Design Insights into the Creation and Evaluation of a Computer Science Educational Game

Britton Horn
Northeastern University
bhorn@ccs.neu.edu

Christopher Clark
Northeastern University
clark.chris@husky.neu.edu

Oskar Strom
Northeastern University
strom.o@husky.neu.edu

Hilery Chao
Brown University
hilery_chao@brown.edu

Amy J. Stahl
Northeastern University
stahl.a@husky.neu.edu

Casper Harteveld
Northeastern University
c.harteveld@neu.edu

Gillian Smith
Northeastern University
gi.smith@neu.edu

ABSTRACT

Computer Science (CS) education at the middle school level using educational games has seen recent growth and shown promising results. Typically these games teach the craft of programming and not the perspectives required for computational thinking, such as abstraction and algorithm design, characteristic of a CS curriculum. This research presents a game designed to teach computational thinking via the problem of minimum spanning trees to middle school students, a set of evaluation instruments, and the results of an experimental pilot study. Results show a moderate increase in minimum spanning tree performance; however, differences between gender, collaboration method, and game genre preference are apparent. Based on these results, we discuss design considerations for future CS educational games focused on computational thinking.

CCS Concepts

•Social and professional topics → Computational thinking; Informal education; K-12 education; •Applied computing → Computer games;

Keywords

educational games; game design; computational thinking

1. INTRODUCTION

The computer science (CS) education community has spent much of its effort on introducing CS education earlier in the curriculum, in hopes to better prepare students for a world where procedural literacy and computational thinking is increasingly important [9, 13] as well as improve chronic issues of a lack of diversity in STEM fields, including CS at the university level and beyond [26]. Many of these efforts to reach younger students are done in informal contexts such as classroom visits, summer camps, and after-school programs [5, 7, 8]. The ACM has also recently proposed a set of computing education guidelines that can be integrated into a standard K-12 curriculum [25], and in the UK there are already efforts in place to teach computational thinking in primary and secondary schools [4].

Games have been a promising and popular component of many efforts to teach CS and broaden participation. At the middle school level, there are programs that encourage students to create games using simplified programming environments such as *Alice* [3] or *Scratch* [18]. Others are educational games especially designed to teach CS. However, the vast majority of these game-related outreach efforts tend to focus on the craft and practice of programming, rather than higher-level CS concepts [10].

With this in mind, we have developed *GrACE*, a puzzle game that aims to teach computational thinking–specifically, concepts of abstraction and algorithms [28]. The game teaches neither the structure nor syntax of programming. Instead, it is designed such that players must solve CS puzzles in an algorithmic manner, thus fostering computational thinking.

Puzzles in *GrACE* are built around the common CS problem of finding a graph's minimum spanning tree (MST). This problem was chosen due to the relative ease of mapping graphs onto spatial puzzles, the existence of many algorithms to solve the problem, and its property that even finding an incorrect solution (such as finding a spanning tree that is not minimal) still can involve computational thinking. Thompson and Bell also find the MST problem useful for mapping to a CS educational game due to the need for players to identify algorithms [23].

With this game, we explore the use of procedural content generation (PCG) to aid in computational thinking. Computer generated puzzles bring advantages including rapid puzzle creation and a guarantee all puzzles meet desired graph structure and aesthetic design constraints. Additionally, the incorporation of PCG means players can request new puzzles on demand increasing the amount of in-game content available to them to practice different strategies [21]. We conjecture that by seeing a larger variety of content,

SIGCSE '16, March 02-05, 2016, Memphis, TN, USA
 ISBN 978-1-4503-3685-7/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2839509.2844656

players receive a better understanding of the underlying concept [22]. We further believe that seeing content variety in a collaborative setting will push discussions among players toward abstracted solutions.

Our long-term goals are to come up with an appropriate puzzle design to foster computational thinking in middle school students and to study the educational benefit of including PCG in educational games. In this paper, we present the evaluation instruments and results of an experimental pilot study of the current design of *GrACE*. Our contribution is a set of design insights into how to construct and evaluate educational games that do not aim to specifically teach programming, but rather focus on high-level concepts.

