The Development of Students' Computational Thinking Practices in Elementary- and Middle-School Classes Using the Learning Game, *Zoombinis*

Asbell-Clarke, J.,[a] Rowe, E.,[a] Almeda, V.,[a] Edwards, T.,[a] Bardar, E.[a], Gasca, S.,[a] Baker, R.S.,[b] & Scruggs, R.[b]

[a]*EdGE at TERC, Cambridge, MA, USA*

[b]*University of Pennsylvania, Philadelphia, PA, USA*

Corresponding Author:

Jodi Asbell-Clarke

EdGE at TERC

Jodi_asbell-clarke@terc.edu

617-873-9716

Abstract

This paper reports on a research study of 45 classes in US schools (grades 3–8) using *Zoombinis,* a popular Computational Thinking (CT) learning game for ages 8 to adult. The study examined the relationship among student gameplay, related classroom activity, and the development of students' CT practices in *Zoombinis* classes. A combination of research methods, including educational data-mining on game data logs, cluster analysis on teacher logs of classroom activity, and multilevel modeling, was used to determine the impact of the duration and nature of student gameplay, as well as the extent and nature of classroom activity, on student CT practices. Automated detectors of gameplay CT practices built for this research were significant predictors of external post-assessment scores, and thus show promise as implicit assessments of CT practices within gameplay. Students with high duration of gameplay and high gameplay CT practices scored highest on external post-assessment of CT practices, when accounting for pre-assessment scores. This research suggests that *Zoombinis* is an effective CT learning tool and CT assessment tool for elementary- and middle-school students.

Keywords: computational thinking, assessment, game-based learning

1. Introduction

The objective of this study was to examine the development of CT Practices among students in 45 classes in grades 3–8 across the US using the learning game *Zoombinis* on tablets or web browsers, along with accompanying teacher-led activities that bridged the implicit *Zoombinis* game-based learning with explicit CT practices in class. We administered a set of pre/post CT assessments

before and after the *Zoombinis* classroom experience to measure the impact of the experience on student CT practices. This paper reports on the relationship among the duration and nature of students' gameplay, the extent and nature of related teacher activity within *Zoombinis* classrooms, and students' development of CT practices as measured by the pre/post assessments. In this study, a set of automated detectors of CT practices from students' *Zoombinis* gameplay logs is compared to external pre/post assessments of the same CT practices. The results of this research provide a better understanding about how CT learning games can improve classroom CT practices and how automated detectors of CT practices within gameplay can be used as a novel form of assessing CT.

1.1 Background on Game-Based Learning and Assessment

Over the past decade and more, Game-based Learning (GBL) has shown to be effective classroom pedagogy for engaging diverse learners in a broad range of complex educational activities including scientific inquiry (Steinkuehler & Duncan, 2008; Asbell-Clarke et al., 2012), argumentation (Bertling, Jackson, Oranje, & Owen, 2015, June), and civics (Stoddard, Banks, Nemacheck, & Wenska, 2016). Digital games engage a broad audience of learners in compelling and often complex play that can foster high-level reasoning, inquiry, persistence, and creativity (Green & Kaufman, 2015; NRC, 2011; Shute et al., 2015). While GBL has attracted many researchers because of the natural motivation and "stickiness" of games, one of the most compelling reasons to use games in classrooms is because of the powerful potential of game-based learning assessments (GBLA). Because digital games allow digital records of players' activity, researchers can use learning analytics to identify common patterns of players' behaviors in the game that are consistent with the intended learning outcomes (Kim, Almond, & Shute, 2016; Fu, Zapata, & Mavronikolas, 2014; Rowe, Asbell-Clarke, & Baker, 2015).

GBLA builds on the idea of stealth assessments (Shute, 2011) where learning is measured within an everyday activity, namely playing a digital game. Plass and colleagues (2015) emphasize that learning and assessment through games rely on a close alignment among: the *game mechanics*—the rules and controls the player interacts with in the game; the *learning mechanics*—the processes through which the designer intends the player to build knowledge; and the *assessment mechanics*—the behaviors that provide evidence of learning in the game. GBLA has often focused on the method of Evidence Centered Design (Mislevy & Riconscente, 2011), where the researchers prescribe a task competency model within the game and collect data to measure student performance directly related to that task.

Our team uses a modified method of Evidence Centered Design to measure implicit knowledge, knowledge that is not articulated by the learner but may be foundational to the development of explicit knowledge (Asbell-Clarke, Rowe, Bardar, & Edwards, 2019; Polanyi, 1966). The measurement of implicit learning is key to understanding how we support and measure learning (Brown, Roediger, & McDaniel, 2014; Underwood, 1996), yet implicit knowledge is not well studied in educational research. Employing novel techniques in learning analytics to measure anticipated and unanticipated evidence of learning help build formative assessments that are more inclusive to a broad and diverse audience of learners (Rowe, Asbell-Clarke, & Baker, 2015; Rowe et al., 2017; Rowe et al., 2019; Shute, Rahimi, & Smith, 2019). The outcome of this method, which is described in detail in Rowe et al. (under review), is a set of validated detectors that can reliably and automatically detect strategies and practices within gameplay that have been identified and coded by teams and researchers through extensive observations.

GBL is particularly powerful in classrooms where teachers actively bridge the implicit learning their students experience in the game with explicit learning through related classroom activities

(Asbell-Clarke et al., 2019; Ash, 2011; Ke, 2009; Lederman & Fumitoshi, 1995). Teachers may build on the games' "aha" moments and help their students make connections between their actions in games and the content being covered in the classroom. This model of bridging draws upon the notion of the "big G" game notion put forth by Gee (2013) that suggests that social game-related experiences occurring outside the game may be critical to the game-based learning, as well as the model of Preparation for Future Learning (PFL) from Bransford and Schwarz (1999), which considers "transfer in" and "transfer out" of knowledge and learning experiences. Transfer in is the prior knowledge learners bring to a learning experience and transfer out is how they apply that learning to other situations. Teachers, peers, or other scaffolds can facilitate the transfer from implicit learning in a game to useful knowledge in the classroom, workplace, or elsewhere, but this is a feat with significant challenges and barriers (Fishman, Riconscente, Snider, Tsai, & Plass, 2014; 2015).

1.2 Background on Computational Thinking

First introduced by Jeanette Wing (2006), the term CT was originally described as the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent (Cuny, Snyder, & Wing, 2010). As the need for a computationally literate workforce grows, CT is attracting increased attention in K–12 education, prompting a call for new models of pedagogy, instruction, and assessment (Barr & Stephenson, 2011; Grover & Pea, 2018; Shute, Sun, & Asbell-Clarke, 2017). The role of CT in K–12 education has been described as laying "the conceptual foundation required to solve problems effectively and efficiently (i.e., algorithmically, with or without the

assistance of computers) with solutions that are reusable in different contexts" (Shute, Sun, & Asbell-Clarke, 2017).

CT has roots in early work such as Seymour Papert's research on procedural thinking exhibited in the early programming environment for children called LOGO (Papert, 1980; Papert & Harel, 1991), but CT encompasses much more than programming in today's technological society. CT is often described as a set of thinking practices that may include: problem decomposition, abstraction, algorithmic thinking, conditional logic, recursive thinking, and debugging. There is evidence that these CT practices may support a variety of other cognitive and non-cognitive activities, especially for learning in STEM subjects (e.g., Barr & Stephenson, 2011; Sneider, Stephenson, Schafer, & Flick, 2014).

The operationalization of CT practices in *Zoombinis* gameplay in this study focus on four CT practices outlined by CSTA (2017) and Shute, Sun, & Asbell-Clarke (2017). While not an exclusive definition of CT, a focus on these practices lays a strong foundation for CT which may also include applications such as coding, debugging, and modeling (CSTA, 2017):

- *Problem Decomposition* is reducing the complexity of a problem by breaking it into smaller, more manageable parts.
- *Pattern Recognition* is seeing trends and groupings in a collection of objects, tasks, or information.
- *Abstraction* is generalizing from observed patterns and making general rules or classifications about objects, tasks, or information by discerning relevant from irrelevant information.
- *Algorithm Design* is establishing reusable procedures that solve sets of problems.

1.3 Background on Computational Thinking and Games

A variety of digital and non-digital games has been used to engage learners in CT practices as well as computer programming. Weintrop and colleagues (2016) found that the CT practices learners developed in construction games, computational problem-solving that was distinct from coding, became central to the way learners reflected and created code and learned programming concepts. Some games have used a robotics context (Berland & Lee, 2011; Kazimoglu, Kiernan, Bacon, & Mackinnon, 2012). Studying computational problem-solving in a racing game, researchers studied procedural thinking (the breaking down of problems into steps), which is strongly related to the CT practice of Problem Decomposition (Holbert and Wilensky, 2010). Researchers have also used video coding to study how players' CT practices developed with during iterative problem-solving in battles (Holbert, 2013). This literature on CT, game-based learning assessments, and CT in games lays the groundwork for the following research study of the game *Zoombinis* with elementary- and middle-school classes (grades 3–8).

2. Description of the Materials

2.1 Description of the Game

In the 1990s, two educational designers (Scot Osterweil and Chris Hancock) saw the need for a learning game that helped build problem-solving skills dealing with computer science, the practices and skills later coined as Computational Thinking (CT) by Wing (2006). *The Logical Journey of the Zoombinis* was the first in a series of three computational thinking games they designed. In 2015, the popular, award-winning, learning game was renamed to simply *Zoombinis* and was relaunched for mobile and desktop platforms, and a web version was developed soon after as part of this research (TERC, 2019).

Zoombinis consists of a series of 12 puzzles, each with four levels of complexity, in which players are charged with bringing packs of Zoombini characters (16 at a time) to safety. Each Zoombini has one of five different types of hair, eyes, nose, and feet (Figures 1). In many of the puzzles, such as Allergic Cliffs (Figure 2) and Bubblewonder Abyss (Figure 3), players use these combinations of attributes to solve puzzles that require sorting, matching, and sequencing of the Zoombinis. Other puzzles in the game apply similar logic and CT practices in different contexts, such as Pizza Pass where players identify the exact combination of pizza toppings to satisfy hungry, but picky, trolls (Figure 4), or Mudball Wall where players complete a multi-dimensional grid with paint balls of different shapes and colors (Figure 5).

