CRESST REPORT 774

Terry P. Vendlinski
Girlie C. Delacruz
Rebecca E. Buschang
Gregory K. W. K. Chung
Eva L. Baker

DEVELOPING HIGH-QUALITY ASSESSMENTS THAT ALIGN WITH INSTRUCTIONAL VIDEO GAMES

OCTOBER 2010

The National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Sciences
UCLA | University of California, Los Angeles

Developing High-Quality Assessments That Align with Instructional Video Games

CRESST Report 774

Terry P. Vendlinski, Girlie C. Delacruz, Rebecca E. Buschang,
Gregory K. W. K. Chung, & Eva L. Baker
CRESST/University of California, Los Angeles

October 2010

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for Advanced Technology in Schools (CATS)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305C080015.

TABLE OF CONTENTS

DEVELOPING HIGH-QUALITY ASSESSMENTS THAT ALIGN WITH INSTRUCTIONAL VIDEO GAMES

Terry P. Vendlinski, Girlie C. Delacruz, Rebecca E. Buschang,
Gregory K.W.K. Chung, & Eva L. Baker
CRESST/University of California, Los Angeles

Abstract

The evaluation of educational interventions requires assessments that consistently (reliably) produce data from which accurate (valid) inferences about the test subjects can be made for some stated purpose. Despite codified definitions of all these terms, there remains vibrant debate about the assessment design process and how measures of technical quality should be determined and reported. More importantly, there seems to be little consistency in the process of developing assessment items that demonstrate consistently high technical quality and that are conceptually aligned with instruction. We present a process, rooted in the work of Baker (1974), in which a theory of learning drives the development of knowledge specifications that then drive the integrated development of assessment items and forms and instructional materials. In this case, the instruction is in the form of an educational video game about adding rational numbers. We discuss the development of the knowledge and item specifications, an initial version of the instructional game based on these specifications, and report the consistently high technical quality (internal consistency and test-retest reliability—Cronbach's alpha of 0.9 to 0.94). We discuss as well the ability of pretest measures to predict subsequent student performance (inferential validity—Spearman's rho of 0.69) in the game and on similarly developed assessment measures. Alignment between the instructional game and assessment items is also discussed.

Background

In most human pursuits, those involved in the endeavor want to know if and when their goal has been attained. This is especially true when the pursuit involves high-stakes outcomes and where large investments of time, money, and energy have supported the effort. In many cases, determining attainment of the goal is simple because it is clear if and when the goal has been reached. However, in some instances, the goal may be less clear and and determining whether or not it has been attained may prove difficult. Educational assessment seems often to fall in the latter category. The causes of the lack of clarity in this particular field seem numerous, including disagreement about what the goal(s) should be, misunderstandings about what a stated goal means, and difficulty measuring attainment of a goal that is only indirectly observable.

While these difficulties are real, and probably long enduring, they need not become an insurmountable barrier in fields like education. Rather, they suggest the need to very clearly define a goal and to be clear about the evidence that will be collected and about how it will be assembled so that observers can accurately and consistently infer whether the stated goal has been achieved. Although this seemingly stresses the summative or accreditation role for testing, we need not limit the enterprise of educational assessment merely to summations of student achievement.

In 1994, Samuel Messick wrote that "the essence of authentic assessment must be sought … in the quest for complete construct representation" (p. 13), and he went on to cogently argue the need to avoid underrepresentation of the construct of interest and to protect against construct-irrelevant variance when designing assessments for any purpose. In particular, he suggests:

> There should be a guiding rationale akin to test specifications that ties the assessment of particular products or performances to the purposes of the testing, to the nature of the substantive domain at issue and to the construct theories of pertinent skills and knowledge (p. 14).

Even prior to Messick, Eva Baker and colleagues (Baker, O'Neil, & Linn, 1993; Linn, Baker, & Dunbar, 1991) suggested that the technical quality of assessments must be integral to a set of learning objectives and instructional design. More recently, she and her colleagues (Baker, Chung, & Delacruz, 2008) have continued to emphasize that high-quality assessments need to

- adequately represent a targeted domain (rather than just a arbitrary subset of items);
- appropriately represent the cognitive demand required for success in the targeted domain;
- be tied to performance categories (or criteria) by empirical evidence;
- provide some evidence that the results are generalizable or transferable;
- demonstrate consistency (between raters or across occasions, etc.); and
- exhibit fair results.

It is essential that, in the words of Baker and colleagues (2008), evidence show "how the measures relate to other measures of the construct and how the measures discriminate between high and low performers" (p. 600). In fact, for over 35 years, Baker (1974) has consistently argued the need for designing assessments around small subsets of essential content that students are expected to learn and then describing how students will demonstrate they have acquired that knowledge.

Others have tried to instantiate these ideas, for various purposes and with varying degrees of success. One of the most notable and successful of these efforts is Mislevy and colleagues'

Evidence-Centered Design (ECD). ECD provides a framework for designing assessments for various purposes from a technical quality perspective. As such, it attempts to model assessment events in terms of the student (knowledge, skills, and abilities of interest), evidence (how what is observed informs inferences about student variables), tasks (the kinds of things that will elicit the evidence required), and assembly (how the student, evidence, and task work together to form the assessment). See Almond, Steinberg, and Mislevy (2002) for additional detail. The key objective of ECD, as stated by these developers is "to bring probability-based reasoning to bear on the problems of modeling and uncertainty that arise naturally in all assessments" (Mislevy, Almond, & Lukas, 2003, p. 1).

Influenced by the development of ECD and firmly rooted in the work of Baker (1997), researchers at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) have approached test design using a similar approach. In particular, during the last decade, CRESST projects have used ontologies as the foundation for a number of educational interventions. These interventions involve not only summative and formative test and test item designs, but also integrated professional development and instructional materials that align and integrate with such test and test items (Phelan, Choi, Vendlinski, Baker, & Herman, in press; Vendlinski, 2009). Ontologies define key conceptual ideas in a specific domain and describe the relationships between these concepts. While ontologies can be expansive, in many cases, the actual breadth and depth of the ontologies can be narrowed for practical reasons such as focusing on particular learning goals. It should be noted, however, that these more narrow ontologies are designed to fit within the more overarching ontologies.

While CRESST researchers have used, and continue to use, ontologies as the foundation for the development of Bayesian networks (in a manner similar to that advocated in ECD), these ontologies have also been used to determine how instruction and assessment might be aligned around key principles that organize understanding in domains as diverse as history (Baker, Freeman, & Clayton, 1991), science (Vendlinski, Niemi, & Wang, 2005), electrical engineering (Chung, Dionne, & Kaiser, 2006), and mathematics (Vendlinski et al., 2009). In this way, ontologies help identify and drive the selection of key big ideas on which instruction, assessment, and even professional development will focus. These efforts align with recent findings in cognitive science, especially the literature describing how experts and novices differ in their attentiveness to problem features, how experts organize knowledge, and how they access knowledge. Consequently, the focus at CRESST has attended to identifying the key ideas that seem to form the essential kernels necessary for novices to organize understanding and develop expertise within a larger domain (e.g., mathematical operations on rational numbers) rather than identifying all atomistic concepts that may be important to performing various tasks.

More recently, CRESST researchers have started applying these same principles to the development and integration of assessments with digital instructional materials like educational games. This paper first describes this process in some detail, tracing the development of key foundational ideas and how these ideas are connected to other ideas. Next, it describes how these ideas are used both as the underpinnings for the development of educational games and for the development of item specifications that are used to assess what students have learned from the game. Finally, the paper addresses the technical quality of the assessments that result from using the integrated process described. It is argued that integrating the identification of key knowledge ideas with such a materials-development process not only focuses the product but also provides an opportunity to quantitatively and qualitatively justify many of the technical quality claims made about the assessment items and forms.

The current project was conducted under the auspices of the Center for Advanced Technology in Schools (CATS). CATS was staffed, in part, by researchers from CRESST, and both centers are located on the UCLA campus.

Methods

Developing Key Foundational Ideas and Knowledge Specifications

The game and assessment development process began with developing the key foundational ideas that would form the learning objectives of instruction, assessment, and game play. Given that the first of the CATS games was intended to address the addition of rational numbers, as specified in the grant proposal, researchers looked at how students should understand rational numbers, how rational number addition should be connected to previous student knowledge of numbers and of addition, and at common student misconceptions that highlight important and common shortcomings in student understanding of both rational numbers and the addition of rational numbers. For example, students will often add both the numerators and denominators of rational numbers (Brown & Quinn, 2006; Driscoll, 1982) or students will often try to add dissimilar units such as miles and miles per hour. The key questions to be answered when defining the knowledge specifications are “What do we want students to learn about a topic,” “How is this knowledge connected to both prior and future learning,” and “Why is this knowledge important to the students’ future success in life and academia?” As ideas for knowledge specifications surfaced, each idea was decomposed into its more atomistic components so that key concepts related to understanding that particular idea were identified and so that these key components could be linked to other concepts that had been previously identified. In this way, recurring themes became clear and connections between the concepts became evident.

The result of this process was the identification of nine key foundational ideas and the ontological connections between these ideas and other ideas in mathematics, especially in the gateway course of algebra (Atanda, 1999; Berkner & Chavez, 1997). Given that the scope of the current study was rational number addition, two of the nine key ideas that emerged from the literature (Lamon, 1999; Wu, 2001) and in discussions between math educators and researchers were selected as particularly important for us to focus on:

- Only identical units can be added to create a single numerical sum.

And

- The size of a rational number is relative to how one whole unit is defined.

We began with these particular specifications for two reasons. First, we wish to connect the concepts we are asking students to (re)learn to their prior knowledge. In this sense, we wish to highlight the similarities between addition of rational numbers and addition of integers. The fact that such similarities are often left implicit for students, or that the tasks are explicitly labeled as being conceptually different, seems to be at least part of the problem students have with rational number addition (Kilpatrick, Swafford, & Findell, 2001; Wu, 2001). Leaving students to make such connections within the set of real numbers on their own seems to engender a number of student misconceptions such as adding numerators and adding denominators of two rational number addends (Brown & Quinn, 2006; Gelman, 1991). Alternatively, explicitly teaching students that fractions are different from integers can cause problems when students are subsequently required to write integers in fractional form or when students try to add integer quantities together that don't share the same units (e.g., 60 miles per hour and 120 miles). Second, we wanted to make the *need* for non-integer rational numbers clear and to help clarify how identical units are defined when dealing with rational numbers.

The two foundational key ideas serve both of these purposes. When young students learn to add small positive integers, it is clear that there is a general learning progression (Fuson, 2003; Carpenter, Fennema, Franke, Empson, & Levi, 1999). Children generally begin to add integers by decomposing each number into units of one (i.e., expressing the number's cardinality) and then adding those units. For example, the number 2 is seen as two units of something (1 + 1). Adding five more is then seen as continuing the count by adding five more of those identical units. Later, students discover that they do not need to decompose both numbers but can just "count on" to the larger quantity. Although the process eventually becomes one of recall, even adults often unconsciously decompose into units and add. For example, adults would intuitively say that adding a two-dollar bill to a five-dollar bill produces seven dollars rather than saying the addition produces two bills. It seems, however, that as integer addition becomes easier to recall,

students do not attend to an important restriction on addition, namely that we only add similar things together. Many teachers apparently do not reinforce this restriction on addition, especially right before rational number addition. Instead, the addition of rational numbers is often seen as unique from what was previously learned, and students struggle to learn a new set of disparate rules about adding these "new" numbers.

The first key idea highlights the connection between integer and rational number addition. In both cases, the goal is that students realize that addition produces a *meaningful* sum only when identical units are added. One can add a goat and a horse, for example, but such an addition only makes sense if each is first "converted" into a common unit such as animal or mammal. Trying to add units such as 70 degrees Fahrenheit and 12:05 p.m. is nonsensical, however, because they are not the same unit, and we have no apparent way to convert them into a unit that is common to both. Unfortunately, the importance of such an understanding is often only made explicit when students are confronted with the need to add fractions (i.e., the need for a common denominator), and the similarity with integer addition is seemingly seldom addressed or is addressed in a perfunctory way (Mack, 1990). While such examples may seem obvious to students at face-value, experience suggests that algebra students will often try to add dissimilar quantities like the number of nickels and the value of dimes or miles and miles per hour when trying to solve problems (De Corte & Verschaffel, 1987).

When students do try to make connections between integer and rational number addition on their own, they often attempt to perform operations they already know, like adding the integer parts of a rational number just as they would add integers. This suggests that there may be a larger problem, namely that students do not really understand what rational numbers mean (Kilpatrick et al., 2001). In this case, even if they know what addition means, they cannot apply such understanding correctly because it is unclear how to decode the "unit" of a rational number. The work of Susan Lamon (1999) and others (Fuson, 2003, and Behr, Harel, Post, & Lesh, 2003, for example) supports such a notion. Other research suggests that student concepts of rational numbers are almost entirely centered on representations of circles, and subtleties like the comparability of the wholes are lost (Mack, 1990; Saxe, Gearhart, & Seltzer, 1999). Clearly, one half of a small unit is not the same size as one half of a larger unit. But if students are asked to place the number one half on a number line that spans the range zero to four, they may place the number ½ to coincide with 2 (half of the number line represented) rather than half the distance between zero and one. Consequently, the second key idea specifies the importance of the relationship of a rational number to the unit when defining a rational number. Like the notion of addition, the importance of connecting the meaning of a rational number to its corresponding

unit seems to be left implicit for many students and seems to result in a number of student misconceptions (Saxe et al., 2007; Wu, 2001).

In order to fully address the notion of rational number addition, both of these knowledge specifications must be addressed. Other key ideas, which flow from these two overarching ideas, were also developed to accommodate the scope of instruction, assessment, and game design envisioned in this phase of the project. The full set of the two knowledge specifications is detailed in Appendix A. A much larger ontology of concepts related to algebra, including those outlined in the full set of knowledge specifications was also developed and served to situate these knowledge specifications within a broader math education context.

