Investigating Student Reflection during Game-Based Learning in Middle Grades Science

Dan Carpenter
North Carolina State University

Elizabeth Cloude
University of Central Florida

Jonathan Rowe
North Carolina State University

Roger Azevedo
University of Central Florida

James Lester
North Carolina State University

ABSTRACT

Reflection plays a critical role in learning by encouraging students to contemplate their knowledge and previous learning experiences to inform their future actions and higher-order thinking, such as reasoning and problem solving. Reflection is particularly important in inquiry-driven learning scenarios where students have the freedom to set goals and regulate their own learning. However, despite the importance of reflection in learning, there are significant theoretical, methodological, and analytical challenges posed by measuring, modeling, and supporting reflection. This paper presents results from a classroom study to investigate middle-school students' reflection during inquiry-driven learning with CRYSTAL ISLAND, a game-based learning environment for middle-school microbiology. To collect evidence of reflection during game-based learning, we used embedded reflection prompts to elicit written reflections during students' interactions with CRYSTAL ISLAND. Results from analysis of data from 105 students highlight relationships between features of students' reflections and learning outcomes related to both science content knowledge and problem solving. We consider implications for building adaptive support in game-based learning environments to foster deep reflection and enhance learning, and we identify key features in students' problem-solving actions and reflections that are predictive of reflection depth. These findings present a foundation for providing adaptive support for reflection during game-based learning.

CCS CONCEPTS

• **Applied computing** → Education; Interactive learning environments; Computers in other domains; Personal computers and PC applications; Computer games.

KEYWORDS

Self-Regulated Learning, Game-Based Learning, Reflection

ACM Reference Format:
Dan Carpenter, Elizabeth Cloude, Jonathan Rowe, Roger Azevedo, and James Lester. 2021. Investigating Student Reflection during Game-Based Learning in Middle Grades Science. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3448139.3448166

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448166

1 INTRODUCTION

Reflection is an essential precursor for developing higher-order thinking skills such as effective problem solving [28], and it can be an effective learning tool when strategies for teaching and encouraging deep reflection are used [22]. During inquiry-driven learning, it is critical for students to effectively reflect on their knowledge, skills, and previous learning experiences to achieve success [11, 28]. A promising direction for supporting students' reflection is game-based learning. With games becoming ubiquitous in students' lives, game-based learning holds significant promise for promoting students' engagement during learning [5, 9]. For instance, game-based learning environments that are strategically designed to incorporate narratives centered around authentic problem scenarios have been shown to simultaneously improve students' motivation, emotional engagement, and learning outcomes [12, 23, 29]. However, challenges remain because students must utilize a range of self-regulation strategies such as monitoring their level of understanding to determine if they are making progress toward successful problem-solving outcomes [2, 31, 36]. Game-based learning environments present a promising opportunity to scaffold higher-order thinking skills, but significant gaps remain as most studies do not incorporate reflective thinking during game-based learning, which serves as the basis for developing higher-order thinking skills such as scientific reasoning.

To address these gaps, we investigated students' reflections during inquiry-driven science problem solving with embedded reflection prompts in CRYSTAL ISLAND, a game-based learning environment for middle school microbiology. Students' written reflections were analyzed to gauge how students reflected during game-based learning and the extent to which features of written reflections during game-based learning were related to learning outcomes. We also modeled the depth at which students reflected over time and identified critical features that were predictive of their reflection depth and learning outcomes. We consider the implications of the findings for building adaptive scaffolding into game-based learning environments to foster students' reflections and enhance higher-order thinking skills and learning.

2 THEORETICAL BACKGROUND AND RELATED WORK

Reflection is defined as deliberate contemplation, and in order to reflect effectively, one needs to be conscious of one's own reflective thinking via introspection, or the process of looking within