This study had students focus on four puzzles in the game: Allergic Cliffs, Pizza Pass, Mudball Wall, and Bubblewonder Abyss, though students were able to play the entire game if desired. Our research team only had time to build CT detectors for the first three of these puzzles.

Figure 1: Image of Zoombinis

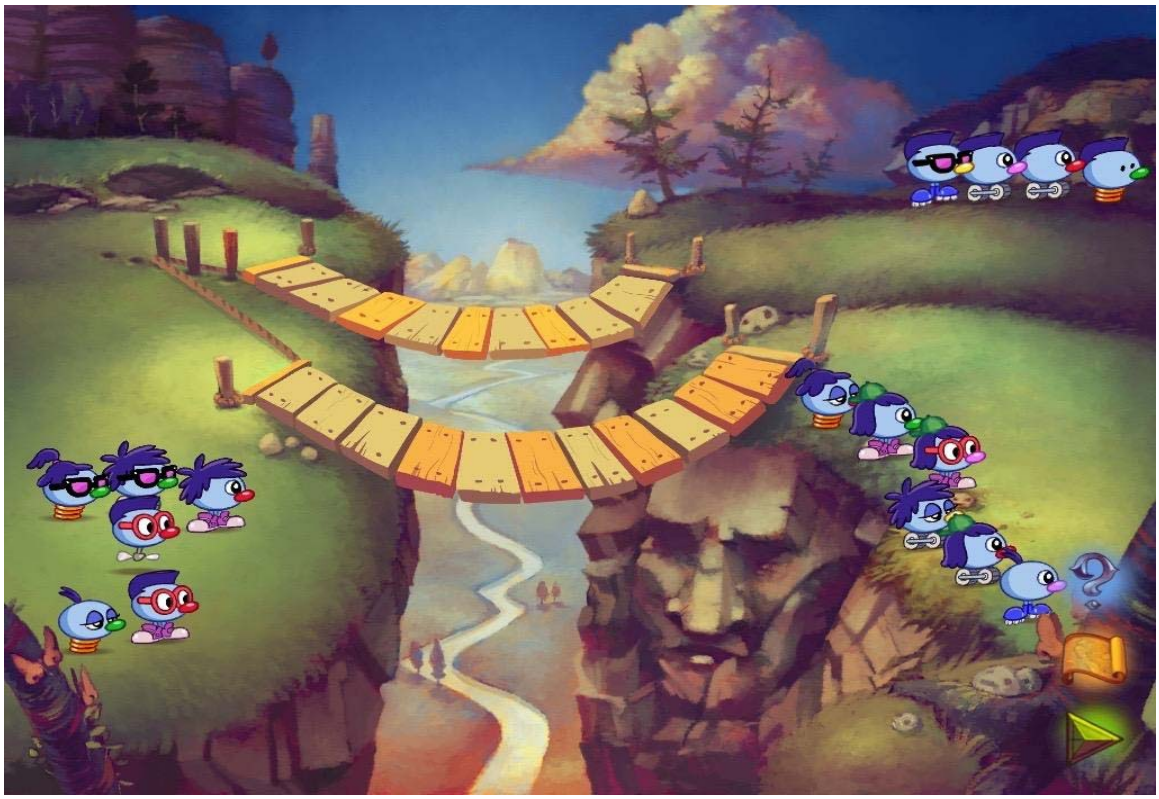

Figure 2: Allergic Cliffs Puzzle where the Zoombinis with flat top hair have all crossed the top bridge and all other values of hair have crossed the bottom bridge. One can infer at this point that the bottom bridge is allergic to flat-top hair.

Figure 3: Bubblewonder Abyss puzzle is a maze with safe and dangerous paths that depend on Zoombinis' attributes, and can sometimes be switched with use.

Figure 4: Pizza Pass Puzzle where each troll likes a specific combination of pizza and sundae toppings. Those in the pit in front are rejected, and those on the rocks in back have some toppings they like but not everything they like.

Figure 5: Mudball Wall Puzzle where paintballs of colors and shapes are launched to hit cells with dots, which then launch the Zoombinis to safety.

2.2 Description of Classroom Bridging Activities

To prepare teachers for bridging during our implementation studies, we gave them examples from gameplay along with discussion activities that leverage ideas from the gameplay as described in detail by Rowe, Bardar, and Asbell-Clarke (2015). Teachers were also encouraged to blend in other CT activities such as coding, as desired. Previous research examples of bridging activities included (excerpted from Asbell-Clarke et al., 2019):

- Showing video clips from gameplay as part of class discussions of strategies
- Playing the game jointly and discussing strategies and concepts

- Physically re-enacting puzzles or scenarios from the game in the classroom
- Explaining classroom content in context of game storylines, characters, and/or strategies
- Connecting game storylines and puzzles to real-world examples
- Playing board games with similar learning mechanics and discussing similarities
- Using scaffolds during gameplay that overlap with tools used in other contexts, such as data tables and charts
- Using clear, consistent terminology across gameplay and non-gameplay
- Taking on the role of the game to more deeply understand the underlying rules and how they relate to classroom content.

In the study, some of the teacher activities were closely connected to the game, such a physical recreation of a *Zoombinis* puzzle where some students enact the rules underlying the puzzle (e.g., which attributes are rejected by which bridge), and other students are the Zoombinis who are attempting to cross. This type of bridging activity allows collaboration, and sharing of practices and information that may not take place in online gameplay and may support some students who need help. Other bridging activities included video walkthroughs of the puzzles where teachers could suggest ways of organizing data in a data table for more effective puzzle-solving.

Bridging activities also encourage conversations in class that allow the teacher to become very explicit with CT, naming the problem decomposition, pattern recognition, abstraction, and algorithm design as they see it. For example (excerpted from Asbell-Clarke et al., 2019), a teacher might ask the following types of questions (with typical student responses in italics).

Problem Decomposition: What is the big problem we want to solve? *To get our Zoombinis across the bridges.* So what part of that problem do we want to solve first? *Let's see if the blue shoes will go over the top bridge.*

Pattern Recognition: Do we have enough information to recognize any patterns in what we've already done? *The blue shoes are all making it over the top bridge.* Teachers may introduce conditional language here. They can structure their questions to lead students towards saying "IF the Zoombini has blue shoes, THEN it will go over the top bridge."

Abstraction: Given the pattern that we see, what do we think the general rule is for this puzzle? *Zoombinis with blue shoes go over the top bridge, and all other Zoombinis go over the bottom bridge.*

Algorithm Design: What types of strategies do you use that help you solve this puzzle? *I always start with shoes and go through all the different types first, then I try noses, then hair*. A teacher can use this as an opportunity to describe this as an example of an algorithm that the student has designed in their head.

Other bridging activities included activities that connected the CT practices in *Zoombinis* to other contexts such as collecting data in a science experiment, solving a math problem, or other puzzle games (e.g., *Mastermind, Guess Who?).* During the study, teachers were also encouraged to use other CT activities of their choosing. Some teachers used online and/or offline activities from *Hour of Code*, or used the *Scratch* programming environment.

2.3 Building Implicit Learning Assessments in *Zoombinis*

To build implicit learning assessment mechanics for the key CT practices in *Zoombinis*, we used the model by Plass and colleagues (2015) to identify specific game mechanics, learning mechanics, and assessment mechanics that are aligned with the CT practices in each puzzle. The CT practices studied are: Problem Decomposition, Pattern Recognition, Abstraction, and Algorithm Design. Table 1 shows how each of these practices are operationalized in the game mechanics (what the player does) with the learning mechanics (what the designer intends the player to learn), and the assessment mechanics (evidence within gameplay that reveals learners' understanding).

Table 1: Operationalization of CT Practices in *Zoombinis* shown through the alignment of the game mechanics, learning mechanics, and assessment mechanics of the game.

	Game Mechanic	Learning Mechanic	Assessment Mechanic
Allergic Cliffs	Figure out which Zoombini attributes can cross which bridge without getting "sneezed" back. In the example shown (Figure 2), the bottom cliff is allergic to flattop hair and the top cliff is allergic to all other hair types.	• Decompose problem into attributes and values. • Systematically test for patterns of values of attributes that cross each bridge. • Abstract patterns to a general solution about the attributes. • Design algorithms to solve similar problems in repeated puzzles.	• Holding one attribute constant while testing others. • Testing one value of each attribute. • Continuing with one value/attribute until all crossed or one is rejected.
Pizza Pass	Figure out what type of pizza (and ice cream sundae in higher levels) will satisfy the hungry trolls (Figure 3). They will each accept only a unique combination of toppings.	• Decompose problem into toppings for each troll. • Systematically test for preferred combinations of patterns. • Abstract patterns to an acceptable recipe for pizza and ice cream. • Design algorithms to solve similar problems in repeated puzzles.	• Testing one topping at a time. • Adding one topping at a time cumulatively. • Starting with all toppings and reducing one at a time.
Mudball Wall	Figure out how colors and shapes of mudballs correspond to rows and columns on the wall (Figure 4), then hit the cells with dots.	• Decompose grid into rows and columns. • Systematically test for pattern of color/shape with row/columns. • Abstract patterns to hit all the cells with dots.	• Testing one row/column at a time. • Using diagonal pattern to get row/column info simultaneously.

		• Design algorithms to solve similar problems in repeated puzzles.	
Bubblewonder Abyss*	Get Zoombinis through a maze with junctions and switches triggered by Zoombini attributes (Figure 5).	• Decompose maze into successful and unsuccessful paths • Predict sequence of each path for constant and variable danger points. • Abstract solutions for safe attributes of Zoombinis for each path. • Design algorithms to solve similar problems in repeated puzzles.	• Grouping Zoombinis by attribute for sequencing • Holding one attribute constant, or sequencing back and forth with a trigger.

* While Bubblewonder Abyss was part of the participants' activity and data were collected, we were unable to build implicit learning detectors of CT practices in Bubblewonder within the scope of this project.

