Tech Know Learn (2016) 21:155–174
DOI 10.1007/s10758-016-9282-z

Measurement in Learning Games Evolution: Review of Methodologies Used in Determining Effectiveness of *Math Snacks* Games and Animations

Karen Trujillo[1] · Barbara Chamberlin[2] · Karin Wiburg[1] · Amanda Armstrong[2]

Published online: 3 May 2016

Abstract This article captures the evolution of research goals and methodologies used to assess the effectiveness and impact of a set of mathematical educational games and animations for middle-school aged students. The researchers initially proposed using a mixed model research design of formative and summative measures, such as user-testing, observations, interviews, and standardized test scores. Over the course of the 5 years, researchers expanded on these qualitative and quantitative methods by adding additional research methods based on additional potential opportunities for formative testing, additional potentials areas of impact, and refining research methodologies. Based on the findings from these methods, the researchers offer recommendations for approaches to evaluating educational game design and animation that support student learning. The authors review the methodologies used (observations, focus groups and panel discussions, pre- and post-tests, self-report surveys and embedded data); environments in which the methods were used (Learning Games Lab Think Tanks, classrooms, summer camps, annual advisory reviews); and three studies (one pilot and two final studies). The researchers close with recommendations for design and evaluation strategies of game- and animation-based learning.

✉ Barbara Chamberlin
bchamber@nmsu.edu

Karen Trujillo
ktrujill@nmsu.edu

Karin Wiburg
kwiburg@nmsu.edu

Amanda Armstrong
aarmstr0@nmsu.edu

[1] College of Education, New Mexico State University, Box 30001, MSC 3AC, Las Cruces, NM 88003, USA

[2] Media Productions and Learning Games Lab, New Mexico State University, Box 30003, MSC 3CUR, Las Cruces, NM 88003, USA

Keywords Mathematics · Learning games · Research methodologies · Animations · Game development

1 Introduction

During a 5-year research and development project, investigators developed a set of math learning games and animations, and conducted original research in testing the effectiveness of the tools with students in grades 5, 6 and 7. The team had anticipated using both formative and summative measures over the course of the project. Formative methods, such as user testing, informed the design of the games and animations, and summative measures, such as pre- and post-tests, were designed to assess effectiveness of the final tools. Over the course of the project, researchers recognized additional opportunities for research, and greatly expanded both their formative and summative measures, refining a wide range of different instruments and methods for both types of research. This paper reviews the original research goals, documents ways in which those goals expanded during the course of the project, and offers specifics in how each research method and measure was used, with a brief summary of findings. It concludes with recommendations for other developers interested in evaluating learning games, which may help guide research plans for long-term research and development projects.

The most significant potential impact of quality *Math Snacks* games and the accurate assessment of their effect on student learning, is how well students are prepared for STEM courses and careers, and specifically if concepts considered core to mathematics learning could be learned by students who played the games. The theoretical basis for the game-based intervention is grounded in constructivist learning principles for building knowledge (Scardamalia and Bereiter 2008). Current theory-based pedagogy supports metaphors that consider the *construction* of knowledge rather than the *transmission* of knowledge (Greeno et al. 1996; Van Meter and Stevens 2000). The animations and games included in the *Math Snacks* suite of products provide an anchor for instruction of critical math content including ratio, equivalence, coordinate plane, fractions, decimals and place value. To design the games, the team used the *Learning Games Design Model* where game developers and content experts work collaboratively throughout the game design process (Chamberlin et al. 2012). The *Math Snacks* project included a mathematician, two mathematics educators, an educational psychologist, an internal evaluator, learning theorists, programmers, artists, instructional designers and other production specialists.

This team worked collectively to develop a different type of game from those often seen for math learning. The *Math Snacks* games were not designed to create automaticity in problem solving but to give students opportunities to explore, question, and understand key concpets related to number sense, ratio, rate, and points on a coordinate plane. This game-based learning was also reinforced by follow-up inquiry activities designed for class instruction. The evaluation tools had to reflect the complexity of learning concepts, assess other potential impacts (such as changes in self-efficacy) and could potentially map specific game-based activities to learning outcomes. By using various assessments of learning activities, the team was able to establish a framework for developing tools based on constant formative assessment during development, assessing student learning, and finally looking at what supported that learning. This is consistent with the Shute et al. framework for assessing serious games, and holds similar assumptions: learning by doing in games (rather than simply practicing) improves learning processes, different types of learning can be verified and measured during play, and feedback during gameplay can support student learning (2009). By integrating both qualitative and quantitative approaches across different stages of the research process, the approach of the team reflects Johnson and Onwuegbuzie's *mixed model* research design, enabling a broader and more complete range of research questions, and using the strengths of one method to overcome the weaknesses in another (2004).

1.1 Products

The *Math Snacks* project began as an investigation of gaps in students' understanding of math concepts in grades 4–7. Researchers analyzed the results of 24,000 student scores on the New Mexico Standards-Based Assessment (NMSBA). On this test, half of the items were open-ended or short answer questions, so it was possible to see student misconceptions in mathematical thinking. Researchers analyzed test results in several different districts and found almost identical patterns across districts, regardless of the economic status, number of English Language learners, or size of the district. Students across all districts had specific trouble with the same content, including number concepts and operations, ratios, fractions, and points on a coordinate grid. Similar patterns of common mistakes were also found across the different districts in all of the mathematics strands, including geometry, data, and algebra. This research served as a road map for developers as they developed games and animations.

The *Math Snacks* suite includes 11 animations and games, all available in English and Spanish. Every *Snack* includes a printable *Teacher Guide* offering recommendations on use as well as additional learning activities, and a *Teaching With* Video showing how teachers can use the game or animation. Each animation also has a printable *Learner Guide* and a *Comic Book Transcript* available in both languages. All *Math Snacks* resources are available online at mathsnacks.org, as well as complete descriptions, screen shots, and sample gameplay videos (Table 1).