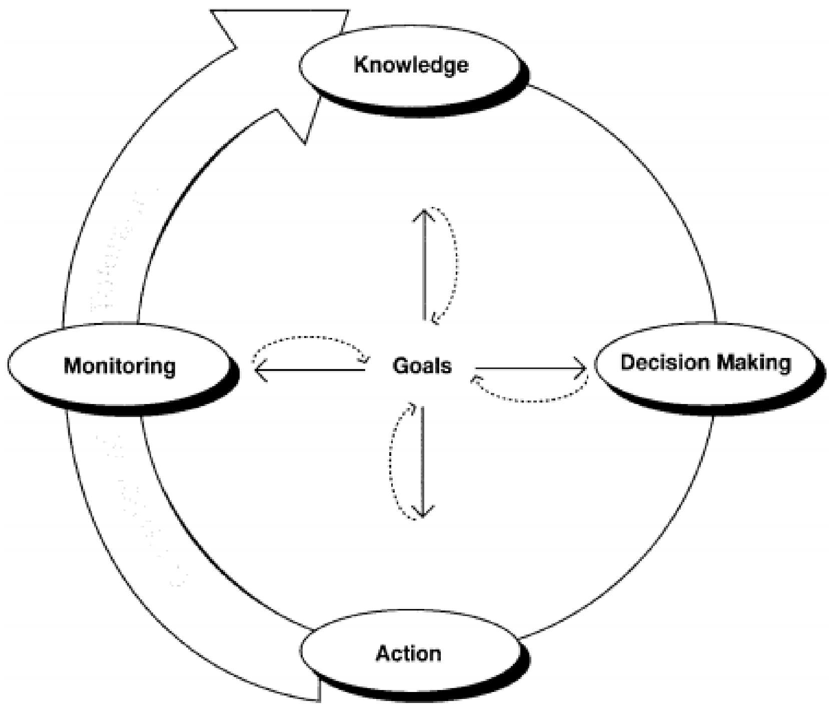

Figure 1: McAlpine et al.'s [16] model of reflection

to observe one's own thoughts, feelings, and reactions to previous experiences or stimuli [28]. The process of both reflective thinking and introspection leads to the development of self-knowledge (e.g., self-efficacy), which fosters the development of metacognition (e.g., monitoring one's understanding during learning activities) [7, 28]. Reflection lays the foundation for developing higher-order thinking skills such as metacognitive monitoring and problem solving [15, 18]. These findings are supported by the model of reflection by McAlpine et al. [16] (Figure 1), where reflection is formulated as continuously and dynamically interacting with both actions and knowledge in service of learning goals (e.g., identifying the cause of an illness). Specifically, the model describes six main components: (1) goals, (2) knowledge, (3) actions, (4) monitoring, (5) decision making, and (6) the corridor of tolerance. In this model, once goals are identified and set, a student will construct a plan to achieve their goals by reflecting on their knowledge, which guides their subsequent actions. Students evaluate and revise their actions by monitoring their progress, where change in actions (e.g., using a new strategy) will be determined by their corridor of tolerance in relation to progress made toward achieving learning goals. Specifically, the corridor of tolerance determines whether monitoring processes will result in a decision to change actions [16]. Significant gaps exist in the literature as McAlpine et al.'s [16] model of reflection has yet to guide the study of reflection with game-based learning environments. We argue that this model could provide insight into the dynamics of reflective thinking within game-based learning environments since these systems are strategically designed around specific learning goals [20].

Game-based learning with narrative elements can simultaneously promote student learning and student engagement. In a study with a previous version of Crystal Island, researchers found a strong positive relationship between students' learning outcomes and engagement [23]. These benefits were seen in students regardless of their prior content knowledge and game playing experience. Another study with a game-based learning environment for introductory computer programming found that including a strong focus on narrative led to increased engagement and therefore improved learning outcomes [12]. A complementary benefit of game-based learning environments is that they can be used to study students' self-regulated learning and higher-order thinking skills. For instance, in a recent study with Decimal Point, a game designed to teach decimals to middle school students, researchers investigated students' ability to regulate their own learning by allowing them to decide which learning activities to complete and in what order [10]. A recent study with Crystal Island investigated the benefits of word embedding representations of reflection for predictive modeling of post-test scores, finding that models that incorporated reflections performed significantly better than those that did not [8]. The current work makes use of interpretable features of reflection rather than word embeddings to predict science content knowledge and also investigates relationships between students' reflections and problem-solving success.

To collect evidence of reflection, free-response reflection prompts are commonly used [32]. Reflections can be collected at certain points during a learning experience as a reflection on progress or after the learning experience has been completed as an overarching reflection on the learning process as a whole [1]. Most research on reflection has explored reflection in learning by post-secondary students and has focused on extended segments of reflective writing, such as essays [13, 14, 33]. However, since higher-order thinking skills such as reflection are important from a young age, there is a need for work that explores and evaluates the relationship between young students' reflections and their learning outcomes. To provide support for reflection during learning, a framework for assessing reflection is necessary. In previous work, written reflection has often been assessed along two dimensions: a dimension of *reflection depth*, which captures the extent to which the writing is reflective; and a dimension of *reflection breadth*, which addresses the range of different topics related to reflection [13, 32]. Depth is often evaluated on an ordinal scale, such as from non-reflective to slightly reflective to highly reflective [14, 34], while breadth may consider aspects such as 'attending to feelings', 'validation' [35], 'justification' [21], 'analysis', and 'perspective' [13]. In this work, we evaluate reflections solely in terms of their depth because middle school students' reflections tend to be short and therefore often lack breadth.

This paper addresses the need to understand how young students reflect on their knowledge and learning processes during game-based learning. Reflection prompts were embedded in a game-based learning environment to collect evidence of reflection over the course of a learning experience. After collecting students' responses to reflection prompts, we scored students' reflections in terms of their depth and extracted linguistic features related to students' reflections and their problem-solving actions to address the following research questions:

- **RQ1:** *Can science content knowledge and science problem-solving learning outcomes be predicted by features of students' reflections and problem-solving actions during learning?* Prior work points to challenges in predicting learning outcomes based on problem-solving actions [24], while other research has demonstrated the importance of reflection features in predicting learning outcomes [8]. We investigate a different set of actions and reflection features, but we expect to find that reflection features offer significant benefits over problem-solving actions alone for predicting learning outcomes. Specifically, we expect problem-solving actions [26] and reflection features to be predictive of problem-solving success based on McAlpine et al.'s [16] model of reflection.
- **RQ2:** *How does the depth of students' reflections change over the course of learning with Crystal Island?* Previous work suggests that students may require substantial guidance on how to reflect effectively [22]. In the absence of such support, we expect that the depth of students' reflections will not increase over the course of their learning experience and may even decrease.
- **RQ3:** *Can reflection depth be predicted by features of students' reflections and problem-solving actions?* Recent work indicates that reflection features can be used to predict depth [13], and we expect to find that problem-solving actions are also predictive.

3 CRYSTAL ISLAND GAME-BASED LEARNING ENVIRONMENT

During learning with Crystal Island, students are asked to investigate the mysterious outbreak of a disease on a remote island. To solve the mystery, students must identify the source of the disease (one of several food items that sick individuals had recently eaten), determine the pathogen that is spreading (either a pathogenic virus or bacteria), and recommend a prevention or treatment for the disease (either vaccination or bed rest). The specific scenario is randomly generated when the game is started, selecting the infected food item and either virus or bacteria as the pathogen. To gather the information needed to solve the mystery, students interact with non-player characters (e.g., Elise the scientist; see Figure 2), read books and research articles covering microbiology topics, view science posters and diagrams, and test objects for pathogens using a virtual scanner. Non-player characters (NPCs) include sick patients who describe their symptoms and recent activities prior to falling ill, bacteria and virus experts, a camp cook who provides information related to what residents have been eating lately, and a camp nurse who provides guidance and support to the student during their investigation. Findings related to the mystery are compiled by students in a virtual diagnosis worksheet which is submitted to the camp nurse as a final diagnosis.

Crystal Island's science problem solving aligns with the Next Generation Science Standards' [27] focus on the nature and practice of scientific inquiry, and text resources align with standards from the Common Core State Standards for English Language Arts on Reading: Informational Text [17].