3. Research Methods

3.1 Research Questions

The research questions explored in this study include:

1. How is student development of CT practices impacted by:
    a. The duration of a student's gameplay in *Zoombinis*?
    b. The CT practices exhibited by the student in their *Zoombinis* gameplay?
    c. The extent and nature of CT classroom activities reported by the teacher during the study period?
    d. Student and classroom characteristics?
2. Do the automated detectors built to study the CT practices exhibited by the student in their *Zoombinis* gameplay significantly predict student development of CT practices?

Based upon previous research, we hypothesize that students' CT practices measured outside the game would be positively impacted by longer duration of student gameplay and by higher exhibition of CT practices in the game. We also hypothesize that greater extent of teachers' bridging activities would positively impact students CT practices, but we do not have reason to believe any one type of bridging activity will be more impactful than others. That is an open

question of this research. Finally, we hypothesize that there may be differences in student outcomes related to gender (as girls are typically underrepresented in computer science), as well as whether or not the student is on an Individual Education Plan (IEP) for academic difficulties.

3.2 Data Sources

To study these questions, we collected digital gameplay logs, teachers' daily logs of classroom activity, student demographic data, and pre/post external assessments of students' CT practices. We used these data to measure:

- Student Gameplay Duration—the total amount of time each player spent playing *Zoombinis*
- Gameplay CT Practices—the strategies and implicit CT practices that were evident in student gameplay logs
- Extent of Bridging in Class—the amount of class-time teachers spent on activities related to CT
- Nature of Bridging in Class—the type of CT-related class activities reported by teachers
- CT Outcomes – Changes in students' CT practices measured independently of the game.

3.3 Research Sample

The initial sample of students consisted of 1,271 students belonging to 36 teachers across 57 classes. To participate, teachers met the following criteria during the study period:

- They are an elementary- or middle-school educator (grades 3–8) in the US.

- They teach at least one class that supports CT through logic, coding, or preparation for coding (e.g., math, science, computer science, tech. ed., etc.).
- Their students have access to Internet-enabled computers to take the pre- and post-assessments required for the study.
- They complete a teacher agreement outlining the study requirements.
- They obtain administrative approval to participate in the study.

From the initial sample, 495 students (39%) were dropped because they were missing either pre- or post-assessment scores. This included cases where the student completed items for all four CT Practices on the pre-assessment, but only completed the first item on the post. An additional 60 students (5%) were excluded because of other missing data such as missing teacher logs or game data. Further details can be found in an online appendix at https://bit.ly/2X19C1H.

The final sample of students consisted of 716 students (476 elementary students, 48% female; 240 middle school students, 40% female) with complete pre- and post-assessments belonging to 32 teachers across 45 classes. The final student sample was about 66% elementary students, 46% female, and 46% enrolled in Title I schools. The majority (76%) of the 32 teachers taught in public schools in 10 states, with 72% reporting more than 10 years of teaching experience.

3.4 Organization and Labeling of Gamelog Data

All student assessment and gamelog data were collected through our team's game data-collection architecture, which was purposefully designed and built to collect, organize, and visualize data collected from game activity. As part of this architecture, an API is built into the game, allowing each player's game activity and every corresponding game event to be logged and associated with

a timestamp and a unique (and anonymous) player ID. Over multiple GBLA studies, the authors have designed a suite of tools with the data architecture to enable:

- Registration of players by classes or individuals
- Synchronization of game data with other sources (e.g., surveys, external pre/post assessments, and multimodal sensory data streams)
- Visualization of game replay using data logs
- Preparation of game data for use by automated data-mining detectors.

As part of previous GBLA studies, we designed a playback tool for visualization of the gameplay data to allow efficient human analysis of gameplay. The playback tool uses the gameplay data log to recreate and display the game in a window with a series of menus below that researchers use for hand-labeling of the data. Unlike on video recordings, with the playback tool researchers can easily scrub through the playback timelines to find events and the playback tool snaps to an event to avoid time-consuming and tedious time synchronization tasks. Researchers can also customize the labeling tool for different puzzles and different games. The playback tool was used for the hand-labeling of all *Zoombinis* gameplay, allowing a broader sample than with video recordings and more extensive hand-labeling. The resulting hand-labeling of the *Zoombinis* data was used as empirical grounding to build automated detectors of the strategies and CT practices evident in student gameplay. Thousands of rounds of gameplay were double-labeled by independent researchers, resulting in a set of kappas for each label (Rowe et al., 2017; 2019). The hand-labeling of the *Zoombinis* data was used as empirical grounding to build automated detectors of the strategies and CT practices evident in student gameplay. Only labels with kappas exceeding 0.70 were used in the automated detectors.

3.5 Measures

3.5.1 Student Measures

3.5.1.1 IACT Assessments

To serve as external pre- and post-assessments of CT, the authors worked with a game-based learning company to design Interactive Assessments of CT (IACT)—a set of logic puzzles to assess CT practices in upper elementary and middle school (see Asbell-Clarke et al., in review). We designed our own assessments because we could not find established instruments in the beginning of the study that focused on the four CT practices of interest. Also for inclusivity, we were particularly interested in building assessments that would look at implicit CT practices without relying on text or coding, which may be a barrier for some students. In a larger study using an augmented sample, we found moderate evidence of concurrent validity (r=0.29 with teacher ratings in the *Zoombinis* sample; r=0.40 with Bebras items among students in grades 5–8 in the RPP samples) and test-retest reliability (r=0.55 and 0.34 for aggregated measure in *Zoombinis* and RPP samples, respectively) for IACT. A more comprehensive summary of these findings and the limitations of this measure are described in detail in Asbell-Clarke et al. (in review).

For the study reported on in this paper, pre/post assessment scale scores were calculated for the IACT logic puzzle items as the means of items per CT practice: Problem Decomposition, Pattern Recognition, Abstraction, and Algorithm Design. Table 2 shows the scoring for each CT practice.

Table 2: Scoring of IACT assessment items for each CT practice

CT Practice	Number of items	Measure used for scoring
Problem Decomposition	4	Mean efficiency (minimum number of moves needed to solve / number of moves taken)

Pattern Recognition	5	Mean number of correct responses
Abstraction	6	Mean percentage of array spaces completed correctly
Algorithm Design	3	Mean efficiency (minimum number of moves needed to solve / number of moves taken)
Aggregated CT	18	Average of Z-scores of 4 CT measures above

The items on the middle-school form of IACT were designed to be more difficult than the items on the elementary form, involving larger grids and longer sequences of moves to complete. For this reason, the standardization of each CT practice was calculated using the means for each form. This ensured each form had a mean of 0 and a standard deviation of 1. Tables 3 and 4 present the means and standard deviations used to standardize the elementary- and middle-school forms on the IACT pre and post-assessments.

Table 3: Means and standard deviations of IACT pre-assessment by form

CT Practice	Elementary Form (N=1460 to 1523)		Middle-School Form (N=784 to 893)	
	Mean	S.D.	Mean	S.D.
Problem Decomposition (mean efficiency)	0.90	0.12	0.91	0.13
Pattern Recognition (% correct)	0.73	0.25	0.33	0.24
Abstraction (% spaces correct)	0.91	0.13	0.66	0.24
Algorithm Design (mean efficiency)	0.86	0.20	0.75	0.26

Table 4: Means and standard deviations of IACT post-assessment by form

CT Practice	Elementary Form (N=1083 to 1174)		Middle School Form (N=546 to 599)	
	Mean	S.D.	Mean	S.D.
Problem Decomposition (mean efficiency)	0.94	0.09	0.95	0.09
Pattern Recognition (% correct)	0.77	0.23	0.36	0.25
Abstraction (% spaces correct)	0.93	0.11	0.68	0.26
Algorithm Design (mean efficiency)	0.90	0.16	0.84	0.23

The first step in creating an aggregate measure of CT was standardizing the means of each item type to produce a Z-score for each CT practice on each form (elementary vs. middle school). This is to take into account that the middle-school form was designed to be more difficult than the elementary form.

The second step in creating an aggregated measure was averaging standardized means from each CT practice. The units shown in Table 4 are the number of standard deviations from the mean Z-score of the four CT practices. There were no significant differences by form in the standardized IACT pre- and post-scores.

To examine measurement invariance across form (elementary vs. middle school) and time (pre and post), we performed 3 types of preliminary analyses. First, the correlations between each CT practice were correlated with the aggregate CT measure. Second, we compared test-retest reliability of the aggregate measures. Third, to assess concurrent validity, we correlated the aggregate measure with external teachers of their students' CT practices. Each of these analyses was done across for each time*form combination.

The correlations of each CT practice with the aggregate CT measure were comparable across forms and times, exceeding 0.60 in all cases. Further details can be found in an online appendix at https://bit.ly/2X19C1H. Only the correlations between Algorithm Design on the post-test significantly differed by form (0.6 for elementary; 0.75 for middle school).

For test-retest reliability and concurrent validity, we found significant differences by form. In both analyses, the correlations were stronger on the Elementary form than the Middle School form suggesting results from these measures might be more valid for elementary students than middle school students (see Limitations).

3.5.1.2 Duration of Student Gameplay

Each time a student opened *Zoombinis,* that time was added to a cumulative duration of Student Gameplay. This does not guarantee students played the entire time the game was open, but our assumption is the majority of this time was spent playing. Students in this study played *Zoombinis* between 19 minutes and 658 hours over the course of the school year (Figure 6), with a mean of 21 hours and median of 4.8 hours. As the difference between the mean and median suggest, there were 8 students with outlying values (above the 99th percentile). To limit the influence of outlier values, we tried mixed models using several different game duration values to split students into groups (quartiles, median, 1 hour, 2 hours, 5 hours). The 2-hour boundary provided the largest improvement in model fit and was retained.