2. RELATED WORK

Many game initiatives for teaching CS exist. Existing CS games frequently focus on teaching programming, either using actual code or a simulated programming-like set of tasks [10]. Those that do support multiplayer activities are typically competitive rather than collaborative, and rhetorically the games often position each player as a solitary character working alone towards a goal. In contrast, *GrACE* casts the player as a helpful character who needs to use their reasoning skills to help a friend, encouraging collaboration even though it is a single-player game.

The focus on competition in educational CS games is seemingly at odds with the value of collaboration in other aspects of CS education. Pair programming [27] is a common practice in introductory programming courses, with benefits to student learning arising from the ability to share strategies and point out potential errors during collaboration. Research in educational games in other fields, such as mathematics, shows competition can be more effective for achieving learning outcomes, but collaboration is still effective and can lead to greater student interest [11, 16].

Most educational games use content that was scripted by the human authors of the game. Procedural content generation (PCG) is the practice of "creating game content automatically, through algorithmic means" [24]. *Refraction* [20] is an example of a math education game that has procedurally generated puzzles. It uses PCG to create its puzzles both to ease authoring effort (by having the computer take on this design role) as well as make hard guarantees about the properties of levels from both an educational and aesthetic perspective. However, *Refraction* simply replaces a typical human-designed level with a computer-designed level; it does not take advantage of the existence of an on-demand puzzle designer during gameplay. *GrACE* uses similar technology to *Refraction* to generate its puzzles, and allows players to request new puzzles at the same difficulty level on demand throughout the game.

Our work examines differences in performance based on whether players had exposure to many different levels, as well as the impact on the use of PCG when players collaborate but are seeing different puzzles from each other. It has been argued that PCG may be useful for educational games through allowing players to explore alternate problems, and communicate abstracted strategies with each other [22].

3. METHODS

Based on our ideas of the potential for PCG for education and the possible mediating role of collaboration, we came up with the following three exploratory propositions:

Proposition 1. Experiencing variety leads to increased learning gains in computational thinking compared to no variety.

Propostion 2. Working in a collaborative context leads to increased learning gains in computational thinking than in an individual context.

Propostion 3. Experiencing variety in a collaborative context leads to increased learning gains in computational thinking compared to no variety in an individual context.

To explore these propositions and, more broadly, our game and evaluation design, we implemented an independent 2x2 factorial design as part of a pilot study. The two 2-level independent variables (IVs) are PCG (PCG vs. No-PCG) and Collaboration (Individual vs. Collaborative).

3.1 Participants

The experiment was implemented as part of a two week summer program at Northeastern University. This program selects 48 middle school talented students each year, and focuses on STEM content. The program historically supports underserved and underrepresented students with limited opportunities and is free of charge.

During the day of implementation, 43 students participated with the consent of their parents, 22 identifying as female and 21 as male. Ages ranged from 10 to 13 (M = 11.9, SD = 0.85). Four participants identified as Hispanic or Latino, 12 reported to be Asian, nine Black or African American, 12 White, and six as "other" (with four participants preferring not to answer).

3.2 Materials

All materials mentioned in this section are publicly archived.[1]

3.2.1 Game

GrACE has a nature-based theme, chosen as a result of focus testing with middle school students to determine the appeal of a variety of metaphors. The game is centered on two characters–a mouse and a rabbit–who are collecting vegetables. Vegetables (nodes) in the ground are connected by cracks (edges) that only the mouse can fit through. Edge weight corresponds to the amount of bunny energy needed to dig along a crack. Players control the mouse as it explores the map and flags cracks for the rabbit. To minimize the bunny's digging energy, players must find and flag the MST. An example can be seen in Fig. 1

Initially, players can only see the starting node and the nodes connected to it. The player chooses an edge to traverse at which point the next node and its connections are revealed. This mechanic discourages players from solving the puzzle visually by examining the entire graph at once. Instead, we designed the game around limited information exploration to encourage stepwise thinking and mimicry of MST solving algorithms.