To collect evidence of students' reflection while they interacted with Crystal Island, the game-based learning environment was augmented with embedded reflection prompts. At critical points during their investigation, students were prompted to reflect on their knowledge and problem-solving actions (see Figure 3). The prompts were designed according to McAlpine et al.'s [16] model of reflection, where students were prompted to contemplate the important information previously learned during their investigation as well as to set goals for solving the outbreak scenario and improving their microbiology content knowledge. Specifically, the prompts asked students, "Please describe the most important things that you've learned so far, and what is your plan moving forward?"

These prompts were triggered by a set of production rules associated with important actions taken by the student as they interacted with Crystal Island (see Table 1). The triggers were designed to align with milestones in the game's narrative (e.g., speaking with the camp nurse), students' acquisition of domain knowledge (e.g., viewing 6 microbiology texts that covered information specific to the illness plaguing the camp), and problem-solving actions (e.g., obtaining a positive test result). Based on which action triggered the reflection prompt, students received a message telling them why this would be a good time to reflect, followed by the prompt (Figure 3). The reflection prompts were queued immediately after each trigger, but the delivery of the prompts was scheduled to minimize disruption to students' gameplay experiences. Specifically, students were only prompted to reflect when they left a building in the game environment. Because reading materials, science diagrams, testing equipment, and virtual characters are all located in virtual buildings

Figure 2: The Crystal Island game-based learning environment

Table 1: Triggers used to prompt reflection

Trigger	Prompt
Briefed by the camp nurse	Agent, it looks like you've spoken with the camp nurse. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?
Viewed 6 microbiology texts	Agent, it looks like you've found several materials that might be useful. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?
Testing contaminated object	Agent, it looks like you found an object that tested positive for pathogenic contaminants. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?
After submitting diagnosis worksheet with the wrong solution	Agent, it looks like you're making progress on diagnosing the illness, but you're not quite there yet. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?
End of game (solved mystery or not)	Please explain how you approached solving the mystery. If you were asked to solve a similar problem in the future, what would you do the same and/or differently?

in Crystal Island, prompting for reflection only when students were outside allowed us to avoid disrupting students when they were directly amid information gathering. Additionally, successive reflection prompts occurred at least 15 minutes apart. This is because, with the triggers being event-based, it is possible that a student may activate several reflection triggers in rapid succession. However, we wanted students to be able to maintain focus while gathering information and working on their investigation.

In addition to the in-game reflection prompts, we also included summary reflection prompts at the end of the learning experience. These prompts asked students to reflect on their overall approach to solving the mystery and to consider what they might do differently when faced with similar problems in the future. Student responses to the post-game summary reflection prompts were not analyzed in this paper because we wanted to focus on how students used reflection to inform adaptations during learning.

Figure 3: Embedded reflection prompt

4 METHOD

4.1 Participants

This paper analyzed data from a classroom study conducted in 2018 and 2019. A total of 153 middle school students participated in the study, but only 118 students reported demographic information. Of these students, 51% identified as female and ranged from 13-14 years of age (M=13.6, SD=0.51), with 43 students identifying as Caucasian/White, 32 as African American, 21 as Hispanic or Latino, and 3 as Asian. None of the students reported that they had previously interacted with Crystal Island. After removing students who had missing pre or post-test data, the dataset contained 105 students. These students were removed to ensure that analysis of students' science content knowledge learning outcomes could control for prior content knowledge. All analyses used this reduced dataset to maintain consistency.

4.2 Procedure

Students interacted with Crystal Island over the course of two or three class periods, on average spending 81.4 minutes in the learning environment. In the week preceding the classroom study, we administered pre-survey instruments and a 17-item multiple-choice microbiology pre-test that covered both factual (e.g., "What is the smallest type of living organism?") and procedural items (e.g., "Your lab partners are examining a pathogen through a microscope and have observed that it is smooth and round in shape. What pathogen are your lab partners probably looking at?"). On the first day of the classroom study, students were introduced to the game by a researcher and then shown a brief video detailing the problem scenario they would face in Crystal Island. Afterward, students interacted with Crystal Island until they completed the mystery or ran out of class time, which roughly allowed for 100 minutes of gameplay over two or three days. On the final day of the classroom study, students took a post-study survey including questionnaires to capture presence and engagement and a 17-item microbiology post-test that was similar but different from the pre-test. Both tests assessed the same knowledge, but questions were presented differently. For example, the pre-test asked, "How do vaccines protect you?", while the post-test asked, "What role do vaccines play in your immune system?" The average pre-test score was 6.90 (SD=2.70) and the average post-test score was 7.30 (SD=3.35).

4.3 Coding and Scoring

4.3.1 Outcome Variables. Crystal Island supports learning goals related to science content knowledge acquisition as well as science reasoning and problem solving. While much of the science content knowledge included in the learning environment is helpful for solving the mystery, some of the information is irrelevant depending on what random problem configuration was selected. That is, if the type of pathogen that is spreading is a virus, information on bacteria is less useful to students whose only goal is to solve the mystery. However, the post-test assesses cumulative knowledge on science content in the game, so students whose goal is to perform well on the post-test should try to acquire as much science content knowledge as possible. Since, according to McAlpine et al.'s [16] model of reflection, students' learning goals guide reflection, we modeled two outcome variables based on the different learning goals. These were (1) science content learning outcomes, which were measured by students' post-test scores, and (2) problem-solving performance, which was based on whether students solved

Table 2: Reflection depth rubric

Rating	Characteristics	Examples
1	Lacks both a hypothesis and plan. The student does not demonstrate awareness of their own knowledge or goals.	"Each clue will help with solving the problem"; "Yeah cool game I learned science"
2	Presents a vague hypothesis or plan, often directly restating information that was presented in the game. The student demonstrates awareness of their own knowledge and goals but does not show that they are evaluating their knowledge to inform their future actions.	"That the illness causing the people being sick might be pathogen"; "I found out that the egg has bacteria"; "I think I am going to talk to other people"
3	Presents a clear hypothesis or plan *without* any reasoning. This demonstrates that the student has evaluated their knowledge and made connections to their goals. However, they have not articulated the reasoning behind the importance of this knowledge or its benefit toward achieving their goals.	"Getting more information off the food I think it has something to do with the food"; "The most important thing is how the illness is spreading"
4	Presents a clear hypothesis *or* plan with reasoning. This demonstrates that the student has evaluated their knowledge and made connections to their goals. However, they have only provided reasoning for the importance of this knowledge *or* its benefit toward achieving their goals, not both.	"I plan on questioning the cook as they know more about the food and how it could be contaminated with viruses or bacteria"; "I need to learn more about what the sick people do on a day to day schedule"
5	Presents both a clear hypothesis *and* plan with reasoning. This demonstrates that the student has evaluated what they have learned and made connections to their goals. Furthermore, they have provided reasoning for the importance of this knowledge, *and* they have indicated how it will help them achieve their goals.	"I think that it might have to do with salmonella because when I tested the milk it was positive with pathogenic bacteria. I think that I will test things that can be contaminated"

the science mystery scenario. We investigate the relationship between students' problem-solving actions and reflection features and both of these outcome variables.

