2016 49th Hawaii International Conference on System Sciences

Standing on the Shoulders of Citizens: Exploring Gameful Collaboration for Creating Social Experiments

Casper Harteveld, Amy Stahl, Gillian Smith, Cigdem Talgar
Northeastern University
{*c.harteveld, gi.smith, c.talgar*}*@neu.edu*
stahl.a@husky.neu.edu

Steven C. Sutherland
University of Houston-Clear Lake
sutherland@uhcl.edu

Abstract

There exists a gap in knowledge between scientists and the larger non-scientist public. Therefore, much of the information provided to the public regarding research that should influence their decisions is often misunderstood. In order to eliminate, or at the very least, minimize this gap, there is a need to educate non-scientists about research methods and experimental design. In order to address this need, we have created a digital game, Mad Science*, that allows non-scientists to create and participate in experiments to better understand research methods. The current study analyses the results of a paper prototyping session, where non-scientists were asked to create experiments using the tools and scaffolding provided in the game. Participants were able to create playable scenarios and testable experiments. However, our results suggest a need for further AI support and scaffolding to address common areas of confusion and to facilitate the experimental design process.*

1. Introduction

Sir Isaac Newton famously wrote, in a letter to his rival Robert Hooke in 1676, that "it is by standing on the shoulders of Giants" who have gone before him that had allowed him to make intellectual progress. Instead of relying on other scholars, in recent decades research has been increasingly relying on the support of the general public to make progress. This type of research is generally labeled citizen science. Currently, citizen scientists [1], [2] are individuals who freely contribute their time and effort to a scientific process that is entirely constructed and framed by professional researchers. This current definition treats citizens as free labor, and does not empower them to truly be independent scientists. The kinds of tasks are often relatively simple and mundane such as categorizing items or registering observations [3]. In this paper, we explore a platform, named *Mad Science*, that enables *citizen-led science*, where citizens will be capable of conceptualizing and constructing experiments independently and building from the knowledge that already exists.

Though new technologies that enable collaborative citizen engagement have likely contributed to recent explorations of citizen science, the use of non-experts as free labor has been practiced for a long time [2], [4]. Some of the earliest foundational work for citizen science explicitly argues for treating participants as inferior. For example, in his Great Tide Experiment, William Whewell described the use of citizen science participants as "subordinate laborers" who are capable of collecting information [5]. He insisted that only professional scientists have the ability to make meaning of what is collected.

Such citizen science has provided value through the production of new, reliable knowledge, e.g. [6], and participants have indicated that they enjoy being helpful to the scientific process [7]. However, the assumption that non-experts can only ever be useful as free labor is highly limiting. We argue that great value exists in empowering citizens to perform science themselves. Whether this is possible remains up for debate, based largely around disputes over the amount of training needed for someone to engage in scientific inquiry [5], [4]. It seems clear that, to enable true citizen science, it is necessary to first research how to design and build appropriate scaffolded and structured environments.

We are thus pursuing a highly iterative design process for *Mad Science*. Our vision for this platform is one in which players can create their own social experiments and participate in other researchers' social experiments, towards collaboratively building an understanding of human behavior. In the long-term, the creation of social experiments will be facilitated by mixed-initiative, AI-based tools [8], [9] that enable a human and computer to act as creative collaborators. However, to design these tools appropriately, we must first understand the kinds of experiments players build and in what areas they most need support. We aim to rigorously analyze the prototypes we construct to better understand requirements for the infrastructure needed for citizen-led science.

Our main design research question in these efforts is: How can we build a game platform that facilitates the creation of social experiments by non-technical, non-scientist players? In our initial efforts, documented in this paper, we are specifically interested in (i) the extent to which participants are able to construct social experiments; (ii) the process of creating social experiments with a specific emphasis on the role of collaboration; (iii) the creativity exhibited by participants; and (iv) the role of the scaffolding we provide. Our aim is to generate insights that can help guide the further

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.18

development of our platform and of others. Our findings are based on an in-depth qualitative analysis of the conversations and drawings of 26 groups in a paper prototype exercise.