2 Research Plans

The initial research plan included formative methods for the development of the animations and games, and a final summative randomized control trial using standardized exams to measure students' knowledge of mathematics content. While the same research questions were addressed as the research evolved, the methods, testing environments, and research designs changed throughout the project.

2.1 Initial Research Plan

2.1.1 Formative Testing

Many development teams integrate formative testing by showing working prototypes to audiences and revising the product. However, the *Math Snacks* team intended to use formative testing much more frequently, and with more groups than most other design teams. Initially, the formative testing for product development and effectiveness was designed to include teachers, learners, and quality assurance experts throughout the design

Table 1 *Math Snacks* animations and games

Type	Name	Description
Animations	*Atlantean Dodgeball/Juego del Quemado De Atlántida*	Ratio errors confuse one of the coaches as two teams face off in an epic dodge ball tournament
	Bad Date/Una Cita Aburrida	Talking too much or not enough gives dating a whole new meaning
	Number Rights/Derechos de los Números	A passionate activist named "¼" rises up and demands equity for all numbers, including fractions, decimals and negative numbers
	Overruled!/¡Rechazado!	Two besotted rulers must embrace proportional reasoning in order to design a bridge to unite their lands
	Ratey the Math Cat/Razonaso el Gato Matemático	Ratey can't resist pointing out the "purr" unit rates in daily life
	Scale Ella/Lola Escala	The villain, Scaleo, transforms the length, width, and height of various objects and our heroine Scale Ella uses the power of scale factor to foil his plans
Games	*Game Over Gopher/La Tuza Instrusa*	Players defend the carrot from hungry space gophers and hunt for buried treasure by placing points on a coordinate grid in each quadrant, on the axes, and at the origin
	Monster School Bus/Escuela de Monstruos	Players pick up monster students and drop them off at school. Using base 10 concepts to get a "Full Load" leads to rewards
	Pearl Diver/Pescador de Perlas	Players navigate the number line while diving among shipwrecks and sunken ruins
	Ratio Rumble/Retumbar de Razones	Players battle opponents by building potions using ratios
	Gate/Sombras	Players restore the balance of nature using number operations and place value

process. The iterative design included interviews and observations conducted with teachers and students using *Math Snacks* products in classrooms and individually throughout development process. Teachers and students would provide feedback on working *Math Snacks* prototypes.

The Learning Games Lab (an in-house testing facility in the development shop) was to be made available during the summer and at least twice a month during the school year for middle-school youth to test products during "think tanks". Formative evaluation tools proposed for the lab included focus group discussions, one-on-one interviews, two-on-one observations, daily blog prompts, video closet confessionals, and think aloud activities. Data collected was to be analyzed using a variety of qualitative methodologies. Finally, the "Quality Assurance Committee" comprised of experts in mathematics, technology, professional development and research was slated to meet annually. This committee provided an independent review of the resources, models, and technologies developed by the project, including a review of the research design, methodologies, and execution of each objective listed in the grant.

2.1.2 Summative Testing

Summative assessments using quantitative methods were proposed to determine the effectiveness of the *Math Snacks* intervention. Random control trials would be designed to measure students' understanding of given mathematical concepts using the *New Mexico Standards Based Assessment (NMSBA).* Hypotheses included:

- Students in experimental classrooms will increase their conceptual understanding of given mathematical concepts at a significantly higher level than students in control classrooms as measured on the New Mexico Standards Based Assessment (NMSBA).
- Students will retain their understanding of math concepts longer in experimental than in control classrooms based on year-end scores over time on the NMSBA.
- Students in experimental classrooms will increase their effective use of math process skills as measured on open-ended portion of the NMSBA at a significantly higher level than students in control classrooms.

The initial plan proposed having 30 experimental teachers and 30 additional teachers randomly selected to serve as control teachers.

2.2 Evolution of Research Plan

While the initial research design guided the entire 5-year project, additional research naturally evolved as part of the iterative design process. The main goals stayed the same (using formative methods to increase the effectiveness of the developed tools with the target population, and testing the impacts of the finished tools with knowledge-gain measures), but the potential for additional research methods increased as the project progressed. The project presented more opportunities for formative testing, revealed potential impacts on students and teachers that researchers wanted to assess, and enabled several methodologies for assessment that were not anticipated in the original design. Most significantly, the team was unable to use the NMSBA as planned because during the duration of the grant, schools in New Mexico were transitioning the curriculum and the assessment from NMSBA to the Partnership for Assessment of Readiness for College and Careers (PARCC). As a result, the *Math Snacks* team developed three assessments using modified released items from similar standardized tests to measure mathematics learning. These measures are described below.

In addition to this change in the summative tests, greater access to learners and teachers revealed additional potential opportunities for formative testing than originally thought. Initially, formative testing was planned for the Learning Games Lab, where students would come for week-long "Think Tank" sessions, participate in creative design activities, and test games in development. As part of their outreach work, the *Math Snacks* team also introduced annual *Math Snacks* Camps, where teachers used *Math Snacks* tools with students during week-long sessions where students and teachers worked together with the tools and teachers created additional activities to reinforce the learning from the games. The camps were originally designed to help researchers better understand the teaching environment and the needs of teachers when using *Math Snacks* products.

With this additional access to teachers and students during development, the team identified additional potential areas of impact. To conduct pilot studies in several different classrooms, and refine research instruments and protocols, the team took full advantage of their network of schools which were established through other mathematics outreach projects, as well as the notoriety *Math Snacks* products had gained through word-of-mouth. These established teacher networks and administrative support gave researchers greater access for classroom testing, observations, and focus groups. In addition to having access to actual classrooms during the school year, the summer camps replicated a classroom-learning environment as well. During these observations, the team witnessed a shift in thinking for both the teachers and students, and realized that—in addition to knowledge—*Math Snacks* lessons appeared to increase students' confidence and enjoyment of math. Similarly, the teachers' approaches to teaching with learning games started to change, with many embracing more inquiry-based ways of using games in the classroom. As researchers had the opportunity to observe how the *Math Snacks* lessons influenced both students and teachers, it became clear that significant impacts of the games could be missed in an assessment strategy that focused only on knowledge change.