From Knowledge to Item Specifications

Each knowledge specification was then used to generate one or more item specifications for each of two categories of cognitive demand. The first cognitive demand addressed procedural fluency with a particular concept, and the second was intended to ascertain how well a student could demonstrate conceptual understanding of the concept. These categorizations are based on the work of numerous researchers (Brown & Quinn, 2006; Carpenter et al., 1999; Fuson, 2003; Lamon, 1999; Mack, 1990; Saxe et al., 1999; Usiskin, 1988; Wu, 2001) and are aligned with the recommendations of the National Math Advisory Panel (U.S. Department of Education, 2008), the National Council of Teachers of Mathematics (NCTM, 2000), and with findings condensed in How People Learn Mathematics (Donovan & Bransford, 2005). While we originally intended to include problem solving as a separate category, this proved difficult since problems can easily become procedural tasks for students who have seen them before; therefore, they are difficult to classify unless the prior experience of a particular student is known. Furthermore, problem solving items can take extended periods of time to administer, and the time we had for testing students was strictly constrained. Consequently, we decided to limit our specifications to two types of items.

For each category of cognitive demand, we detailed what type of stimulus we would present to the students and then specified what the students should be able to do given such a stimulus. Essentially, the stimulus details the kind of prompt a student would receive, and the expected response gives some detail about what the student should be expected to do. Often, prompts for the procedural and the conceptual cognitive demands were similar or even identical; however, the expected responses were quite different for each type of cognitive demand. For procedural items, students were often asked to determine, show, identify, or label something, whereas items intended to assess conceptual understanding nearly always asked students to explain why something was the case.

Once we developed the item specifications, we used them to identify or develop specific items addressing each of the specifications and for each cognitive demand. Multiple items were generated for each specification, and these were edited for clarity and then reviewed by the research team for fidelity to the particular item specification and level of cognitive demand. Although such a process is a first step to assure the technical quality of an item and of the test it will be part of, it is arguably a very weak measure. Consequently, more empirically based methods were used to assure the technical quality of each item and the test forms.

Technical Quality of Assessment Items

As part of our validation effort, we assumed that student performance on items designed around interrelated key ideas should correlate (i.e. convergent validity). As suggested by researchers as far back as Cronbach and Meehl (1955), this should be true whenever a trait or construct is being measured, regardless of the probe being used to test that trait or construct, as long as each probe is intended to measure that particular trait or construct. In fact, the writings of Cronbach and others suggest that inferences from various measures which are designed to measure the same things should converge and all support the same inference (Campbell & Fiske, 1959). A cautionary note is important: The above should not be interpreted to mean that whenever assessment outcomes correlate, both assessments measure the same thing (Serci, 2009, in Lissitz). While such a phrase can be cited from the early testing literature (Guilford, 1946) and even in recent articles (Borsboom, Cramer, Kievit, Zand Scholten, & Franic, 2009), we feel such a belief is problematic for two reasons. First, there is always a chance that two outcomes will correlate by happenstance even though the two are not truly related. Second, and more important, two outcomes could be mediated by another variable and not be directly related at all (e.g., everyone who ate tomatoes in 1800 is dead so there must be a connection between eating tomatoes and human mortality). Consequently, it is important that the theoretical relationships between variables such as items and the construct of interest be specified before the items are empirically tested. We believe that the knowledge and item specifications help structure and detail such a relationship.

This notion leads to two similar but distinct ideas concerning item and assessment quality. First, the inferences about the existence of a particular construct or trait in a student made from data generated by the assessment items should be accurate (valid). If a particular trait is present, inferences about the construct based on responses to that item should support a conclusion that a student possesses that trait or construct. If the trait is not present, it should support the opposite inference. Second, the inferences made about whether or not the construct is present in a student should be consistent across items that supposedly measure the same construct, and that construct should be apparent across measurement occasions. In other words, if a test-taker responds to an

item multiple times within the same test or on different testing occasions, the inference about the presence of the construct should not change, given that the construct itself has not changed. The same should be true about multiple items that supposedly assess the same construct. Each item should provide the same evidence about whether the construct is present or not. In other words, the test should provide consistent (reliable) results.

Here again, it is important to caveat the preceding paragraph. Over 50 years ago, Campbell and Fiske (1959) demonstrated the importance that the *items* used to measure the trait or construct of interest not *all* be the same or even very similar. This point was emphasized again by Messick (1989). Once again, the process of developing various item specifications to measure similar concepts using various prompts and at varying levels of cognitive demand helps ensure that this type of item quality inheres in the process of item development. We test student understanding using a number of different pencil-and-paper items. In addition, we test students' ability to successfully complete the levels of a math video game that requires them to understand rational numbers and add rational numbers together. This video game was also developed from the same knowledge specifications used to develop the item specifications.

For the present study, we investigated a number of lines of evidence to determine that the developed assessments had sufficient technical quality to allow us to accurately and consistently infer that students understood the meaning of rational numbers and could add them. To that end, we investigated the validity (accuracy) of the inferences we wished to make about the students' understanding of rational numbers and rational number addition based on the assessment items we developed. In addition, we determined the consistency (reliability) of the items and forms that resulted from our development process.

We investigated two sources of evidence to evaluate the validity of the inferences we were making:

- **Evidence about the content of the assessments.** Specifically, did the assessment cover the breadth of the construct (understanding of rational number addition) and at sufficient depth to support our claim that it adequately measures this construct? Was the construct adequately represented?
- **Evidence from other activities requiring the same construct.** Specifically, did performance on the pretest (ostensibly a measure of student understanding of rational number addition) adequately predict performance on the game (a different measure of student understanding of the construct of rational number addition)?

We also calculated three different measures of reliability of the items and forms:

- **Interitem reliability.** This comparison answered the question of whether items on a form were consistently measuring the same construct, namely student understanding of rational number addition, in relationship to other items and to the test as a whole.

- **Item test-retest reliability.** This comparison determined whether students performed the same on identical items from pretest to the posttest, given that they had no intentional, intervening instruction on the assessed content.
- **Form test-retest reliability.** This comparison looked at overall score correlations for individual students on the pretest and the posttest to determine if both provided the same (i.e., a consistent) evaluation of student ability given that no intentional instruction was provided during the interim.

Evaluating the Evidence of Technical Quality

To ensure that the evidence we collected assessed the content at the depth and breadth necessary to draw valid conclusions, we generated key foundational ("big") ideas as previously described and developed or found items for each of the item specifications associated with one of the foundational ideas. These ideas were then used to create various levels in the game. This assured that the game and the assessments would be aligned. Consequently, it was hypothesized that a percent correct score on pretest items would predict the quality of game performance (as measured by the maximum level a student reached in game play). Since percentage correct on the pretest is a scale variable and success at game levels requiring these components of understanding is an ordinal variable, we used Spearman's rho to determine the degree and significance of the correlation between these measures.

Interitem reliability, item consistency from pretest to posttest, and form consistency between pretest and posttest, as a whole, were calculated using Cronbach's alpha. The correlation between the pretest (before game play) score and the posttest (after game play) score served as a measure of test-retest reliability since the students in this study played a version of the game with minimal math instruction (as described in the following subsection). We also computed the point-biserial correlation to determine the correlation between a dichotomous item score (a nominal variable) and total test score without the item (a scale variable).

Description of the Rational Number Video Game (PuppetMan) Task

In the rational number addition video game (called PuppetMan), students are presented with the challenge of bouncing a small sack-like doll over various hazards in order to get it safely to the other side. To do so, students place small trampolines at various fixed locations along a one- or two-dimensional grid. Each trampoline is made "bouncy" by dragging coils onto the trampoline. The distance each coil will cause PuppetMan to bounce is commensurate with its length. Therefore, if you add a coil of 1 unit to a trampoline, that trampoline will cause PuppetMan to bounce exactly one unit. In PuppetMan, one whole unit is always the distance between two lines. It is this unit that becomes the referent for coils of fractional bounce later on. Coils can be added to a trampoline to increase the distance PuppetMan will bounce; however,

only identical coils can be added together. While any size coil can be placed on the trampoline initially, subsequent coils can only be added to the trampoline if they are the same size. Initially, whole unit (integer) coils can be added one at a time, reinforcing the *meaning* of addition even with integers.

The game exploits the fact that real numbers can be broken into smaller, identical parts (decomposed), if necessary, to facilitate addition and to demonstrate that this process is similar in both integer and fractional addition. The intent is to make explicit connections between integer addition (with which many students have confidence) and fractional addition (with which many students struggle). Moreover, the game play requires that players (students) be attentive to the size of a unit they are adding. Unlike many previous games designed to teach mathematics, however, fluency with the basic ideas (the learning goals as specified in the knowledge specifications) in PuppetMan is integral, not ancillary, to game play.

As game play proceeds, the trampolines must be placed at distances along the grid that are fractional parts of the whole unit. Consequently, students are given a set of coils, then shown how to break coils into fractional units. Since only identical units can be added together (in agreement with the foundational idea of addition), students must be attentive to what the rational number means, to what units are being added, to what units are already on the trampoline, and to how they will break coils into different size pieces. So while students can add a coil that is a 1/2 unit to another coil that is also a 1/2 unit, they are not allowed to add a coil that is a 1/2 unit to a coil that is a whole unit until the whole unit is broken into two 1/2-unit coils. At the time all three of these coils are added to the trampoline, the trampoline will show that it has 3/2 (rather than 1 1/2) units of bounce (see Knowledge Specification 2.3.0 in Appendix A for further explanation). As noted previously, this is intended to reinforce both the meaning of addition and to reinforce the player's understanding of the meaning of rational numbers.

In the first two versions of PuppetMan, the procedure for converting fractions of different sizes (i.e., fractions with different denominators) is not accomplished through multiplication since that was beyond the specified learning goals (knowledge specifications) around which the game was designed. Rather, students were shown how they could use the mouse to click on a coil and then scroll up or down to break the coil into more pieces (each smaller in size) or fewer pieces (each larger in size), respectively. The fractional representation of the coil was shown alongside each coil as the student scrolled on the coil. For example, if a student clicked on a coil that was one whole unit in length and scrolled up, the coil broke into two halves, then three thirds, etc. If the student used the same procedure with a 1/2 coil, then the coil broke into two fourths, three sixths, etc. As long as students did not click somewhere else on the game, they could also scroll down on these same coils to make fewer pieces that were larger in size.

As shown in Appendix B, the grid representation was also used to convey the meaning and use of rational numbers. As mentioned previously, one whole unit was always the space between two red lines. In the one-dimensional game, the red lines denoting *unit* were vertical, and in the two-dimensional game these *unit* lines were both vertical and horizontal. Fractional parts of that distance were represented as the distance between green dots placed equidistantly between red lines along the grid.

Two versions of the game were developed to test the impact of design variations in type of instruction on math and game outcomes. The control condition was a game that provided instruction about the mechanics of playing the game. For example, this game condition taught students how to drag coils onto the trampolines and how to move the trampolines onto the grid. The control condition did provide small amounts of conceptual instruction because the math was so strongly integrated into the actual game play. In some cases, such as scrolling the coils, explaining game mechanics was synonymous with representing a basic math concept. This instruction was minimized in the control condition as much as possible and was intended to teach students how to play the game rather than to increase understanding of how a unit is defined, of rational numbers and their relationship to that unit, or of addition. The second version of the game—the treatment—contained a great deal more instruction and help for the students. It helped students, for example, understand the meaning of the denominator and its importance in choosing the correct size pieces to add, the numerator's importance for choosing the number of those pieces to combine, and the meaning of addition as related to both integer and non-integer rational numbers. In this way, the treatment and control conditions were very different.

The two versions of the game allow us the opportunity to test various hypotheses. For example, by comparing the treatment and control conditions, we can determine if students learned better with feedback designed to improve understanding of rational number addition than without such feedback. More important to this paper, however, is that we can use the control (no significant math instruction) condition to empirically test the quality of the measures (pretest, game, and posttest) of each student's understanding of rational number addition.

The Sample

In the present study, four samples of students were drawn from a high school summer school population. Our purpose was to measure student understanding of rational number addition. The students were a convenience sample based on subject availability at the time the first prototype of the game was completed. They were drawn from groups of students enrolled in summer school at two large, public high schools in a southern California school district. The

total sample consisted of 186 students. Each of these students took a pretest, played one of the two versions of the game, and took the posttest.

The students were enrolled in summer school for various reasons, which contributed to the diversity of the sample. The first group in the sample consisted of students trying to pass a high-stakes state test to prove proficiency in Introductory Algebra (n = 49). The second group of students were studying to retake a different high-stakes test—the California High School Exit Exam (CAHSEE)—on which a passing score is necessary to earn a high school diploma (n = 27). Another group of students were taking a keyboard class so that they could take more desirable elective classes the following year (n = 49). The final group was composed of students (n = 61) who were taking a course to make up for less-than-proficient performance in their English and Language Arts class the previous school year. Consequently, the sample contained students likely to have varying math abilities and, more importantly, likely to have disparate understandings of rational numbers.

After the pretest, the students were randomly assigned to one of the two versions of the game. Students in both groups were then given 30 to 50 minutes to play the game (depending on the class schedule at the local school). Unfortunately, we encountered technical problems with the student log files in the Introductory Algebra group that prevented us from determining the game level they had achieved. Consequently, the sample size for this aspect of the study (comparing pretest results to game level achieved) was reduced to 137 students.

All students then took the posttest. While all items on the posttest were developed using the procedure for developing knowledge and item specs (as described in the methods section), not all items exactly duplicated items that were on the pretest. Several new items were introduced on the posttest for various reasons. First, we added items to the posttest that directly related to the game environment so that we could determine if the format of these items made them easier or harder to answer than items that were presented using a more typical math symbology. In part, this was an effort to ensure that the posttest would consist of at least some items that were sensitive to the game environment and to see if those items detected learning that more traditional items did not. These items also allowed us to assess knowledge of rational number addition using a different format as Campbell and Fiske (1959) suggest.