We used two game versions in accordance with the PCG manipulation: one version with PCG and one without. The standard version contains 11 difficulty levels, each with a single associated puzzle pre-generated by the computer and the same across all instances of the game. These puzzles were each chosen at random from the set of levels used in the PCG version, to limit accidental biasing from including

[1] http://hdl.handle.net/2047/D20199448

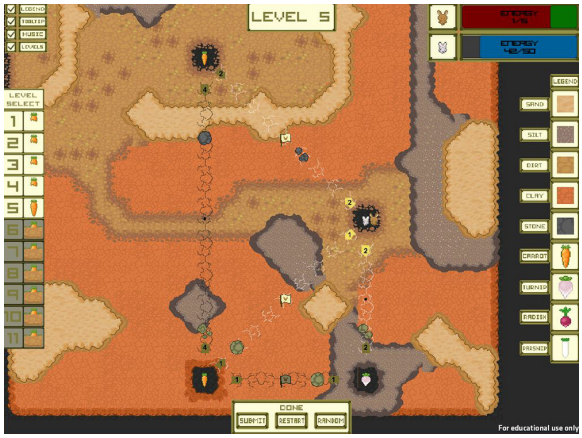

Figure 1: Screenshot ©Northeastern University.

human-authored puzzles. Players can restart their puzzle and access previously played puzzles once completed. In addition, the PCG version allows for the possibility at anytime to press a "random" button to generate a new puzzle with the same difficulty level. We generated 100 puzzles for each level in advance, making the PCG version contain 1,100 puzzles in total. Level one contained two nodes and one edge to demonstrate basic game mechanics. Level 11 contained nine nodes and 16 edges.

Maps were created using a constraint-based, Answer Set Programming (ASP) method [19]. Graph constraints included the number of nodes and edges, minimum and maximum weights for edges, node connectivity, and a valid MST. Aesthetic constraints were the map size, distance between nodes, valid edge intersection angles, and various art tile layout rules. Constraints were separated so that future art implementations could have their own set of constraints swapped with the current set and not impact the core map generation.

The game was instrumented to track all player actions (e.g. placing or removing a flag, submitting an answer) and game states (e.g. mouse and bunny energy).

3.2.2 *Questionnaires*

The pre-questionnaire involved an adaptation of the 22-item New Computer Game Attitude Scale (NCGAS), which was developed to evaluate middle school students' attitudes toward educational games [12]. In our adaptation, we used the *Liking* and *Leisure* constructs, and adapted it to a 7-point Likert scale for consistency with our other measures. Liking (3 items; $\alpha = .760$) measures how much the students like playing video games; Leisure (5 items; $\alpha = .856$) measures the degree to which video games are incorporated in the leisure time of students.

On the post-questionnaire we included a scale on Game Experience to assess what students' opinions are of the game. We created two anticipated constructs, Enjoyment ($\alpha = .93$) and Difficulty ($\alpha = .78$), each with four 7-point Likert items. The post-questionnaire ended with demographics questions and questions about game-playing and school subject preferences, where students were requested to check any genre or topic that they liked.

3.2.3 *Comprehension Test*

The comprehension test was developed to measure conceptual understanding of the MST. Previous studies on computational thinking have included tests that made effective use of measuring change in students' conceptual understanding through design scenarios [2] or concept exams [14] specific to the software being tested. To allow for objective comparison across students, we provided a multiple choice assessment along the lines of the concept exams. In contrast to previous studies, we opted for questions not only about solving MST problems in the context of the specific software but also about solving MST problems with another metaphor and in an abstract manner. The goal with this is to measure knowledge transfer from the game to other contexts.