Finally, the team had 4 years to refine their research methodologies, and chose to explore some embedded assessment data that was collected in-game—such as time spent at each level of game, repeated attempts to succeed levels after failure, and replay of levels that were successfully passed. Previously, the team had planned to observe student gameplay, but the embedded data gave a fuller picture of how the games were played.

3 Research Environments and Studies

This was a large-scale project, capturing both formative and summative data in a wide variety of different locations. Before reviewing the specific methods used, it is helpful to specify where and how research took place.

3.1 Learning Games Lab Think Tanks

In the Learning Games Lab, researchers used Game Design Think Tanks where middle school students used games during the design process. The researchers conducted interviews, observations, and focus groups. In addition, students used a video closet to provide individual feedback. With *Math Snacks*, youth "consultants" in grades 5–8 participated in summer sessions, with returning 'experts' throughout the school year. Because the Think Tanks occurred almost every week in the summer, developers were able to test products with the kids throughout development. This enables more frequent user testing at earlier stages of game development than often seen with other user testing strategies.

Math Snacks products were tested throughout development: from early concept formation (where consultants viewed narrated storyboards, early character sketches, and even giving feedback on potential game titles and theming), through usability testing, and final game balancing. The detailed process and documentation (as described specifically for *Math Snacks* work in Chamberlin et al. 2014) meant that researchers could identify which students had seen games at which point, bringing fresh learners in when needed, or testing specific problems with repeat consultants. It was not uncommon for a game to go through 40 or more different testing sessions in the Learning Games Lab.

3.2 Classrooms

Throughout the project, researchers spent time collecting data through classroom observations. Teacher Advisors were recruited during year 1 and year 2 of the project and were asked to use *Math Snacks* products during the development phase so that developers and researchers could observe students using the tools in real environments. In these sessions, researchers observed students struggling with particular questions on the learner guide and even discovered bugs in the games that weren't obvious when small groups played. In the Pilot Study and final studies, classroom observations were used to determine fidelity of implementation, to note student interaction with *Math Snacks* materials, and to rate teacher effectiveness using the *Observation of Learning Environment II* (OLE 2). These observations were also critically important in supporting other forms of data collection, such as the quantitative gameplay data. For example, in the final games study, there were certain classrooms where gameplay data was missing for certain students even though they participated in all testing sessions. When reviewing the observation data, it became clear that students were playing the games *in pairs* and only one student was logging on to the computer, therefore the only record of these students playing the games was through observations.

3.3 *Math Snacks* Camps

Math Snacks camps started during year 1 of the grant and continued for 3 consecutive summers. In each instance, *Math Snacks* camps were held for 5 days from 8:30 am to 2:30 pm where teachers worked with students until noon dismissal, and teachers then stayed an additional 2 h. Table 2 provides a summary of each *Math Snacks* camp and which *Math Snacks* products were used in each camp.

Even though the *Math Snacks* Camps evolved based on feedback from both students and teachers, the overall structure remained the same. The key elements for each camp included four components:

1. *Teachers worked in teams of two*: Teachers met 4–6 weeks prior to each camp and were assigned 1–2 *Math Snacks* tools depending on what had been developed and what needed to be tested. These teachers were given the freedom to design activities that complimented the mathematics content being covered in the assigned *Math Snacks*. For example, if the teacher team was working with *Ratio Rumble* or *Bad Date*, they had to develop hands-on activities that would enhance the learning of ratio.

Table 2 *Math Snacks* camp participants and content

Number of teachers	Number of students	Student grade levels	*Math Snacks* used
8	35	Gr 4–6	Animations: Scale Ella, Number Rights, Bad Date, Overruled
13	97	Gr 4–8	Animations: Scale Ella, Atlantean Dodgeball, Number Rights Games: Gate, Monster School Bus, Pearl Diver
9	30	Gr 6–7	Animations: Scale Ella, Atlantean Dodgeball, Number Rights, Games: Gate, Monster School Bus, Pearl Diver
16	90	Gr 4–8	Animations: Scale Ella, Atlantean Dodgeball, Bad Date, Number Rights Games: Gate, Monster School Bus, Pearl Diver, Game Over Gopher, Ratey the Math Cat
6	45	Gr 6–7	Animations: Scale Ella, Atlantean Dodgeball, Bad Date, Ratey the Math Cat Number Rights Games: Gate, Monster School Bus, Pearl Diver, Game Over Gopher,
10	54	Gr 4–7	Games: Gate, Game Over Gopher, Monster School Bus, Ratio Rumble

2. *Teachers worked with students in a classroom environment:* In each camp, the students were divided into groups of 10–15 based on age. In the first 2 years of the camp, the teacher teams spent each day with the same group of students and focused on particular concepts. In years 3 and 4, each teacher presented his or her lessons to multiple groups of students using a middle school model. This shift allowed the research team to observe multiple groups of students interacting with the same materials, and students to be exposed to multiple *Math Snacks*.
3. *Teachers and researched debriefed daily:* After students were dismissed for the day, the teachers stayed for an additional 2 h to debrief about the daily activities. Teachers and researchers shared what they observed, and the teams worked together to change the lessons for the next day.
4. *Teachers developed materials iteratively throughout the camp Math Snacks:* Camps gave teachers the opportunity to develop complementary classroom learning activities for the *Math Snacks* games and animations, and test and refine those activities designed by the development team. Research team members and teachers had the opportunity to observe the lessons and provided feedback throughout the camp. This process resulted in classroom lessons and inquiry-based activities designed by teachers, tested in classroom settings, and improved through feedback.

