitoring the implementation of the intervention that includes data collection, observation of the game in practice, and planning for ways to address off-target implementation (Perlman & Redding, 2010).

In addition to the measure of academic achievement, it is logical that a measure of student motivation and efficacy would be an important factor to investigate. Success in playing educational games might boost student's self-efficacy regarding academic competence and transfer it to other academic learning (Dai & Wind, 2011). Self-efficacy and motivation are likely to be more proximal outcomes than achievement that is more distal (Bandura & Schunk, 1981). Additional outcome measures such as standardized test scores would clarify if lack of student motivation or effort was responsible for low scores on the study assessment because the results were personally inconsequential.

Subsequent studies would allow for revision of the current assessment or selection of new standardized assessments that are currently available. As alluded to before, in future research, the current assessment would likely be changed from constructed response to multiple-choice in order to automate scoring. In addition, items showing consistent misfit would be substituted for other items in the item pool reviewed by content experts. Given the continued state adoption of the CCSS, common K-12 standardized assessments are being developed in English and Mathematics that are anchored to the new standards and due to be implemented in the 2014-2015 school year. Therefore, future research could also use new standardized assessments in order to measure changes in academic achievement.

Given the important role of teachers in the successful implementation of digital games, future studies would benefit from more information about teachers' skills,

comfort, and beliefs about technology. Sivin-Kachala and Bialo (2000) stated that the teacher is the most important factor in determining student attitudes about technology. The technology proficiency of the teacher, beliefs about technology, and current use of technology in the classroom would be useful information in future analyses. Familiarity with use of computers and technology would likely be related to comfort and greater integration of the game in the classroom (Wood et al., 2005). Regardless, given the importance of the teacher in successful implementation, future studies must ensure that teachers are well-trained to use the game, technological problems are fixed, and technical support is available if needed.

The addition of stealth assessment data, or learner performance data gathered continuously during the course of play, would strengthen the game and also the ability to assess level of student performance in a future evaluation study (Shute, 2011). Simply knowing the exact amount of time students spent playing the digital game and the number of times students attempted answers before selecting the right answer would provide much needed information about student performance within the game. Shute (2011) noted that with more sophisticated stealth assessment that automates data collection to collect valid evidence of student's competency and support learning, the teacher's workload in relation to managing student's work is decreased and teachers are more likely to implement digital games to support learning. However, such sophisticated assessment strategies are just beginning to be fully implemented into new games and would require major computer programming to accomplish.

Summary

Given the potential promise of digital games and the unique properties of Ko's Journey, this study empirically evaluated the effects of the game on seventh grade students' mathematics achievement. The effects of Ko's Journey were measured by a researcher-constructed test of the Common Core Mathematics Standards (NGA, 2010). The pretest-posttest control group design study found no effects of the game on academic achievement; however, implementation was significantly affected by technical difficulties in the classroom.

REFERENCES

21-6 Productions. (2013). *DimensionM*. Retrieved from http://www.21-6.com/dimensionm.asp

Adams, D. M., Mayer, R. E., MacNamara, A., Koenig, A., & Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of Educational Psychology*, *104*(1), 235-249. doi: 10.1037/a0025595

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Annetta, L. A. (2010). The "I's" have it: A framework for serious educational game design. *Review of General Psychology*, *14*(2), 105-112. doi: 10.1037/a0018985

Bai, H., Pan, W., Hirumi, A., & Kebritchi, M. (2012). Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. *British Journal of Educational Technology*, *43*(6), 993-1003. doi:10.1111/j.1467-8535.2011.01269.x

Balfanz, R., & Byrnes, V. (2006). Closing the mathematics achievement gap in high-poverty middle schools: Enablers and constraints. *Journal of Education for Students Placed at Risk*, *11*(2), 143-159.

Balfanz, R., Ruby, A., & MacIver, D. (2002). Essential components and next steps for comprehensive whole school reform in high poverty middle schools. In S. Stringfield & D. Land (Eds.), *Educating at-risk students* (pp. 128-147). Chicago: National Society for the Study of Education.

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology, 41*(3), 586-598.

Barab, S., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play: Using games to position person, context and content. *Educational Researcher, 39*(7), 525-536.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, A game without guns. *Educational Technology, Research and Development, 53*(1), 86-107.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA'S Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York: The Guilford Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom*. Washington, DC: ERIC Clearinghouse on Higher Education.

Bourgonjon, J., Valcke, M., Soetaert, R., & Schellens, T. (2010). Students' perceptions about the use of video games in the classroom. *Computers & Education, 54*(4), 1145-1156.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42. doi:10.3102/0013189x018001032

Bruckman, A. (1999). *Can educational be fun?* Paper presented at the Game Developers Conference, San Jose, CA.

Bryk, A. S., Raudenbush, S. W., & Cogdon, R. T. (1996). HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs. Chicago, IL: Scientific Software International.

Butler, D., & Sellbom, M. (2002). Barriers to adopting technology for teaching and learning. *EDUcause Review, 2*, 22-27.

Callingham, R., & Bond, T. G. (2006). Research in mathematics education and Rasch measurement. *Mathematics Education Research Journal, 18*(2), 1-10.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). E*ffectiveness of reading and mathematics software products: Findings from two student cohorts*. Washington, DC: U.S. Department of Education.

Cannon Bowers, J. A., Bowers, C. A., & Procci, K. (2011). Using video games as educational tools in healthcare. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 47-84). Charlotte, NC: Information Age Publishing.

Center for Advanced Technologies. (2002). *ASTRA EAGLE*. St. Petersburg, FL: Author.

Chatham, R. E. (2011). After the revolution: Game-informed training in the U.S. military. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 73-99). Charlotte, NC: New Age Publishing.

Cheung, A., & Slavin, R. E. (2011). *The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis*. Baltimore, MD: Johns Hopkins University, Center for Research and Reform in Education.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445-459.

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The research-based early maths assessment. *Educational Psychology*, *28*(4), 457-482. doi: 10.1080/01443410701777272

Coe, R. (2002). *It's the effect size, stupid: What effect size is and why it is important*. Paper presented at the British Educational Research Association, Exeter.