We created two tests, each with a different but similar puzzle for all three contexts: game, metaphor, and abstract. Each puzzle has three associated questions with four choice options for each question. Consistently across the puzzles, one question concerned the problem of adding a single edge, one about removing a single edge, and one about correcting an incorrect spanning tree. The questions were developed independently, validated by the three test question authors, and error-checked by a fourth researcher on the team.

Game. For these questions, we edited screenshots of random puzzles from the PCG version and constructed questions using similar language to that used in the game.

Metaphor. Here we adapted the "Muddy City" exercise from CS Unplugged [1]. In our adaptation, we simplified the language of this roads network problem because the students had to read it to themselves, and redid the artwork to make it clearer and easier to modify.

Abstract. These puzzles were based on a typical graph with nodes and edges that we referred to as circles and lines.

3.3 Procedure

This study was run as part of a "computational thinking" activity in the summer program. We randomly assigned students to a computer, and evenly distributed them across the conditions; however, due to a slightly lower turnout and one student opt-out, we had fewer students in both PCG conditions. We counterbalanced the comprehension test to rule out the difficulty of the test as a bias.

The experiment started in a single classroom where students received their pre-questionnaire, pre-test and lab assignment. Instructions for the game were then provided by a facilitator without mentioning the MST concept or the game's relationship to CS. One lab held all students in the individual condition, the other had students in the collaborative condition. We divided both spaces such that one half was assigned to the PCG condition and the other half to the No-PCG condition to minimize interference. Assigned pairs played next to each other and were encouraged to communicate and look at each other's screens. In the individual condition, students were discouraged from interacting. All facilitators followed a script instructing them how to respond during the questionnaires, tests and gameplay.

Students completed the post-questionnaire and post-test after the game, followed by a semi-structured discussion on the topic of CS and the MST concept in particular. The experiment took three hours, with one hour of gameplay.

4. RESULTS

For the purposes of this paper, we limit ourselves to reporting the main results that we derived on demographics, gameplay, and the performance on the comprehension test.

4.1 Demographics

The majority (61.9%) reported playing games more than 1-2 days a week. There were no strong preferences or dislikes of genres. Based on the preferences for school subjects, with 32 (74.4%) indicating a liking for math and 33 (76.6%) a liking for science, it is clear that this is a group with a bias towards STEM content. In terms of gender differences, boys reported to play more than girls, $\chi^2(1) = 4.39$, $p = .036$, $V = .371$. Boys further reported to like playing adventure and action games more so than girls, for both $\chi^2(1) = 10.5$, $p = .001$, $V = .495$. For school subjects, girls liked art and language and literature better, respectively $\chi^2(1) = 7.21$, $p = .001$, $V = .409$ and $\chi^2(1) = 6.89$, $p = .007$, $V = .400$. Boys, on the other hand, report liking technology better, $\chi^2(1) = 4.06$, $p = .044$, $V = .309$. No racial differences were observed in game and school subject preferences.

The results regarding game attitude suggest that the majority like games but are more spread in their opinion when it comes to the role of games in their leisure time. The variety in Leisure is partially a result of a difference in gender, $t(40) = -5.67$, $p < .001$, $r = .73$, with girls disagreeing more about its importance. This finding is consistent with the gender difference in frequency of playing.

4.2 Gameplay

Investigating the items associated with Enjoyment, it becomes clear that a strong majority agreed or strongly agreed they had fun and majorities further indicated they would recommend this to a friend, play it at home, and learned from it. Despite the enjoyment, in exploring the items associated with Difficulty, it shows that majorities thought it was frustrating, challenging, and hard. In contrast, the majority disagreed that they felt bored. No differences are found on gender and race, except for the Enjoyment item "I thought it was fun". It turns out girls ($Mdn = 5$, $IQR = 4$-6) agreed with this less than boys ($Mdn = 6.5$, $IQR = 5$-7), $U = 144$, $p = .047$, $r = .31$.