Our contributions are twofold. First, our analysis will provide useful insights for others who want to pursue citizen-led science projects, especially in the area of social science for which few citizen science projects currently exist. Although the insights are derived within the context of the *Mad Science* project, the paper prototype used is open and little informed by the design of the platform. This openness allows our insights to extend beyond our specific project. Second, our approach for rigorously analyzing paper prototypes provides an exemplar to the field for attaining a better understanding of building collaborative scientific platforms and contrasts from the discount usability engineering [10] approaches typically used in paper prototyping.

2. Background

In this section we provide the background for the work presented in this paper. We argue specifically why more research is needed on games for learning and citizen science, that assessment and scaffolding are two main issues for implementing science inquiry successfully in a game-based environment, and that complex game design projects may benefit from a more rigorous study of paper prototypes.

2.1. Games for Learning and Citizen Science

The potential of games for transforming education and advancing learning has been identified in the past decade [11], [12]. In terms of education, games are especially unique in their capacity to engage learners [13]. Players may voluntarily invest countless hours in a good game. In this way, theoretically, learners spend more time on the subject matter and are encouraged to learn more. This affordance to engage is of importance to the success of citizen science as this stands or falls with the participation of people. Issues of engagement are of significance because the kinds of tasks that citizen scientists undertake are sometimes mundane or repetitive, or they may be complex, requiring specialized training or knowledge, such as in *Mad Science*. This need for engagement may explain why games are often considered for citizen science projects. However, applied in the context of citizen science, engagement (or motivation) is still an emerging topic and one that requires further study [14]. In fact, according to Prestopnik and Crowston [15] motivation is one of the key aspects of citizen science that require further study. Our efforts are to look into engaging citizen scientists in learning a complex skill and in projects that are led by the citizen scientists themselves, which is something we refer to as *citizen-led science*. This model is different from other identified models, where although citizen scientists may participate in all steps of the scientific process, there is still a formal scientist involved [3].

2.2. Need for Assessment and Scaffolding

The educational objectives of *Mad Science* are aimed at increasing both an understanding of the research methods content (e.g., independent and dependent variable, hypotheses, and types of statistical analyses) and science inquiry. Ketelhut [16] found that students were able to learn science inquiry skills in virtual environments; however, they found that assessing inquiry learning might not be adequately accomplished using simple testing procedures. In addition to the need for a sophisticated assessment of learner's progress, scaffolding is important in building scientific inquiry in virtual environments [17]. Scaffolding is an educational technique where supportive strategies are incrementally removed when they are no longer needed. It is a widely used technique and specifically aimed to bridge the learning gaps between novices and experts. Both assessment as well scaffolding are of crucial importance for having non-technical, non-scientist users authentically participate in scientific research. Unfortunately, little research has been done on game-based assessment [18]. Although models for scaffolding exist, not much research has been done on applying this to games and especially not by using mixed-initiative systems, which is what we are interested in exploring. Mixed-initiative systems were actually originally created in an educational context [8], and cast the act of designing an artifact as a conversation between human and computer. In *Mad Science*, we aim to develop the creation tools as mixed-initiative in that the human and computer can collaborate to produce characters, environments, and scenarios. However, to do this successfully we first need to get an understanding of how the computer can be of help.

2.3. Rigor in Game Design

Designing games necessitates an iterative process because user-software interaction cannot be accurately predicted [19]. Because game designers can only indirectly determine how players experience the game, Salen and Zimmerman [20] call game design a *second-order design problem*. This design process becomes more complex with games for impact, where designers need to take aesthetics into consideration in addition to the meaningful purpose the game attempts to achieve. It has been observed that successful serious game design processes require an appropriate consideration of purpose, content, and play [12]. There is no magic formula for this and so designers are forced to iteratively design their games until they find the right balance.