Cognition and Technology Group at Vanderbilt. (1990). Anchored instruction and its relationship to situated cognition. *Educational Researcher*, *19*(6), 2-10.

Cognition and Technology Group at Vanderbilt. (1992a). The Jasper experiment: An exploration of issues in learning and instructional design. *Educational Technology Research and Development, 40*(1), 65-80. doi: 10.1007/bf02296707

Cognition and Technology Group at Vanderbilt. (1992b). The Jasper Series as an example of anchored instruction: Theory, program description, and assessment data. *Educational Psychologist, 27*(3), 291-315.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661-686. doi:10.1016/j.compedu.2012.03.004

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. O., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco: Jossey-Bass.

Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology, 67*(6), 717-724. doi:10.1037/0022-0663.67.6.717

Csikszentmihalyi, M. (1990). *Flow*. New York, NY: Harper and Row.

Cunningham, J. D., & Bradley, K. D. (2010). *Applying the Rasch model to measure change in student performance over time*. Paper presented at the American Educational Research Association Annual Meeting, Denver, CO.

Dai, D. Y., & Wind, A. P. (2011). Computer games and opportunity to learn: Implications for teaching students from low socioeconomic backgrounds. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 477-502). Charlotte, NC: Information Age Publishing.

Davidson. (1983). Math Blaster! [computer program].

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*, 105-115.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinisic motivation. *Psychological Bulletin, 125,* 627-668.

Dede, C. J. (2011). Developing a research agenda for educational games and simulations. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 233-247). Charlotte, NC: Information Age Publishing.

De Jean, J., Upitis, R., Koch, C., & Young, J. (1999). The story of Phoenix Quest: How girls respond to a prototype language and mathematics computer game. *Gender & Education, 11,* 207-223. doi:10.1080/09540259920708

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68,* 179-202.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*(39), 1-8.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Assoication, 76*(374), 341-353.

Dickey, M. D. (2006). Game design narrative for learning: Appropriating adventure game design narrative devices and techniques for the design of interactive learning environments. *Educational Technology Research and Development, 54*(3), 245-263. doi:10.2307/30221219

DiSalvo, B. J., Crowley, K., & Norwood, R. (2008). Learning in context: Digital games and young Black men. *Games and Culture, 3*(2), 131-141. doi:10.1177/1555412008314130

diSessa, A. (2000). *Changing minds: Computers, learning and literacy*. Cambridge, MA: MIT Press.

diSessa, A., & Lay, E. H. (1986). *Boxer* [computer program]. Cambridge, MA :MIT Press.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*(2), 170-177. doi:10.1037/1082-989x.1.2.170

Egenfeldt-Nielsen, S. (2007). Third generation educational use of computer games. *Journal of Educational Multimedia and Hypermedia, 16*(3), 263-281.

Ellison, G., & Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the American mathematics competitions. *Journal of Economic Perspectives, 24*(2), 109-128.

Entertainment Software Association. (2013). *Industry facts*. Retrieved from http://www.theesa.com/facts/index.asp

Federation of American Scientists. (2002). *National summit on educational games: Fact sheet*. Retrieved from www.fas.org/gamessummit/Resources/Factsheet.pdf

Federation of American Scientists. (2006). Summit on educational games: Harnessing the power of video games for learning. Washington, DC: Author.

Finckh, A., & Tramèr, M. R. (2010). Small studies overestimate the benefit of therapies for OA. *Nature Reviews Rheumatology, 6*(11), 617-618. doi:10.1038/nrrheum.2010.162

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Retrieved from http://www.fpg.unc.edu/~nirn/resources/publications/Monograph/pdf/Monograph_full.pdf

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Freeman, C. E. (2004). Trends in educational equity of girls & women: 2004 (NCES 2005-016). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational research: An introduction* (7th ed.). Boston: Pearson Education.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming, 33*(4), 441-467. doi:10.1177/1046878102238607

Gee, J. P. (2003). Opportunities to learn: A language-based perspective on assessment. *Assessment in Education, 10*(1), 27-46.

Gee, J. P. (2006). *Game-like learning: An example of situated learning and implications for opportunity to learn*. Madison, WI: Academic Advanced Distributed Learning.

Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (2nd ed.). New York, NY: Palgrave Macmillian.

Ghani, J. A., & Deshpande, S. P. (1994). Task characteristics and the experience of optimal flow in human-computer interaction. *Journal of Psychology, 128*(4), 381.

Girard, C., Ecalle, J., & Magnan, A. (2012). Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning, 29*(3), 207-219. doi: 10.1111/j.1365-2729.2012.00489.x

Graesser, A. C., Hauft-Smith, K., Cohen, A. D., & Pyles, L. D. (1980). Advanced outlines, familiarity, and text genre on retention of prose. *The Journal of Experimental Education, 48*(4), 281-290. doi:10.2307/20151355

Great Schools. (2013). Test scores and students and teachers. Retrieved from www.greatschools.org

Gredler, M. E. (1996). Educational games and simulations: A technology in search of a research paradigm. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 521-540). New York: Simon & Schuster Macmillan.

Greenberg, B. S., Sherry, J., Lachlan, K., Lucas, K., & Holmstrom, A. (2010). Orientations to video games among gender and age groups. *Simulation & Gaming, 41*(2), 238-259.

Gunter, G., Kenny, R., & Vick, E. (2007). Taking educational games seriously: Using the RETAIN model to design endogenous fantasy into standalone educational games. *Educational Technology Research and Development, 56*, 511-537.

Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educationa games. *Journal of the Learning Sciences, 20*(2), 169-206.

Habgood, M. P. J., Ainsworth, S. E., & Benford, S. (2005). Endogenous fantasy and learning in digital games. *Simulation & Gaming, 36*(4), 483-498.

Hacker, M., & Kiggens, J. (2011). Gaming to learn: A promising approach using educational games to stimulate STEM learning. In M. Barak & M. Hacker (Eds.), *Fostering Human Development Through Engineering and Technology Education* (pp. 257-279). Rotterdam: Sense Publishers.

Hays, R. T. (2005). *The effectiveness of instructional games: A literature review and discussion* (Technical Report No 2005-004). Orlando, FL: Naval Air Warfare Training Systems Division.