Both Liking and Leisure correlate with Enjoyment, respectively $r = .40$, $p = .009$ and $r = .309$, $p = .049$, which further suggests an initial predisposition is of influence in how the game is experienced. It is then not surprising to find that students who play more than 1-2 days a week enjoyed the game more than those who play less, $t(39) = 2.91$, $p = .006$, $r = .42$. Liking, leisure, and frequency of play do not relate to Difficulty.

There are aspects that do relate to Difficulty. Those who like puzzle and strategy games perceived less difficulty than those who do not, respectively $t(41) = 2.31$, $p = .026$, $r = .34$ and $t(40) = 2.73$, $p = .009$, $r = .40$. Interestingly, the same is true for students with a preference for sports games and physical education, respectively $t(21) = 2.54$, $p = .019$, $r = .48$ and $t(41) = 2.11$, $p = .041$, $r = .31$. This may suggest that these students like a challenge and have a different perception of difficulty. Surprisingly, it is students with an interest in simulation and role-playing games that scored higher on Enjoyment, respectively $t(40) = 2.41$, $p = .021$, $r = .36$ and $t(40) = 2.22$, $p = .032$, $r = .33$.

Regarding the conditions, it appears students in the PCG conditions agreed more with the Difficulty item "I felt frustrated", $t(41) = 2.47$, $p = .018$, $r = .36$. For the collaborative conditions it is interesting to note that it approaches significance for agreeing more on the Enjoyment item "I would recommend this game to a friend"(*.061*) and on the Difficulty item "I thought it was hard".

The game data revealed that students in the collaboration condition took longer to play the game with an average level time of 111 seconds ($SD = 33.3$) compared to individual participants' average of just 85 seconds ($SD = 28.5$), $t(40) = 2.67$, $p = .01$, $r = .39$. Somewhat related and nearing significance is that students in the individual condition ($Mdn = 11$, $IQR = 9$-11) were more likely to complete the final level than the collaboration condition ($Mdn = 9$, $IQR = 7.5$-10.5), $\chi^2(1) = 3.93$, $p = .056$. Every student was able to complete at least level 5 (4 nodes and 5 edges).

The "random" button was used a total of 125 times by the 19 students in the PCG group ($M = 6.58$, $SD = 6.74$). No statistically significant difference was seen between the collaboration and individual groups. The button was used most on Level 6 at 31 times (24.8%), followed by Level 8 at 25 (20%) and level seven at 23 (18.4%). The "random" was not used by seven students and—surprisingly—on Level 11.

4.3 Comprehension

Independent Samples T-tests show that there is no significant difference between the two versions on the total scores before and after the game, suggesting that the test versions are equal in difficulty. Test scores were calculated by providing one point per correctly answered question. Thus, with nine questions students could get a maximum score of nine points. Even though MST problem was likely unfamiliar to students, they performed quite well on the pre-test ($M_{pre} = 5.71$, $SD_{pre} = 2.13$) and marginally but significantly improved their performance on the post-test ($M_{pre} = 6.37$, $SD_{pre} = 2.33$), $t(41) = 2.40$, $p = .021$, $r = .35$. Of note is that the number of students with a perfect score increased from two (4.8%) to nine students (21%).

Table 1 provides an overview of the average improvement across conditions on the overall test and on the specific puzzles as well as kinds of questions. The table implies there are differences across the conditions. The collaborative conditions seem to improve less so than the individual ones, and this seems especially true for the collaborative condition with PCG, suggesting an interaction effect exists. No strong significant difference appears at first; however, in exploring the data a surprising finding was that students who liked the action genre ($M_{pre} = 5.61$, $SD_{pre} = 1.97$; $M_{post} = 6.88$, $SD_{post} = 1.92$) improved more than students who did not ($M_{pre} = 5.84$, $SD_{pre} = 2.36$; $M_{post} = 5.74$, $SD_{post} = 2.68$), $t(40) = 2.73$, $p = .009$, $r = .40$. None of the other demographics had such an influence. As boys liked the action genre more so than girls, we performed 2x2 ANCOVA analyses with PCG (PCG vs. No-PCG) and Collaboration (Ind. vs. Coll.) as between-subjects factors and with gender and action as covariates. On the total improvement this revealed a main effect of action, $F(1, 36) = 10.1$, $p = .003$, $\eta_p^2 = .22$, and Collaboration, $F(1, 36) = 4.69$, $p = .037$, $\eta_p^2 = .12$. The effects of gender ($p = .20$) and PCG ($p = .67$) were insignificant; however, the interaction effect between PCG and Collaboration approximates significance, $F(1, 36) = 3.64$, $p = .064$, $\eta_p^2 = .09$. The interaction effect is significant on the abstract puzzles specifically, $F(1, 37) = 5.36$, $p = .026$, $\eta_p^2 = .13$. As exemplified in Table 1, we can see that students in the individual condition with PCG improved more so than in the No-PCG conditions, whereas in the collaborative condition with PCG the performance seems to be less than in the No-PCG conditions.