For various parts of the present study, we narrowed the sample size even further. Since our goal in much of the present study was to assure that inferences from the pretest are highly correlated with inferences from the game and with inferences from the posttest, we excluded students who received instruction that was intended to improve their mathematical understanding and performance on the game (i.e., students in the treatment condition). While the students in the

remaining (control) condition sample did receive instruction on how to play the game, providing such instruction to this group was necessary in order to control for mediating factors unrelated to the construct of interest (e.g., understanding the interface in order to play the game versus how to use rational number addition to play the game). Furthermore, providing instruction minimized criterion-irrelevant variance (e.g., students who could determine how to use the mouse scroll wheel to produce parts of coils on their own from those who could not) in game play. Of the original sample, then, 68 students remained—14 students in the CAHSEE class, 23 students in the Keyboarding class, and 31 students in the English class.

Finally, we limited our analysis (where applicable) to students who answered all items on both the pretest and the posttest. Several students either made no attempt to complete any items on a test, or they left parts of the test completely blank. In many cases, it was difficult to determine if the student did not know an answer or if they just decided not to write an answer to a particular question even though they were asked to write "don't know" on items that they did not know the answer to. Consequently, we made the decision to analyze only those students who attempted every answer. Combined with the previous restrictions, this left 58 students in our sample of control students who had complete pre-tests and data on their level in the game after 30 minutes of playing the game. Of these students, 7 in the CAHSEE class, 17 in the Keyboarding class, and 22 in the English class (a total of 46) had complete posttests.

Determining the Technical Quality of the Game and Other Measures

The first step in determining the technical quality of the pre- and posttest items was to evaluate the accuracy of the inferences made from the test questions. We hypothesized that a student's pretest score should predict game performance as determined by how far in the game a student could get. Obviously, such an analysis presupposes that the students of interest are not exposed to actual math instruction during the game. We contend that since both of these measures were designed to assess the construct described by the knowledge specifications, a strong correlation between the pretest and game level would be evidence that inferences from each are valid (accurate) indicators of student understanding of rational number addition. The breadth of such a claim would be determined by an analysis of the content actually assessed and used in the game. This would be determined by a content analysis and alignment study.

Based on previous literature indicating that female students in secondary school are less likely to demonstrate proficiency in math and are less likely to play video games, we also tested the inferences made from the pretest, the posttest, and the game for gender bias. Moreover, since it seems logical that prior experience playing video games could have a significant effect on student performance in subsequent game play (including PuppetMan), we expected this variable

to also be correlated with game performance. Consequently, the correlation between video game experience and performance in PuppetMan (and the interaction with gender) was also evaluated to ascertain their effects as mediators or moderators of game achievement.

Next, we analyzed how well the items together measure the same construct (the interitem reliability of the pre- and the posttest forms) and how well each item predicts its respective overall test score (the point-biserial correlation). A high interitem reliability coefficient suggests that the items, as a whole, are measuring the same construct. This would be expected if the items are measuring the knowledge specified and if the knowledge specifications adhere together as expected. A high point-biserial correlation, on the other hand, suggests that items are functioning as expected— students that score higher on the overall test are more likely to answer a question correctly than students who score lower on the test. If students with higher overall test scores are answering individual items *incorrectly* with greater frequency than students with lower test scores, this could be an indicator that something in the item itself (e.g., a diagram or something else in the item's presentation), rather than the construct being measured, could be causing students to miss the item. Alternatively, when students who perform lower on the test overall get an item correct more often than would be expected based on overall test score, this could be an indication that something in the item (e.g., wording, etc.) rather than understanding the intended construct might be the source of student success.

Although Scott (1960) suggested that interitem and point-biserial correlation are mathematically synonymous, we used both to measure how each item contributes to the overall score and to ensure that each individual item is behaving as anticipated (SPSS, 2000). Moreover, since we and others have seen high interitem reliability on tests that contain items with low, or even negative, point-biserial correlations (Sturmey, Matson, and Sevin, 1992; Vendlinski and Phelan, 2009), we analyze both of these statistics. Finally, we evaluate the interitem correlation coefficient if each item were removed from the test. This analysis provides us a measure of which item(s) might be removed if either reliability needed to be improved or the test needed to be shortened and reliability maintained. Such a statistic can identify items that might be removed even if each of the item's other parameters described above are acceptable.

To complete our analysis of item consistency, we also investigated the correlation between how an item functions on the pretest and how the same item functions on the posttest, that is after the student has a reasonable time to forget their initial response to the item. The correlation between these items should be high if the items are consistently measuring the same construct. A pitfall with such a procedure, however, is that students might learn from the pretest itself and, as a result, answer items on the posttest more correctly than they did previously. Consequently, we evaluated our data to isolate such learning effects.

Results

Content Coverage

Table 1 shows the coverage of the first two knowledge specifications after the items were drafted, reviewed, selected, and edited for test use. Additional items (approximately twice as many items as needed) were developed from the test and item specifications (shown in Appendix A) but were not initially used.

Three things are immediately evident in Table 1. First, the vast majority of items on both the pretest (80%) and on the posttest (69%) assess more than one, single knowledge specification, even though the knowledge specification process focused on writing or finding items that addressed a particular knowledge specification. For example, Test Item 10 addressed Knowledge Specifications 1.3.2, 1.3.3, 1.3.4, and 2.1.3. Nevertheless, unlike Item 10, most items do seem to address only one of the two key ideas. On the pretest, 75% of the items address only the first or only the second key idea (out of 20 items, 6 address the first key idea, 9 address the second key concept, and 5 address both); on the posttest, approximately 86% of the items draw on knowledge from a single key idea (out of 36 items, 14 address the first key concept, 17 address the second key idea, and 5 address both). So while most items may be assessing the key ideas broadly, most are doing so within a defined conceptual area rather than across the two key conceptual areas.

A second point to note from Table 1 is that the items broadly test most content across all game levels. In other words, while the game was built to use progressively more challenging concepts—integer addition, rational number addition with common denominators, and rational number addition with dissimilar denominators—in one and then in two dimensions, the test seems to measure the concepts necessary for success on these concepts broadly and across levels. Consequently, specific knowledge specifications are not associated with specific levels of the game, and specific items do not test specific game levels. In fact, with the exception of levels designed largely to teach game mechanics (Levels 1, 2, 3, 10, 11, and 12), each level uses both key conceptual ideas.

A final point to note from Table 1 is that some specific concepts are not being addressed by the assessments. In particular, the following Knowledge Specifications are addressed by the game, but are not assessed by any of the assessment items:

- 1.2.0—In mathematics, one unit is understood to be one of some quantity (intervals, areas, volumes, etc.).

- 2.1.2—Positive integers can be broken (decomposed) into parts that are each one unit in quantity. These single (identical) units can be added to create a single numerical sum).
- 2.3.0—Dissimilar quantities can be represented as an expression or using some other characterization, but are not typically expressed as a single sum.[1]

Moreover, the Knowledge Specification that addresses additive identity (Specification 2.4.0) and the Knowledge Specification that addresses the concepts of positive numbers, negative numbers, and the additive inverse (Specifications 2.5.0 through 2.7.0) are not addressed at all (either by the assessments or in the game).

Table 1

Comparison of Knowledge Specifications to Game Level, Pretest Items, and Posttest Items

Game Level	Knowledge Specs.	Items on Pretest (Form 1)	Items on Pretest (Form 2)	Items on Posttest
1 – One jump of one whole unit with one whole-unit coil	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
2 – One jump of two whole units with two whole-unit coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.1	19	19	19
	1.3.2	10	10	10, 22
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.1.2	Not assessed	Not assessed	Not assessed
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
3 – Two jumps of one whole unit with one whole-unit coil each	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
4 – Two jumps of two whole units	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22

[1] We are considering numbers like 2¾ to have an implied addition. In other words, 2¾ = 2 + ¾, whereas 11/4 is a single sum.

Game Level	Knowledge Specs.	Items on Pretest (Form 1)	Items on Pretest (Form 2)	Items on Posttest
with two whole-unit coils each	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.1	19	19	19
	1.3.2	10	10	10, 22
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.1.2	Not assessed	Not assessed	Not assessed
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
5 – One jump of one whole unit jumps with two 1/2 unit coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
6 – One jump of 1 1/2 units with one whole-unit coil and one 1/2 unit coil	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
	2.3.0	Not assessed	Not assessed	Not assessed
7 – Two jumps (one of 1 1/2 units and one of 1/2 unit) with two whole coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22

Game Level	Knowledge Specs.	Items on Pretest (Form 1)	Items on Pretest (Form 2)	Items on Posttest
	1.3.1	19	19	19
	1.3.2	10	10	10, 22
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.1.2	Not assessed	Not assessed	Not assessed
	2.1.3	10, 14, 15	10, 13, 15	10, 13, 14, 15, 24
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
	2.3.0	Not assessed	Not assessed	Not assessed

Game Level	Knowledge Specs.	Items on Pretest (Form 1)	Items on Pretest (Form 2)	Items on Posttest
8 – Two jumps (one of one whole unit and one of 2/3 unit) with 1 whole and three 1/3 coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
9 – Two jumps (one of 4/6 unit and one of 2/6 units) with one 1/2 and three 1/3 coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.1	19	19	19
	1.3.2	10	10	10, 22
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.1.2	Not assessed	Not assessed	Not assessed
	2.1.3	10, 14, 15	10, 13, 15	10, 13, 14, 15, 24
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
	2.3.0	Not assessed	Not assessed	Not assessed
10 – Two-dimensional grid. Whole units and whole coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
11 – Two-dimensional grid. Whole units and whole coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
12 – Two-dimensional grid.	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22

Game Level	Knowledge Specs.	Items on Pretest (Form 1)	Items on Pretest (Form 2)	Items on Posttest
1/2 units and whole coils	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.2	10	10	10
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
13 – Two-dimensional grid. 1/2 units and whole and 1/3 coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.1	19	19	19
	1.3.2	10	10	10, 22
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.1.2	Not assessed	Not assessed	Not assessed
	2.1.3	10, 14, 15	10, 13, 15	10, 13, 14, 15, 24
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
	2.3.0	Not assessed	Not assessed	Not assessed
14 – Two-dimensional grid. 1/3 units and whole and 1/3 coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.1	19	19	19
	1.3.2	10	10	10, 22
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.1.2	Not assessed	Not assessed	Not assessed
	2.1.3	10, 14, 15	10, 13, 15	10, 13, 14, 15, 24
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
	2.3.0	Not assessed	Not assessed	Not assessed

Game Level	Knowledge Specs.	Items on Pretest (Form 1)	Items on Pretest (Form 2)	Items on Posttest
15 – Two-dimensional grid. 1/3 units and whole, 1/2, 1/3, 1/4 and 1/5 coils	1.1.0	2, 8, 16, 17, 19, 20, 21	2, 9, 16, 17, 19, 20, 21	2, 8, 9, 16, 17, 19, 20, 21, 22
	1.2.0	Not assessed	Not assessed	Not assessed
	1.3.0	20	20	20, 22
	1.3.1	19	19	19
	1.3.2	10	10	10, 22
	1.3.3	10, 21	10, 21	10, 21, 22, 25, 27, 29, 30
	1.3.4	10	10	10, 22, 26, 28
	2.1.0	11	12	11, 12, 23
	2.1.1	11	12	11, 12, 23
	2.1.2	Not assessed	Not assessed	Not assessed
	2.1.3	10, 14, 15	10, 13, 15	10, 13, 14, 15, 24
	2.2.0	11, 14, 15, 19	12, 13, 15, 19	11, 12, 13, 14, 15, 19, 23, 24
	2.3.0	Not assessed	Not assessed	Not assessed

Note. Missing numbers are for items developed but not used on the tests.

Relationship between Tasks Requiring the Same Content Knowledge

To ensure that the pretest is accurately reflecting student knowledge, we confirmed that students scored similarly on tasks requiring the same content knowledge. As was seen in Table 1, although most levels required the integration of the two key conceptual ideas, the levels became increasingly more difficult in one dimension and then increasingly more difficult in two dimensions. In particular, after being asked to add integers, students were asked to add fractions of the same size, and then fractions and integers. Ultimately, students had to add dissimilar fractions. As expected, scores on the pretest correlated with the level a student completed in the non-instructional version of the game. In particular, we noted a significant correlation between pretest score and game level achieved after 30 minutes of play (rho = .691, $p < .001$). Similar significant correlations were noted between a sum score of the items that just measured concepts from the first knowledge specification (Items 2, 8, 9, 16, 17, 20, and 21) and game level achieved at 30 minutes of play (rho = 0.580, $p < 0.001$) and a sum score of the pretest items that just measured concepts from the second knowledge specification (Items 11, 12, 13, 14, and 15) and game level achieved at 30 minutes of play (rho = 0.583, $p < 0.001$). The strongest correlation was between the sum score of the two items that were expected to measure both knowledge specification (Items 10 and 19) and game level achieved after 30 minutes of play (rho = 0.608,

$p < 0.001$). As might be expected from previous discussions and from the correlations just reported, the sum score of the items that just measure concepts from the first knowledge specification and the sum score of the items that just measure the second knowledge specification are also significantly correlated as (rho = 0.570, $p < 0.001$).

We recorded the maximum level each student had attained in the game at ten-minute intervals ranging from 10 minutes to 90 minutes. Because of school schedules, however, most students only had approximately 40 minutes to play the game. While all the students were given about the same time to play the game, slight variations (up to 10 minutes) in play time occurred because of variations in school site class periods. Consequently, not every student had the full 40 minutes to play. Students were also told that both the game and assessments were no-stakes. Consequently, most students played at the speed with which they felt comfortable or that allowed them to accomplish each level. With few exceptions, the maximum level the student achieved at the end of 30 minutes was predictive of (if not identical to) the maximum level the student had at 40 minutes, at 50 minutes, etc., if they were allowed to play that long. The correlation between game level reached at 10 minutes and each subsequent 10-minute interval is shown in Table 2.

Table 2

Correlation Between Levels Reached after 10 Minutes and Subsequent 10-Minute Intervals

	10 minutes	20 minutes	30 minutes	40 minutes
20 minutes	0.786			
30 minutes	0.771	0.917		
40 minutes	0.771	0.898	0.954	
50 minutes	0.773	0.895	0.948	0.995

Note. All correlations are significant at the $p < 0.001$ level

The correlations in Table 2 suggest that using the game level achieved at 30 minutes rather than at 40 or 50 minutes is unlikely to dramatically alter the results reported above since the correlation between game levels achieved at 30 minutes and at a later time is very high.