Hefner, D., Klimmt, C., & Vorderer, P. (2007). Identification with the player character as determinant of video game enjoyment. *Lecture Notes in Computer Science, 4740,* 39-48.

Herrmann-Abell, C., Flanagan, J., & Roseman, J. (2012, March). *Results from a pilot study of a curriculum unit designed to help middle school students understand chemical reactions in living systems*. Paper presented at the NARST Annual International Conference, Indianapolis, IN.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 65-97). New York: Macmillan Publishing Company.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist*. London: Sage.

Hwang, G., Wu, P., & Chen, C. (2012). An online game approach for improving students' learning performance in web-based problem-solving activities. *Computers & Education, 59*(4), 1246-1256. doi: 10.1016/j.compedu.2012.05.009

Imagine Education. (2011). *Ko's journey* [computer software]. Taos, NM: Author.

Imagine Education. (2012). *Ko's journey on-line math game: Next generation learning challenges*. Taos, NM: Author.

Imagine Education. (2013). *Ko's journey educator guide.* Retrieved from http://www.kosjourney.com/downloads/educators_guide.pdf

Inal, Y., Sancar, H., & Cagiltay, K. (2006). *Children's avatar preferences and their personalities.* Paper presented at the Society for Information Technology & Teacher Education International Conference 2006, Orlando, FL.

Kastner, M., & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia--Social and Behavioral Sciences, 12*(0), 263-273. doi:http://dx.doi.org/10.1016/j.sbspro.2011.02.035

Ke, F. (2008a). A case study of computer gaming for math: Engaged learning from gameplay? *Computers & Education, 51*(4), 1609-1620. doi:10.1016/j.compedu.2008.03.003

Ke, F. (2008b). Computer games application within alternative classroom goal structures: Cognitive, metacognitive, and affective evaluation. *Educational Technology Research & Development, 56*(5/6), 539-556. doi:10.1007/s11423-008-9086-5

Ke, F. (2013). Computer-game-based tutoring of mathematics. *Computers & Education, 60*(1), 448-457. doi:http://dx.doi.org/10.1016/j.compedu.2012.08.012

Ke, F., & Grabowski, B. (2007). Gameplaying for maths learning: Cooperative or not? *British Journal of Educational Technology, 38*(2), 249-259. doi:10.1111/j.1467-8535.2006.00593.x

Kebritchi, M. (2008). *Effects of a computer game on mathematics achievement and class motivation: An experimental study.* Retrieved from http://0-search.proquest.com.source.unco.edu/docview/251432553?accountid=12832

Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education, 55*(2), 427-443. doi:10.1016/j.compedu.2010.02.007

King, J. (2004). *Software solutions for obtaining a kappa-type statistic for use with multiple raters*. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Kirriemuir, J., & McFarlane, A. (2004). Literature review in games and learning. *Futurelab*. Retrieved from http://archive.futurelab.org.uk/resources/publications-reports-articles/literature-reviews/Literature-Review378

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16*(3), 385-402. doi: 10.1007/bf03173189

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London: Routledge.

Krathwohl D. L. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41,* 212-218.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished fissertation, California State University, Los Angeles.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Kulik, C. L. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior, 7*(1–2), 75-94. doi:10.1016/0747-5632(91)90030-5

Kulik, J. A. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say.* Arlington, VA: SRI International.

Kulik, J. A., Kulik, C. L. C., & Bangert-Drowns, R. L. (1985). Effectiveness of computer-based education in elementary schools. *Computers in Human Behavior, 1*(1), 59-74. doi:10.1016/0747-5632(85)90007-X

Kyriacou, C. (1992). Active learning in secondary school mathematics. *British Educational Research Journal, 18*(3), 309.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Lee, K. M., Park, N., & Jin, S. (2006). Narrative and interactivity in computer games. In R. Vorderer & J. Bryant (Eds.), *Playing video games* (pp. 259-274). Mahwah, NJ: Erlbaum.

Leemkuil, H., & de Jong, T. (2011). Instructional support in games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 353-369). Charlotte, NC: Information Age Publishing.

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A. R., Evans, C., & Vitak, J. (2008). *Teens, video games, and civics: Teens' gaming experiences are diverse, and include significant social interaction and civic engagement*. Washington, DC: Pew Internet & American Life Project.

Lepper, M. R., & Malone, T. W. (1987). Intrinsic motivation and instructional effectiveness in computer-based education. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning and instruction: III. Cognitive and affective process analyses* (pp. 255-286). Hillsdale, NJ: Erlbaum.

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review, 22*(3), 215-243. doi:10.1007/s10648-010-9125-8

Liao, Y. (1998). Effects of hypermedia versus traditional instruction on students' achievement: A meta-analysis. *Journal of Research on Computing in Education, 30*(4), 341.

Liao, Y. (2007). Effects of computer-assisted instruction on students' achievement in Taiwan: A meta-analysis. *Computers & Education, 48*(2), 216-233. doi: http://dx.doi.org/10.1016/j.compedu.2004.12.005

Linacre, J. M. (2012). *Winsteps help for Rasch analysis*. Retrieved from http://www.winsteps.com/winman/

Linacre, J. M., & Wright, B. D. (2004). Winsteps: Multiple choice, rating scale, and partial credit Rasch analysis [Computer software]. Chicago: MESA Press.

Lindberg, S., Hyde, J., Petersen, J., & Linn, M. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). Factor: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*(1), 88-91.

Lou, Y., Abrami, P. C., & d'Apollonia, S. (2001). Small group and individual learning with technology: a meta-analysis. *Review of Educational Research, 71*(3), 449-521. doi:10.3102/00346543071003449

Lowrie, T., & Jorgensen, R. (2011). Gender differences in students' mathematics game playing. *Computers & Education, 57*(4), 2244-2248. doi: 10.1016/j.compedu.2011.06.010

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1-27.

Maas, C., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 5*(4), 333-369. doi:10.1016/S0364-0213(81)80017-1

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press.