4.3.2 Scoring Reflection Depth. To assess the depth of students' reflections during their interaction with Crystal Island, a rubric that one of the authors previously helped to develop [4] was used (see Table 2). This rubric scores reflection depth on a scale from 1 (not reflective) to 5 (highly reflective). While other work that has assessed written reflections has used complementary dimensions of depth and breadth [13, 14, 33], these reflections were assessed solely in terms of their depth because the reflections were often short (M=18.6 words, SD=14.0) and therefore limited in reflective breadth.

The rubric was developed by two researchers using a grounded theory approach [25]. First, the researchers worked together to identify reflections that were particularly weak and discussed what stood out about them. They found that these reflections lacked any commentary on knowledge or a plan of action, were too abstract to be meaningful, or were largely unactionable. These insights formed the basis for a reflection depth score of 1 (see Table 2 for examples). Next, to inform a reflection depth score of 5, the researchers identified some reflections that were particularly strong and discussed them. These reflections presented both a clear hypothesis regarding the current problem and a plan that was supported by reasoning, or laid out a high-quality sequence of abstract plans, thereby demonstrating significant information processing by the student. Qualities for the remaining reflection depth scores were similarly determined.

Each of the remaining reflections was annotated by both researchers and a final reflection depth score was obtained by averaging the two scores. An intraclass correlation of 0.669 was achieved, indicating moderate inter-rater reliability. Across all reflections, the average depth score was 2.41 (SD=0.86).

4.3.3 Feature Extraction. Features related to students' problem-solving actions and their reflections during Crystal Island were used in this work. Problem-solving actions capture students' progress through the game's narrative, their acquisition of science content knowledge, and their problem-solving processes. A narrative-related feature that we explored was the number of plot points completed by the student. Plot points included speaking with a character for the first time, learning about viruses and bacteria, finding out what symptoms patients had and what foods they had recently eaten, and testing objects for pathogens. For science content knowledge features, we looked at the number of microbiology books that the student read since books are a primary source of microbiology information in the game. Conversations with NPCs were also considered a major source of science content knowledge since students can speak with a virus expert, a bacteria expert, and an expert on how diseases spread between people. Finally, features related to science problem-solving processes included the number of times items were tested for pathogens, the number of times the diagnosis worksheet was submitted with a final diagnosis, and whether a positive test result was obtained by the end of the learning experience. The number of tests and diagnosis worksheet submissions are related to problem-solving efficiency [30], and the positive test result is related to problem-solving performance.

Features based on students' reflections were also derived. First, the number of words in each reflection was calculated and the amount of time between when a student was prompted for reflection and when they submitted the reflection was extracted from Crystal Island's trace logs. Next, the number of domain-specific words used in each reflection was calculated. The dictionary of keywords used for this feature was determined by first extracting all text content from Crystal Island (i.e., books, articles, and conversations with non-player characters). Then, stop words were filtered out according to the Natural Language Toolkit [3] of English stop words, words were stemmed to reduce them to their root form, and the 100 most common terms were selected. Candidate keywords were then curated by the authors and finally cross-referenced with an educational standards document that was used to develop Crystal Island's educational content, resulting in a final set of 36 keywords. The top 5 most common keywords were bacteria, disease, virus, cell, and infection. This approach was taken to ensure that domain-specific keywords were identified systematically.

In addition to these reflection features, a subset of the features included in the LIWC text analysis tool [19] was used to capture information related to the content of each reflection. Rather than using all the available LIWC features, we used the five most predictive features according to a recent study on automated reflection depth assessment [13]. Following Jung & Wise [13], the features that we used were *Focuspast* and *Focusfuture* (the extent to which the text focuses on the past or the future), *Authentic* (how honest, humble, and vulnerable the text is), *I* (the extent to which the text includes first person pronouns), and *Analytic* (how much the text focuses on formal, logical, and hierarchical thinking patterns).

4.4 Statistical Analysis

To answer our research questions, we utilized hierarchical regression analysis and latent growth curve modeling. Hierarchical regression analysis provides a framework for determining which variables explain statistically significant variance in a dependent variable after accounting for all predictors. This allows us to see which predictors are related to the outcome variable. In this paper, we investigate whether features related to students' problem-solving actions and reflection responses were related to post-test scores, and whether students successfully solved the mystery.

Latent growth curve modeling provides a framework for modeling trajectories of repeated measures over time for a group and can control for factors that may account for different trajectories. Growth models estimate an initial value for the variable being modeled as well as an estimate of how that variable changes over time. We investigated the growth curves of students' reflection depth ratings over subsequent prompts for reflection during Crystal Island. In addition, we accounted for students' pre-test scores to determine whether students' prior content knowledge impacted their depth of reflection.

5 RESULTS

RQ1: Can science content knowledge and science problem-solving learning outcomes be predicted by features of students' reflections and problem-solving actions during learning? We performed hierarchical regression analysis to explore whether features related to students' in-game reflections were predictive of students' science content knowledge and science problem-solving outcomes, as defined above.

The hierarchical regression analysis was conducted in three steps, with results shown in Table 3. First, we predicted the outcome variables using only students' pre test scores and found this to be a significant predictor of post test score (β=0.464, $p < .001$) and a significant predictor of mystery solution (β=0.010, $p < .05$). Next, we incorporated features related to students' problem-solving actions that were based on our reflection prompting triggers. When predicting post test score, none of these features were found to be significant. When predicting whether the student solved the mystery, the number of plot points completed (β=0.024, $p < .001$) and a positive test result (β=0.389, $p < .001$) were both found to be significantly predictive.