As noted above, we also decided to use only data from students in the control condition that completed all items on the pretest. Fifty-eight (58) students randomly assigned to the control condition completed the pretest. Nine (9) students in the control condition left one or more items blank on the pretest. We analyzed game performance for both of these groups to see if game level was significantly different for students included (i.e., those who completed all items) versus students excluded from the sample (i.e., those who left one or more test items blank). As shown

in Table 3, these two groups did not differ significantly in their game performance (t[11] = -1.51, p = 0.159). Consequently, the decision to exclude participants who did not fully complete the pre-test from our analysis is not expected to influence the results reported below.

Table 3

Mean Game Level Achieved after 30 Minutes by Students Who Did and Who Did Not Complete All Pre- and Posttest Items

	N	Mean	*SD*
Students who DID NOT complete all items on the pre- test	9	12.11	2.98
Students who DID complete all items on the pre-test	58	13.74	3.29

Given that percent score on the pretest was a scale variable and game level was ordinal in nature, we used Spearman's rho to determine the degree and significance of the correlation between these measures.

While the correlation between pretest and game level achieved after 30 minutes of play explains roughly half (48%) of the variance in the game level a student achieved[2], we expect that other variables could be moderating this result, as previously hypothesized. In particular, since the task in which a student must demonstrate knowledge of rational number addition was playing a video game, we hypothesized that the amount of experience a student had with video games in general, how much they actually played video games, and the student's self-reported ability in playing video games, in addition to math ability, might all account for variability in success in our video game task. In essence, such variables would account for the variability in game success that was attributable to video game play and not precisely to math ability. As expected, these variables all show a significant correlation with the highest game level a student achieved after 30 minutes of play. The amount of time a student played video games each week had the greatest correlation with the game level a student achieved (rho = 0.521, $p < 0.001$); student self-perceived game play ability (rho = 0.404, p = 0.002) was also strongly correlated with getting farther in PuppetMan. While still significant, the correlation between the number of years a student reported playing video games and level achieved after 30 minutes (rho = 0.345, p = 0.008) was not as strong as the two other measures of video game experience.

Given the foundational nature of both addition and rational numbers (i.e., the Knowledge Specifications), we also expected results on the pretest to be correlated with general math

[2] This relationship describes students playing the game version with no significant embedded math instruction.

performance and student self-perception of general math ability. While human subjects restrictions prevented us from obtaining actual student math grades, we did ask students in the control condition to report their overall grades in math since sixth grade (using the scale "Mostly A's," "Mostly B's," etc.) as well as their perception of their own mathematical ability. Once again, as expected, student scores on the pretest correlated with both self-reported overall math grades (rho = 0.489, $p < 0.001$) and student perception of math ability (rho = 0.477, $p < .001$). Each of these measures was also correlated with the highest level a student reached in the game after playing for 30 minutes (rho = 0.343, $p = 0.009$ for self reported math grades and rho = 0.394, $p = 0.002$ for perceived math ability).

Since the literature suggests significant differences in the amount of video game play between girls and boys and knowing that such differences, rather than differences in conceptual knowledge, might affect the level a particular student achieved in the game, we wanted to test this source of construct irrelevant variance. As expected, gender (male = 0 and female = 1) is correlated with self-reported game play ability (rho = 0.427, $p = 0.001$), with self-reported total years of game play (rho = 0.297, $p = 0.025$), and with self-reported weekly amount of game play (rho = 0.363, $p = 0.006$). Given that each of these game play variables is also significantly correlated with the level a student reached in the game, it seems that gender might be a mediating factor in predicting game level outcome. In fact, the highest game level achieved is significantly correlated with gender (rho = 0.308, $p = 0.021$). Neither pretest score nor self-reported math grades, however, are significantly correlated with gender.

Based on these results, we constructed a linear regression model to determine which of the variables (percent score on the pretest, weekly game play, or gender) best accounted for the differences in the game level a student ultimately achieved in PuppetMan. We began with a base model that included only pretest score and added in the other components in terms of their significance in predicting level in game after 30 minutes. Interaction terms were added after component terms in each case. As can be seen in Table 4, after the effects of pretest and weekly game play are included in the model, the effects of gender are no longer significant. The interaction term between pretest percentage score and amount of weekly game play were insignificant, and so they were dropped from the model before the gender term was added. In this model, pretest and weekly game play together explain over 64% of the variance in game performance as measured by level achieved after 30 minutes of play[3].

[3] This relationship describes students playing the game version with no significant embedded math instruction.

Table 4

Regression Model of Pretest and Weekly Game Play to Predict Level of Performance in Game (Control Group)

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	0.709[a]	0.503	0.493	2.35369	0.503	54.564	1	54	0.000
2	0.801[b]	0.641	0.628	2.01701	0.139	20.532	1	53	0.000
3	0.811[c]	0.658	0.638	1.98828	0.017	2.543	1	52	0.117

Note. [a]Predictors: (Constant), Pretest Percent (Complete). [b]Predictors: (Constant), Pretest Percent (Complete), GAMEPLAY_WEEKLY/Amount of weekly video game play. [c]Predictors: (Constant), Pretest Percent (Complete), GAMEPLAY_WEEKLY/Amount of weekly video game play, BKGD_GENDER/Gender.

Relationship between Items on a Particular Test Form (Interitem Reliability)

A high degree of interitem reliability is some assurance that all the items on a test are measuring the same construct (Cronbach & Meehl, 1955). While the Knowledge Specifications detail two key concepts underlying student performance on this test, the analysis in Table 1 suggests that these concepts are related in the present task. In particular, it suggests both are central to the concept of adding rational numbers.

Various forms of the pretest and the posttest were used during the initial study. We report on each of these in order. All 27 CAHSEE students, 24 of the Keyboarding students, and 31 of the English students took Form 1 of the pretest. The interitem reliability was very high for this form ($\alpha = 0.918$). Furthermore, as can be seen in Table 5, the point-biserial correlations were generally high (mean point-biserial is 0.541). A high point-biserial correlation suggests that performing well on an item is strongly correlated with scoring well on the overall test without that item included and is an indication of individual item quality within a test. Note that some items in Table 5 required students to supply a numerator and a denominator. Each of these items has three entries in the table: an overall, dichotomous item score; a dichotomous numerator score (item number with "N" suffix); and a dichotomous denominator score (item number with "D" suffix). In addition, Item 10 required students to properly order four fractions on the number line. The suffixes of these items represent their order from lowest (A) to highest (D) value.

Table 5

Item Statistics by Item for Items Appearing on Pretest Form 1

Item	Point-biserial correlation	Cronbach's alpha if item deleted
2	0.433	0.915
8	-0.167	0.921
10A	0.724	0.908
10B	0.656	0.909
10C	0.637	0.910
10D	0.696	0.908
11	0.478	0.914
11N	0.140	0.918
11D	0.478	0.914
14	0.729	0.908
14N	0.748	0.907
14D	0.808	0.906
15	0.737	0.908
15N	0.780	0.907
15D	0.748	0.907
16	0.139	0.918
17	0.681	0.909
19	0.403	0.916
20	0.429	0.915
21	0.548	0.912

Only one item displayed poor quality. The point-biserial for Item 8 suggests that students who do poorly on the overall test are more likely to get this item correct than students who do well on the test. In addition, the alpha statistic suggests that deleting Item 8 would result in a better correlation among the items. A further analysis suggests that this item was very easy for students. Approximately 95% of the students who attempted this item on the pretest answered the item correctly. Two other items (11N and 16) had marginal point-biserial correlation coefficients, but the overall form reliability was not estimated to change appreciably if these items were removed.

Form 2 of the pretest was only used on the second day of testing and was taken by 25 of the Keyboarding students and 30 of the students in the English and Language Arts review class. Once again, the form displayed very high interitem reliability ($\alpha = 0.909$). In addition, as can be

seen in Table 6, the point-biserial correlations were again generally high (mean point-biserial is 0.532). Note that the same item coding scheme was used in this table as was done in Table 5.

Table 6

Item Statistics by Item for Items Appearing on Pretest Form 2

Item	Point-biserial correlation	Cronbach's alpha if item deleted
2	0.523	0.905
9	0.484	0.906
10A	0.591	0.903
10B	0.531	0.905
10C	0.741	0.899
10D	0.785	0.898
12	0.170	0.911
12N	0.123	0.912
12D	0.250	0.910
13	0.800	0.898
13N	0.800	0.898
13D	0.650	0.902
15	0.740	0.899
15N	0.740	0.899
15D	0.685	0.901
16	0.215	0.910
17	0.530	0.905
19	0.400	0.908
20	0.425	0.907
21	0.464	0.907

Every item in this test displayed acceptable item quality. Four items (12, 12N, 12D, and 16) had marginal point-biserial correlations, and the alpha coefficients suggest that the interitem correlation for the test as a whole would marginally improve if these items were dropped. As an aside, it should also be noted that Item 12 was the only rational number addition item that resulted in an improper fraction (i.e., the numerator is larger than the denominator).

The posttest was given to each participant on the same day after each completed the pretest and then had a period of dedicated game play. The same posttest was used for each of the three groups and was composed largely of items on the pretest. As can be seen in Table 7, and as

previously discussed, the posttest included all the items from both pretests, as well as nine items (Questions 22–30) that asked questions in the context of the game. As with the pretests, the form displayed very high interitem reliability ($\alpha = 0.940$). In addition, as can be seen in Table 7, the point-biserial correlations were again generally high (mean point-biserial is 0.515). Note that the same item coding scheme was used in this table as was done in Table 5 and Table 6.

Table 7

Item Statistics by Item for Items Appearing on Posttest

Item	Point-biserial correlation	Cronbach's alpha if item deleted
2	0.509	0.939
8	0.017	0.942
9	0.505	0.939
10A	0.557	0.938
10B	0.558	0.938
10C	0.649	0.937
10D	0.643	0.937
11	0.544	0.938
11N	0.130	0.941
11D	0.544	0.938
12	0.476	0.939
12N	0.060	0.942
12D	0.571	0.938
13	0.782	0.936
13N	0.741	0.936
13D	0.728	0.937
14	0.822	0.936
14N	0.760	0.936
14D	0.759	0.936
15	0.749	0.936
15N	0.802	0.936
15D	0.731	0.937
16	0.252	0.940
17	0.534	0.938
19	0.405	0.940
20	0.479	0.939
21	0.502	0.939
22	0.555	0.938
23	0.491	0.939
24	0.687	0.937
25	0.521	0.939
26	0.501	0.939
27	0.164	0.941
28	0.174	0.941
29	0.208	0.941
30	0.435	0.939

As might be suspected from the pretest results, item quality was overwhelmingly good on the posttest. Only one item displayed poor quality. Once again, the point-biserial for Item 8 suggests that students who do poorly on the overall test are more likely to get this item correct than students who do well on the test. In addition, the alpha statistic suggests that deleting Item 8 would result in a better correlation among the items. Once again, however, these statistics are undoubtedly affected by the fact that over 91% of the students who attempted this item answered it correctly. Three other items (11N, 12N, 16) had marginal point-biserial correlation coefficients both on their respective pretests and on the posttest and the Cronbach's alpha calculations suggest that the interitem reliability coefficient would improve if these items were dropped. Finally, three of the nine posttest items that included the game context had marginal point-biserial coefficients. Here again, the Cronbach's alpha calculations suggest the overall interitem correlation would be improved slightly if these items were deleted from the posttest. Given the high overall interitem reliability, these items could be dropped from the test without compromising reliability; however, new items to test Concepts 2.1.0 and 2.1.1 would need to be substituted for these items since, at present, these concepts are only tested by those items.

Item Test-Retest Reliability

Determining whether students perform the same on identical items from one test to the next, given that they had no intentional, intervening instruction on the assessed content is a further assurance that the items are measuring one or more constructs consistently (reliably). Students were given identical items on the pretest and the posttest. In some cases, all students had items on both the pre- and posttest (Items 2, 10, 15, 16, 17, 19, 20, and 21). In other cases, some students saw items on both the pre- and the posttest, while other students only saw the items on the posttest. For example, students taking Form 1 of the pre-test saw Items 8, 11, and 14 on the pre- and posttest while students taking Form 2 of the pre-test only saw these items on the posttest. Students taking Form 2, on the other hand, saw Items 9, 12, and 13 on both tests, while students taking Form 1 only saw these items on the posttest.

A chi-square analysis was completed for each of the pretest/posttest question pairs for the students who received the non-instructional version of the game. In every case, this analysis indicated that a student's outcome on the pretest version of an item was a significant predictor of the student's outcome on the identical posttest measure. Taken as a whole, the correlation between items that appear on both the pretest and the posttest for this entire group of students was significant ($\alpha = 0.940$, $p < 0.001$). The correlation for items that appeared only on Form 1 of the pretest and the posttest and the items that appeared only on Form 2 of the pretest and the posttest were also significant ($\alpha = 0.867$, $p < 0.001$ and $\alpha = 0.827$, $p < 0.001$, respectively).

Item Test-Retest Effects

Our study design also allowed us to measure the effect of students seeing an item for the first time on the pretest and then seeing an item again (on the posttest) versus just seeing an item for the first time on the posttest. We investigated whether students learn from and perform better on an item just because they have been asked to answer it before. To ascertain these differences, we explored how students taking Form 1 of the pretest differed in their posttest responses to Items 8, 11, and 14 from their peers who took Form 2 of the pretest and who, therefore, had not seen these items before. Similarly we explored how students taking Form 2 of the pretest differed in their posttest responses to Items 9, 12, and 13 from their peers who took Form 1 of the pretest and who, therefore, had not seen these items before. Finally, we explored how all the students scored on items that appeared both on the pretest and on the posttest (Items 2, 10, 15, 16, 17, 19, 20, and 21) by comparing mean total score on these pretest items to mean total score on the same posttest items.