Mayer, R. E., Griffith, E., Jurkowitz, I. T. N., & Rothman, D. (2008). Increased interestingness of extraneous details in a multimedia science presentation leads to decreased learning. *Journal of Experimental Psychology: Applied, 14*(4), 329-339. doi:10.1037/a001383510.1037/a0013835.supp

McFarlane, A., Sparrowhawk, A., & Heald, Y. (2002). *Report on the educational use of games*. Retrieved from www.teem.org.uk

Meluso, A., Zheng, M., Spires, H. A., & Lester, J. (2012). Enhancing 5th graders' science content knowledge and self-efficacy through game-based learning. *Computers & Education, 59*(2), 497-504. doi:org/10.1016/j.compedu.2011.12.019

Mitchell, A., & Savill-Smith, C. (2004). The use of computer and video games for learning: A review of the literature. London: Learning and Skills Development Agency.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.

National Assessment of Educational Progress. (2011). Mathematics assessments. Washington, DC: U.S. Department of Education

National Center for Educational Statistics. (2012). Indicator 13: Concentration of public school students eligible for free or reduced-price lunch. Retrieved from http://nces.ed.gov/programs/coe/pdf/coe_pcp.pdf

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Retrieved from http://standardstrial.nctm.org/document/

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards for mathematics.* Washington, DC: Author.

National Mathematics Advisory Panel. (2008). *Foundations for success.* Washington, DC: U.S. Department of Education.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *Curriculum Journal, 16*(4), 455-474. doi: 10.1080/09585170500384529

Patton, M. Q. (2002). Two decades of developments in qualitative inquiry. *Qualitative Social Work, 1*(3), 261-283.

Perlman, C. L., & Redding, S. (Eds). (2010). *Handbook on effective implementation of school improvement grants*. Lincoln, IL: Center on Innovation & Improvement. Retrieved from http://www.centerii.org/survey

Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. New York: Basic Books.

Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.

Prensky, M. (2011). Comments on research comparing games to other instructional methods. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 251-280). Charlotte, NC: Information Age Publishing.

Prentice Hall. (2010). *Mathematics Course 2: All-In-One Student Workbook--Version A.* Boston, MA: Author.

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology, 14*(2), 154-166. doi:10.1037/a0019440

Randel, J. M., Morris, B. A., Wetzel, C. D., & Whitehill, B. V. (1992). The effectiveness of games for educational purposes: A review of recent research. *Simulation & Gaming, 23*(3), 261-276. doi: 10.1177/1046878192233001

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). MLwiN Version 2.02. Bristol, UK: Center for Multilevel Modelling, University of Bristol.

Rasch, G. (1960). *Probablistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). London: Sage Publications.

Ricci, K. E., Salas, E., & Cannon-Bowers, J. A. (1996). Do computer-based games facilitate knowledge acquisition and retention? *Military Psychology, 8*(4), 295-307. doi: 10.1207/s15327876mp0804_3

Richy, F., Ethgen, O., Bruyere, O., Deceulaer, F., & Reginster, J.-Y. (2003). From sample size to effect-size: Small study effect investigation (SSEi) [Article]. *Internet Journal of Epidemiology, 1*(2), 12-16.

Rideout, V. J., Foehr, U. G., & Roberts, D. (2010). *Generation M2: Media in the lives of 8 to 18-year olds.* Menlo Park, CA: Kaiser Family Foundation.

Rieber, L. P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research and Development, 44*(2), 43-58.

Ritzhaupt, A., Higgins, H., & Allred, B. (2011). Effects of modern educational game play on attitudes towards mathematics, mathematics self-efficacy, and mathematics achievement [Article]. *Journal of Interactive Learning Research, 22*(2), 277-297.

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities: A Contempory Journal*, *2*(1), 30-38.

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268-302. doi:10.3102/0002831210372249

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., …Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*(4), 833-878. doi: 10.3102/0002831210367426

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641. doi:10.1037/0033-2909.86.3.638

Rudner, L. M., Glass Gene, V., Evartt, D. L., & Emery, P. J. (2002). *A user's guide to the meta-analysis of research studies: Meta-Stat: software to aid in the meta-analysis of research findings*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED471519

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*, 68-78.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. Cambridge, MA: The MIT Press.

Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher, 20*(3), 2-9.

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disaster of well taught mathematics classes. *Educational Psychologist, 23*(2), 145-166.

Sfard, A. (2003). Balancing the unbalancebale: The NCTM standards in light of theories of learning mathematics. In J. Kilpatrick, G. W. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp.353-385). Reston, VA: National Council of Teachers of Mathematics, Inc.

Shih, T. H. (2008). *Adequate sample sizes for viable 2-level hierarchical linear modeling analysis: A study on sample size requirement in HLM in relation to different intraclass correlations.* Unpublished doctoral dissertation, University of Virginia, Virginia. Retrieved from http://0-search.proquest.com.source.unco.edu/docview/304435724?accountid=12832.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. doi: 10.2307/40071124

Shute, V. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishing, Inc.

Sick, J. (2008). Rasch measurement in language education: Part 1. *JALT Testing & Evaluation SIG Newsletter, 12*(1), 1-6.

Sick, J. (2010). Assumptions and requirements of the Rasch measurement. *JALT Testing & Evaluation SIG Newsletter, 14*(2), 23-29.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*(4), 323-355.

Sivin-Kachala, J., & Bialo, E. R. (2000). *Research report on the effectiveness of technology in schools* (7th ed.). Washington, DC: Software and Information Industry Association.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*(3), 427-515.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*(4), 1391-1466. doi:10.3102/0034654309341374

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research, 79*(2), 839-911.

Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31*(4), 500-506. doi:10.3102/0162373709352369

Snijders, T., & Bosker, R. (1996). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences. In P. L. Ackerman, R. L. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research* (pp. 13-59). New York: Freeman.

Squire, K. (2006). From content to context: Videogames as designed experience. *Educational Researcher, 35*(8), 19-29. doi: 10.3102/0013189x035008019

SPSS Inc. (2007). SPSS for Windows, Version 17. Chicago: SPSS Inc.

Squire, K., Barnett, M., Grant, J. M., & Higginbotham, T. (2004). *Electromagnetism supercharged!: Learning physics with digital simulation games.* Paper presented at the 6th International Conference on of the Learning Sciences, Santa Monica, CA.