A common approach to both software development in general and game design in particular is to use paper

prototyping for this second-order design problem [19]. Paper prototypes are quick-and-dirty mockups to get quick feedback on design ideas from users without needing to fully implement them. In addition to iterative design and paper prototyping, the complexity involved with games has necessitated the development of game user research techniques [21], [22], which are qualitative and quantitative research methods that help to get an understanding of user behavior in the context of games. By iteratively designing with game user research, designers receive a substantiated understanding of how they need to proceed to the next iteration cycle. However, game user research is still not a widespread practice and what is used often follows the principles of discount usability engineering [10]. Paper prototypes are predominantly evaluated through facilitator observations or short surveys only. Although this provides for quick findings, this may not necessarily lead to the grounded insights needed for further development, and in the case of complex projects such insights may be very valuable. With this paper, where we applied a rigorous qualitative approach, we aim to contribute to maturing the use of game user research by highlighting the value of incorporating systematic research of paper prototypes.

3. Mad Science Project

Mad Science is an ongoing game design project with the aim to crowd-source research regarding how people respond to different manipulations. In this 2-D digital game, players join the corporation Mad Science Inc. as one of their new mad scientists. Mad Science Inc.'s mission is to "understand why people do what they do" through social experiments. The scope of experiments that players will be able to create in the future involves a wide variety of social situations and decision-making scenarios, from ethical and personal dilemmas to international, political conflicts and complex societal challenges to goofy hypothetical situations. In essence, players are immersed into situations where they need to make a decision that involves taking an action or responding to a conversation. Figure 1 shows an experiment in the current digital prototype.

Before players can contribute to the mission, they are familiarized with the core rules of Mad Science Inc., such as "We may be mad but we are mad together," to emphasize the collaborative nature of the corporation and that, in the vein of Sir Isaac Newton, research is a collaborative effort where players should adopt and apply the knowledge generated by other players to make scientific progress. In the game, this will be facilitated by allowing players to copy experiments they have participated in and providing them access to a library of assets. Players can run a new experiment that builds forth on a previous experiment with just a few button clicks, which is a major advantage over traditional laboratory experiments. Players are also gradually familiarized with

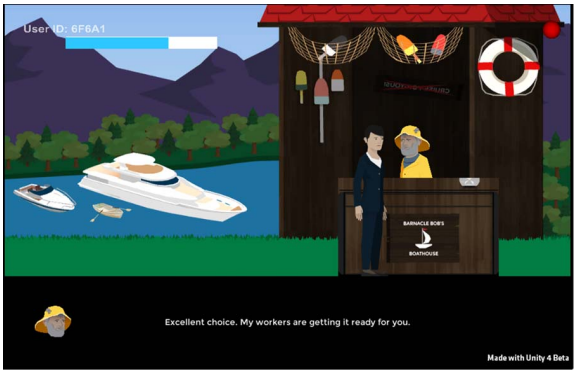

Figure 1. Example of an experiment. Dialogue and choice options are provided at the bottom.

the corporation's proprietary machinery to create social experiments. This machinery includes the following set of creation tools that are relevant to this paper: (i) a *character creator* to build new characters or customize existing ones; (ii) a *scene creator* where objects and characters can be placed in a scene; (iii) a *scripter* that allows players to make a sequence of dialogue, choice options, and actions (e.g., sound or movement); and (iv) a *manipulator* to set independent and dependent variables.

To achieve the aim of crowdsourcing social scientific research, the project's first major challenge is to teach players with no programming or research background the complex skill of creating research experiments with the scripter and other tools. In order to have players learn these complex skills, we are pursuing to study how people acquire the necessary science inquiry skills to create experiments, understand what scaffolding is necessary to facilitate their learning, and how this can be accomplished in a gameful manner. The main goal of the present study was to learn whether participants would be able to create social experiments using *Mad Science* and where they struggled with the process. Identifying misunderstandings of research design and difficulties in creating experiments allows us to better understand where scaffolding will be necessary in the game itself. Therefore, we were less concerned with the impact of the minimal scaffolding provided in this study and more concerned with using this study to inform what and where scaffolding would be necessary going forward.