3.4 Annual Advisory Reviews

Developers found it helpful to get advice from other game developers, teachers, and fellow researchers *before* usable prototypes were ready for testing with teachers and students. An external evaluation team convened annual panels to review proposed content for the *Math Snacks*, identify potential problem areas, recommend areas for expanded development, and review measures.

3.5 Pilot Study

In 2012, a pilot study was conducted to validate the *Measure of Mathematics Learning I* and to determine whether or not teachers required a scripted lesson plan for the randomized controlled trials during years 3 and 4. Nine teachers and their students (225 sixth and 125 seventh grade students) were randomly assigned to one of two groups: a scripted lesson group, and an open lesson group. Both groups of teachers were asked to teach six *Math Snacks* lessons over the same 6-week time period. Teachers in the scripted lesson group were provided with scripted lesson plans with step by step instructions for each portion of the lesson. Teachers in the open lesson group were simply directed to the *Math Snacks* website to find the appropriate teacher support materials. The instruments used during this study included: (1) the *Measure of Mathematics Learning I*, which was given to students as a pre-test and a post-test to measure learning gains; (2) observations; and (3) teacher surveys.

3.6 Final Studies

3.6.1 Math Snacks Animation Study

The Animation Study assessed the impact of *Math Snacks* animations (*Blind Date*, *Atlantian Dodgeball*, *Overruled!*, *Ratey the Math Cat*, *Number Rights*, and *Scale Ella*), and

Table 3 Animation study

	Test 1	Intervention (*8-week period*)	Test 2
Group A	Pre-test	*Six 90 min Math Snacks* lessons *with* district-approved curriculum	Post-test
Group B	Pre-test	Instruction with district-approved curriculum only (No *Math Snacks*)	Post-test

one game (*Pearl Diver*, which was included as a bonus activity because it integrated perfectly with the animation, *Number Rights*). During the fall of 2012, 41 sixth grade teachers from different school districts were recruited. Three teachers dropped from the study due to schedule changes or invalid pretesting conditions. The final group consisted of 38 teachers, 19 in the control group and 19 in the experimental group.

Teachers were assigned to two groups using a stratified random sampling technique. The experimental group integrated *Math Snacks* animations into existing district-approved curriculums, and the control group used only the district-approved curriculum over the same period of time to teach the target concepts of ratio, proportion, scale factor and number line. The research team provided training for the teachers in the experimental group on how to use the snacks. These teachers were also given a set of lesson plan protocols, a set of *Learner Guides* and the *Teacher Guides* for each *Math Snack*, access to the *Math Snacks* Website, and instructional videos on how to teach the animations lessons. The research team was also available when teachers had questions about the delivery of the *Math Snacks* lessons. The control group teachers agreed to teach the concepts related to ratio, proportion, scale factor and number line during the study period, but were not provided with any additional training or support.

There were various types of data collected during this study. All students were given the *Measure of Mathematics Learning I* (Revised to become *MMLII*) and pre-post self efficacy measure. Teachers in the experimental group were observed doing one of the *Math Snacks* lessons and completed a survey after each of the *Math Snacks* lessons. Although the research team collected some valuable information, the findings were limited due to the design of the study.

3.6.2 Math Snacks Game Study

The *Math Snacks* Games Study built on lessons learned in the Animation study. First, in some districts sixth grade classes were self -contained in an elementary setting and in other districts sixth grade classes were in a middle school setting. Second, curriculum maps varied depending on the district, which required teachers to make adjustments. Third, the research design only included one treatment interval. Fourth, observations were only done with experimental teachers to measure fidelity of implementation. Finally, due to the adoption of the Common Core, the concepts covered in *Math Snacks* products shifted slightly from one grade level to another.

The research team decided that the *Math Snacks* Games Study should take place in one school district in order to eliminate the differences in the classroom environments; to ensure that each teacher was following the same curriculum map; while maintaining access to a diverse population of students and teachers. Table 3 illustrates the approach to the games intervention study. The final study initially included 50 fifth grade classrooms from 14 different elementary schools in a low-income, urban school district. Classes were assigned to one of two groups using a stratified random sample, with 25 classes in Group A

Table 4 *Math Snacks* Games Intervention Study: Delayed Intervention Model

	Test 1	First Intervention (*5-week period*)	Test 2	Delayed Intervention (*5-week period*)	Test 3
Group A	Pre-test 1	*Math Snacks* with district-approved curriculum	Post-test 1	Continued instruction with district-approved curriculum	Post-test 2
Group B	Pre-test 1	Instruction with district-approved curriculum only (No *Math Snacks*)	Post-test 1	*Math Snacks* with district-approved curriculum	Post-test 2

and 25 classes in Group B with over 800 students. In the end, 741 students completed the study and two classrooms were eliminated from Group B for not completing the activities.

In addition to changing the grade level, the research team also wanted all classrooms to have access to *Math Snacks* in order to lessen the impact of teacher's teaching style as a variable. The delayed intervention model was used where teachers in Group A were asked to use four lesson plan protocols during the first 5 weeks, while teachers in Group B adhered to the traditional district mathematics curriculum. During the second 5-week period, teachers in Group B used the *Math Snacks* tools and protocols while teachers in Group A returned to the district curriculum. Each lesson protocol included a 30-min game play session with group discussion, a hands-on activity related to game play, and a second game-play session with a final discussion for a total of 90 min per lesson. Teachers and students in Group A had full access to the *Math Snacks* tools during this second five-week period, but *Math Snacks* were not part of in class instruction during that time. Embedded data showed that students did some playing of games out-of class, but not to the extent they played during the initial intervention. This delayed intervention model allowed for three rounds of pre- and post-testing of students using the expanded *Measure of Mathematics Learning II* (Revised the the third version, *MMLIII*), allowing researchers to both compare data from an intervention group to a control group and if the control group would show similar gains when exposed to *Math Snacks*. It also enabled measurement of longer-term retention for students in Group A as seen in Table 4.