In both cases, there were no significant differences between scores on the pretest and identical posttest items[4]. The 52 students who took Form 1 of the pretest (containing Items 8, 11, and 14) had a mean score of 4.808 for these items on the pretest and a mean score of 4.79 for these items on the posttest. The 53 students who took Form 2 of the pretest (containing Items 9, 12, and 13) had a mean score of 5.13 for these items on the pretest and a mean score of 5.02 for these items on the posttest. Note that since two of the items had three parts each, the maximum score for the three items is 7 in each case. Neither of these differences is significant, however. As might be expected, student outcomes on these items on both tests are significantly correlated. Student performance on Items 8, 11, and 13 (Form 1 of the pretest) strongly predicts performance on those identical items on the posttest ($\alpha = 0.816$, $p < 0.001$), while student performance on Items 9, 12, and 13 (Form 2 of the pretest) strongly predicts performance on those identical items on the posttest ($\alpha = 0.874$, $p < 0.001$). There were also no significant differences between mean total scores on items that students saw on both the pre- and on the posttest (Items 2, 10, 15, 16, 17, 19, 20, and 21). On average, students scored 6.9 on these items on the pretest and 7.1 on these items on the posttest. The slight increase in average score was not enough to be significant ($t[66] = 1.117$, $p = 0.268$)

Form Test-Retest Reliability

Our final analysis looked at the consistency of the pretest and posttest measures of student ability to master these knowledge specifications. In this comparison, we looked at overall

[4] This relationship describes students playing the game version with no significant embedded math instruction between pre- and post-test.

percentage scores on the pretest and on the posttest for individual students playing the non-instructional game to determine if both tests provided similar evaluations of student ability. As expected, the correlation between a student's percentage correct score on the pretest and that student's percentage correct score on the posttest was very strong and significant ($\alpha = 0.903$, $p < .001$).[5]

Conclusions

We set out to determine if an integrated method of building assessment and instruction (in this case, an instructional video game) would produce assessments and instruction of high technical quality. This paper documents that process and reports on the technical quality of the assessments ultimately used to evaluate student learning.

The broad conclusion is that the process worked well to structure the process and generate assessments of outstanding technical quality. Arguably, the assessments demonstrated both weaknesses and strengths. In particular, the two weaknesses we identified concerned content coverage and an item with low discrimination. It was evident that, while certain items measured one of the two key ideas, a subset of items could not be used to assess the presence of a particular construct *within* either of those ideas. This was somewhat surprising as particular items were generated to test each particular construct. In addition, there were three cases (regarding Knowledge Specifications 1.2.0, 2.1.2, and 2.3.0) where a student's grasp of a concept was only addressed in the game. There were four cases (regarding Knowledge Specifications 2.4.0–2.7.0) where the knowledge specification was not tested at all. This suggests that it is not possible to make statements about student ability regarding these concepts. Finally, one item (Item 8) demonstrated low point-biserial correlations with the larger test score. Although these appear to slightly weaken the test reliability, they should not be considered a flaw in the process. In fact, the process suggested these concerns might arise during item development.

The knowledge specifications researchers actually tested were limited for two reasons. First, the initial version of the game was not designed to address certain knowledge specifications (2.4.0–2.7.0) because the game was developed in an incremental nature. For example, while negative numbers were addressed in the knowledge specifications, the initial version of the game did not allow for the use of negative numbers. Subsequent versions of the game were designed to include this topic. In this way, complete knowledge specifications provided an overview of the full featured game and provided a road map for game development even though the initial version of the game did not allow for the use of such knowledge. Consequently, the assessments were limited so that they would align to the game by assessing

[5] This relationship describes students playing the game version with no significant embedded math instruction.

only the knowledge a player had the opportunity to use and demonstrate in the game. The procedure described allows both for subsequent game development (i.e., provides a road map for future development of the game and assessments) and ensures alignment between the tasks students are asked to perform in both the game and on the assessments. In this way, the specifications performed their role exactly as intended.

The second reason that not all knowledge specifications were covered by the assessments involved classroom time constraints. Because the time available for testing and game play was limited to a single class period, researchers had to allow adequate time for each activity. Furthermore, researchers did not want either the pre- or posttest to become a timed (speeded) event, and researchers made every attempt not to fatigue students with pretesting prior to game play so as not to adversely affect game performance. This decision mandated a similar restriction on items tested on the posttest so that the pre- and posttests would be largely parallel. Consequently, while the game used the decomposition and addition of integers to teach students the mechanics of the game, items assessing these concepts were not used on either test. For this reason, it could be argued that the assessments and game were not completely aligned and that the assessments did not completely sample the conceptual areas of interest. Nevertheless, it should be noted that our process for creating and selecting items identified these shortcomings prior to testing and that researchers made an informed decision to drop those items prior to field testing. As such, the exclusion of the items testing these concepts was intentional rather than an oversight. Moreover, given that items were designed to measure these concepts as part of the development process, the assessments could easily be modified to include such items if field testing warranted. By linking the specifications with the game, it became easy to both prioritize items for elimination and to eliminate them in a logical manner. Here again, the process performed exactly as intended.

With one exception, the process also generated items that had outstanding technical quality. Not only did performance on the pretest predict subsequent performance on a dissimilar task designed from the same specifications, it also accounted for the largest part of variability in that performance. Moreover, the process of item development yielded test forms (both pre- and post-) that demonstrated high interitem reliability (both above 0.9) and, overwhelmingly, items that were strongly correlated with overall game (task) performance. The items and forms also demonstrated strong test-retest reliability, while showing little evidence that students learned the concepts being assessed merely by being exposed to the items tested on the pre-test.

A limitation of this research is that the only items actually researched with students were items that tested understanding of the two key ideas at a procedural level. While items were developed to test student conceptual understanding, they were not used with the student

population identified in this paper. In part, this was because such items require more time to take and more time to score.

Further research continues on expanding the use of items testing conceptual understanding, as well as expanding the concepts tested. Currently, the PuppetMan game and assessments have been expanded to test the knowledge specifications not included in this first field trial. In addition, this development process has been applied to the domain of solving equations. New knowledge specifications have again driven game and assessment development in that domain.

Together, these experiences and the research cited here suggest that developing knowledge specifications that drive the development of item specifications and assessment items, as well as instruction (in this case, an instructional game), can produce tasks and assessments that demonstrate outstanding technical quality.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). Retrieved from http://escholarship.bc.edu/jtla/vol1/5

Atanda, R. (1999). Do gatekeeper courses expand education options? *Education Statistics Quarterly, 1*(1).

Baker, E. L. (1974). Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Technology, 14*(6), 10–16.

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice, 36*, 247–254.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2008). Design and validation of technology-based performance assessments. In J. M. Spector, M. D. Merrill, J. J. G. van Merrionnboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 595–604). Mahwah, NJ: Erlbaum.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131–153). Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*, 1210–1218.

Behr, M. J., Harel, G., Post, T., & Lesh, R. (2003). Rational number, ratio, and proportion. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 296–333). Reston, VA: National Council of Teachers of Mathematics.

Berkner, L., & Chavez, L. (1997). *Access to postsecondary education for the 1992 high school graduates* (National Center for Education Statistics 98–105). Washington, DC: U.S. Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=98105

Borsboom, D., Cramer, A. Kievit, R. A., Zand Scholten, A. & Franic, S. (2009) The end of construct validity. In R. W. Lissitz, (Ed.), *The concept of validity*. Information Age Publishers.

Brown, G., & Quinn, R. J. (2006). Algebra students' difficulty with fractions. *Australian Mathematics Teacher, 62*(4), 28–40.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Carpenter, T. P., Fennema, E., Franke, M. L., Empson, S. B., & Levi, L. W. (1999). *Children's mathematics: Cognitively guided instruction*. Portsmouth, NH: Heinemann.

Chung, G. K. W. K., Dionne, G. B., & Kaiser, W. J. (2006, April). *An exploratory study examining the feasibility of using Bayesian networks to predict circuit analysis understanding*. Paper presented at the National Council on Measurement in Education National Meeting, San Francisco, CA.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

De Corte, E., & Verschaffel, L. (1987). First graders' eye movements during elementary addition and subtraction word problem solving. In G. Luez & V. Lass (Eds.), *Fourth European Conference on Eye Movements, Vol. 1: Proceedings* (pp. 148 – 150). Gottingen: Hogzefe.

Donovan, M. S., & Bransford, J. D. (Eds.). (2005). *How students learn: Mathematics in the classroom*. Washington, DC: National Academies Press.

Driscoll, M. (1982). *Research within reach: Secondary school mathematics. A research guided response to the concerns of educators.* St. Louis, Missouri: CEMREL, Inc. Retrieved from ERIC database. (ED225842)

Fuson, K. C. (2003). Research on whole number addition and subtraction. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 243-275). Reston, VA: National Council of Teachers of Mathematics.

Gelman, R. (1991). Epigenetic foundations of knowledge structures: Initial and transcendent constructions. In Carey & Gelman (Eds.) *The epigenesis of mind: Essays on biology and cognition,* (pp. 293-322). Hillsdale, NJ: Lawrence Erlbaum.

Guilford, J. P. (1946). New Standards for test evaluation. *Educational psychology measurement, 6*, 427–439.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up*. Washington, DC: National Academy Press.

Lamon, S. J. (1999). *Teaching fractions and ratios for understanding: Essential knowledge and instructional strategies for teachers.* Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Mack, N. K. (1990). Learning fractions with understanding: Building on informal knowledge. *Journal for Research in Mathematics Education, 21*(1), 16–32.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report 03-16). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Phelan, J. C., Choi., K. C., Vendlinski, T. P., Baker, E., & Herman, J. (in press). Differential improvement in student understanding of mathematical principle following formative assessment intervention. *Journal of Education Research.*

Saxe, G. B., Shaughnessy, M. M., Earnest, D., Cremer, S., Platas, L. M., Sitabkhan, Y., & Young, A. (2007). *Fractions on the number line: The travel of ideas.* In T. Lamberg, L. R. Wiest (Eds.), 29th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education. Stateline (Lake Tahoe), NV: University of Nevada, Reno.

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction, 17*, 1–24.

Scott, W.A. (1960). Measures of test homogeneity. *Educational and Psychological Measurement, 20*(4), 751–757.

Serci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 19–38). Information Age Publishers.

SPSS. (2000). *Improve your written tests using item analysis* (Technical Report). Retrieved from ftp://ftp.spss.com/pub/web/wp/IAWP-02001.pdf

Sturmey, P., Matson, J. L., & Sevin, J. A. (1992). Brief report: Analysis of the internal consistency of three autism scales. *Journal of Autism and Developmental Disorders, 22*(2), 321–328.

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

Usiskin, Z. (1988). Conceptions of school algebra and uses of variables. In A.F. Coxford & A.P. Schlte (Eds.), *The ideas of algebra* (pp. 8–19). Reston, VA: National Council of Teachers of Mathematics.

Vendlinski, T. P. (2009). *The importance of intention and order: Teaching for conceptual understanding using handheld technology.* Dallas, TX: Texas Instruments.

Vendlinski, T. P., Hemberg, B. C., Mundy, C., Baker, E. L., Herman, J. L. Phelan, J., … Kang, T. (2009, March). *Designing professional development around key principles and formative assessments to improve teachers' knowledge to teach mathematics.* Meeting of the Society for Research on Educational Effectiveness, Crystal City, VA.

Vendlinski, T. P., Niemi, D., & Wang, J. (2005). Learning assessment by designing assessments: An on-line formative assessment design tool. In C. Crawford, R. Carlsen, L. Gibson, K. McFerrin, J. Price, & R. Webber (Eds.), *Proceedings of the Society for Information Technology and Teacher Education International Conference 2005* (pp. 228–240). Norfolk, VA: Association for the Advancement of Computing in Education.

Vendlinski, T. P., & Phelan, J. (2009, April). *Can teacher use of technical quality data about benchmark assessments make benchmarks as effective as formative assessments in improving student achievement?* Paper presented at the Annual Meeting of the American Education Research Association, San Diego, CA.

Wu, H. (2001, Summer). How to prepare students for algebra. *American Educator, 25*(2), 10–17.

APPENDIX A
KNOWLEDGE SPECIFICATIONS

Knowledge Specs		**Computational Fluency:** Students can execute procedures in the domain without the need to create or derive the procedure. Fluid performance is based on recall of patterns or other well established procedures and is fast, automatic, and error-free. *How is something done?*		**Conceptual Understanding:** Captures demonstration of understanding of the mathematical concepts. *Why is something done?*	
		When presented with… *(Assessment Stimulus)*	**Students should be able to…**	**When presented with… *(Assessment Stimulus)***	**Students should be able to…**
1.0.0. Does the student understand the importance of the unit whole or amount?					
	1.1.0. The size of a rational number is relative to how one Whole Unit is defined.	Any rational number…	Place it on a number line relative to the whole interval explicitly (0 and 1 labeled) or implicitly (0 and an integer other than 1 labeled) defined.	Apparent contradictions involving rational number such as ¾ < ½ or ½ does not equal ½.	Explain that the contradiction can be resolved if their relative wholes must be equal when comparing.
		Given a unit whole (interval, volume, area, etc.)…	Show how much of the whole must be shaded to represent a fractional amount.		
	1.2.0. In mathematics, one unit is understood to be one of some quantity (intervals, areas, volumes, etc.).	A histogram of a certain quantity represented by discrete objects…	Identity the unit that each single discrete object represents (e.g., each rose represents thousands of flowers sold on Valentine's Day).	Given a relationship between a real world measure and a scale model…	Explain how what size of unit to use on the model to accurately represent the real world quantity (e.g., 1 inch equals 25 feet since the real-world measure is 100 feet and the model can be up to 4 inches in length).
	1.3.0. In our number system, the unit can be represented as one whole interval on a number line.	Given a number line labeled with consecutive integers that may or may not include zero…	Show the unit interval that fits with the given number line or accurately place another non-consecutive integer on the number line.	A number line that is labeled by skip units (2, 4, 6, etc.) or a line labeled by ½ units that may or may not include zero	Explain how to determine where other integer and rational values should be placed.