4. Methods

For our digital prototype, we developed three digital experiments based on existing, classical studies in the decision sciences [23]. The digital prototypes for the in-game tools were not ready for a large scale study; instead, we opted for a paper prototyping session to explore how the creation of experiments can be facilitated. We developed a course module that included the three digital experiments, a

homework assignment, and a paper prototype as part of an activity spanning two, 60-minute classes. On the first day students played the experiments and received the homework assignment. On the second day they engaged with the paper prototype. In this section we provide an overview of the module implementation with a focus on the paper prototype. Details of the first day are reported elsewhere [23].

4.1. Participants

The module was implemented in the course Sex, Relationships and Communication, with extra credit for students who completed the homework assignment. This elective offered by the Department of Communication Studies focuses on communication as it occurs in sexual and romantic relationships, specifically on the role of verbal and nonverbal communication in these relationships. Although research is used to inform students about what is known, students in this class do not conduct any empirical research. They are therefore relevant subjects for our study. It should be noted that we did not inquire upfront what experiences the students had with research prior to the module implementation so some may have been more experienced than others.

It should also be noted that students are not representative of citizen scientists in general. We chose this audience because students are a captive audience that provides us the necessary input to iterate and refine our work until it is ready for distribution to a broader audience. In addition, in terms of level of education, college students may not be dissimilar. Although further study is needed in identifying who citizen scientists are [14], and demographics may likely differ per project, the current understanding is that most are older individuals who are highly educated, with at least some college degree [24].

About 80 students participated, divided over 26 groups. Each group was provided with a USB audio recorder to record their conversations during the session and were instructed to turn their recorders on after the instructions were read. Only 14 groups had their audio recorders on throughout their entire discussion. Two groups did not have recordings and one group forgot to turn on their recorder for most of the session. The remainder of the groups, 9 in total, only recorded a summary of their experiment at the end, an error which is most likely due to facilitation as all of those recordings came from a single classroom. Although this limited our analysis of the audio recordings, the 14 full recordings and 26 visual narratives served us with enough data to provide for initial insight into how to facilitate the creation of social experiments.

4.2. Material

The three digital experiments have been developed with the Unity game engine and students accessed them on the project website. They served to provide an understanding of what social experiments are. The homework assignment requested students to think about and document an experiment that they would be interested in creating using *Mad Science*. To help students, the assignment included a set of informally written questions that are relevant for the design of an experiment (e.g., what will you vary for the players in your scenario?).

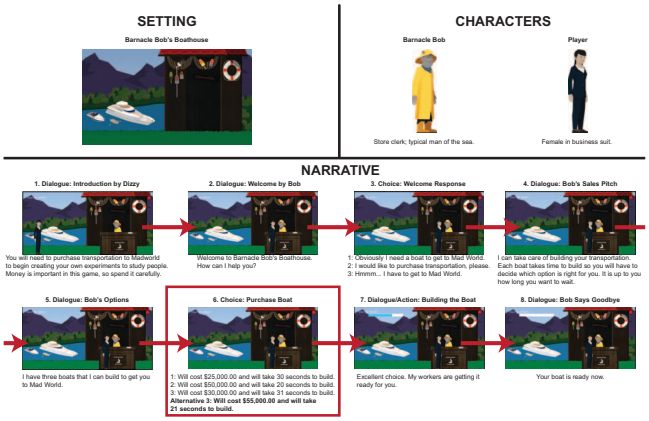

Figure 2. The template for making a visual narrative.

The paper prototype was based on the digital authoring tools, in particular the scene creator, character creator, and scripter. For the prototype, students were tasked to brainstorm in a group about an experiment and illustrate this on an 18 in x 24 in blank sheet of paper. To scaffold this process, we asked them to illustrate this visual narrative according to a specific template, which used one of the three digital experiments as an example (see Figure 2). In the upper left corner of the sheet they had to draw and describe the setting, a process that is similar to the scene creator. In the upper right corner students had to draw and describe the characters. This would be how the character creator would work. The rest of the sheet was used to describe the sequence of dialogue and events, which is what the scripter would facilitate. The manipulator was simplified to circling the scene where the manipulation would take place.

4.3. Procedure

We divided the students randomly over four classrooms. Each classroom was assigned two facilitators, with at least one with expertise in research methods and one with knowledge about the project. At the start the students were placed into groups of three or four. Groups were handed a blank sheet, the visual narrative template, and detailed instructions with questions for each relevant aspect of the design of an experiment. Each group also received 8GB USB voice recorder and were asked to turn it on at the beginning of their group work.