Steiner, C., Kickmeier-Rust, M., & Albert, D. (2009). Little big difference: Gender aspects and gender-based adaptation in educational games. In M. Chang, R. Kuo, Kinshuk, G. Chen, & M. Hirose (Eds.), Learning by playing. Game-based education system design and development (Vol. 5670, pp. 150-161). Berlin: Heidelberg Springer.

Stylianides, A. J., & Stylianides, G. J. (2007). Learning mathematics with understanding: A critical consideration of the learning principle in the principles and standards for school mathematics. *TMME, 4*(1), 103-114.

Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning. *Review of Educational Research, 81*(1), 4-28. doi:10.3102/0034654310393361

Thomas, S., Heck, R., & Bauer, K. (2005). Weighting and adjusting for design effects in secondary data analysis. *New Directions for Institutional Reseach, 127*, 51-72.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. doi:10.1037/a0023353

Tobias, S., Fletcher, J. D., Dai, D. Y., & Wind, A. P. (2011). Review of research on computer games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 127-222). Charlotte, NC: Information Age Publishing.

U.S. Department of Education. (1997). *Mathematics equals opportunity: White paper prepared for U.S. Secretary of Education Richard W. Riley*. Retrieved from http://mathpl.us/docs/mathemat.pdf

U.S. Department of Education. (2002). *Guidance on the comprehensive school reform program.* Retrieved from http://www.ed.gov/programs/compreform/guidance/index.html

Van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless. *EDUcause Review, 41*(2), 1-16.

Vogel, C. (2008). Algebra: Changing the equation. *District Adminstration, 44*, 34-40.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J. A. N., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis [Article]. *Journal of Educational Computing Research, 34*(3), 229-243.

Vygotsky, L. (1987). *Mind in society*. Cambridge, MA: Harvard University Press.

Wang, W. C., & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65*(4), 376-404.

Watkins, M. W. (2000). Monte Carlo PCA for parallel analysis [computer program]. Retrieved from http://edpsychassociates.com/Watkins3.html

Wenglinsky, H. (1998). Does it compute? The relationship between educational technology and student achievement in mathematics. Policy Information Report. Princeton, NJ: Educational Testing Service.

Whitehead, A. N. (1929). *The aims of education*. New York, NY: MacMillian.

Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America, 17*, 684-688.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement & Evaluation in Counseling & Development, 45*(3), 197-210. doi:10.1177/0748175612440286

Woltman, H., Feldstain, J., MacKay, C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. Tutorials in Quantitative Methods for Psychology, 8(1), 52-69.

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychiatry and Psychology, 17*(2), 89-100.

Wood, E., Mueller, J., Willoughby, T., Specht, J., & Deyoung, T. (2005). Teachers' perceptions: Barriers and supports to using technology in theclassroom. *Education, Communication & Information, 5*(2), 187-206.

Wouters, P., & van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education, 60*, 412-425. doi: 10.1016/j.compedu.2012.07.018

Wright, B. D. (2003). Rack and stack: Time 1 vs. Time 2 or pre-test vs. post-test. *Rasch Measurement Transactions, 17*(1), 905-906.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Panchapakesean, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Wu, W. H., Hsiao, H. C., Wu, P. L., Lin, C. H., & Huang, S. H. (2012). Investigating the learning-theory foundations of game-based learning: A meta-analysis. *Journal of Computer Assisted Learning, 28*(3), 265-279.
doi: 10.1111/j.1365-2729.2011.00437.x

Yelland, N., & Masters, J. (2007). Rethinking scaffolding in the information age. *Computers & Education, 48*(3), 362-382. doi:http://dx.doi.org/10.1016/j.compedu.2005.01.010

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., …Yukhymenko, M. (2012). Our princess is in another castle. *Review of Educational Research, 82*(1), 61-89. doi:10.3102/0034654312436980

APPENDIX A

FINAL TWENTY QUESTION ASSESSMENT WITH COMMON CORE STATE MATHEMATICS STANDARDS

Final 20 Question Assessment with CCSS Mathematics Standards

1. **If a person walks 2/5 mile in each 20 minutes, what is the rate in miles per hour? - Short_Answer**
    - Correct Answer: 1.2
    - **7.RP1** Compute unit rates associated with ratios of fractions, including ratios of lengths, areas and other quantities measured in like or different units. For example, if a person walks 1/2 mile in each 1/4 hour, compute the unit rate as the complex fraction 1/2/1/4 miles per hour, equivalently 2 miles per hour.

2. **Three seventh grade classes were asked if they wanted pizza or hamburgers for their special Friday lunch. Combine the totals for all three rooms. What is the ratio of the number of students who prefer pizza to the number of students who prefer hamburgers? - Short_Answer**
    - Correct Answer: 38:43
    - **7.RP2.b** Recognize and represent proportional relationships between quantities.
        - b. Identify the constant of proportionality (unit rate) in tables, graphs, equations, diagrams, and verbal descriptions of proportional relationships.

3. **If you purchase 5 items for $10, how would you calculate the unit price? - Multiple_Choice**
    - 5 ÷ 5
    - **10 ÷ 5**
    - 10 x 5
    - 5 + 10
    - **7.RP2.c** Recognize and represent proportional relationships between quantities.
        - c. Represent proportional relationships by equations. *For example, if total cost t is proportional to the number n of items purchased at a constant price p, the relationship between the total cost and the number of items can be expressed as t = pn.*

4. **According to your smoothie recipe, you need 2 cups of ice and 3 cups fruit for 5 smoothie servings. How many servings can you make with 15 cups of fruit and an unlimited amount of ice? - Short_Answer**
    - Correct Answer: 25
    - **7.RP2.c** Recognize and represent proportional relationships between quantities.
        - c. Represent proportional relationships by equations. *For example, if total cost t is proportional to the number n of items purchased at a constant price p, the relationship between the total cost and the number of items can be expressed as t = pn.*

5. **To make curtains in your home, you purchase 7 1/2 yards of fabric at $13 per yard. If there is a 7% sales tax, what is the total cost of the fabric? Round to the nearest cent. - Multiple_Choice**
    - **$104.33**
    - $27.50
    - $102.59
    - $99.99
    - **7.RP3** Use proportional relationships to solve multistep ratio and percent problems. *Examples: simple interest, tax, markups and markdowns, gratuities and commissions, fees, percent increase and decrease, percent error.*