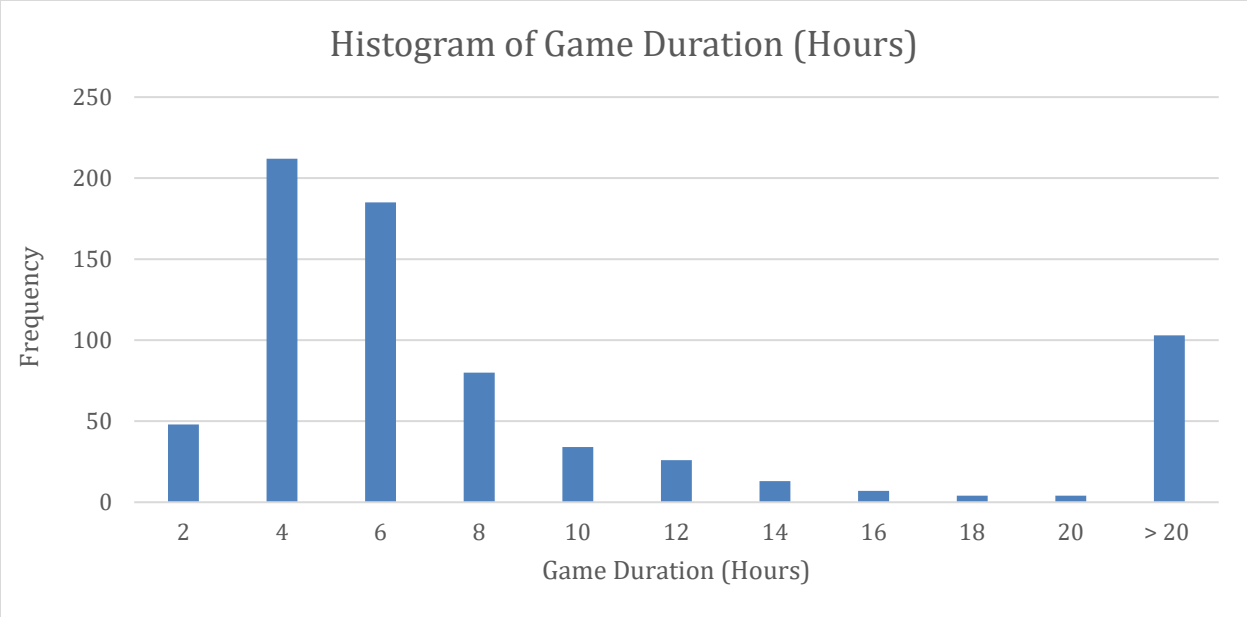

Figure 6: Histogram of Game Duration

3.5.1.3 Student Gameplay CT Practices

As described in 1.3, 2.2, and 2.3, researchers created implicit measures of Students' Gameplay CT Practices using an emergent GBLA method developed and refined with other games (Rowe et al., 2019). There are six key steps in this process (Figure 7):

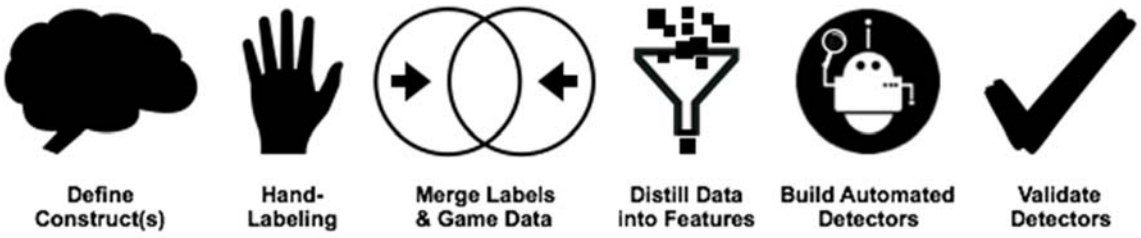

Figure 7. Graphical representation of the six-step emergent approach to GBLA (Source: Rowe et al., 2019)

Although the bridging materials and classroom implementation of *Zoombinis* included four puzzles, the research team only hand labeled the data from three of these puzzles. As described in more detail in Rowe et al. (under review), researchers reliably labeled the four CT practices of

interest in students' gameplay in the three puzzles: Allergic Cliffs, Pizza Pass, and Mudball Wall. We also reliably labeled instances when students' behavior suggested they were still learning the game mechanic for each puzzle (e.g., repeating the same pizza in Pizza Pass or same mudball in Mudball Wall). We combined the evidence of strategies, CT practices, and learning game mechanic from human-labeling with salient features from the game logs to build automated detectors for each CT practice and Learning Game Mechanic across these three puzzles.

For each game behavior, we attempted to fit the detectors using four common classification algorithms (W-J48, W-J-Rip, Step Regression, Naïve Bayes) used in detecting affect and engagement in computer-based learning environments (Kai et al., 2015; Paquette et al., 2016) and in our prior work detecting implicit physics learning (Rowe et al., 2017). To the best of our knowledge, no other study has reported automated CT practice detectors from gameplay. These classification algorithms allowed us to predict whether or not a student is demonstrating CT practices as identified with the human-applied labels. Early results with the Mudball Wall detectors performed at level of quality higher than seen in many medical applications (Almeda et al., 2019; Rowe et al., 2019), suggesting that these could be used to deploy as in-game assessments to reveal students' implicit learning. Subsequent correlations of all Student Gameplay CT Practices detectors and external CT assessments were found to provide moderate to strong evidence of convergent validity (Rowe et al., under review). The Learning Game Mechanic detector is negatively correlated with the CT practice detectors ($r=-0.36$, $p<0.001$), suggesting students who were still learning the game mechanic were less likely to display evidence of CT in their game behaviors.

The detectors of Student Gameplay CT Practices were applied to the final sample of *Zoombinis* gameplay logs from 716 students in this study. Our validated detectors do not only produce an

inference of whether implicit CT practices are present or absent in student gameplay logs, but also produce a confidence in that inference. For example, if the Problem Decomposition detector has a confidence of 80% for a round of *Zoombinis* gameplay, this indicates that there is an 80% probability that the student was demonstrating this CT practice in that round. When averaged across all of their play, average confidences indicate higher prevalence of the CT practices in their gameplay.

For this study, we created aggregated measures of each CT practice using the confidence levels from the detectors in the following ways.

Problem Decomposition: We averaged across the three puzzles, including the two detectors within Mudball Wall (implicit and explicit problem decomposition).

Pattern Recognition: We averaged across detectors, one per puzzle.

Abstraction: We averaged across detectors, one per puzzle.

Algorithm Design: For each puzzle, we found the maximum detector confidence from all strategies, with the exception of two detectors. 2-D Pattern Completer, a Mudball Wall detector, was discarded based on results presented in Rowe et al. (under review). Hold Attribute and Hold Value Constant, an Allergic Cliffs detector, was also not included because it did not reliably distinguish which students were demonstrating this strategy. We then averaged those confidences across the three puzzles as an indicator of their strategy (algorithm) used across puzzles during gameplay.

Because of the significant negative correlation between Learning Game Mechanic and the four CT Practice detectors, an aggregate Learning Game Mechanic measure was created by averaging the detectors across the three puzzles and then subtracting from 1 (i.e., reverse coding). This results in a fifth CT Practices Detector where 0 means CT practices are absent and Learning Game Mechanic

is present and 1 means CT practices are present and Learning Game Mechanic is absent. These will be referred to as the Student Gameplay CT Practices.

We created a mean confidence level (mean=0.63, S.D.=0.09) across all five detectors based on the average confidence of each CT practice. The greater the average confidence, the higher the likelihood that students were exhibiting in-game behaviors consistent with CT practices and not exhibiting evidence they were still learning the game mechanic.

3.5.2 Measures of Student, Classroom, and School Demographics

We collected the following types of data from teacher application surveys and the Common Core of Data (Common Core of Data, 2020):

Student demographics: Gender

Classroom demographics: Subject Area (computer science, technology/robotics, math, science)

School demographics: School Type (elementary or middle school), Title I school status.

These measures were included in descriptive and multi-level analyses as control variables that may have potential relationships to the student outcome measures. Most important among them is the school type because a more difficult form was administered to middle-school students and easier forms were administered to elementary students. As described above, the metrics were different across forms, so the standardization was done separately for elementary- and middle-school forms.

3.5.3 Measures of Classroom Activity

Guided by our prior experience with teachers logging about their instructional use of a physics game (Rowe, Bardar, Asbell-Clarke, Shane-Simpson, & Roberts, 2016), we adopted a quantitative approach to measuring classroom activity through multiple-choice questions with predefined answers about the amount of time spent on each type of activity. Teachers were asked to complete these logs for each day they were implementing *Zoombinis*.

We used these teacher logs to identify clusters of classes based upon the extent of their activity related to *Zoombinis*, which we call Bridging, and the extent of their use of Other CT Activities. Teachers completed 938 daily logs for their 57 classes in NoviSurvey software. Twelve of these classes were in the clustering analyses but did not have enough students with complete data to remain in the multi-level analyses. To identify groups of classes, we used a K-means clustering process where those objects belonging to the same cluster share similarities in attributes (Witten & Frank, 2002). Here, those objects are *classes* and attributes are the teachers' reported use of *specific instructional activities* in those classes. The ultimate goal of clustering is to partition our sample into clusters of classrooms that reported use of specific instructional activities in a similar pattern, and to characterize this response pattern that distinguishes each cluster. K-means clustering creates a protype class and classifies each class according to the prototype it most resembles. This procedure maximizes differences between clusters and minimizes within-cluster variance, helping avoid any potential multicollinearity issues in using raw data.

We initially anticipated clusters would form around the nature and extent of the teachers' instructional activities (e.g., classes would differ on both the type and duration of specific *Zoombinis* bridging activities that were used), but this was not born out in the clustering analyses. Instead, classes were grouped more by their extent (duration) of *Zoombinis* bridging activity and less by the nature of that bridging activity. This means teachers distributed their class time across the various types of bridging activities in roughly the same proportion of class days, regardless of how many classes in which they used *Zoombinis*.

To capture the extent that students participated in classroom activities related to CT practices, we created two clusters of classroom activities: (1) *Zoombinis* Bridging and (2) Other CT Activities. *Zoombinis Bridging* includes all classroom use of *Zoombinis* activities as well as discussions of

gameplay, be they whole class or small group. *Other CT Activities* is a measure of the extent to which teachers reported using CT classroom activities other than the *Zoombinis* game or *Zoombinis* bridging activities.