The team used the *MMLIII* to measure knowledge gain; a pre and post self efficacy survey to measure student math efficacy; surveyed teachers and students on their perceptions, use and enjoyment of the games; observed classroom instruction for all teachers on two separate occasions; and used embedded game play data. Significant findings are discussed below with overviews of each tool used.

4 Methods, Instruments and Findings

While specific findings have been published in more extensive articles (Wiburg et al. 2016), this section summarizes general findings specific to each method. All of the methods and instruments used in both formative and summative research were used in one of the research environments or studies listed above. Table 5 shows the methods used, where they were used and their role as formative (informing design and development of the *Math Snacks* tools or testing instruments) or summative (assessing the impact of the final tools).

Table 5 Methods in *Math Snacks* research: where and how they were used

Methods	Where methods were used						
	Formative: inform design and development					Summative: assess impact	
	Games Lab "Think Tanks"	Classroom visits	*Math Snacks* camps	Annual panel review	Pilot studies	Animation study	Games study
Observations							
Teacher instruction: OLE 2*			✔		✔		✔
Student gameplay	✔	✔	✔		✔	✔	✔
Focus group and panel discussions							
Teacher: Focus groups			✔				
Student: Focus groups	✔		✔				
Experts: Panel review				✔			
Pre- and post-tests							
Student knowledge gain: (*MML I*, *II* and *III*)			✔		✔	✔	✔
Student: Self-efficacy survey			✔		✔		✔
Self-report surveys							
Teacher: Use of *Math Snacks*			✔		✔	✔	✔
Teacher: Value of *Math Snacks*			✔		✔	✔	✔
Student: Enjoyment of *Math Snacks*			✔		✔	✔	✔
Embedded game data							
Game play: Length of time, location of play							✔
Achievement: Levels achieved and failed							✔

4.1 Observations

4.1.1 Gameplay Observations in the Learning Games Lab

Observations of *Math Snacks* games ranged from watching kids interacting with paper prototypes to working with playable video-game prototypes, through watching students' full versions of games. Most of the observations were short sessions (10–20 min), and involved specific issues, not playability of a full game. Researchers met before each session to clarify specific questions (These inquiries included, “do users now understand how to get through the first level” or “do they know what to do with the half-potions”). By having constant access to the testers, developers were able to test small parts of the game, making changes as needed, then testing again. Most *Math Snacks* games went through 30–50 observation sessions. As the games neared completion, the observations became longer, and included watching how kids progressed through the levels, assessing level difficulty, and testing overall usability.

4.1.2 Classroom Observations

During development, the team conducted observations of gameplay in the classroom often watching students play full working prototypes during class time. These observations worked similarly to those used in the Learning Games Lab, but included whole classes of students. These observations helped developers create tools to support teachers using the games in the classroom including: practice levels where a teacher could play through and show concepts, guidelines to help students with gameplay, and tips on how to manage a classroom of students playing a game.

In the pilot study, observations were done to determine the level in which the teachers in both groups were implementing the *Math Snacks* lessons. Each lesson was scripted and evaluated to determine how teachers were using online tools and whether or not detailed lesson plan protocols were required to maximize student learning.

In the Animation study, researchers only observed teachers who were in the treatment group to measure the fidelity of implementation, but teachers in the control group were not formally observed. Once the study was over, the research team concluded that it would be beneficial to observe all teachers to provide a more accurate picture of the learning environment in all classrooms, regardless of whether or not they were in the treatment group.

During the Games Study, it was determined that each teacher would be observed on two separate occasions. To determine the quality of inquiry-based instruction, researchers used the *OLE2* (*Observation of Learning Environments*, IEMSE 2014) observation tool to assess quality of interaction of the teacher with students, based on an inquiry-based framework. Observers also noted ways in which teachers followed the pre-defined protocol, noting how much game play was allowed, and which of the bonus activities were completed. Each teacher was observed once during the intervention phase, and once during a traditional instruction period.

In addition to providing evidence on teacher quality, the classroom observations were valuable in interpreting other findings, and in anticipating possible limitations to other methods. For example, when interpreting embedded data in the Games Study, researchers found that in some classrooms there was evidence of only half of the students playing the games. Classroom observers reported that in these classrooms, the teacher paired the students on a computer, with only one of the pair able to sign on. Without the observations, researchers would have assumed that students were not exposed to the games at all: because of the observations, researchers can assume that students in class were exposed to the games and either played individually, played in a team, or observed gameplay while another student played. Observations also led to additional research questions. For example, in some classrooms teachers required students to wear headphones or asked them not to talk, which led to very little interaction between students and limited their ability to assist each other. In other classrooms students were sitting next to each other helping each other with strategy and in some cases students were serving as experts walking around giving advice to other students on gameplay and strategy. This information would have been lost without the observations.

4.2 Focus Groups

In the Learning Games Lab, students participated in focus groups during the Think Tanks. Students were asked to provide feedback on graphics or character sketches, scripts, general theming, and content knowledge. Students were asked to share their thoughts, feelings, attitudes and ideas on animations, games, and learner guides throughout development.