The paper prototype activity was self-directed. The facilitators were only there to answer questions, not to guide the process itself. Students were asked to present the experiments they designed for the homework and then decide to pick one of the experiments to continue to work on or work on a new idea. In the second phase the groups were asked to draw the visual narrative. Step 1 was to create a setting and characters. Groups were asked to draw the setting where the environment takes place (e.g., desks and doors) and to draw and describe their characters. We guided this by asking to think about the attributes of the characters (e.g., gender and skin color) and roles in the scenario (e.g., salesperson and boyfriend). Step 2 was to create a visual narrative of the experiment, with dialogue and annotations. Groups were asked to show scene by scene how the experiment progresses, to number and organize these scenes in chronological order, and provide a concise heading for each scene. Furthermore, they were asked to label scenes with what is occurring using the following three labels used in the scripter and associated questions:

1) Dialogue: What is the character saying? Which character is speaking?
2) Choice: Is the player required to make a choice? How many options are there and what are the options to choose between?
3) Action: What is the action being completed? Who or what is completing the action?

The final instruction was to circle the scene where differences would exist for different players. In other words, here we asked groups to illustrate where their manipulation would take place. This instruction was accompanied with the questions "What will vary?" and "What is the alternate version from the circled scene?"

We encouraged groups to take about 20 minutes for the first phase. However, groups were able to manage the process themselves and could leave the class after they finalized the visual narrative and spoke to the facilitators.

5. Analysis

Because we asked the groups to create experiments using *Mad Science*, two raters first independently scored whether the final scenarios were true experiments and playable by evaluating the visual narratives and the transcripts of the groups interactions. In order to meet the requirements of a testable experiment, there had to be two or more conditions of an independent variable (i.e., gender was treated as an independent variable although it was not being manipulated) and a clear measure for the dependent variable. In order to be considered a playable scenario, the design had to meet the capabilities of the authoring tools in *Mad Science*.

Of the 26 visual narratives, there were three scenarios that the raters did not initially agree upon. Two of the three were because one of the raters was not sure that the design could be implemented in *Mad Science*. After a discussion, it was clarified that they could. The third was because the group was looking at gender differences. Because the gender of the player could not be manipulated, it would not be a true experiment; however, there was agreement that the design offered a testable hypothesis. After discussing the narratives between the two raters, 16 of the 26 scenarios were considered both testable experiments and playable scenarios, while 10 scenarios were considered playable scenarios but not testable experiments.

5.1. Breakdown of the Experiments Created

To understand the types of experiments groups created, we will describe the independent variables (IV) and dependent variables (DV) players used in their scenarios. Of the 16 testable experiments, nine tested hypotheses about relationships. This was not surprising because the class was a Sex, Relationships, and Communication course. Additionally, the homework assignment included examples for creating experiments, which included examples about disclosure. Five of the nine relationship experiments tested the impact of levels of disclosure on how much the individual disclosing was liked.

Figure 3 shows an example of one of the group's visual narratives for an experiment testing the effect of level of disclosure on whether the character is liked. One tested how players respond to various levels of face threat when meeting their significant others parents for the first time. Another tested how players would approach a potential romantic interest based on with whom the player was playing beer pong at a party. One experiment tested which person players would tend to pursue based on whether they played hard-to-get or not. And, two tested how players would respond to flirting, one based on the setting in which the flirting occurred and the other based on the players gender (responding to aggressive flirting).