		Knowledge Specs	**Computational Fluency:** Students can execute procedures in the domain without the need to create or derive the procedure. Fluid performance is based on recall of patterns or other well established procedures and is fast, automatic, and error-free. *How is something done?*		**Conceptual Understanding:** Captures demonstration of understanding of the mathematical concepts. *Why is something done?*	
			When presented with… *(Assessment Stimulus)*	**Students should be able to…**	**When presented with… *(Assessment Stimulus)***	**Students should be able to…**
		1.3.1. Positive integers are represented by successive whole intervals on the positive side of zero.	An integer number line labeled with 0 and 1…	Label other integer values.	An integer number line labeled with 0 and some integer other than 1…	Explain how to identify the whole interval and then use it to accurately place the integers between 0 and the given integer on the number line.
		1.3.2. The interval between each integer is constant once it is established.	Given a number line with marks between integers (e.g., every *x* mark equals 1 interval)…	Label other integer values.	Given a number line with at least zero and one other integer labeled…	Explain how to accurately place other larger or smaller integers on the number line.
		1.3.3. Positive non-integers are represented by fractional parts of the interval between whole numbers.	Given a number line with marks between integers (e.g., every *x* mark equals 1 interval)…	Label rational values including rational numbers greater than 1.	Given a number line with at least zero and one other integer labeled…	Explain how to label or place other rational values on the number line.
		1.3.4. All Rational Numbers can be represented as additions of integers or fractions.	Given any non-unitary integer or rational number…	Show the addition of ones (in the case of an integer) or unitary fractions (in the case of the rational) that would produce the given number.		

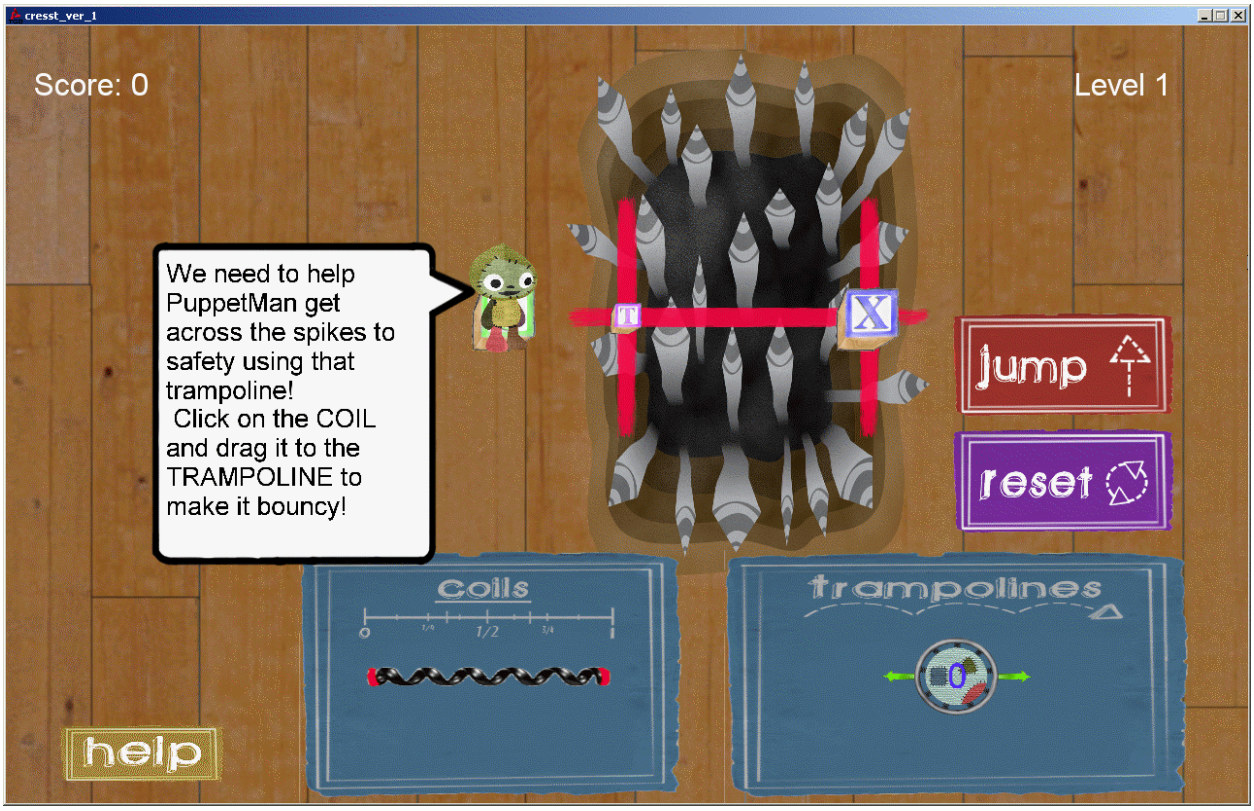
Knowledge Specs			**Computational Fluency:** Students can execute procedures in the domain without the need to create or derive the procedure. Fluid performance is based on recall of patterns or other well established procedures, and is fast, automatic, and error-free. *How is something done?*		**Conceptual Understanding:** Captures demonstration of understanding of the mathematical concepts. *Why is something done?*	
			When presented with… *(Assessment Stimulus)*	**Students should be able to…**	**When presented with… *(Assessment Stimulus)***	**Students should be able to…**
2.0.0. Does the student understand the meaning of addition?						
	2.1.0. To add quantities, the units (or parts of units) must be identical.		Given a fraction and a sum with a similar or dissimilar denominator…	Determine fraction that must be added to produce the indicated sum.		
		2.1.1. Identical (or common) units can be descriptive (e.g., apples, oranges, and fruit) or they can be quantitative (e.g., identical lengths, identical areas, etc.).	Given quantities with similar units…	Determine the common unit that will "allow" addition.		
		2.1.2. Positive integers can be broken (decomposed) into parts that are each one unit in quantity. These single (identical) units can be added to create a single numerical sum.	Given objects that represent a collection of ones (e.g., a $5 bill and a $2 bill or a 3-gallon bottle of anti-freeze and a 2-quart bottle of anti-freeze)…	Determine the number of a given unit in each quantity by decomposing into equal numbers of unitary units (e.g., five $1 and two $1).		
		2.1.3. Each Whole Unit or part of a Whole Unit (fractions) can be further broken into smaller, identical parts, if necessary.	Given integers or fractions with dissimilar units (denominators)…	Break integers or fractions into an equivalent number of common units that will allow the numbers to be added.		
	2.2.0. Identical (common) units can be added to create a single numerical sum.		Given a certain number of integer or fractions with the same denominator…	Determine the sum of those integers or fractions.	Given a certain number of integers or fractions with the same denominator…	Explain what common unit would be used to add (e.g., 3 + 5 would be three ones plus 5 ones, or ¾ + ¼ would be three fourths + one fourth).

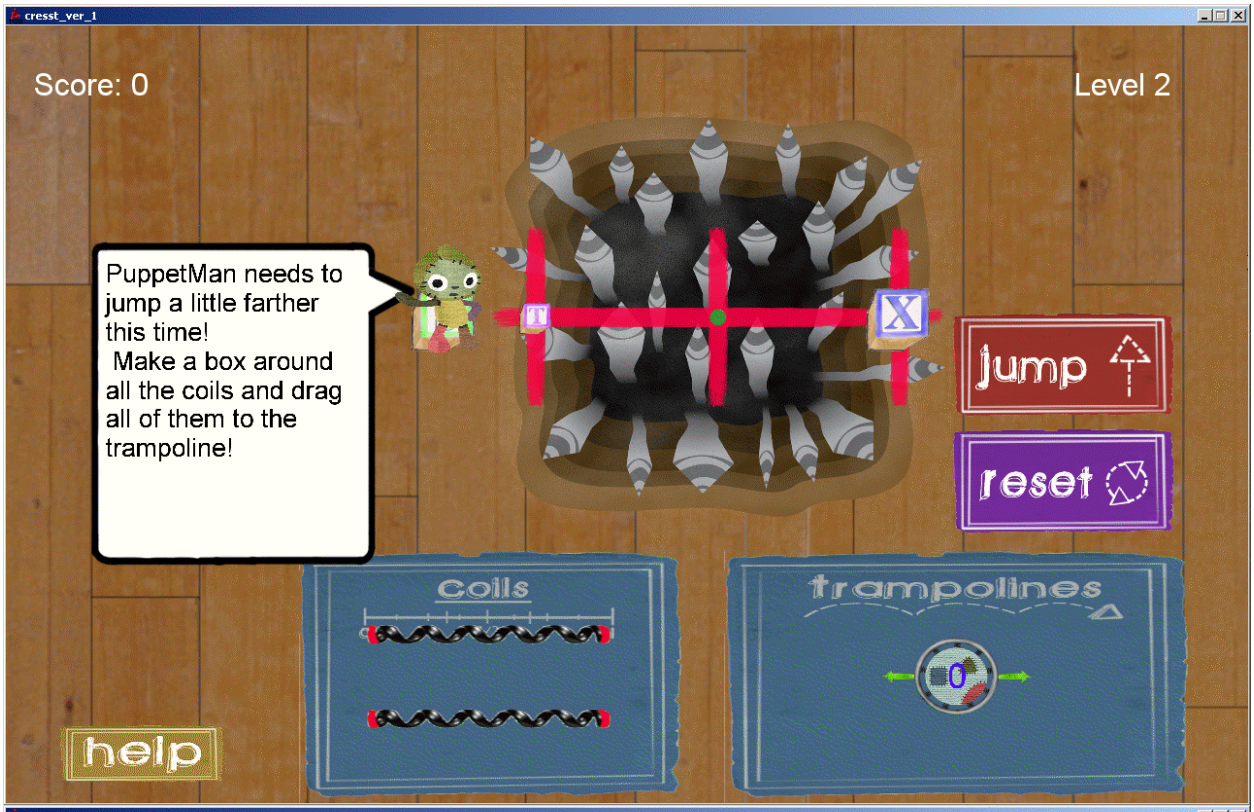
Knowledge Specs		**Computational Fluency:** Students can execute procedures in the domain without the need to create or derive the procedure. Fluid performance is based on recall of patterns or other well established procedures, and is fast, automatic, and error-free. *How is something done?*		**Conceptual Understanding:** Captures demonstration of understanding of the mathematical concepts. *Why is something done?*	
		When presented with… *(Assessment Stimulus)*	**Students should be able to…**	**When presented with… *(Assessment Stimulus)***	**Students should be able to…**
	2.3.0. Dissimilar quantities can be represented as an expression or using some other characterization but are not typically expressed as a single sum (Note: We are considering numbers like 2¾ to have an implied addition. In other words, 2¾ = 2 + ¾, whereas 11/4 is a single sum.	Given at least two dissimilar integers (such as 2 ¼ or 60 miles per hour and 3 miles) or fractions…	Determine if they can be added (2¼ can be expressed as 2 + ¼; miles per hour and miles cannot be added).	Given at least two dissimilar integers or fractions	Explain why they cannot be added or, in the case of complex fractions like 3½, that there is an implied addition.
	2.4.0. Zero can be added to any quantity. When zero is added to any quantity, the value of the quantity remains unchanged (Additive Identity).	Given any integer or fraction…	Determine the sum when adding zero or some form of zero (e.g., 3 + -3) to the original integer or fraction.	Given any integer or fraction…	Explain why adding zero will leave the number unchanged. Explanation can be given in a variety of ways (e.g., logic, diagrammatic, etc.).
	2.5.0. Adding two positive numbers will always produce a sum that is greater (more positive) than either number.	Given any combination of positive integers and positive fractions (including variables that must be positive)…	Determine that the sum must be positive.	Given any combination of positive integers and positive fractions (including variables that must be positive)…	Explain why the sum must always be positive. Explanation can be shown in a variety of ways (e.g., written, diagrammatic, etc.).
	2.6.0. Adding two negative numbers will always produce a sum that is less than (more negative) either number.	Given any combination of negative integers and negative fractions (including variables that must be negative)…	Determine that the sum must be negative.	Given any combination of negative integers and negative fractions (including variables that must be negative)…	Explain why the sum must always be negative. Explanation can be shown in a variety of ways (e.g., written, diagrammatic, etc.).

Knowledge Specs		**Computational Fluency:** Students can execute procedures in the domain without the need to create or derive the procedure. Fluid performance is based on recall of patterns or other well established procedures, and is fast, automatic, and error-free. *How is something done?*		**Conceptual Understanding:** Captures demonstration of understanding of the mathematical concepts. *Why is something done?*	
		When presented with… ***(Assessment Stimulus)***	**Students should be able to…**	**When presented with…** ***(Assessment Stimulus)***	**Students should be able to…**
	2.7.0. Since they are opposites, adding a number and its opposite (two numbers of the same absolute value but opposite in sign) will result in a sum of zero (the additive inverse).	Given any number (integer or fraction) and its opposite…	Determine that the sum will be zero.	Given any number (integer or fraction) and its opposite or a number that is the sum of the opposite of the original number and another number of the same sign (e.g., 3 + -5 can be seen as 3 + -3 + -2)…	Explain how this is a case of adding a number and its opposite (e.g., 3 + -5 = 3 + -3 + -2 and is the same as 0 + -2).