For the third step, we extracted features that summarize how students reflected over the course of their interaction with Crystal Island. For predicting post test score, average reflection depth rating (β=1.455, $p < .01$) and the average amount of time spent on each reflection (β= -0.029, $p < .05$) were found to be significant. Looking at whether the student solved the mystery, these two features were again found to be significant (β=0.096, $p < .05$; β= -0.004, $p < .05$). We also incorporated a subset of LIWC features, which capture emotional, cognitive, and structural components of natural language. For predicting post test score, none of the LIWC features were found to be significant. For predicting whether the student solved the mystery, *Focuspast* was nearly found to be significant (β=0.033, $p < .1$).

These results align with Sabourin et al. [24], where relationships between problem-solving actions and science content knowledge learning outcomes were not observed, as well as with prior research that has demonstrated the importance of reflection features for predictive student modeling [8].

RQ2: How does the depth of students' reflections change over the course of learning in Crystal Island? Among the 105 students who had complete data, students had different numbers of in-game reflections. Four students only completed one in-game reflection, so they were omitted from latent growth curve analysis because a single observation is insufficient to model the trajectory of reflection depth over time. The students who completed exactly two in-game reflections (N=20), exactly three in-game reflections (N=43), and exactly four in-game reflections (N=38) were modeled by a single growth curve model, since growth curves can account for partially missing data [6].

Results for a growth model accounting for linear growth and a baseline no growth model are presented in Table 4. According to a chi-square goodness of fit test, the linear growth model very nearly fit the data ($p < .1$) but the no growth model did not. We found that students' reflection depth over the course of several prompts for reflection could be described by a line with an intercept of 2.263 and a slope of -0.003. That is, for students' first reflections during their interaction with Crystal Island, they typically received a reflection depth rating of 2.263. The slope of -0.003 indicates that the depth of students' reflections barely changed over the course of subsequent reflections. Although the average reflection depth for students' fourth in-game reflection (*M*=2.42, *SD*=0.75) was higher than the average reflection depth for their first reflection (*M*=2.30,

Table 3: Hierarchical regression for predicting post test score and whether the student solved the mystery. All regression coefficients are from the final step in the analyses. *** p < .001, ** p < .01, * p < .05, † p < .1

Predictor	Post-Test Score			Mystery Solved		
	β	ΔR^2	F	β	ΔR^2	F
Step 1: Prior Content Knowledge		0.240	34.824***		0.0318	5.1907*
Pre-Test Score	0.464	0.240***		0.010	0.0318*	
Step 2: Problem-Solving Actions		0.016	0.4599		0.3037	9.9129***
Number of Plot Points Completed	0.022	0.002		0.024	0.1043***	
Number of Books Read	0.001	0.004		-0.008	0.0205†	
Number of Worksheet Submissions	0.062	0.003		0.029	0.0003	
Number of Lab Tests	-0.012	0.003		-0.002	0.0025	
Obtained a Positive Test Result	-0.637	0.005		0.389	0.1761***	
Step 3: Reflection Features		0.138	2.2469*		0.1191	2.1604*
Reflection Depth Rating	1.455	0.075**		0.096	0.0346*	
Number of Words	0.024	0.003		0.008	0.0003	
Number of Keywords	0.313	0.006		0.044	0.0065	
Time Spent Reflecting	-0.029	0.029*		-0.004	0.0408*	
Focuspast	0.242	0.013		0.033	0.0210†	
Focusfuture	0.260	0.005		-0.016	0.0022	
Authentic	-0.010	0.003		-0.002	0.0042	
I	-0.042	0.002		0.017	0.0034	
Analytic	0.010	0.001		0.002	0.0059	

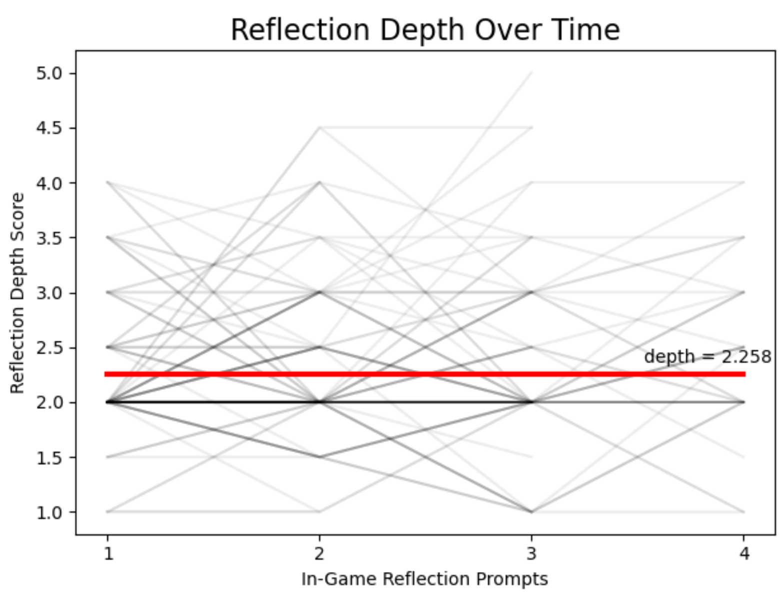

Figure 4: Students' reflection depth over the course of their interaction with Crystal Island. Darker lines represent more common trajectories, and the red line represents the model implied trajectory over all students.

SD=0.65), reflection depth does not change significantly. To further explore, we compared the linear growth model to the baseline no growth model, where students' reflection depth was found to remain at a constant score of 2.258, using a Chi-square difference test. We found that the two were not significantly different, leading to the conclusion that students' reflection depth does not change over the course of several reflection responses. The no growth trajectory compared to students' individual trajectories can be seen in Figure 4. The most common student trajectories are represented by darker lines in the graph.