During the *Math Snacks* camps, teachers participated in focus groups daily. Once students were excused for the day, the teachers remained to speak with the researchers about the daily activities, plan for the following day, and to discuss how their participation in the summer camp was impacting their feelings about using animations and games for learning. The data from these focus groups were used for multiple reasons. First, the data was used to make adjustments to the camps from day to day and also from year to year. For example, in 2011 teachers said they wanted to teach multiple groups of students instead of interacting with a single group of students for the entire week. In 2012, the camps' structure was adjusted to reflect this request. Second, the data was evaluated to identify patterns in teacher practice. Teachers expressed a shift in their thinking about using games to teach content. They started to see how games could be used as a launch, a key learning activity, and a summary instead of only as a supplemental activity. Finally, teacher focus group data was used to identify student learning patterns. For example, teachers observed that when students were using sound during the games they were more interactive and talked more than when they were using headphones. This led the team to discourage the use of headphones during game play.

In addition to teacher focus groups, during the final summer camp in year 4, focus groups were also held with all of the students by grade level. Although these focus groups provided some interesting feedback, the data was primarily used to triangulate the survey and observation data.

4.3 Expert Panel Review

As planned in the initial research proposal, the *Math Snacks* team convened an annual expert review to look at design plans, prototypes, and final versions of tools and research instruments. This did not change from the original plans, but it was the first time the development team had employed this strategy. The members of the annual panel changed each year, but the panel always included mathematics educators, researchers, reviewers of children's media, and other game developers. The panel reviewed products at various levels of development, including paper prototypes, concept videos, and teacher support tools (including the website, and print guides). An external evaluation partner managed the review process, and compiled findings after each session. Panelists were paid a small stipend to attend. Feedback from panelists offered specifics on the developed products, but through different lenses than other experts. The panelists provided recommendations as gatekeepers, media analysts, and external developers. The team was helpful in identifying specific aspects of game designs that could be controversial or difficult to understand (such as an animation that subtly conveyed all girls are princesses and needed to be rescued), provided outside opinion on if the concept was well addressed by the tool (for example, *Game Over Gopher* needed additional work to reflect constructivist principles of learning), and recommended eliminating development of products that did not seem to *fit* with the existing collection (such as a number line tool that wasn't an animation *or* a game).

Though using an external review panel can be costly, the input was tremendously helpful. Having perspectives from outside math educators and game developers helped the *Math Snacks* team really analyze if their game development aligned with their intentions. The entire *Math Snacks* team (both developers and researchers) agreed the input was valuable, both for the products in development and for the continued professional development of the team.

4.4 Pre- and Post-tests

4.4.1 Knowledge Gain: Measures of Mathematics Learning

As noted earlier, three *Measures of Mathematics Learning* were constructed to measure student math learning.

Measure of Mathematics Learning I: A twenty-seven item measure was constructed from released items that were publically available from national test databases. Three members of the research team located released mathematics items from a variety of sources that included the National Assessment of Educational Progress (NAEP), the Massachusetts Comprehensive Assessment System (MCAS), the Florida Comprehensive Assessment Test (FCAT) and the California Standards Test (CST). Items were selected from grades five through eight and included recognition (i.e., multiple-choice), as well as inference and explanation (i.e., open-ended) items. Project evaluators modified released items after reviewing and addressing issues that may have led to their release. Score reliability was determined during the pilot study using coefficient alpha.

Measure of Mathematics Learning II: The *MMLI* was modified slightly after the pilot study and was used during the Animation Study in 2013. One question was eliminated and another question was revised after it was found to contain multiple correct answers. The *MMLII* was also deemed reliable using coefficient alpha, but after the *MMLII* pretest was evaluated, there was no pretest equivalency between the control group and the experimental group. Even though the students in the treatment group shows significant learning gains over time, the difference when compared with the control group was not significant.

Measure of Mathematics Learning III: The *MMLIII* was developed using the same methodology described for the *MMLI*. In fact, some of the test items from the *MMLII* that covered ratio were used to test the effectiveness of *Ratio Rumble*. Additional questions were added in order to test student understanding of the coordinate plane for *Game Over Gopher*, place value for *Gate*, and base 10 concepts for *Monster School Bus* and piloted during the final *Math Snacks* camp.

4.4.2 Findings from the Games Study on the MMLIII

Three times during the study, students completed the *MMLIII.* This measure was intended to assess short-term progress in discrete areas of mathematical knowledge (ratios, coordinate plane, and number systems including fractions and decimals). In order to test reliability, Cronbach's Alpha was calculated on the pre-test sample at .89. Alpha coefficients above .80 are considered good and values above .90 are considered excellent (Nunnally and Bernstein 1994)

Scores on the *MMLIII* were aggregated for both the control and experimental groups in Table 6. Figure 1 shows a visual display of these same results.

Table 6 Means, standard deviations, and sample size for Groups A and B on tests one, two and three

	Test 1	Phase I (*5-week period*)	Test 2	Phase II (*5-week period*)	Test 3
Group A n = 361	23.64 (8.47)	*Math Snacks* intervention	28.91 (8.99)	District math curriculum	31.58 (8.80)
Group B n = 380	23.84 (8.36)	District math curriculum	26.37 (8.71)	*Math Snacks* intervention	31.10 (8.43)

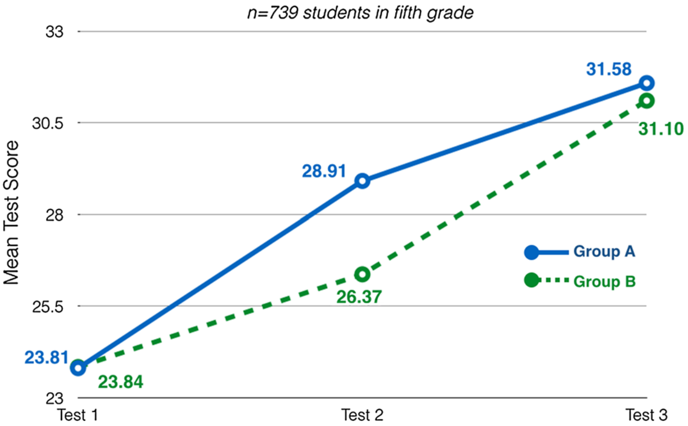

Fig. 1 Mean scores for groups A and B on tests 1, 2 and 3

As Table 6 shows and as Fig. 1 illustrates, student scores from Groups A and B were nearly identical for Test 1. However, following the *Math Snacks Intervention*, student scores from Group A showed significantly greater gains on the dependent measure, *MMLIII,* than those students in Group B, who served as a control group. Test 3 was administered after Group B was given the *Math Snacks* Intervention. On Test 3, students in Group B performed at a level that statistically matched Group A, showing they were able to catch up even though they received the delayed intervention. Test three also showed that Group A maintained their skill levels from the first intervention and continued to make progress.