Test Measure	Ind. x No-PCG (N=12)	Coll. x No-PCG (N=12)	Ind. x PCG (N=8)	Coll. x PCG (N=10)	Overall (N=42)
Total	1.08 (1.44)	0.67 (1.30)	1.13 (1.73)	-0.30 (2.31)	0.64 (1.73)
Game	0.50 (1.09)	0.17 (0.94)	0.63 (1.51)	0.30 (1.49)	0.38 (1.21)
Roads network	0.50 (0.52)	0.33 (0.78)	-0.25 (1.04)	-0.30 (1.33)	0.12 (0.97)
Abstract	0.08 (1.00)	0.17 (1.03)	0.75 (1.03)	-0.30 (0.95)	0.14 (1.03)
Addition	0.17 (0.86)	-0.08 (0.90)	0.50 (0.76)	0.00 (0.94)	0.12 (0.86)
Deletion	0.42 (0.90)	0.42 (0.90)	0.25 (1.16)	-0.10 (0.99)	0.26 (0.96)
Correction	0.50 (0.80)	0.33 (0.89)	0.38 (1.30)	-0.20 (0.92)	0.26 (0.96)

Table 1: Comprehension test improvement across conditions, in M (SD). The test is separated into three representations (original game aesthetic, road network metaphor and abstract) and three concepts (edge addition, edge deletion and completed tree correction). Conditions are broken down by Individual (Ind.) vs Collaboration (Coll.) and PCG vs No-PCG.

5. DISCUSSION

Our experience with designing the game, constructing instruments to evaluate it, and performing a pilot study has led to a number of design insights and areas for future study. It should be noted that the current study is limited by using a biased group of diverse but well performing students who are interested in STEM.

The potential of PCG and collaboration. With the results in this instance of *GrACE*, we must outright reject Proposition 2 since the opposite happened. The students in the collaboration conditions improved less. The possibility to discuss with partners does not outweigh playing more exercises, and it was noticeable that pairs played less and got less far. Future research should look into the communication between pairs and ways to better support collaboration through the game. Although an interaction effect seems to happen, it is also different than expected. PCG is only beneficial for those who play alone. For that reason we need to reject Proposition 3 about the multiplier effect of PCG and Collaboration too. However, we can partially accept Proposition 1. The current data provides an indication that when PCG is used in an individual context, increased learning gains are a result. The fact this is especially noticeable for the abstract puzzles is a promising result. Teaching abstract thinking is a difficult yet important skill to master, especially for students aspiring to a CS career.

Designing for collaboration. In keeping with other findings for educational games in non-CS STEM fields [11, 16], our findings suggest that players are more likely to recommend the game to a friend in the collaboration condition, yet perform better in the game in the individual condition. The goal of informal learning interventions is to be both engaging and educational. The Collaboration condition in *GrACE* does not involve students collaborating on the same puzzle, but rather collaborating to help each other with their own puzzles outside the game. In future versions of the game, we intend to explore ways to make this strategy-level collaboration more explicit and occur within the game in order to achieve a positive collaborative learning effect.