Table 4: Growth model parameter estimates with standard error in parentheses and fit indices. * p < .05

	Intercept	Slope	df	Chi-Square	AIC	BIC
Linear Growth	2.263 (0.147)*	-0.003 (0.061)	10	17.099	314.19	327.29
No Growth	2.258 (0.134)*	0	14	20.233	309.32	315.87

Table 5: Hierarchical regression results for predicting average reflection depth. All regression coefficients are from the final step in the analyses. *** p < .001, ** p < .01, * p < .05, † p < .1

Predictor	Reflection Depth		
	β	ΔR^2	F
Step 1: Prior Content Knowledge		0.0453	13.3242***
Pre Test Score	0.001	0.0453***	
Step 2: Problem-Solving Actions		0.1690	9.9371***
Number of Plot Points Completed	0.001	0.0044	
Number of Books Read	-0.005	0.0000	
Number of Worksheet Submissions	-0.029	0.0951***	
Number of Lab Tests	0.000	0.0050	
Obtained a Positive Test Result	0.204	0.0644***	
Step 3: Reflection Features		0.4796	17.6252***
Number of Words	0.036	0.4331***	
Number of Keywords	0.074	0.0111†	
Time Spent Reflecting	-0.003	0.0069	
Focuspast	0.036	0.0051	
Focusfuture	0.027	0.0001	
Authentic	0.000	0.0001	
I	-0.008	0.0069	
Analytic	0.005	0.0163*	

To examine the effect that having a high or low pre-test score, as determined by a median split (median=7, N_{low}=45, N_{high}=56), has on students' initial reflection depth and changes in reflection depth over time, we included pre-test as a time-invariant covariate in our latent growth curve model. We found that there were no significant effects of students' pre-test performance on reflection depth during Crystal Island.

RQ3: Can reflection depth be predicted by features of students' reflections and problem-solving actions? To identify features that are predictive of students' reflection depth, we conducted another hierarchical regression analysis (Table 5). The same features were investigated as in RQ1, although reflection depth was removed as an independent variable and instead used as the outcome variable. While we did not find an effect of pre-test score on students' reflection depth trajectories, it was a significant predictor of average reflection depth (β= .001, p <.001). Problem-solving actions that were found to predict students' average reflection rating were the number of times the diagnosis worksheet was submitted (β= -.029, p <.001) and whether a positive test result was obtained (β=0.204, p < .001). For features related to students' reflections, the average number of words per reflection was predictive of reflection depth (β=0.036, p < .001). Also, students' use of analytic language, as measured by the LIWC Analytic feature, was predictive of reflection depth (β=0.005, p < .05).

6 DISCUSSION

Reflection is an essential precursor for developing higher-order thinking skills such as effective problem solving and knowledge acquisition [28]. This study investigated middle-school students' reflections during inquiry-driven learning within Crystal Island in science classrooms, and we consider each of the three research questions and the implications of this work below.

RQ1: Can science content knowledge and science problem-solving learning outcomes be predicted by features of students' reflections and problem-solving actions during learning? To examine the extent to which learning and performance outcomes were predicted by features of students' reflections and problem-solving actions during learning with Crystal Island, we conducted a hierarchical regression analysis. Results suggested that features extracted from students' reflections were predictive of post-test scores, but problem-solving actions were not. Based on previous work with data from another Crystal Island study [24], we had expected students' problem-solving actions to offer insufficient information for predicting science content knowledge learning outcomes. Yet, while students' problem-solving actions were not predictive of post-test scores, we found that some features of their reflections were, which is consistent with prior findings by Geden et al. [8]. Over the course of their interaction with Crystal Island, students' average reflection depth and the average amount of time spent on each reflection were predictive of post-test scores. The results suggest that students who are deeper or more efficient reflectors may be more successful in this learning environment. Interestingly, the average number of words per reflection and the number of domain-specific words used in students' reflections were not predictive of post-test scores. Thus, the amount of externalized reflection seems to be less important than the *depth* and *efficiency* of reflection for learning outcomes. In future work, we aim to investigate whether students who performed well on the post-test were both deep and efficient reflectors, or if there is an observable tradeoff between depth and efficiency as they relate to science content knowledge learning outcomes.

Next, we analyzed whether there were relationships between successful problem solving (i.e., correctly solving the mystery) and problem-solving actions and reflection features. The models suggested that some problem-solving actions were predictive of successful problem solving. Specifically, the number of plot points completed and whether the student received a positive test result on one of their virtual tests for a pathogen were positively related to solving the mystery. With the goal of Crystal Island requiring students to identify the type of pathogen that is spreading and the food item through which it is spreading, it is clear why receiving a positive test result would be predictive of solving the mystery - after getting the positive result, all students need to do is determine a viable treatment or prevention plan and present this information to the camp nurse. As for the number of plot points completed, this feature indicates how widely students have explored the virtual environment. Thus, it seems reasonable that students who have explored the game more completely would be able to use the information they collected to solve the mystery. This relationship between problem-solving actions, especially those related to testing hypotheses, aligns with prior findings [26]. As for features related to students' reflections, average reflection depth and the average amount of time per reflection were found to be predictive of successful problem solving, such that deeper and more efficient reflection was positively related to problem-solving success.

RQ2: How does the depth of students' reflections change over the course of learning in Crystal Island? In this study, we found that students' reflection depth did not change significantly over the course of their learning experience in Crystal Island. Based on Riedinger [22], we had hypothesized that, in the absence of any support for reflection, reflections would not increase in depth over time and might even decrease in depth. We hypothesized that students might become less motivated to engage in deep reflection over time and would instead quickly provide a shallow reflection to satisfy the bare minimum requirement, leading to a decrease in reflection depth scores. Additionally, since we have observed that some students get stuck in their investigations and have difficulty making progress toward solving the mystery, we hypothesized that students might engage in shallower reflection over time as a combined result of increased frustration and a lack of new information to reflect on.