3.5.3.1 Extent of Zoombinis *Bridging*

To create the bridging clusters, we used K-means clustering with two clusters of instructional activity features derived from the teacher logs. Table 5 presents the mean of the final 16 features used to create the bridging clusters.

Table 5: *Zoombinis* Bridging Clustering Results

Zoombinis Bridging Activity Feature	Cluster Means	
	Low Duration (N=29)	High Duration (N=25)
Sum of Minimum *Zoombinis* Class time Activities (Hours)	3	8
# Class Days with *Zoombinis* Walkthrough Videos	2	4
# Class Days with *Zoombinis* Bridging Activities	4	11
# Specific *Zoombinis* Activities (videos & activities)	6	10
# Walkthrough Videos Watched	2	3
# Specific *Zoombinis* Activities	3	7
Percent of *Zoombinis* Walkthrough Videos Watched	56%	75%
Percent of Specific *Zoombinis* Activities Used	19%	41%
# Class Days with Whole Class *Zoombinis* Gameplay	2	4
# Class Days with Small Group *Zoombinis* Gameplay	1	5
# Class Days with *Zoombinis* Gameplay Busywork	0	1

# Class Days with *Zoombinis* Gameplay Assigned by Teachers	2	5
# Class Days with no *Zoombinis* Gameplay	4	5
# Class Days with Whole Class Discussion of *Zoombinis* Gameplay	3	5
# Class Days with No *Zoombinis* Discussion	4	6
# Class Days with *Zoombinis* Activities	4	8

The main distinction between the bridging clusters is the number of class days spent doing each of the bridging activities (*Zoombinis* activities, walkthrough videos, discussions of gameplay, assigned gameplay). All t-test results were statistically significant at $p<0.05$. Students in High *Zoombinis* Bridging classes participated in a higher proportion of the *Zoombinis* activities, watched a larger proportion of the walkthrough videos, and, on average, had twice as much class time spent on *Zoombinis* bridging and gameplay than students in Low *Zoombinis* Bridging classes.

We hypothesize that students in High *Zoombinis* Bridging classes will show greater improvement in their CT practices (as measured by IACT) than students in Low *Zoombinis* Bridging classes because of the importance of bridging we have found in prior research (Asbell-Clarke et al., 2019) and exit interviews with study teachers conducted by external evaluators (Barchas-Lichtenstein et al., 2019).

3.5.3.2 Extent of Other CT Activities

Most of the non-*Zoombinis* related CT activities reported by teachers centered around coding or programming activities. Teachers reported an approximate time of each class spent on coding or programming. The total amount of time spent across classes was summed using the minimum time in each range (i.e., 10–20 minutes became 10 minutes) to provide a conservative estimate. Each class was classified by whether or not it had less than 10 minutes devoted to coding across all

classes. Similarly, each class was labeled as having ever integrating block-based programming or Code.org activities. The remaining features used in this measure included the logging of the absence of *Zoombinis* bridging and gameplay.

Table 6 presents the 10 final classroom features used to create two other CT activity clusters using K-Means. The Low Other CT Activity group consisted of 40 classrooms with relatively little time spent coding or programming, relatively little block-based programming, and very few classes without *Zoombinis* gameplay or bridging. The High Other CT Activity group included 14 classrooms where teachers reported greater use of other CT activities, such as primarily coding/programming.

Table 6: Other CT Activity Clustering Results

Other CT Activity Feature	Cluster Means	
	Low (N=40)	High (N=14)
Sum of coding class time across all classes is less than 10 minutes	50%	0%
Any Block-Based programming Ever (1=Yes)	38%	79%
Any *Hour of Code* or *Code.Org* Activities Ever (1=Yes)	5%	36%
Sum of minimum class time on coding/programming activities (hours)	0.4	3.5
Sum of minimum class time on non-*Zoombinis* CT discussion (hours)	0.3	0.5
Computer Science course (1=Yes)	10%	64%
Percent class days no *Zoombinis* discussion	22%	51%
Percent class days no *Zoombinis* gameplay	20%	49%
# class days no *Zoombinis* activity (bridging or gameplay)	2	9

Percent class days no *Zoombinis* activity	14%	45%

All t-test results were statistically significant at $p<0.05$ except the amount of class time spent on non-*Zoombinis* CT discussions where both clusters averaged less than an hour. Because of the strong association that is inherent between CT practices and coding/programming, we anticipate students in High Other CT Activity classrooms will improve more in their CT practices than students with more limited other CT activities such as coding/programming.

In summary, cluster analysis identified two groups (High/Low) of classes along two different dimensions: the extent of *Zoombinis* Bridging, and the extent of Other CT Activity.

3.6 Data Analysis

Descriptive analyses examined relationships between the following sets of measures:

- Student pre- and post-assessment scores (IACT Z-scores)
- Student, classroom, teacher, and school demographics
- The Duration of Student Gameplay
- Student Gameplay CT Practices
- Extent (high/low clustering) of *Zoombinis* bridging and other CT activities in classes.

Multilevel models were chosen to account for any common variance in IACT post-assessment scores due to the clustering of students within classrooms. Using the SPSS MIXED linear models procedure, we estimated unconditional 2-level models with students nested within classrooms, using Restricted Maximum Likelihood (REML) and unstructured covariances. In unconditional 2-level models, a statistically significant percentage of the variance was attributable to classroom level variation (16 percent).

Sets of covariates were added to the unconditional 2-level models in this order:

Set 1. Pre-assessment score (IACT Z-score) & Grade Level (elementary or middle school).

Set 2. Duration of Gameplay (High Student Gameplay Duration ($\geq$ 2 hours) (1=Yes); and Average confidence of 5 detectors of Gameplay CT Practices from 3 puzzles (confidence between 0 and 1).

Set 3. Classroom Activities: Extent of *Zoombinis* Bridging (1=High) and Extent of Other CT Activity (1=High).

Set 4. Student, classroom, teacher, and school demographics: Student gender (1=Female) and whether or not the students were enrolled in a Title I school (1=Yes).

Set 5. Interactions between Classroom Activity Clusters (Set 3).

Set 6. Interactions between Classroom Activity Clusters and measures in the Sets 2 and 4.

The variance explained by the model was compared as each covariate or interaction was added. Only covariates and interactions that significantly improved the fit of the model were retained in the results presented in this paper.

4. Results

4.1 Descriptive Results

We conducted a series of independent t-tests to compare means between groups related to gender, student gameplay, and classroom activities. The Benjamini and Hochberg correction (Benjamini & Hochberg, 2000) was applied to control for multiple comparisons. As previously mentioned, we split students into the following two groups according to Student Gameplay Duration: High Student Gameplay Duration $\geq$ 2 hours or Low Student Gameplay Duration < 2 hours (see Table 7). There were proportionately few of the 668 students (22%) with High Student Gameplay Duration in classrooms with High Other CT Activities compared to almost three-quarters (73%) of the 48 students with Low Student Gameplay Duration who were in classrooms with High Other

CT Activities ($p < 0.0001$). There were no significant differences between the High and Low Student Gameplay Duration groups in the percentage of students belonging to the High *Zoombinis* Bridging clusters and Title 1 schools, both *p*s > 0.05.

Table 7: Independent T-tests for Student Gameplay Duration

	Student Gameplay Duration	
Covariate	Low < 2 hours (N=48)	High ≥ 2 hours (N=668)
Classroom Activity		
% High *Zoombinis* Bridging	35%	41%
% High Other CT Activity*	73%	22%
School		
% Enrolled in Title I Schools	38%	46%

*Significant after Benjamini and Hochberg correction for multiple comparisons.

Table 8 summarizes the results comparing means between Low and High *Zoombinis* Bridging groups. The High *Zoombinis* Bridging group had significantly fewer students enrolled in Title 1 schools, $p < 0.0001$. Within the High *Zoombinis* Bridging group, the same proportion of students had Student Gameplay Duration < 2 hours as has Student Gameplay Duration ≥2 hours ($p>0.05$). Similarly, roughly a quarter of students in each *Zoombinis* Bridging group was also in a class with High Other CT Activities ($p > 0.05$).

Table 8: Independent T-tests for Extent of *Zoombinis* Bridging

	Extent of *Zoombinis* Bridging	
Covariate	Low (N=427)	High (N=289)
Student Gameplay		
Student Gameplay Duration (Hours)*	15.54	29.30
% High Student Gameplay Duration	93%	94%
Classroom Activity		
% High Other CT Activity	24%	27%
School		
% Enrolled in Title I Schools*	52%	35%

*Significant after Benjamini and Hochberg correction for multiple comparisons.

We compared means between groups of Low and High Other CT Activity (see Table 9). The High Other CT Activity group had significantly fewer hours of Student Gameplay Duration, $p < 0.0001$. The proportion of students belonging to High Student Gameplay Duration and Title 1 categories was significantly lower in the High Other CT Activity than the Low Other CT Activity group, all ps < 0.05. There were no significant differences between groups in the percentage of students in the High *Zoombinis* Bridging group, $p = 0.29$.

Table 9. Independent T-tests for Other CT Activity

	Other CT Activity	
Covariate	Limited (N=534)	Extensive (N=182)
Student Gameplay Duration		
Student Gameplay Duration (Hours)*	26.34	5.71
% High Student Gameplay Duration*	98%	81%
Classroom Activity		
% High Bridging	40%	43%
School		
% Enrolled in Title I school*	43%	52%

*Significant after Benjamini and Hochberg correction for multiple comparisons.

4.2 Multilevel Modeling Results

The best fitting multilevel model is presented in Table 10. The intercept indicates the IACT post-score would be -0.78 when all covariates are 0—this would be a student who scored the overall mean IACT pre-assessment and all CT detectors were absent (confidence=0), played *Zoombinis* < 2 hours, and was in a non-Title I school. The best-fitting model suggest that for every standard deviation increase in students' scores on the IACT pre-assessment, their IACT post-assessment was 0.81 standard deviations higher.