Placing content generation in context. A common argument in the game design and PCG community is that additional content leads to replayability and enjoyment [21]. However, our findings show that playing in the PCG condition leads to greater frustration. One possible explanation for this is that students who are struggling with one particular puzzle may generate new content hoping for an easier puzzle, but instead find one at the same difficulty level that is also challenging. While frustration may be linked to game enjoyment overall, this effect is something that requires more dedicated study. It is clear that the context in which PCG is used and how it is integrated into the game's overall mechanics must be carefully considered.

Fun vs. frustration Students reported a high rate of enjoyment for the game but also found it frustrating, challenging, and hard. This is perhaps tied to Papert's concept of "hard fun" in educational games: not that players find a game fun despite it being hard, but that they find it fun because it is hard [15]. While many educational games work to simplify concepts and offer extremely gentle introductions, with *GrACE* it seems that its difficulty is important for the enjoyment of the game. Our future work will include examining what specifically it is about the game that students find frustrating by looking at patterns in behavior from the game data, analyzing player demographics more in-depth, and testing additional iterations of the game.

Beyond gender: Understanding player profiles. Researchers have often discussed the importance of designing games that are accessible to girls, either by attempting to understand the design preferences exhibited by girls [6, 5] or by incorporating stereotypically feminine play styles and preferences into a game's design to make it "gender-neutral" [17]. However, our results show that while gender identity can influence player enjoyment, it is not alone responsible. Indeed, genre preferences (such as a preference for puzzle games, action games, or competitive sports games) have an impact on both experience and outcome performance, independent from gender. This points to the need for a deeper understanding of a target audience, refocusing design efforts away from looking at gender alone.

6. CONCLUSION

In this paper, we discussed the results from an experimental pilot study of *GrACE*, a game designed to foster computational thinking through PCG and collaboration, with the aim of retrieving design insights into the creation and evaluation of this CS educational game. While the game achieved a moderate improvement in the conceptual understanding of the MST problem, multiple unexpected insights were found such as the need for designing explicitly for collaboration, and the need to more deeply understand player profiles. These design insights are guiding our future work with *GrACE*, and can serve as guidelines for others interested in developing CS educational games.

7. ACKNOWLEDGMENTS

We want to thank the Northeastern STEM Center for Education for integrating our pilot study in their summer program and our external evaluators from TERC: Jim Hammerman and Audrey Martinez-Gudapakkam. This material is based upon work supported by the National Science Foundation under Grant No. 1422750.

8. REFERENCES

[1] T. Bell, I. H. Witten, and M. Fellows. *Computer science unplugged: An enrichment and extension programme for primary-aged children*. 2010. http://csunplugged.org/books.

[2] K. Brennan and M. Resnick. New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American Educational Research Association, Vancouver, Canada*, 2012.

[3] S. Cooper, W. Dann, and R. Pausch. Alice: A 3-D tool for introductory programming concepts. *Journal of Computing Sciences in Colleges*, 15(5):107–116, 2000.

[4] A. Csizmadia, P. Curzon, M. Dorling, S. Humphreys, T. Ng, C. Selby, and J. Woollard. Computational thinking: A guide for teachers. Technical report, Computing At School, 2015.

[5] J. Denner, L. Werner, S. Bean, and S. Campe. The girls creating games program: Strategies for engaging middle-school girls in information technology. *Frontiers: A Journal of Women Studies*, 26(1):90–98, Jan. 2005.

[6] M. D. Dickey. Girl gamers: The controversy of girl games and the relevance of female-oriented game design for instructional design. *British Journal of Educational Technology*, 37(5):785–793, 2006.

[7] P. Doerschuk, J. Liu, and J. Mann. Pilot summer camps in computing for middle school girls: From organization through assessment. *SIGCSE Bull.*, 39(3):4–8, June 2007.

[8] C. Frieze. Diversifying the images of computer science: Undergraduate women take on the challenge! *SIGCSE Bull.*, 37(1):397–400, Feb. 2005.