Using latent growth curve modeling, we found that students' reflection depth remained mostly constant over the course of their learning experience. According to the growth model of students' reflection depth, depth scores tended to stay at approximately 2.3. This roughly corresponds to a score of 2 on the rubric, which indicates that students' reflections were either too vague to be very useful or were mostly direct restatements of information found in the game without any processing by the student. This demonstrates that students were often aware of their knowledge and learning goals, but did not evaluate this knowledge to help drive adaptations for the future. Moreover, we found that there was not a significant impact of students' prior science content knowledge, as indicated by their pre-test scores, on reflection depth. Thus, since all students' reflections were consistently shallow, there is clearly a need to provide support for reflection during game-based learning.

RQ3: Can reflection depth be predicted by features of students' reflections and problem-solving actions? Building toward the future goal of adaptively supporting reflection during game-based learning, we identified several problem-solving actions and reflection features that were predictive of average reflection depth. For problem-solving actions, the number of times that students submitted the diagnosis worksheet and whether they received a positive test result were predictive of reflection depth. These actions, which can be viewed as measures of science problem solving efficiency and science problem solving ability, respectively, align with previous work suggesting that problem-solving processes are related to higher-order thinking skills like goal setting and adaptation, which are related to reflection [30]. However, since these actions often occur at the end of students' interactions with Crystal Island, they do not hold much promise for adaptively identifying students who are not engaging in deep reflection.

Some features extracted from students' reflections were found to be predictive of reflection depth, as was the case in Jung & Wise [13]. First, the number of words per reflection was predictive of reflection depth, with longer reflections tending to have higher scores. This is not very surprising since it seems likely that a student's reflection would need to exceed some minimum length for them to adequately evaluate their knowledge and consider adaptations for the future. However, it may be the case that the "correct length" for a reflection could be a range rather than a lower bound. That is, if a student's reflection covers too much information, they may fail to derive actionable insights regarding the changes they should make to their learning processes. It will be important to identify guidelines for determining when a reflection is too short or too long to support meaningful adaptation during game-based learning. Nevertheless, the number of words in a reflection may be useful as a baseline indicator to encourage students to engage in deeper reflection. For example, when students try to submit a reflection that is too short, the system can simply ask them to provide some more information.

Specifically looking at the features borrowed from Jung & Wise [13], the LIWC Analytic feature was predictive of reflection depth. This feature presents an opportunity to ask students to focus more on logic and reasoning when reflecting. Leveraging these insights, future versions of Crystal Island can adaptively support reflection by providing feedback based on how students respond to the embedded reflection prompts.

7 CONCLUSION

Reflection on knowledge and past actions is critical for students to adapt their goals and actions to achieve desired learning outcomes.

Game-based learning, with the distinctive capacity to situate students in authentic problem-solving scenarios that require the use of higher-order thinking skills, is a promising setting for investigating reflection. We have presented a study with a game-based learning environment for middle school microbiology that features embedded prompts to elicit reflections from students. Results showed that students' reflections were predictive of learning outcomes more than problem-solving actions alone. It was also found that students consistently exhibited a lack of deep reflection and that their reflection depth did not change during the learning experience. The study also revealed that students' problem-solving actions and linguistic features of their reflections were predictive of overall reflection depth. The findings demonstrate the importance of collecting evidence of reflection to predict learning outcomes, as well as the potential to inform adaptive reflection scaffolding.

There are several promising directions for future research on reflection in game-based learning. First, it may be beneficial to explore reflections in relation to students' problem-solving actions to more accurately model predicted learning outcomes. For instance, it may be informative to look at the relationship between what students reflected on and the actions they took in the learning environment. Second, it will be important to investigate alternative methods for prompting students to reflect during game-based learning to promote deep reflection. For example, it may be beneficial to engage students in more directed reflection on specific aspects of their learning processes that are critical to their success rather than providing them with open-ended prompts. Additionally, it will be important to explore whether the timing of a reflection prompt during game-based learning impacts depth of reflection, and whether a system can learn when the best time to prompt for reflection is. Finally, it will be important to develop techniques for providing adaptive scaffolding for reflection using formative feedback. This will provide insight into methods that foster deeper reflection during inquiry-driven learning in game-based learning environments and whether such improvements benefit student learning outcomes.

ACKNOWLEDGMENTS

This research was supported by funding from the National Science Foundation under Grant DRL-1661202. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. Thanks to Robert Taylor, Michael Geden, Michelle Taub, Megan Wiedbusch, and Daryn Dever for their contributions to the study.

REFERENCES

[1] Ramin Akbari. 2007. Reflections on reflection: A critical appraisal of reflective practices in L2 teacher education. System 35, 2 (June 2007), 192-207. https://doi.org/10.1016/j.system.2006.12.008

[2] Brian R. Belland, Andrew E. Walker, Nam Ju Kim, and Mason Lefler. 2017. Synthesizing results from empirical research on computer-based scaffolding in STEM education: A meta-analysis. Review of Educational Research 87, 2 (October 2016), 309-344. https://doi.org/10.3102%2F0034654316670999

[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc.

[4] Dan Carpenter, Michael Geden, Jonathan Rowe, Roger Azevedo, and James Lester. 2020. Automated analysis of middle school students' written reflections during game-based learning. In Proceedings of the Twenty-First International Conference on Artificial Intelligence in Education. Ifrane, Morocco, 67-78.

[5] Geoffrey T. Crisp. 2014. Assessment in next generation learning spaces. The Future of Learning and Teaching in Next Generation Learning Spaces, 12 (July 2014), 85-100. https://doi.org/10.1108/S1479-362820140000012009

[6] Patrick J. Curran, Khawla Obeidat, and Diane Losardo. 2010. Twelve frequently asked questions about growth curve modeling. Journal of Cognition and Development 11, 2 (April 2010), 121-136. https://doi.org/10.1080/15248371003699969

[7] John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist 34, 10, 906-911. https://doi.org/10.1037/0003-066X.34.10.906

[8] Michael Geden, Andrew Emerson, Dan Carpenter, Jonathan Rowe, Roger Azevedo, and James Lester. 2020. Predictive student modeling in game-based learning environments with word embedding representations of reflection. International Journal of Artificial Intelligence in Education (October 2020), 1-23. https://doi.org/10.1007/s40593-020-00220-4

[9] Juho Hamari, David J. Shernoff, Elizabeth Rowe, Brianno Coller, Jodi Asbell-Clarke, and Teon Edwards. 2016. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. Computers in Human Behavior 54 (January 2016), 170-179. https://doi.org/10.1016/j.chb.2015.07.045