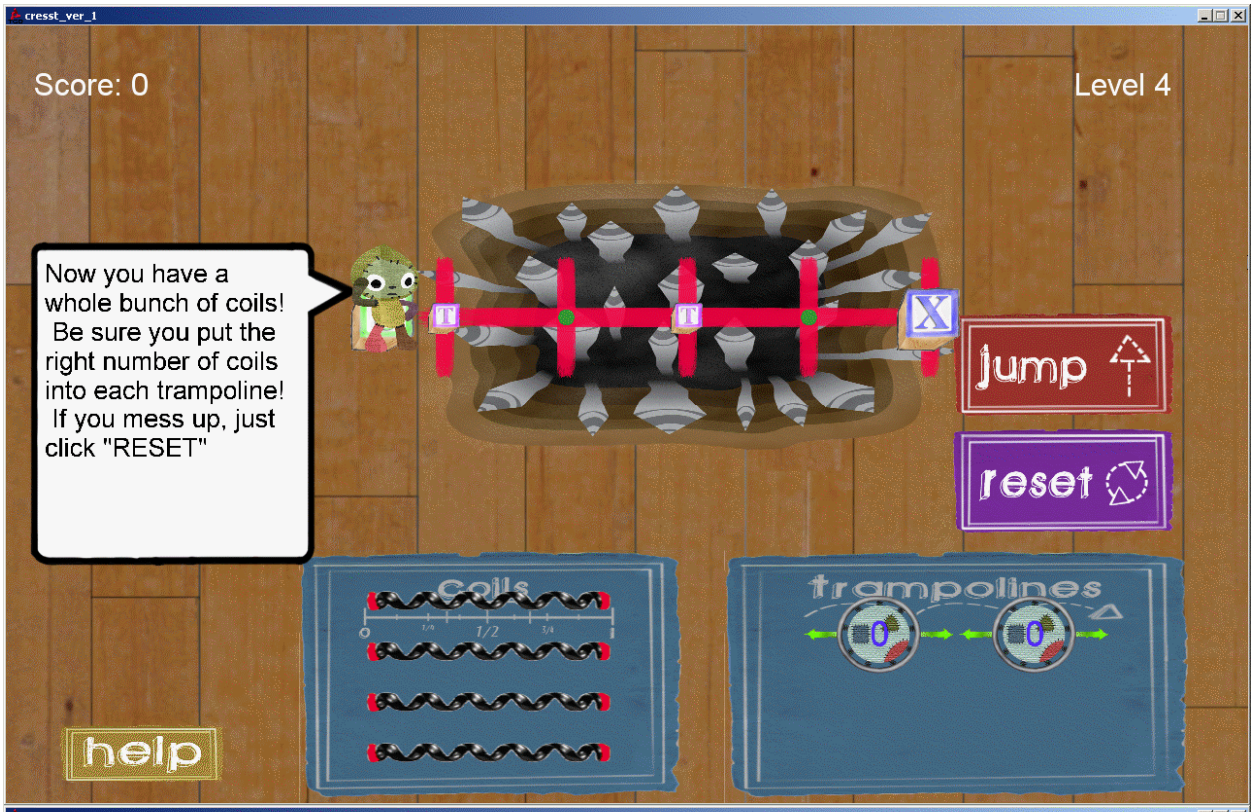
APPENDIX B

PUPPETMAN SCREEN SHOTS FROM THE FIRST VERSION TESTED WITH STUDENTS

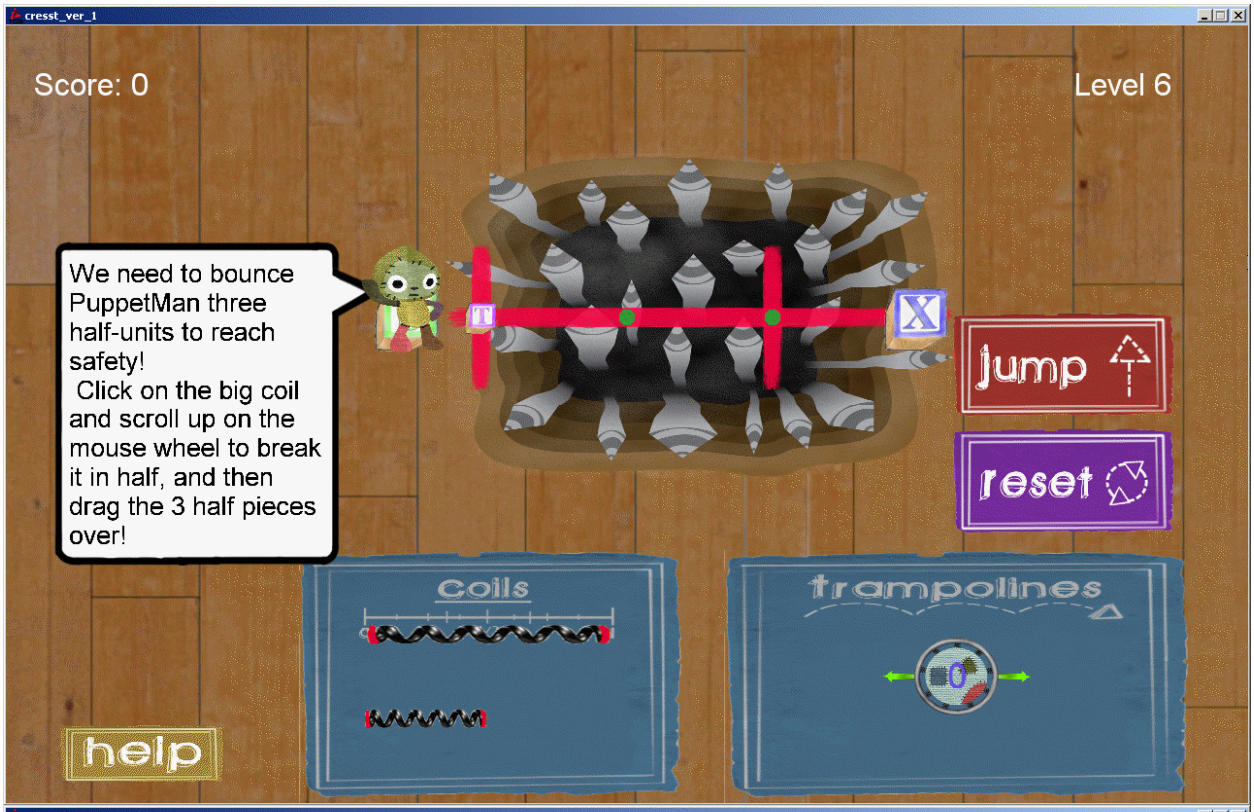


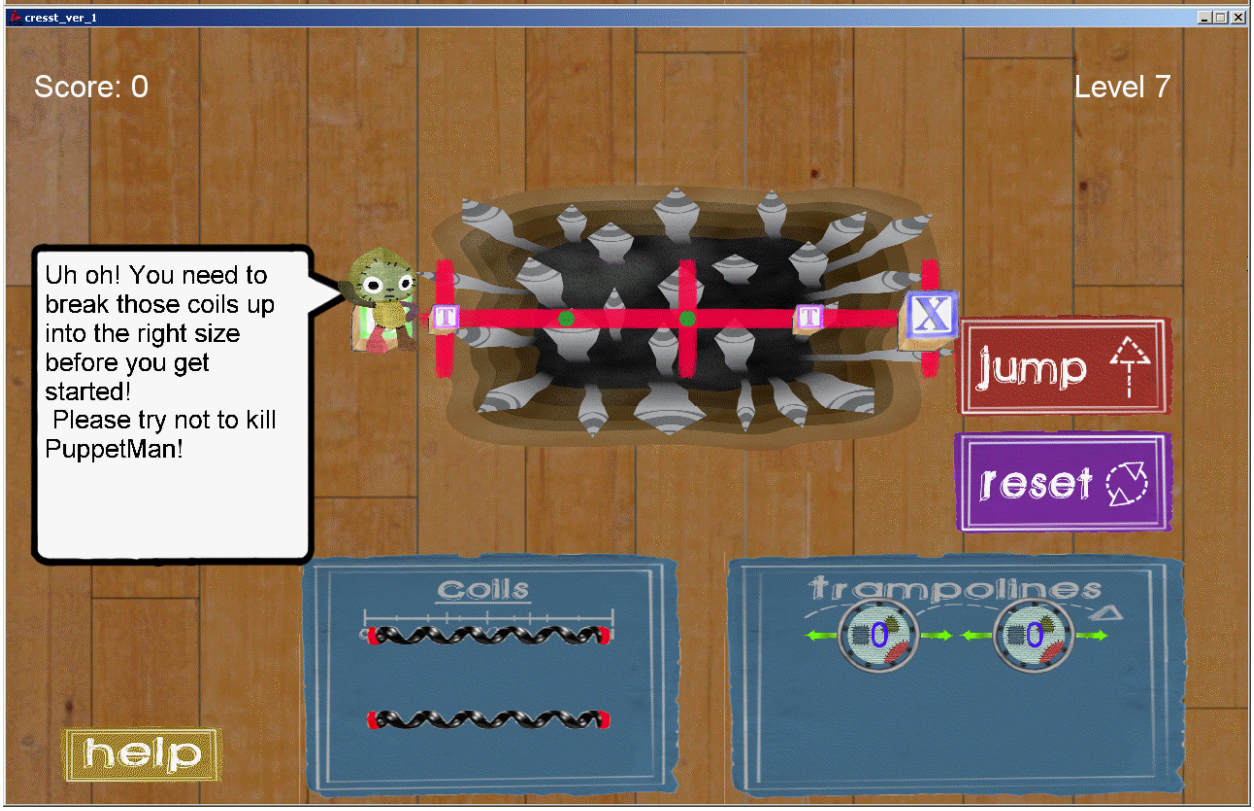

cresst_ver_1
Score: 0
Level 2
PuppetMan needs to jump a little farther this time!
Make a box around all the coils and drag all of them to the trampoline!
jump
reset
coils
trampolines
0
help

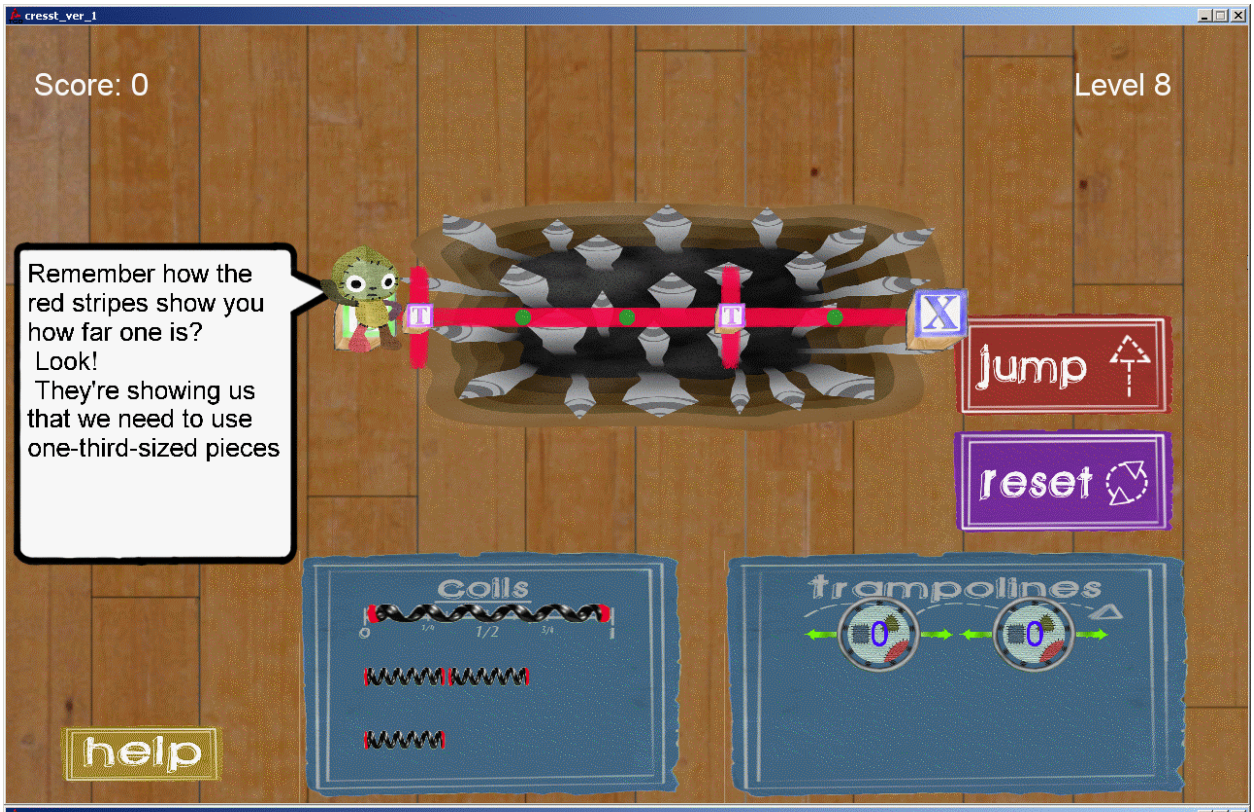

cresst_ver_1
Score: 0
Level 3
Sometimes you'll have multiple trampolines!
Drag one coil into each trampoline.
jump
reset
coils
trampolines
0
0
help