Table 10: Best fitting multilevel model of estimated fixed effects with interactions on IACT Post Z-scores

						95% Confidence Interval	
Parameter	Estimate	Std. Error	df	t	Sig.	Lower Bound	Upper Bound
Intercept	-0.78	0.16	432	-4.87	0.00	-1.09	-0.46
IACT Pre-Assessment Z-Score	0.81	0.19	709	4.24	0.00	0.43	1.19
High Student Gameplay Duration (1=Yes)	-1.04	0.31	688	3.41	0.00	-1.64	-0.44
Detector Confidence for Gameplay CT Practices	1.35	0.25	515	5.34	0.00	0.85	1.85
High Student Gameplay Duration (1=Yes) * Detector Confidence for Gameplay CT Practices	1.83	0.47	708	-3.88	0.00	0.90	2.76
Pre-IACT Score (Z score) * Detector Confidence for Gameplay CT Practices	-0.63	0.31	709	-2.03	0.046	-1.24	-0.01
Title I School (1=Yes)	-0.13	0.05	708	2.70	0.01	-0.23	-0.03

N=716 students, 45 classes

4.2.1 Student, Class, and School Demographics Results

Students in Title I schools scored 0.13 standard deviations lower on their IACT post-assessment than students not in Title I schools. Once students' gameplay duration and in-game behaviors were taken into account, there were no differences in IACT post-scores by student gender, grade level (elementary vs. middle school), duration of Bridging activities, or duration of other CT activities.

4.2.2 Duration of Student Gameplay Results

Students who played *Zoombinis* ≥ 2 hours performed worse than students who played <2 hours, although this relationship was moderated by their in-game behaviors. Figure 8 shows the difference in IACT post-scores between students with High and Low Student Gameplay Duration and those students: (1) whose CT behaviors were absent but exhibited evidence they were still learning the game mechanic; (2) whose CT behaviors and learning game mechanic confidences were at the mean level (confidence=0.62); and (3) whose CT behaviors were present and exhibited little evidence they were still learning the game mechanic.

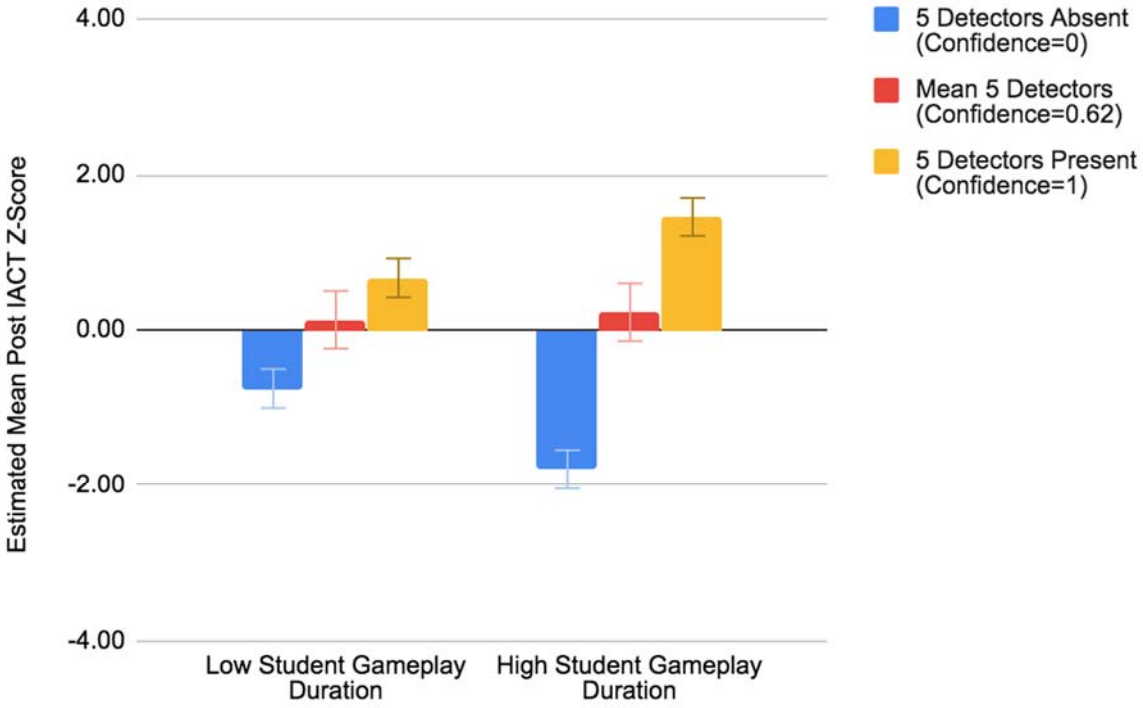

Figure 8: Interaction between Student Gameplay Duration and Student In-Game CT Detectors on students' IACT Post Z-scores

Note: Estimated means were evaluated at the following values: IACT Pre Z-Score = .0116.

Regardless of Student Gameplay Duration, students with no evidence of CT practices performed worse than students with mean confidences of 0.62 and 1. Among students with gameplay durations greater than 2 hours, this effect is more pronounced. These results suggest a mutually reinforcing relationship between the nature and extent of gameplay—the benefit of high levels of CT practices exhibited in their gameplay is enhanced with longer durations of gameplay.

4.2.3 Student Gameplay CT Practices Results

Students who demonstrated more evidence of CT Practices in their *Zoombinis* gameplay (as evidenced by higher average detector confidences) scored higher on the IACT post-assessment. This addresses research question 2 about the confidence of these detectors as predictors of CT Practices. If the Gameplay CT Practice detectors go from absent (confidence=0) to present (confidence=1), students' IACT post-scores increased by 1.35 standard deviations.

In addition to the significant interaction between the Duration of Student Gameplay and Gameplay CT Practices in Figure 6, Gameplay CT Practices were also moderated by student performance on the Pre-IACT assessment. Figure 9 shows the interaction between Student Pre-IACT scores and the average detector confidence of the Gameplay CT Practices on estimated mean IACT post-scores.

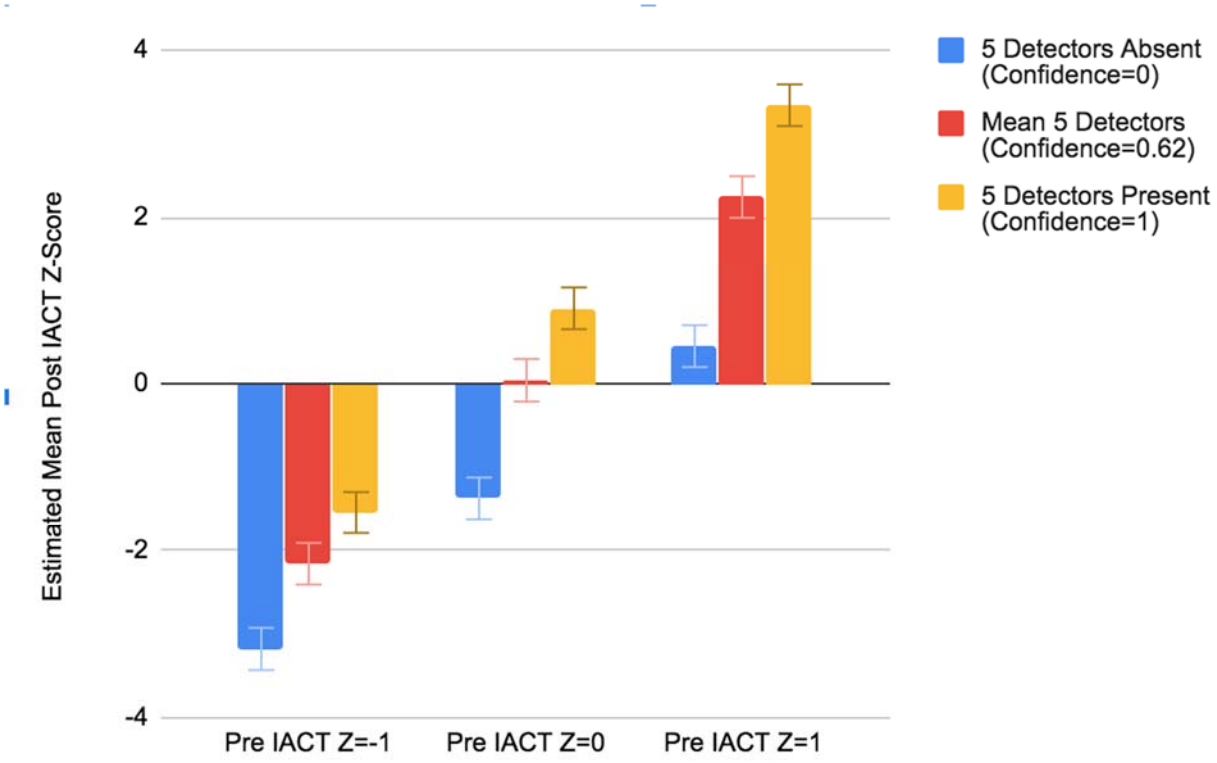

Figure 9: Interaction between Student Pre-IACT Score and Gameplay CT Practices

Note: Estimated means assumed Student Gameplay Duration and Title I school had a value of 0.5 to estimate the mean for each group.

We calculated estimated means at three levels of students' Gameplay CT Practices:

- Absent Gameplay CT Practices—where the mean confidence of students' demonstration of behaviors consistent with CT practices was 0.
- Mean Gameplay CT Practices—where students' demonstration of behaviors consistent with CT practices was at the group mean confidence (0.62).
- Present Gameplay CT Practices—where the mean confidence of students' demonstration of behaviors consistent with CT practices was 1.

Regardless of students' Pre-IACT scores, students who exhibited more Gameplay CT Practices performed better than those who exhibited less Gameplay CT practices. It is worth noting that

students who began the study with Pre-IACT scores close to the group mean yet exhibited high levels of Gameplay CT practices had higher Post-IACT scores than students who began the study with Pre-IACT scores 1 standard deviation above the mean who exhibited no Gameplay CT Practices. This is also true of students who scored 1 standard deviation below the mean on the Pre-IACT. They performed just as well as students with Pre-IACT scores at the mean who exhibited no Gameplay CT Practices.