[10] Erik Harpstead, J. Elizabeth Richey, Huy Nguyen, and Bruce M. McLaren. 2019. Exploring the sublteties of agency and indirect control in digital learning games. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge. Tempe, Arizona, 121-129. https://doi.org/10.1145/3303772.3303797

[11] Cruz Izu and Brad Alexander. 2018. Using unstructured practice plus reflection to develop programming/problem-solving fluency. In Proceedings of the 20th Australasian Computing Education Conference. 25-34. https://doi.org/10.1145/3160489.3160496

[12] Chaima Jemmali, Sara Bunian, Andrea Mambretti, and Magy Seif El-Nasr. 2018. Educational game design: An empirical study of the effects of narrative. In Proceedings of the 13th International Conference on the Foundations of Digital Games. 1-10. https://doi.org/10.1145/3235765.3235783

[13] Yeonji Jung and Alyssa Wise. 2020. How and how well do students reflect? Multi-dimensional automated reflection assessment in health professions education. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. Frankfurt, Germany, 595-604. https://doi.org/10.1145/3375462.3375528

[14] Vitomir Kovanović, Srećko Joksimović, Negin Mirriahi, Ellen Blaine, Dragan Gašević, George Siemens, and Shane Dawson. 2018. Understand students' self-reflections through learning analytics. In Proceedings of the 8th International Conference on Learning Analytics & Knowledge. Tempe, Arizona, 389-398. https://doi.org/10.1145/3170358.3170374

[15] Tian Luo and John Baaki. 2019. Scaffolding problem-solving and instructional design processes: Engaging students in reflection-in-action and external representations in three online courses. Student-Centered Virtual Learning Environments in Higher Education, 40-69. https://doi.org/10.4018/978-1-5225-5769-2.ch003

[16] L. McAlpine, C. Weston, J. Beauchamp, C. Wiseman, and C. Beauchamp. 1999. Building a metacognitive model of reflection. Higher Education 37 (March 1999), 105-131. https://doi.org/10.1023/A:1003548425626

[17] National Governors Association Center for Best Practices & Council of Chief State School Officers. 2010. Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

[18] Niraj Patel, S. Glenn Baker, and Laura D. Scherer. 2019. Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs. Journal of Experimental Psychology: General 148, 12, 2129-2153. https://doi.org/10.1037/xge0000592

[19] James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2001. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates.

[20] Jan L. Plass, Richard E. Mayer, and Bruce D. Homer. 2020. Handbook of Game-Based Learning. MIT Press.

[21] Eric Poldner, Marieke Van der Schaaf, P. Robert-Jan Simons, Jan Van Tartwijk, and Guus Wijngaards. 2014. Assessing student teachers' reflective writing through quantitative content analysis. European Journal of Teacher Education 37, 3 (March 2014), 348-373. https://doi.org/10.1080/02619768.2014.892479

[22] Bonnie Riedinger. 2006. Mining for meaning: Teaching students how to reflect. In Handbook of Research on ePortfolios, 90-101. IGI Global.

[23] Jonathan P. Rowe, Lucy R. Shores, Bradford W. Mott, and James C. Lester. 2011. Integrating learning, problem solving, and engagement in narrative-centered learning environments. International Journal of Artificial Intelligence in Education 21, 1-2, 115-133. https://doi.org/10.3233/JAI-2011-019

[24] Jennifer Sabourin, Jonathan Rowe, Bradford W. Mott, and James C. Lester. 2012. Exploring inquiry-based problem-solving strategies in game-based learning environments. International Conference on Intelligent Tutoring Systems. Berlin, Heidelberg, 470-475.

[25] Johnny Saldaña. 2009. The Coding Manual for Qualitative Researchers. Sage. Thousand Oaks, CA.

[26] Hiller Spires, Jonathan P. Rowe, Bradford W. Mott and James C. Lester. 2011. Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes. Journal of Educational Computing Research 44, 4 (August 2011), 453-472. https://doi.org/10.2190%2FEC.44.4.e
[27] NGSS Lead States. 2013. Next Generation Science Standards: For States, By States. The National Academies Press. Washington, DC.
[28] Pina Tarricone. 2011. The Taxonomy of Metacognition. Psychology Press.
[29] Michelle Taub, Roger Azevedo, Amanda E. Bradbury, and Nicholas V. Mudrick. 2020. Self-regulation and reflection during game-based learning. In Handbook of Game-Based Learning, 239-262. MIT Press.
[30] Michelle Taub, Roger Azevedo, Amanda E. Bradbury, Garrett C. Millar, and James Lester. 2018. Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. Learning and Instruction 54 (April 2018), 93-103. https://doi.org/10.1016/j.learninstruc.2017.08.005
[31] Michelle Taub, Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. 2020. The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. Computers & Education 147 (April 2020). https://doi.org/10.1016/j.compedu.2019.103781
[32] Thomas Daniel Ullmann. 2015. Automated Detection of Reflection in Texts - A Machine Learning Based Approach. Doctoral dissertation, The Open University.
[33] Thomas Daniel Ullmann. 2019. Automated analysis of reflection in writing: Validating machine learning approaches. International Journal of Artificial Intelligence in Education 29, 2 (May 2019), 217-257. https://doi.org/10.1007/s40593-019-00174-2
[34] Max Van Manen. 1977. Linking ways of knowing with ways of being practical. Curriculum Inquiry 6, 3 (December 2014), 205-228. https://doi.org/10.1080/03626784.1977.11075533
[35] Frances K. Y. Wong, David Kember, Loretta Y. F. Chung, and Louisa Yan. 1995. Assessing the level of student reflection from reflective journals. Journal of Advanced Nursing 22, 1 (July 1995), 48-57. https://doi.org/10.1046/j.1365-2648.1995.22010048.x
[36] Elaine H. J. Yew and Karen Goh. 2016. Problem-based learning: An overview of its process and impact on learning. Health Professions Education 2, 2 (December 2016), 75-79. https://doi.org/10.1016/j.hpe.2016.01.004