6. **The waiter arrives with the bill after dinner. The total is $25.75. You tip at the rate of 18%. What is the total bill? - Short_Answer**
    - Correct Answer: $30.39
    - **7.RP3** Use proportional relationships to solve multistep ratio and percent problems. *Examples: simple interest, tax, markups and markdowns, gratuities and commissions, fees, percent increase and decrease, percent error.*

7. **Find the value of n. Write your answer as a decimal. - Short_Answer**
    - Correct Answer: .7
    - **7.NS1.b** 1. Apply and extend previous understandings of addition and subtraction to add and subtract rational numbers; represent addition and subtraction on a horizontal or vertical number line diagram.
        - b. Understand $p + q$ as the number located a distance $|q|$ from p, in the positive or negative direction depending on whether q is positive or negative. Show that a number and its opposite have a sum of 0 (are additive inverses). Interpret sums of rational numbers by describing real-world contexts.

8. **Lily has 2 1/2 cups of Neapolitan ice cream. A serving size is 1/3 of a cup. How many friends can she serve full servings? - Short_Answer**
    - Correct Answer: 7
    - **7.NS2.b** 2. Apply and extend previous understandings of multiplication and division and of fractions to multiply and divide rational numbers.
        - b. Understand that integers can be divided, provided that the divisor is not zero, and every quotient of integers (with non-zero divisor) is a rational number. If p and q are integers, then $-(p/q) = (-p)/q = p/(-q)$. Interpret quotients of rational numbers by describing real world contexts.

9. **Recorded temperatures at the South Pole Station in Antarctica have ranged from a high of -24° F to a low of -115° F in the month of June. What is the difference in temperature? - Short_Answer**
    - Correct Answer: 91
    - **7.NS1.c** 1. Apply and extend previous understandings of addition and subtraction to add and subtract rational numbers; represent addition and subtraction on a horizontal or vertical number line diagram.
        - c. Understand subtraction of rational numbers as adding the additive inverse, $p - q = p + (-q)$. Show that the distance between two rational numbers on the number line is the absolute value of their difference, and apply this principle in real-world contexts.

10. **Three gallons of paint cost $33.12. How much would a pint cost? - Short_Answer**
    - Correct Answer: $1.38
    - **7.EE3** Solve real-life and mathematical problems using numerical and algebraic expressions and equations.
        - 3. Solve multi-step real-life and mathematical problems posed with positive and negative rational numbers in any form (whole numbers, fractions, and decimals), using tools strategically. Apply properties of operations to calculate with numbers in any form; convert between forms as appropriate; and assess the reasonableness of answers usingmental computation and estimation strategies. *For example: If a woman making $25 an hour gets a 10% raise, she will make an additional 1/10 of her salary an hour, or $2.50, for a new salary of $27.50. If you want to place a towel bar 9 3/4 inches long in the center of a door that is 27 1/2 inches wide, you will need to place the bar about 9 inches from each edge; this estimate can be used as a check on the exact computation.*

11. **11 - 3n = 5. Solve for n. - Short_Answer**
    - Correct Answer: 2
    - **7.EE4.a** 4. Use variables to represent quantities in a real-world or mathematical problem, and construct simple equations and inequalities to solve problems by reasoning about the quantities.
        - a. Solve word problems leading to equations of the form $px + q = r$ and $p(x + q) = r$, where p, q, and r are specific rational numbers. Solve equations of these forms fluently. Compare an algebraic solution to an arithmetic solution, identifying the sequence of the operations used in each approach. *For example, the perimeter of a rectangle is 54 cm. Its length is 6 cm. What is its width?*

12. **Emma runs 8 miles per hour in cross country. She has already run 4 miles. Her goal is to run a total of 64 miles per week. How much more time, in hours, does she have to run to reach her goal? Write your answer as a decimal. - Short_Answer**
    - Correct Answer: 7.5
    - **7.EE4.b** 4. Use variables to represent quantities in a real-world or mathematical problem, and construct simple equations and inequalities to solve problems by reasoning about the quantities.
        - b. Solve word problems leading to inequalities of the form $px + q > r$ or $px + q < r$, where p, q, and r are specific rational numbers. Graph the solution set of the inequality and interpret it in the context of the problem. *For example: As a salesperson, you are paid $50 per week plus $3 per sale. This week you want your pay to be at least $100. Write an inequality for the number of sales you need to make, describe the solutions.*

13. **Keisha drew a scale drawing of the local swimming pool. In real life, the pool is 164 feet long. It is 4 inches on the drawing. How many feet does 1 inch represent (x)? - Short_Answer**
    - Correct Answer: 41
    - **7.G1** Draw, construct, and describe geometrical figures and describe the relationships between them.
        - 1. Solve problems involving scale drawings of geometric figures, including computing actual lengths and areas from a scale drawing and reproducing a scale drawing at a different scale.

14. **A pizza has a diameter of 12 inches. Find the circumference of the pizza. Use 3.14 for π (Pi). Round your answer to the nearest hundredth. - Short_Answer**
    - Correct Answer: 37.68
    - **7.G4** Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.
        - 4. Know the formulas for the area and circumference of a circle and use them to solve problems; give an informal derivation of the relationship between the circumference and area of a circle.

15. **In the diagram below suppose that Angle 1 = 4x and Angle 2 = 2x +10. Note: Angle 1 and Angle 3, and Angle 3 and Angle 2 are supplementary angles and add up to 180°. Solve to find the measure of Angle 3. - Short_Answer**
    - Correct Answer: 160
    - **7.G5** Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.
        - 5. Use facts about supplementary, complementary, vertical, and adjacent angles in a multi-step problem to write and solve simple equations for an unknown angle in a figure

16. **You want to paint the outside of this cube. What is the surface area? - Short_Answer**
    - Correct Answer: 216
    - **7.G6** Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.
        - 6. Solve real-world and mathematical problems involving area, volume and surface area of two- and three-dimensional objects composed of triangles, quadrilaterals, polygons, cubes, and right prisms.