[9] M. Guzdial. Education: Paving the way for computational thinking. *Commun. ACM*, 51(8):25–27, Aug. 2008.

[10] C. Harteveld, G. Smith, G. Carmichael, E. Gee, and C. Stewart-Gardiner. A design-focused analysis of games teaching computer science. In *Proceedings of Games+Learning+Society 10*, 2014.

[11] F. Ke and B. Grabowski. Gameplaying for maths learning: Cooperative or not? *British Journal of Educational Technology*, 38(2):249–259, 2007.

[12] E. Z.-F. Liu, C.-Y. Lee, and J.-H. Chen. Developing a new computer game attitude scale for Taiwanese early adolescents. *Journal of Educational Technology & Society*, 16(1):183–193, 2013.

[13] M. Mateas. Procedural literacy: Educating the new media practitioner. In D. Davidson, editor, *Beyond Fun*, pages 67–83. ETC Press, Pittsburgh, PA, USA, 2008.

[14] A. Munson, B. Moskal, A. Harriger, T. Lauriski-Karriker, and D. Heersink. Computing at the high school level: Changing what teachers and students know and believe. *Computers & Education*, 57(2):1836–1849, 2011.

[15] S. Papert. Does easy do it? Children, games, and learning. *Game Developer*, June 1998.

[16] J. L. Plass, P. A. O'Keefe, B. D. Homer, J. Case, E. O. Hayward, M. Stein, and K. Perlin. The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. *Journal of Educational Psychology*, 105(4):1050, 2013.

[17] S. G. Ray. *Gender inclusive game design: Expanding the market*. Cengage Learning, 2004.

[18] M. Resnick, J. Maloney, A. Monroy-Hernández, N. Rusk, E. Eastmond, K. Brennan, A. Millner, E. Rosenbaum, J. Silver, B. Silverman, and Y. Kafai. Scratch: Programming for all. *Commun. ACM*, 52(11):60–67, Nov. 2009.

[19] A. Smith and M. Mateas. Answer set programming for procedural content generation: A design space approach. *Computational Intelligence and AI in Games, IEEE Transactions on*, 3(3):187 –200, Sept. 2011.

[20] A. M. Smith, E. Andersen, M. Mateas, and Z. Popovic. A case study of expressively constrainable level design automation tools for a puzzle game. In *Proceedings of the 2012 Conference on the Foundations of Digital Games*, Raleigh, NC, June 2012.

[21] G. Smith. Understanding procedural content generation: A design-centric analysis of the role of PCG in games. In *Proceedings of the 2014 ACM Conference on Computer-Human Interaction*, Toronto, Canada, Apr. 2014.

[22] G. Smith and C. Harteveld. Procedural content generation as an opportunity to foster collaborative mindful learning. In *Workshop on Games and Learning, co-located with Foundations of Digital Games 2013*, Chania, Greece, May 2013.

[23] D. Thompson and T. . Virtually unplugged: Rich data capture to evaluate CS pedagogy in 3D virtual worlds. In *Learning and Teaching in Computing and Engineering (LaTiCE), 2015 International Conference on*, pages 156–163, Apr. 2015.

[24] J. Togelius, G. Yannakakis, K. Stanley, and C. Browne. Search-based procedural content generation: A taxonomy and survey. *Computational Intelligence and AI in Games, IEEE Transactions on*, 3(3):172 –186, Sept. 2011.

[25] A. Tucker, F. Deek, J. Jones, D. McCowan, C. Stephenson, and A. Verno. ACM K-12 CS model curriculum, 2nd edition. Technical report, Computer Science Teachers Association, 2006.

[26] United States Department of Commerce. Women in STEM: A gender gap to innovation. Technical report, Aug. 2011.

[27] L. A. Williams and R. R. Kessler. All I really need to know about pair programming I learned in kindergarten. *Commun. ACM*, 43(5):108–114, May 2000.

[28] J. M. Wing. Computational thinking. *Communications of the ACM*, 49(3):33–35, 2006.