Once both interactions with Gameplay CT Practices are combined into one model (Figure 10), it becomes clear that students who spent more than 2 hours playing *Zoombinis* but had no evidence of Gameplay CT Practices performed worse than other groups, regardless of their IACT Pre-Score. The only groups of students with positive changes in IACT scores were those who exhibited Gameplay CT Practices, with longer Student Gameplay Durations enhancing the impact of their gameplay. Among students scoring 1 standard deviation below the mean on their IACT Pre-Score, students who exhibited Gameplay CT Practices and played for more than 2 hours had scores close to the mean performance (Estimated Mean=0).

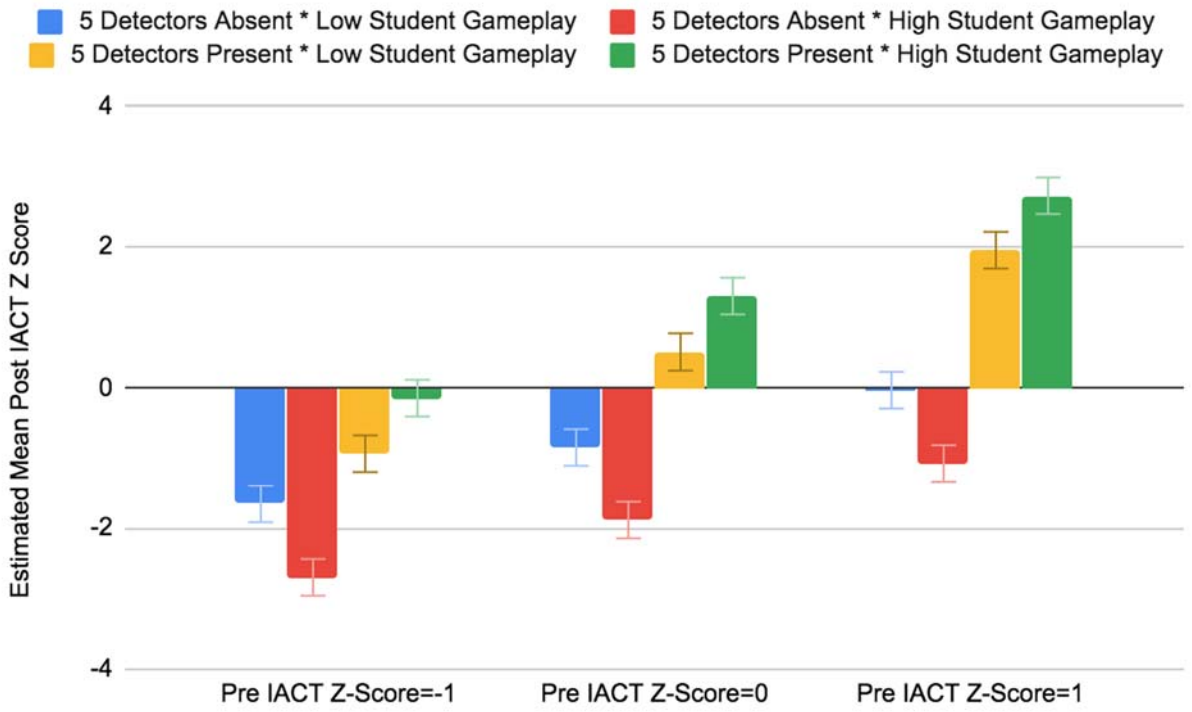

Figure 10: Interaction between students IACT Pre Z-Scores, Student Gameplay Duration, and Gameplay CT Practices on students' IACT Post Z-scores

Note: Estimated means were evaluated with Title I=0.5.

4.2.4 Classroom Activity Results

Once all of these covariates and interactions were taken into account, there were no differences in IACT post-scores by extent of *Zoombinis* Bridging. The lack of a bridging impact is contrary to our previous research (Asbell-Clarke et al., 2019) and to findings from teacher interviews by the external evaluator of the *Zoombinis* implementation study, where teachers perceived a beneficial effect on CT from gameplay and bridging activities.

5. Discussion

This study investigated the impact of the CT learning game *Zoombinis* and related classroom activities on the CT practices of students in grades 3–8. The study also examined the ability of the

automated detectors of Gameplay CT Practices to confidently predict student development of CT practices. As expected, we found significant relationships between the duration of *Zoombinis* gameplay, the CT practices demonstrated during gameplay, and improvements in students' IACT scores. This is consistent with literature that shows that well-crafted learning games can be effective in promoting student learning (Clark, Tanner-Smith, & Killingsworth, 2016), particularly in areas of complex problem-solving (Eseryel, Law, Ifenthaler, Ge, & Miller, 2014; Kang, Liu, & Qu, 2017). We also found many factors that did not influence student performance, including student gender, grade level (elementary vs. middle school), and surprisingly, the extent of *Zoombinis* bridging activities used in class, as well as the extent of other CT activities. This last finding counters our previous research that showed the importance of teacher activity in making implicit game-based science learning explicit outside the game (Asbell-Clarke et al., 2019).

Contributing to the growing literature in game-based learning assessments (Mislevy et al., 2016; Shute, Ke, & Wang, 2017), we built and studied automated detectors of implicit CT practices based upon extensive human analysis of *Zoombinis* gameplay. These detectors were indeed able to predict student development of CT practices with confidence. Those students who demonstrated high gameplay CT practices as measured by the detectors also had higher scores on the external post-assessments of CT practices when accounting for pre-assessment scores.

As hypothesized, this study shows that students who had longer durations of gameplay in *Zoombinis* performed better than those with shorter durations of gameplay. In addition, those students who showed a higher level of CT practices within their gameplay also performed better than those demonstrating lower level of CT practices in their gameplay. Moreover, there is an amplifying interaction between these two effects. Students who played for longer durations and exhibited more CT practices in their gameplay performed significantly better than all other groups.

Among students who exhibited relatively fewer CT practices in their gameplay, there was little difference in post-assessment scores regardless of how long they played.

We also examined the types of classroom activities that may influence the impact of *Zoombinis* on students' CT practices. Classroom activity was grouped along two dimensions: extent of *Zoombinis* Bridging (using specified *Zoombinis* bridging activities), and extent of Other CT Activity. There were no differences in students' scores because of the extent of *Zoombinis* Bridging activity or Other CT Activity. This is contrary to findings in our previous studies (Asbell-Clarke et al., 2019) where teacher bridging was key to the impact of game-based learning in the classroom. It is possible that *Zoombinis* is a more "self-contained" learning game and so it is able to support students' transfer of implicit game-based learning to explicit classroom learning without the necessity of teacher bridging. It also may be that the classroom bridging activities have the greatest impact on students' Gameplay CT Practices, and duration of bridging matters less once student gameplay behaviors are taken into account. The assessments we used in this study were game-like enough that students were able to make that transfer of CT practices from one application to another more easily. Further study is necessary to see if the implicit CT practices demonstrated in *Zoombinis* can translate into other CT applications such as improved coding, without teacher bridging activities. In summary, overall the students who played *Zoombinis* more and played it using CT practices performed better on post assessments, after accounting for pre-assessment scores. These results reveal that automated detectors of CT practices in student gameplay logs show promise as formative learning assessments to measure the development of student CT practices.

6. Risks and Limitations

We faced several limitations in the design of our study because of the young state of the CT field at the point of this research. Because CT was not taught commonly in schools at the time of the study, it was impossible to find a control group with an adequate number of classes that would consistently cover the same CT practices within a different context.

Also because of the immaturity of the field of CT education at these grade bands, we could not rely on an established set of measures for our pre/post assessments of CT practices. We therefore designed our own set of assessments, IACT, and conducted appropriate validity and reliability analyses (Asbell-Clarke et al., in review). These assessments, however, are not optimal and could use improvement for future research.

The assessment data was collected through an online set of logic puzzles. Only data for students who completed the pre- and post-assessments are included. This removed about 39% of the initial sample. We cannot distinguish in our data logs if an assessment timed-out because there were connectivity issues or if a student was struggling. Because we had to remove those data from the sample, we may have biased the sample towards students who did not struggle on the assessment. This is an important concern because in teacher study exit interviews, our external evaluators found many teachers reported that it was their students with academic struggles who became leaders in *Zoombinis* activities (Barchas-Lichtenstein, et al., 2019). The students who may benefit most from *Zoombinis* may be underrepresented in our final sample.

Preliminary examination of measurement invariance of these CT assessments suggest the elementary and middle school forms have a similar structure across time. The test-retest and concurrent validity analyses suggest the forms may be valid and reliable for elementary students. There are several limitations of these analyses. First, there was an intervention between the pre

and post assessment. This intervention may have been differentially effective for elementary vs. middle school students. This might explain why there was more variability across time in one form than the other. Second, elementary teachers spend more time with their students than middle school teachers. This might allow the elementary ratings to be more accurate (more highly correlated) than the middle school ratings. These limitations should be born in mind when interpreting these results.

Finally, we were reliant on teachers' self-reporting to understand the extent and nature of classroom activities that may have influenced the study. While the teacher reports we received were detailed, we assumed teachers who did not report Other CT Activities did not do them. This may not be the case if teachers assumed, contrary to instructions, we only wanted to know about *Zoombinis*-related Bridging activities. Thus, the lack of relationship between Other CT Activities and students' IACT post-scores may be related to differences in teachers reporting and not the nature of the Other CT Activity itself.

7. Conclusion

In a field as young as CT, there are many moving parts to establishing productive lines of research. This study of the learning game *Zoombinis*, in 45 elementary- and middle-school classrooms in the US, advances knowledge in this exciting research field in a number of ways.

This paper not only reports in detail on the extent and nature of a multitude of CT activities taking place in these classrooms, we also have shown a novel form of implicit learning assessments deploying automated data-mining detectors can be used to measure how students build CT practices within the game. We were able to show that it is not only how much children play this powerful learning game that matters, it is also *how* they play the game.

We were surprised to find that *Zoombinis* gameplay alone without teacher bridging was related to improved CT practices. This finding, however, is contrary to what teachers reported to external evaluators in their study exit interviews (Barchas-Lichtenstein et al., 2019). We believe further research is needed to develop easy-to-administer, standardized measures of bridging, and to better understand the role of games and teachers in the classroom at scale as there is considerable evidence that teachers make a key difference in game-based learning classrooms (Asbell-Clarke et al., 2019).