17. **An aquarium is built in the shape of a triangular prism. The volume of a triangular prism can be found by the formula: volume=1/2*length*width*height. What is the volume of the aquarium? - Short_Answer**
    - Correct Answer: 36
    - **7.G6** Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.
        - 6. Solve real-world and mathematical problems involving area, volume and surface area of two- and three-dimensional objects composed of triangles, quadrilaterals, polygons, cubes, and right prisms.

18. **The Mars Company website states that each bag of original milk chocolate M&M's contains 1.69 ounces and has an average of 55 M&M's. A random sampling of 45 packages of M&M's found the following percentages of colors. What is the best estimate for the number of blue M&M's in the next bag? Round to the nearest whole number. - Short_Answer**
    - Correct Answer: 10
    - **7.SP2** Use random sampling to draw inferences about a population.
        - 2. Use data from a random sample to draw inferences about a population with an unknown characteristic of interest. Generate multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions. *For example, estimate the mean word length in a book by randomly sampling words from the book; predict the winner of a school election based on randomly sampled survey data. Gauge how far off the estimate or prediction might be.*

19. **Jose's parents kept records of the number of text messages he sent per day of the week. What is the median number of text messages for the week? - Short_Answer**
    - Correct Answer: 12
    - **7.SP4** Draw informal comparative inferences about two populations.
        - 4. Use measures of center and measures of variability for numerical data from random samples to draw informal comparative inferences about two populations. *For example, decide whether the words in a chapter of a seventh-grade science book are generally longer than the words in a chapter of a fourth-grade science book.*

20. **You spin a spinner numbered 1 through 6. Each number is equally likely. Find the probability of it landing on an even number. Write the probability as a decimal? - Short_Answer**
    - Correct Answer: .5
    - **7.SP5** Investigate chance processes and develop, use, and evaluate probability models.
        - 5. Understand that the probability of a chance event is a number between 0 and 1 that expresses the likelihood of the event occurring. Larger numbers indicate greater likelihood. A probability near 0indicates an unlikely event, a probability around 1/2 indicates an event that is neither unlikely nor likely, and a probability near 1 indicates a likely event.

APPENDIX B

COMMON CORE STATE MATHEMATICS STANDARDS: SEVENTH GRADE CONTENT AREAS AND PRETEST ITEMS

Common Core State Mathematics Standards: Seventh Grade Content Areas and Pretest Items

- **Ratios and Proportional Relationships**
  - Analyze proportional relationships and use them to solve real-world and mathematical problems.

- 7.RP1 **PTQ1**
- 7.RP2.c **PTQ4**
- 7.RP2.c **PTQ3**
- 7.RP3 **PTQ6**
- 7.RP3 **PTQ5**
- 7.RP2.b **PTQ2**

- **The Number System**
  - Apply and extend previous understandings of operations with fractions to add, subtract, multiply, and divide rational numbers.

- 7.NS1.b **PTQ7**
- 7.NS1.c **PTQ9**
- 7.NS2.b **PTQ8**

- **Expressions and Equations**
  - Use properties of operations to generate equivalent expressions.
  - Solve real-life and mathematical problems using numerical and algebraic expressions and equations.

- 7.EE4.a **PTQ11**
- 7.EE4.b **PTQ12**
- 7.EE3 **PTQ10**

- **Geometry**
  - Draw, construct and describe geometrical figures and describe the relationships between them.
  - Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.

- 7.G4 **PTQ14**
- 7.G1 **PTQ13**
- 7.G5 **PTQ15**
- 7.G6 **PTQ16**
- 7.G6 **PTQ17**

- **Statistics and Probability**
  - Use random sampling to draw inferences about a population.
  - Draw informal comparative inferences about two populations.
  - Investigate chance processes and develop, use, and evaluate probability models.

- 7.SP2 **PTQ18**
- 7.SP5 **PTQ20**
- 7.SP4 **PTQ19**

APPENDIX C

INSTITUTIONAL REVIEW BOARD APPLICATION, APPROVAL, AND PERMISSION TO USE ARCHIVAL DATA

Research Involving Human Participants
Coversheet for UNC IRB Application

Important note: **You must use Adobe Acrobat to complete this form**. Do not use Preview (the default Mac OS X application for displaying PDF documents). There is a compatibility problem, and PDF forms filled out in Preview do not display the form data when opened in Acrobat. If you choose not to use Acrobat, you will likely encounter a delay in processing of your IRB application.

If you do not have Acrobat Reader, click on this button to download a free copy:

Project Title: Evaluation of the Effects of Digital Mathematics game on Academic Achievement

Contact Information (reviewers will communicate via IRBNet)

Principal Investigator: Christine Marie Wale Phone #: 970-381-1232

School/Department: Educational Psychology UNC e-mail: christi.wale@hotmail.com

Research Advisor: Marilyn Welsh UNC e-mail: Marilyn.Welsh@unco.edu
(*required for students*)

CERTIFICATION OF PRINCIPAL INVESTIGATOR (PI)

I certify that this application accurately reflects the proposed research and that I and all researchers who will have contact with the participants or access to the data have reviewed this application and the Guidelines of the UNC IRB, and will comply with the letter and spirit of these policies. I understand that any changes in procedure which affect participants must be submitted to the IRB (using the Request for Change in Protocol Form) for written approval prior to their implementation. I further understand that any adverse events and significant changes in risk for participants must be immediately reported in writing to the UNC IRB.

The signature of the PI must be completed on IRBNet.

CERTIFICATION OF RESEARCH ADVISOR

I certify that I have thoroughly reviewed this application, confirm its accuracy, and accept responsibility for monitoring the conduct of this research, the maintenance of any consent documents as required by the IRB; and, in the case of expedited reviews, the continuation review of this project in approximately one year.

The signature of the Research Advisor (if applicable) must be completed on IRBNet.

Summary Information (to be completed by the Lead Investigator)

Review Category: ☑ Exempt (2-3 weeks) ☐ Expedited (3-4 weeks) ☐ Full-Board (4-6 weeks)

Research participants will be: (e.g., adults, elderly, children, healthy, unhealthy, etc.)	Middle school students
Type of data collected will be: (e.g., survey responses, interviews, blood samples, existing data, etc.)	Secondary data analysis
Location of collected data:	Middle schools

Is standard consent documentation used: ☐ YES ☑ NO If NO, you must be addressed within application.