For the non-relationship experiments, there were several different theories being tested. One group was interested in the impact of the number and location of other customers in a bar on player preferences for where to sit. One group was interested in whether players earning money for an intelligence test (higher scores led to more money) would help a homeless person when money and time would be lost. This group was interested in whether having more money would increase peoples willingness to give, but also how emotional intelligence was related to giving. Another group looked at the impact of a low or high quiz average affected decisions about the weight a future unknown exam should receive. While another looked at how the cost of an additional bottle of wine on the menu at a restaurant (more or less expensive than all the others) would impact how much players would spend on a bottle with dinner. One tested how

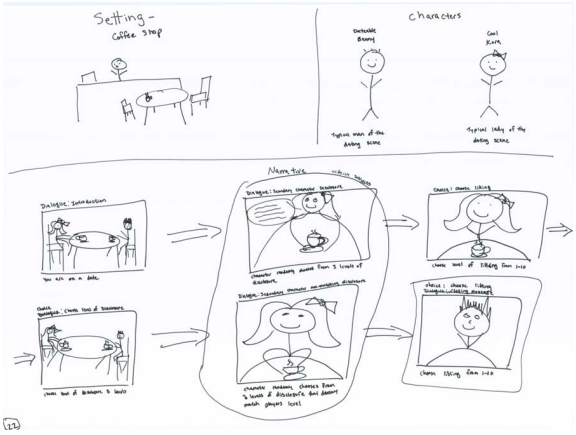

Figure 3. An example of a visual narrative (by C2G22).

long players would wait at a restaurant and whether players would leave feedback about their waitress based on whether the waitress had been polite or rude during a number of interactions. And the final experiment tested whether players would pick up a lost wallet and keep it or turn it in to the authorities (with or without having taken the money out) if an authority figure was present and based on whether the authority figure was a moral or legal authority figure.

5.2. Breakdown of the Codes

To gain a better insight into the process that the groups went through to design the scenario, we analyzed the transcriptions of each groups conversations. However, because we were unable to completely differentiate between speakers, all analyses are at the group level and we do not report frequencies for any individual speakers. This, unfortunately, does not permit us to determine whether conversations were largely dominated by only one or a few individuals, which is a limitation of using audio recorders rather than video recorders. For our analyses, we used the R statistical package and the RQDA package, specifically, for coding the transcripts and completing our qualitative analyses. As our interest lies in analyzing the collaborative process, we uploaded the 15 full transcriptions into a RQDA project. Within RQDA, we created descriptive codes with which to explore how the groups brainstormed ideas, discussed experimental design, and decided how to draw out their narratives.

We had 19 individual codes that were used to categorize the conversations. Conversations were assigned one or more codes from the onset through the end of the conversation, regardless of the length of the conversation. Only when the conversation changed to another topic, the codes were changed. In the next sections, we will break down the codes and, for the experimental design, by-products of collaboration, and confusion codes, we will discuss the major themes. When the number of times a code was used is reported, the percentage of the total number of codes for the reported code will appear in parentheses. For each subcategory reported, the percentage of the total number of subcategory observations for that code will appear in parentheses.

5.2.1. Creating the Experiment. To better understand the process the groups went through to create their experiments, we coded conversations specifically for instances where the independent variables, dependent variable, or hypotheses were discussed. Each of these codes were further analyzed for major themes and subcategories were created. Here we will further discuss the codes as a whole and the subcategories that were identified. It is possible for discussions to belong to multiple subcategories.

Independent Variables. The independent variables were discussed 58 (5.63%) times across 13 groups. An example of the IV code is:

> *I don't know how to work the game, so I wouldn't know how like random it is, but they would either be saying, like revealing a lot of information, a normal amount of information, or no information.*—C2G22

Subcategories included identifying and justifying the IV, how to differentiate levels and how many levels of the IV should be used, how the IV would be translated into the *Mad Science* game world, and controlling for confounding variables. Identifying and justifying the IV was discussed 23 (33.33%) times and included discussions about what the different levels of disclosure are and providing background for why they chose specific levels. Differentiating and identifying how many levels of the DV were discussed 9 (13.04%) times and included how many levels are necessary or required and how to manipulate the IV between players. Translating the IV into the game world was discussed 28 (40.58%) times and included making the IV matter to the player in the game and scripting the different levels of the IV. Controlling confounds was identified 9 (13.04%) times and included removing the social desirability bias randomized assignment to conditions.