4.5 Student Surveys

At the beginning of the Animation Study and the Game Study, students completed a survey that included a self-efficacy measure designed by researchers (adapted from Aiken 2002) asking students to rank their certainty in their mathematics skill in specific areas (e.g., addition, division, coordinate plane, etc.) and their opinions about mathematics. At the end of the Game Study, an extended survey was also given to students with questions specifically related to the *Math Snacks* materials. Table 7 reveals findings from the student surveys. Although the responses to the first part of the survey cannot be attributed 100 % to the *Math Snacks* intervention, the questions that refer specifically to the *Math Snacks* games seem to indicate that students felt that the games helped them understand these concepts better.

Table 7 Sample findings from student surveys in games study

How confident are you in solving problems related to the concept? (N = 741 students)

Concept	Pre test (% answered certain or very certain)	Post test (% answered certain or very certain)	% increase
Ratio	32	57	25
Proportion	28	40	12
Coordinate plane	41	73	32

Which *Math Snacks* Concepts did the *Math Snacks* games help you understand better?

Specific content	Percentage of students reporting
Coordinate grid (Game Over Gopher)	79
Decimals (Monster School Bus and Gate)	56
Fractions (Game Over Gopher and Ratio Rumble)	46
Adding and subtracting (Gate)	43
Ratio (Ratio Rumble)	53
Place value (Gate)	41

How do *Math Snacks* lessons differ from regular math lessons? (Themes defined)

"You are using the computer"

"You are learning by playing games"

"They are more engaging/fun"

Did you share *Math Snacks* with others?

68 % of the students reported sharing *Math Snacks* games with parents

66 % of the students reported sharing *Math Snacks* games with friends

4.6 Teacher Surveys

Teacher surveys were conducted during *Math Snacks* camps: the pilot study, the Animation study and the Games study. During the camp, the survey data was used to determine whether or not the summer camp model was effective, solicit feedback on the *Math Snacks* materials, and determine if the teachers had a shift in their thinking about using games in the classroom. During the Pilot Study, the Animation Study, and the Games Study, teachers completed an online survey with both multiple choice and short answer questions after finishing each lesson. This survey was used to triangulate the observations, determine the fidelity of implementation, and to gather data about teacher attitudes toward each lesson. In the Games Study, additional questions were asked about recommendations, future use and a ha moments. Table 8 shows the results for these questions.

Self-reported feedback from teachers in the Games Study revealed that most teachers played each game for less than an hour in preparation for using the game in class with their students. The survey revealed that most teachers followed the protocol as intended, however there were some instances where the teachers were unable to complete the bonus activity. Teachers reported using the bonus activities on average of 86 % of the time.

Table 8 Selected Findings from Self-Report Surveys in Games Study (n = 49 teachers)

Math Snacks Game	Would you recommend this game to colleagues?	Would you use this game in the future?	Did your students have any "a ha" Moments?
Game Over Gopher	98 %	96 % Yes 4 % Maybe	59 % Yes
Gate	100 % Yes	96 % Yes 4 % Maybe	51 % Yes
Monster School Bus	100 % Yes	96 % Yes 4 % Maybe	73 % Yes
Ratio Rumble	100 % Yes	90 % Yes 6 % Maybe 4 % No	71 % Yes

4.7 Embedded Data

In the final Games Study, all *Math Snacks* games were played online through a web browser. During the period of the study, all players could self-identify as part of the study, giving their name, their teacher's name, and noting if they were playing in class or out of class. Data was collected from 690 of the 741 students in the study (93 %). Data collection focused on two specific categories: game play data, which included the length of game play, where students played (in class or out of class); and player achievement data, which included levels achieved or replayed, number of times any level was *not completed*, and the use of specific mechanics within each game.

Embedded data collected during game play revealed students played an average of 4 h and 51 min on all games, though duration ranged from 42 min to almost 14 h, including time students spent playing games at home. Total time spent in games by students in both groups was similar (Group A: 4 h, 53 min; Group B: 4 h, 49 min). Of the students who played games outside of class, on average, 34 % of that time was spent at home (114 of 218 min). The data also revealed a surprising range of game play lengths: students self-reported playing the game for as few as 23 min, and as long as 29 h, 43 min. This data provided quantitative evidence that confirmed data collected using observations, interviews and surveys of students which showed that students enjoyed the games. The embedded data quantified student enjoyment by showing students voluntarily played the game in their spare time, when not asked to by teachers.

Initial analysis indicates a modest correlation with highest level achieved in some games and test scores on the *MMLIII,* with little or no correlation to total time spent on game. This speaks to the type of data that should be collected in future research to determine whether or not level achievement or total time spent playing leads to student learning as measured on an assessment.

Use of the embedded data included complications on identifying *which child is playing,* particularly in situations where students partnered on a single computer. The use of embedded data highlighted the value of classroom observations during the study. As stated earlier, researchers were confused as to why some students had not logged into the system, particularly in one class where about one-third of the students who completed the pre and post test measures had no game play data recorded. In reviewing observation notes, researchers realized that students shared computers every time they played the games, thus capturing name and log in information for only one student in each pair. A required, single login for all games, while not feasible for this project, would greatly strengthen validity of data in tying performance to a specific subject and capturing gameplay from every session.