Statement on open data, ethics and conflict of interest

The data used in this study may be obtained by written request to the corresponding author. The Institutional Review Board (IRB) at the authors' organization reviewed and approved this research study following its regulations. The IRB adheres to the US National Science Foundation policies that are committed to the highest standards of ethical integrity in research and scholarship. There is no conflict of interest resulting from the research work in this study.

Acknowledgements

The authors are grateful to the National Science Foundation for generous funding for this project (#1502882) from the Division of Research on Learning. In addition, we thank the rest of our team and the many teachers and students who contributed to this study.

References

Almeda, M., Rowe, E., Asbell-Clarke, J., Baker, R., Scruggs, R., Bardar, E., & Gasca, S. (2019, October). Modeling Implicit Computational Thinking in *Zoombini's* Mudball Wall Gameplay. Paper presented at the Technology, Mind, and Society conference, October, Washington D.C.

Asbell-Clarke, J., Edwards, T., Rowe, E., Larsen, J., Sylvan, E., & Hewitt, J. (2012). Martian boneyards: Scientific inquiry in an MMO game. *International Journal of Game-Based Learning (IJGBL), 2*(1), 52–76.

Asbell-Clarke, J., Rowe, E., Bardar, E., & Edwards, T. (2019). The importance of teacher bridging in game-based learning classrooms. In M. Farber (Ed.), *Global Perspectives on Gameful and Playful Teaching and Learning*. IGI Global.
Asbell-Clarke, J., Rowe, E., Almeda, V., Gasca, S., Edwards, T., Bardar, E., Shute, V., & Ventura, M. (in review). Interactive assessments of CT (IACT): digital interactive logic puzzles to assess computational thinking in grades 3–8.
Ash, K. (2011). Gaming goes academic. *Education Week*, 30(25), 24–28.
Barchas-Lichtenstein, J., Brucker, J. L., Voiklis, J., Thomas, U. G., Fraser, J., Shane-Simpson, C., & Field, S. (2019). Cultivating computational thinking in elementary & middle school learners. Knology Publication #NSF.051.213.05. New York: Knology.
Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12: what is involved and what is the role of the computer science education community? *Inroads*, *2*(1), 48–54.
Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*(1), 60–83.
Berland, M., & Lee, V. R. (2011). Collaborative strategic board games as a site for distributed computational thinking. *International Journal of Game-Based Learning (IJGBL)*, *1*(2), 65–81.
Bertling, M., Jackson, G. T., Oranje, A., & Owen, V. E. (2015, June). Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning. In *International Conference on Artificial Intelligence in Education* (pp. 545–549). Springer, Cham.
Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24, 61–100.
Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). *Make it stick*. Harvard University Press.
Clark, D. B., Tanner-Smith, E., & Killingsworth, S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122.
Common Core of Data (CCD) (2020). Retrieved from https://nces.ed.gov/ccd/.
CSTA (2017). Retrieved from https://www.csteachers.org/page/standards.
Cuny, J., Snyder, L., & Wing, J. M. (2010). Demystifying computational thinking for non-computer scientists. Unpublished manuscript in progress, referenced in http://www.cs.cmu.edu/~CompThink/resources/TheLinkWing.pdf
Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. (2014). An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Educational Technology & Society*, *17*(1), 42–53.
Fishman, B., Riconscente, M., Snider, R., Tsai, T., & Plass, J. (2014). Empowering educators: Supporting student progress in the classroom with digital games. Ann Arbor, MI: University of Michigan. Retrieved from http://gamesandlearning.umich.edu/agames.
Fishman, B., Riconscente, M., Snider, R., Tsai, T., & Plass, J. (2015). Empowering educators: Supporting student progress in the classroom with digital games (Part 2). Ann Arbor: University of Michigan, gamesandlearning.umich.edu/agames.
Fu, J., Zapata, D., & Mavronikolas, E. (2014). Statistical methods for assessments in simulations and serious games. *ETS Research Report Series*, *2014*(2), 1–17.
Gee, J. P. (2013). Learning systems, not games. *Texas Education Review*, *1*.
Green, G., & Kaufman, J. C. (2015). *Video games and creativity*. Academic Press.

Grover, S., & Pea, R. (2018). Computational thinking: A competency whose time has come. In *Computer Science Education: Perspectives on teaching and learning*. S. Sentance, S. Carsten, & E. Barendsen. (Eds.). Bloomsbury.
Holbert, N. (2013). *Reimagining Game Design: Exploring the Design of Constructible Authentic Representations for Science Reasoning*. Northwestern University.
Holbert, N., & Wilensky, U. (2010). *FormulaT Racing*. Evanston, IL: Center for Connected Learning and Computer-based Modeling.
Kai, S., Paquette, L., Baker, R. S., Bosch, N., D'Mello, S., Ocumpaugh, J., ... & Ventura, M. (2015). A Comparison of video-based and interaction-based affect detectors in physics playground. *International Educational Data Mining Society*.
Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior*, *72*, 757–770.
Kazimoglu, C., Kiernan, M., Bacon, L., & Mackinnon, L. (2012). A serious game for developing computational thinking and learning introductory computer programming. *Procedia-Social and Behavioral Sciences*, *47*, 1991–1999.
Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In R. E. Furdig (Ed.), *Handbook of Research on Effective Electronic Gaming in Education* (pp. 1–32). New York: IGI Global.
Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing*, *16*(2), 142–163.
Lederman, L. C., & Fumitoshi, K. (1995). Debriefing the Debriefing Process: A new look. In D. C. K. Arai (Ed.), *Simulation and gaming across disciplines and cultures*. London: Sage Publications.
Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2016). Psychometrics and game-based assessment. *Technology and testing: Improving Educational and Psychological Measurement*, 23–48.
Mislevy, R. J., & Riconscente, M. M. (2011). Evidence-centered assessment design. In *Handbook of test development* (pp. 75–104). Routledge.
National Research Council (NRC) (2011). *Learning Science Through Computer Games and Simulations.* M.A. Honey & M. L. Hilton (Eds.). Washington, DC: National Academies.
Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc.
Papert, S., & Harel, I. (1991). Situating constructionism. *Constructionism*, *36*(2), 1–11.
Paquette, L., Rowe, J., Baker, R., Mott, B., Lester, J., DeFalco, J., ... & Georgoulas, V. (2016). Sensor-free or sensor-full: A comparison of data modalities in multi-channel affect detection. *International Educational Data Mining Society*.
Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, *50*(4), 258–283.
Polanyi, M. (1966). *The Tacit Dimension*. University of Chicago Press. Chicago, IL. USA.
Rowe, E., Asbell-Clarke, J., Almeda, M., Bardar, E., Baker, R. S., & Scruggs, R., (2019). Advancing research in game-based learning assessment: Tools and methods for measuring implicit learning. In E. Kennedy & J. Qian (Eds.), *Advancing Educational Research with Emerging Technology*. IGI Global.
Rowe, E., Asbell-Clarke, J., Almeda, M., Scruggs, R., Baker, R.S., Bardar, E. & Gasca, S. (under review). Assessing implicit computational thinking in Zoombinis puzzle gameplay. Submitted to a special issue of *Computers & Human Behavior on Learning Analytics and Assessment*.

Rowe, E., Asbell-Clarke, J. & Baker, R. (2015). Serious game analytics to measure implicit science learning. In C. S. Loh, Y. Sheng & D. Ifenthaler (Eds.), *Serious Game Analytics: Methodologies for Performance Measurement, Assessment, and Improvement* (343–360). Springer Science+Business.
Rowe, E., Asbell-Clarke, J., Baker, R., Eagle, M., Hicks, A., Barnes, T., Brown, R., & Edwards, T., (2017). Assessing implicit science learning in digital games. *Computers in Human Behavior, 76*, 617–630. DOI: 10.1016/j.chb.2017.03.043.
Rowe, E., Asbell-Clarke, J., Cunningham, K. & Gasca, S. (2017, October). Assessing implicit computational thinking in Zoombinis gameplay: Pizza Pass, Fleens, and Bubblewonder Abyss. Work-in-progress presented at the ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play, Amsterdam.
Rowe, E., Asbell-Clarke, J., Gasca, S., & Cunningham, K. (2017, August). Assessing implicit computational thinking in Zoombinis gameplay. Poster presented at the International Conference on the Foundations of Digital Games in Hyannis, MA.
Rowe, E., Bardar, E., Asbell-Clarke, J., Shane-Simpson, C., & Roberts, S. (2016). Building bridges: Teachers leveraging game-based implicit science learning in physics classrooms. In D. Russell & J. Laffey (Eds.), *Handbook of Research on Gaming Trends in P-12 Education* (499–525*)*. Hershey, PA: IGI-Global.

Shute, V. J., Rahimi, S., & Smith, G. (2019). Game-based learning analytics in physics playground. In M. Chang & A. Tlili (Eds.), *Data analytics approaches in educational games and gamification systems* (pp. 69–93). New York: Springer.
Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction, 55*(2), 503–524.
Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., ... & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education*, *86*, 224–235.
Shute, V. J., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In *Instructional Techniques to Facilitate Learning and Motivation of Serious Games*. Springer, Cham. 59–78.
Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review, 22,* 142–158.
Sneider, C., Stephenson, C., Schafer, B., & Flick, L. (2014). Exploring the science framework and NGSS: Computational thinking in the science classroom. *Science Scope*, *38*(3), 10.
Steinkuehler, C., & Duncan, S. (2008). Scientific habits of mind in virtual worlds. *Journal of Science Education and Technology*, *17*(6), 530–543.
Stoddard, J., Banks, A. M., Nemacheck, C., & Wenska, E. (2016). The challenges of gaming for democratic education: The case of iCivics. *Democracy and Education*, *24*(2), 2.
Underwood, G. D. (1996). *Implicit cognition*. Oxford University Press.
TERC (2019). *Zoombinis. Retrieved from* http://zoombinis.com.
Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35.
Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), 76–77.