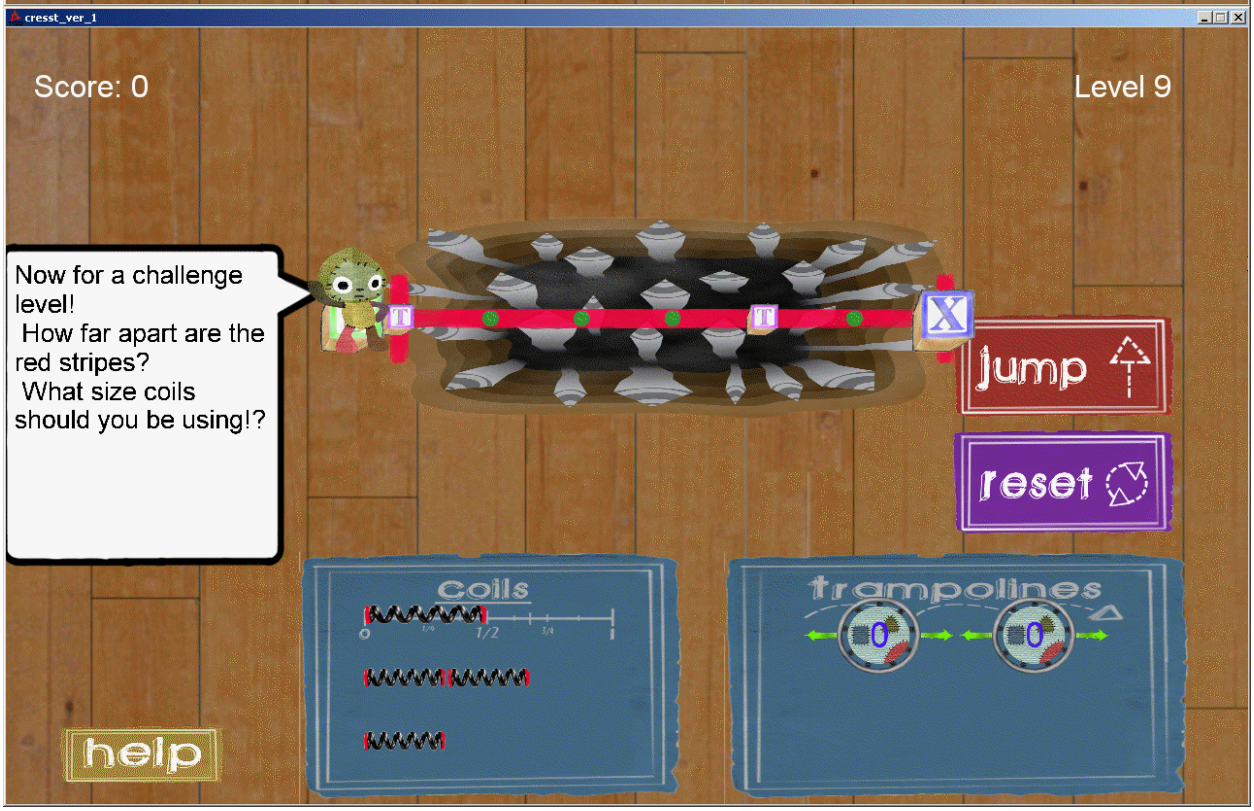

cresst_ver_1
Score: 0
Level 4
Now you have a whole bunch of coils! Be sure you put the right number of coils into each trampoline! If you mess up, just click "RESET"
jump
reset
Coils
trampolines
help


cresst_ver_1
Score: 0
Level 5
Sometimes you'll want smaller coils for your trampolines. Here, we have two coils, but each is half normal size! Drag both of them into the trampoline!
jump
reset
Coils
trampolines
help


cresst_ver_1
Score: 0
Level 6
We need to bounce PuppetMan three half-units to reach safety!
Click on the big coil and scroll up on the mouse wheel to break it in half, and then drag the 3 half pieces over!
jump
reset
coils
trampolines
0
help

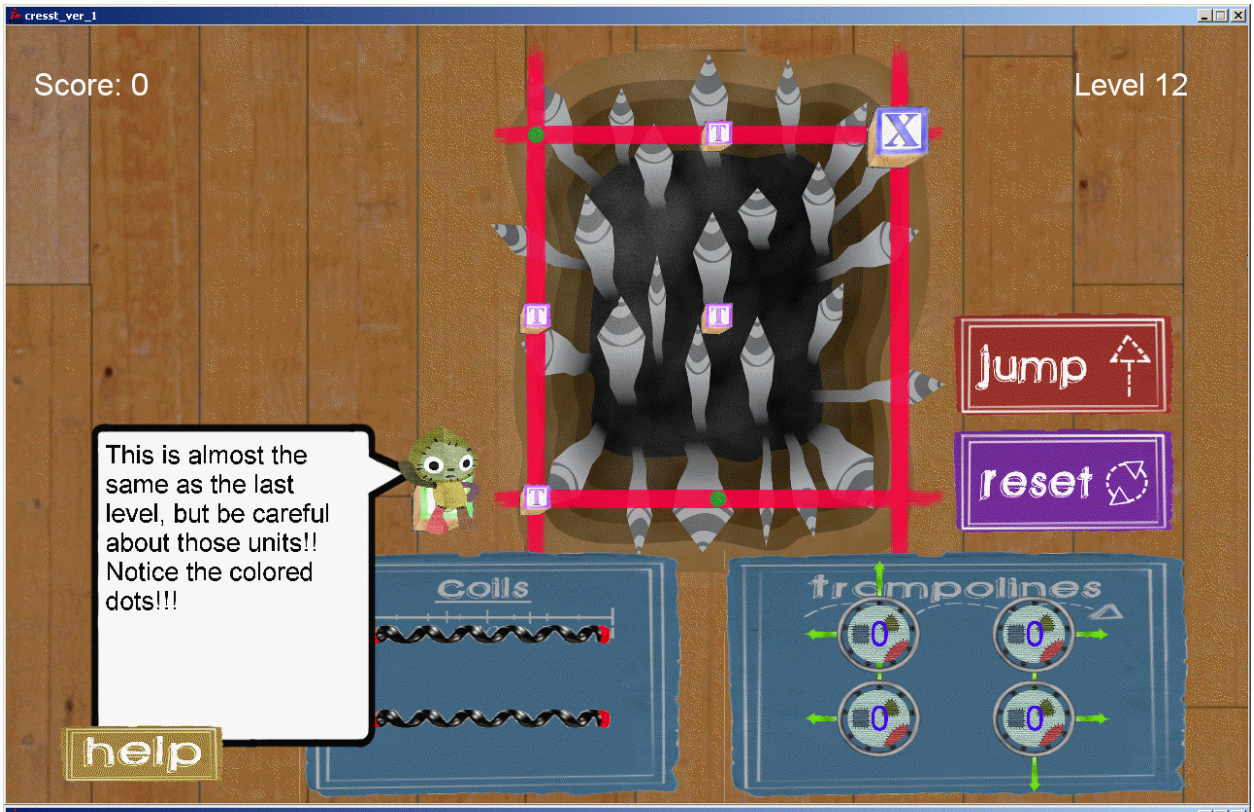

cresst_ver_1
Score: 0
Level 7
Uh oh! You need to break those coils up into the right size before you get started!
Please try not to kill PuppetMan!
jump
reset
coils
trampolines
0
0
help

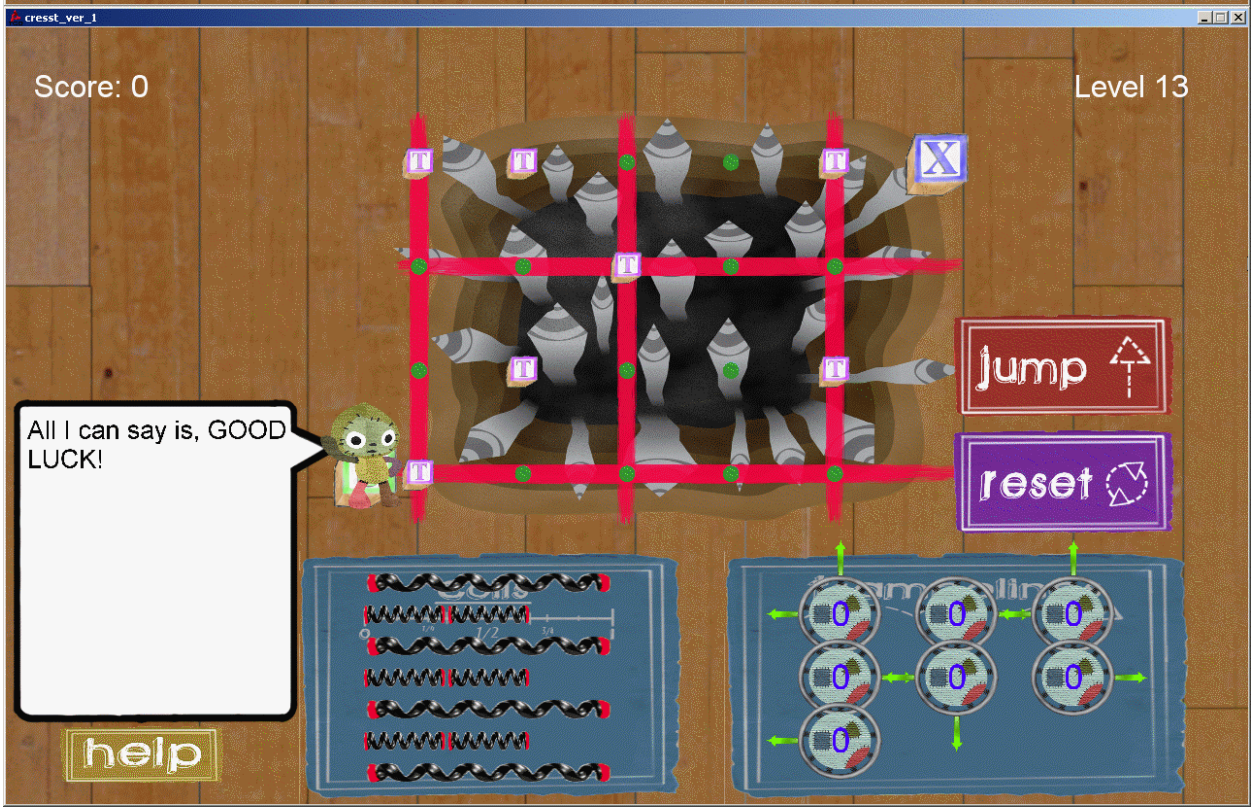

cresst_ver_1
Score: 0
Level 8
Remember how the red stripes show you how far one is?
Look!
They're showing us that we need to use one-third-sized pieces
jump
reset
coils
trampolines
help

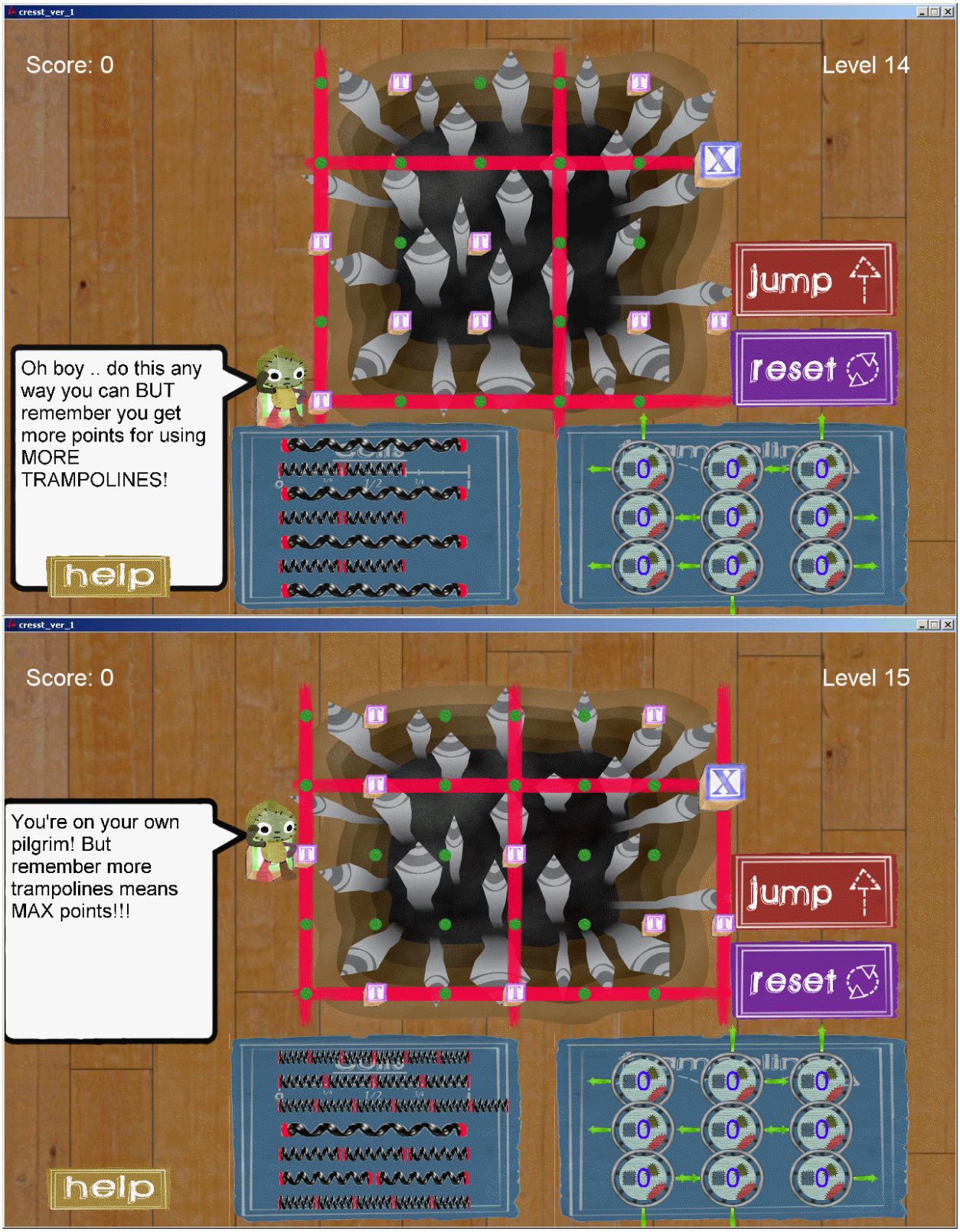

cresst_ver_1
Score: 0
Level 9
Now for a challenge level!
How far apart are the red stripes?
What size coils should you be using!?
jump
reset
coils
trampolines
help


cresst_ver_1
Score: 0
Level 10
GASP, two dimensions! Fortunately, everything works exactly the same, but trampolines can change bounce direction! Drag one coil into each trampoline.
Jump
reset
coils
trampolines
help


cresst_ver_1
Score: 0
Level 11
This is the exact same as the last level, just more trampolines and different directions!
Jump
reset
coils
trampolines
help


cresst_ver_1
Score: 0
Level 12
Jump
reset
This is almost the same as the last level, but be careful about those units!! Notice the colored dots!!!
help
Coils
trampolines
0
0
0
0


cresst_ver_1
Score: 0
Level 13
Jump
reset
All I can say is, GOOD LUCK!
help
0
0
0
0
0
0
0


cresst_ver_1
Score: 0
Level 14
Oh boy .. do this any way you can BUT remember you get more points for using MORE TRAMPOLINES!
help
Jump
reset
cresst_ver_1
Score: 0
Level 15
You're on your own pilgrim! But remember more trampolines means MAX points!!!
help
Jump
reset