5 Recommendations

Although the *Math Snacks* team had a defined research protocol, it was necessary to adapt the research plan based on circumstances and opportunities for new data collection methods. After analyzing the changes throughout the project, the team would make the following recommendations for extensive multi-year research and development projects.

5.1 Use Multiple Approaches

While the team originally planned to use both formative and summative methods, the scale of each grew considerably, and the results of these expanded methods helped to provide a complete picture of the data. Beyond simply blending formative and summative assessment design, research in this study was strengthened by using multiple approaches to assess the same intervention: observations can reinforce quantitative data, self-report surveys can validate embedded gameplay data, and tests can confirm focus group recommendations. Throughout this study, researchers utilized multiple approaches, and analyzed data from multiple sources to identify trends, unanticipated outcomes, and findings.

5.2 Secure Subjects and Conduct Research Frequently

Perhaps the most significant change was the increased number of observations, surveys, panels, and other measures used. Conducting formative research was facilitated because the team was able to gain access to teachers and students through Learning Games Lab activities, *Math Snacks* camps, and the classrooms of Teacher Advisors. The relationship with partner school districts was fostered over time and was crucial throughout development and testing. Throughout development, when researchers questioned a strategy, potential impact, or possible use of the tools, they were able to test with students and teachers in informal settings and in classrooms. As developers questioned specific design issues, they were always able to "take it to the students". When you enable access to subjects, formative research is easier, requires less planning, and can be much more responsive to identified needs. All subjects were consented upon recruitment so that any findings could be included in publications.

5.3 Encourage Iterative Development

A significant benefit of frequent formative assessment is the impact it has on iterative development. When testing is frequent, developers can test small parts at a time, making changes and then test again. The Learning Games Lab uses this formative testing extensively in developing the games, but the *Math Snacks* team embraced this in the development of additional learning tools and in creating the teacher support materials. Teachers were given the ability to iteratively develop companion activities and learning tools in *Math Snacks* camps. They also created their own best lesson plan for using the games, then tested them with their students. The game developers used feedback from the teacher observations to modify specific aspects of the game that would not have been noticed without class observations of the teacher practices. The researchers and developers also observed how teachers made modifications to using the game and to additional post-game playing activities. In some cases, observers were able to watch the teachers teach the game for more than one day and were able to notice what changes they made from one day to the next in supervising and facilitating game play, game play discussions, and instructional activities that occurred after the game.

This kind of iteration in design is not new in the design or research fields, but it was a hallmark of progress in the research agenda of this project, and one that was not previously anticipated. In upcoming *Math Snacks* development, the team has planned for similar iterative development opportunities for all products.

5.4 Challenges

It is tempting (and often a requirement of funded programs) that evaluation tools be established at the beginning of projects. While this gives developers a clear idea of the outcomes of the educational games, it limits the iteration that formative testing can bring the tools. Given the success of the research on this project, and in addition to the limitations of each individual method outlined above, this comprehensive research methodology offered challenges as well.

This team enjoyed a surprisingly low turnover rate: the principle investigators, lead researchers, and project managers stayed actively involved with the project for the duration. With this consistent team, it was easier to maintain institutional knowledge of the project, as all members shared the learning curve of what was learned at each stage. Because of the overwhelming amount of data generated by the methods, the team had to streamline documentation for formative work, capturing important takeaways that could be immediately used. This constrained the full analysis often required for publication of formative findings: in short—much of the data collected has not been shared with a larger audience. The team has extensive amounts of survey data, self-report data, embedded data, and observation data that has not yet been analyzed in full.

For other researchers designing an evaluation plan for game based transformations, it may not be realistic to plan for the same 5-year cycle of development and testing which this team enjoyed. However, the same strategies can be effectively used in smaller programs. Game evaluation is strengthened by using multiple measures and assessment approaches, by testing frequently, and by planning for iterative development of research methods.

Acknowledgments This research was supported, in part, by the National Science Foundation (NSF) grant, award number 0918794, *Math Snacks: Addressing Gaps in Conceptual Mathematics Understanding with Innovative Media*. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work reflects the significant efforts of researchers, educators and game developers. The authors wish to acknowledge the efforts and expertise of everyone on the team in developing the intervention; Theodore Stanford, Alfred Valdez and Ken Korn who designed and analyzed much of the measures and findings; and the students, teachers and school districts that participated in extensive user testing and the final research studies.

References

Aiken, L. R. (2002). *Attitudes and related psychosocial constructs: Theories, assessment, and research*. Thousand Oaks, CA: Sage Publications, Inc.

Chamberlin, B., Trespalacios, J., & Gallagher, R. R. (2012). The learning games design model: Immersion, collaboration, and outcomes-driven development. *International Journal of Game-Based Learning, 2*(3), 87–110.

Chamberlin, B., Trespalacios, J., & Gallagher, R. (2014). Bridging research and game development: A learning games design model for multi-game projects. In Mehdi Khosrow-Pour (Ed.), *Educational technology use and design for improved learning opportunities* (pp. 151–171). Hershey, PA: IGI Global.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15–46). New York: Macmillan.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Scardamalia, M., & Bereiter, C. (2008). Pedagogical biases in educational technologies. *Educational Technology, 48*(3), 3–11.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). *Melding the power of serious games and embedded assessment to monitor and foster learning* (pp. 295–321). Serious games: Mechanisms and effects.

Van Meter, P., & Stevens, R. J. (2000). The role of theory in the study of peer collaboration. *Journal of Experiemental Education, 69*(1), 113–127.

Wiburg, K., Chamberlin, B. A., Valdez, A., Trujillo, K. M., & Stanford, T. B. (2016). Impact of *Math Snacks* games on students' conceptual understanding. *Journal of Computers in Mathematics and Science Teaching, 35*(2), 173–193.