Is permission required (e.g., school district)? ☑ YES ☐ NO If YES, you must include letter (this is not consent).

Is this a funded research project? ☐ YES ☑ NO If YES, you must provide source within application.

2.22.15

Clear the Form

Narrative: Evaluation of the Effect of Digital Mathematics Game on Academic Achievement

Christine Wale

A. Purpose

1. The relatively few empirical studies of digital games, the contradictory results in mathematics education, and methodological flaws in empirical studies indicate a clear need for further rigorous empirical investigation of digital games to better understand if the promise of the use of digital games in education is warranted and how to best implement them in the classroom. This study will address this need and empirically evaluate the effect of a digital mathematics game, Ko's Journey, on 7th grade students mathematical achievement. Ko's Journey is digital online computer game that follows Ko, a young girl in an ancient wilderness who must make her way back to her kin. Students' progress through the game by using the guidebook and story based math modules targeting critical areas of the 7th grade Common Core State Standards. The mathematics topics encountered are anchored to the game, and not superfluous to the overall story, such as helping Ko set a compass to the proper degree or mix medicine into ratios for a sick wolf pup (Imagine Education, 2012). Mathematics achievement is defined by a researcher-constructed test aligned with the Common Core Mathematics Standards (Common Core State Standards, 2012) and measured on a unidimensional equal interval scale (Rasch, 1960). Additionally, the Rasch measurement theory (Rasch, 1960) will be used to identify the differential impact of Ko's Journey on the assessment items, such as what items are learned, and to evaluate need for further refinement of the current assessment.

This research will be conducted through of the use of an archival data set from a program evaluation of Ko's Journey.

2. Justify selection of category type: Exempt

I have permission from Imagine Education and Dr. John Cooney, University of Northern Colorado Emeritus faculty, to use this data set for this project (letter attached). This qualifies as exempt because this research involves the study of existing data and students are identified by randomly generated identification numbers.

B. Methods – Be specific when addressing the following items.

1. Participants

The participants were 1152 students from 9 different schools. Students were selected for control and experimental groups by their regular classroom teachers. The Ko's Journey computer game was added to the regular curriculum for the experimental groups.

2. Data Collection Procedures

The data was collected by Imagine Education I have received permission from Imagine Education and Dr. John Cooney to analyze this data.

Data were collected from middle school students enrolled in multiple school districts throughout the U.S. in accordance with the school districts' policies and procedures for conducting program evaluation on curriculum innovations. The research was conducted in commonly accepted educational settings such as the students' regular classroom or home school setting to evaluate the effectiveness of adding the educational software to the students' normal mathematics curriculum. The data are comprised of an educational achievement test (Attached) intended to

measure students' knowledge of mathematical concepts introduced in the middle school mathematics curriculum, students' gender and their English language proficiency. The information is maintained in a manner that students cannot be identified, directly or through identifiers linked to the students.

3. Data Analysis Procedures

Data analysis will start by using Rasch measurement theory (Rasch, 1960) using Winsteps (Linacre & Wright, 2004) to construct a true equal interval scale of measurement for items and the subset of persons that completed the pretest and posttest. Classical test theory and Rasch modeling will used to assess the model. The basic Rasch assumptions of unidimensionality, local independence, no error due to guessing, equal discrimination will be assessed through analyzing fit statistics and a Rasch Principal Component Analysis (PCA) of residuals and a parallel analysis that will contrast residual eigenvalues from this analysis with randomly generated residual eigenvalues (Watkins, 2000). Additional assumptions will be examined using residual based fit statistics, MNSQ and ZSTD Infit and Outfit statistics.

Given the hierarchical structure of the data, students nested in classrooms, and the varying implementation of Ko's Journey, hierarchical linear modeling will be used and the a priori model will be analyzed in HLM7 (Bryk, Raudenbush, & Cogdon, 1996).

4. Data Handling Procedures

Data will be stored on a password protected computer and all individuals are identified by a randomly generated identification number and identifiers that cannot be linked back to participants.

.

C. Risks, Discomforts and Benefits

The risks inherent in this study are no greater than those normally encountered during regular classroom participation.

Students in the study benefited from the opportunity to use new technology to improve mathematics achievement. Student in the control group will be given the opportunity to use the computer program in the future.

Information gained about the usefulness of the current assessment as well as refining it for future use will also be beneficial.

D. Costs and Compensations

E. Grant Information (if applicable)

Attach all relevant materials to the application.
These materials may include, but are not limited to:
 Consent Documents – Follow the guidelines for construction of consent documents.
 Letters of Permission – Attach written permission from site of data collection if external to UNC. Letters or forwarded e-mails should document the permission of appropriate officials to recruit participation from and collect data in schools, child care centers, hospitals, clinics, and other universities.
 Survey Instruments – Copies of widely used standardized tests are not necessary.

 Questionnaires
 Interview Questions/Potential Questions/Protocols/Range of Topics
 Debriefing Materials (if applicable)
 Documentation of IRB Training (required for federally funded research and for full board review protocols)

Submit the original and one copy of the cover page, narrative, and supplementary materials to the Office of
Sponsored Programs (OSP), Campus Box 143 (Kepner #25), Attn: Sherry May.

UNIVERSITY of
NORTHERN COLORADO

Institutional Review Board

DATE:	June 19, 2013
TO:	Christine Wale
FROM:	University of Northern Colorado (UNCO) IRB
PROJECT TITLE:	[468156-1] Evaluation of the effect of a digital mathematics game on academic achievement
SUBMISSION TYPE:	New Project
ACTION:	APPROVAL/VERIFICATION OF EXEMPT STATUS
DECISION DATE:	June 13, 2013

Thank you for your submission of New Project materials for this project. The University of Northern Colorado (UNCO) IRB approves this project and verifies its status as EXEMPT according to federal IRB regulations.

We will retain a copy of this correspondence within our records for a duration of 4 years.

If you have any questions, please contact Sherry May at 970-351-1910 or Sherry.May@unco.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Northern Colorado (UNCO) IRB's records.