Dependent Variables. Dependent variables were discussed 61 (5.92%) times across 13 groups. The subcategories identified for this code included identifying and justifying the DV, operationalizing the DV, eliminating issues with the DV, and determining when to measure the DV. Operationalizing the DV was by far the most common, with 38 codings (53.52%) falling into this category. Examples of operationalizing the DV included discussions of how the choices that the player can make should be expressed, and how to measure a concept like "kindness" within the game. We found four instances of groups eliminating issues with the DV (5.63%). For example, more than one group discussed the possibility of social desirability bias influencing

their players choices. Justifying and identifying the DV was discussed 25 times (35.21%), often when groups were using a question from the instructions. Groups discussed when to measure the DV in four cases (5.63%). In one case, the group quoted below discussed measuring the DV more than once within their experiment.

> *So I think it kinda tests like if the person is really open at the beginning but then changes, do you still like them when they change. Or if they're really closed off, but then they change later, do you end up liking them after, now?*—C2G22

Hypotheses. Twelve groups discussed hypotheses and there were 25 (2.43%) instances of this code across the groups. Subcategories were potential confounds/issues, discussing/comparing hypotheses, using the hypotheses to drive variable selection, and confusion about what a hypothesis is or how to create one. Group members discussed or compared hypotheses with each other 16 times (55.17%). There were often multiple hypotheses within a group, and topics debated included how much disclosure is ideal and how players would react to positive versus negative attitudes. Discussions under the other three categories were far less frequent. Potential confounds/issues were observed 6 times (20.69%); using hypotheses to drive variable selection, 5 (17.24%); and confusion about what a hypothesis is or how to hypothesize, 2 (6.90%).

5.2.2. By-product of Working Together. A main theme in *Mad Science* is how non-scientists create experiments through collaboration throughout the design process. Collaboration offers benefits beyond the sharing of workloads, ideas, and responsibility; players are afforded the opportunity to learn from, teach, and receive feedback from others. Depending on the environment where this collaboration takes place, group members also have the opportunity to have fun through interacting with others in a playful manner. We coded the number of times group members were recorded laughing. Laughter was observed 165 (16.02%) times across 14 out of the 15 groups, which was the highest number of unique observations for any code. This supports our goal of providing a fun environment for individuals to engage with the experimental design process. In order to capture the additional by-products of collaboration, we analyzed the codes for conversations revolving around group members teaching each other and receiving feedback through rejected ideas and compliments.

5.2.3. Teaching. An expected benefit of collaboration is the opportunity to share knowledge with other members of your group. Across 10 of the 15 groups, there were 23 (2.23%) conversations where group members were observed to be teaching one another. The following is an example of teaching:

> *So basically the amount of information that they reveal about themselves lets the player or the player could tell how much they like them. And that would show oh yeah, people who—what's the right word?*
> *Disclose*.—C2G22

As for subcategories, there was 1 (3.85%) instance of teaching about the reason for the hypothesis, based on concepts learned in another class. Seven (26.92%) conversations were teaching about how to manipulate an IV, 3 (11.54%) conversations were teaching about the DV, 1 (3.85%) conversation was to teach about creating hypotheses, 9 (34.62%) conversations were about creating dialogue flows and choice options in the context of the game, and 5 (19.23%) conversation were coded as "other". The "other" category accounted for teaching that was not directly related to the task, but where group members were sharing explaining concepts and theories from their classes.

5.2.4. Complimenting or Rejecting. Because collaborations are social interactions, there are opportunities for individuals to receive both positive and negative feedback from their peers. Therefore, we coded compliments and rejections. Only 3 groups were observed rejecting any ideas and there were only 4 (0.39%) instances across the groups. All rejections were in the form of rejecting an idea and were always accompanied with an explanation for why they did not want to pursue a particular plan. This suggests that no individuals prohibited others from contributing (a concern for group work) because prohibiting others from contributing would have resulted in rejections of ideas. However, there may be issues of groupthink [25] occurring, another major concern for groups and one that is difficult to capture in the current study. More important was the higher number of compliments observed. There were 52 (5.05%) complimenting conversations across 14 groups. To better understand the nature of the interactions, we identified the major subthemes of the compliments: individual assignments, drawing skills, groups effort, experiment design and art ideas, and another groups art. Compliments from one group member to another on his/her drawing skills were observed 24 times (42.11%). Compliments on ideas were the next most common, with 18 codings (31.58%), and those exchanges included:

> *And we should also vary how nice the waiter or waitress is to them.*
> *So good!*
> *Like when the waitress is like, "Don't worry it'll be few more minutes!" or the waitress is like, "No, you have to keep waiting."*
> *That's genius*.—C4G3

Group members encouraged or praised their team as a whole on 7 occasions (12.28%). Compliments on individual assignments were observed 6 times (10.53%), usually towards the beginning of the groups work, when members

were each describing the experiments they had designed on their own the day before. On 2 occasions (3.51%), a participant complimented the work of another group.

5.2.5. Confusion and Clarification. Through the process of having players create visual narratives for their experiments, using the provided scaffolding and the framework shown in *Mad Science*, we hoped to learn where individuals became confused or required additional interaction with the session facilitator. Learning where players became confused allows us to modify the in-game tools and AI support provided to players to better facilitate the experimental design process. We captured instances where groups expressed confusion about the assignment and when the groups interacted with the facilitator. However, we also identified conversations where players did not express confusion, but where the coder observed confusion or misinformation being agreed upon.

5.2.6. Expressed Confusion. There were 39 (3.79%) conversations across 11 groups where players expressed confusion. It was divided into the following categories: manipulating the independent variable, Dialogue/Choice/Action, who the player is, how to draw out the storyboard, constant changing of ideas, and other. Confusion about manipulating the independent variable was expressed 10 times (22.73%) and included discussions such as:

Who should hit on who?
No, how should the conversation flow?
Okay, who's hitting on who?—C3G1

That group designed their experiment to study reactions to different flirting styles, but struggled to pinpoint their IV and manipulate it in a controlled manner. The following interaction was categorized under two categories–who the player is (6.82%) and Dialogue/Choice/Action (13.64%):

Dialogue, wait wait is that what it is? No no, choice. Disclosing too much info
But who are you playing as?—C2G5

Discussions under constant changing of ideas, the category with the highest frequency (14, 31.84%), generally resembled this:

So wait, going back to the original idea?
I don't know
What was the original idea?—C2G69

Only four Expressed Confusion codings fell outside of those categories and were labelled "other" (9.90%).

5.2.7. Observed Confusion. Eighteen (1.75%) conversations were coded across nine groups where the coder observed confusion. An example of observed confusion involved a failure to understand hypothesis testing:

And formulate analysis or...hypothesis.
Well we're not doing that. Well, sorry. The computer's doing that for us. It's bringing the data and we're formulating the idea afterwards—C3G1

To identify where groups tended to agree upon incorrect information, we broke down this code into six subcategories. There were 6 (33.33%) instances where groups incorrectly discussed manipulating an IV, 1 (5.56%) instance of incorrectly differentiating dialogues from actions and choices, 4 (22.22%) instances of incorrectly identifying their DV, 2 (11.11%) instances of incorrectly creating a hypothesis, 4 (22.22%) instances of errors involving dialogue flows and choice options, and 1 (5.56%) instance listed as "other" because it was not related to their visual narrative.

6. Discussion

Our goals for the present study were to identify (i) whether participants were able to create playable and testable experiments and (ii) how they accomplished this goal through collaboration, (iii) what creativity participants exhibited, and (iv) where scaffolding would benefit players in a digital game, based on where the groups struggled in their designs. For the most part, groups were able to create playable and testable experiments. Of the 26 groups, all created playable scenarios, 16 of which were also testable experiments. These findings are encouraging for our development of *Mad Science* because it suggests that non-scientists are able to create experiments with appropriate manipulations of IVs and operationalized DVs for testing their hypotheses in the game.

Unfortunately, not all of the groups were able to create testable experiments. Seven of the 10 groups that did not create testable experiments were also of the 11 groups that did not turn on their recording devices during the design process. Therefore, we were unable to gain deeper insight into where there may have been confusion about the process. Despite that facilitators met up prior to discuss the process, all facilitators were requested to read out loud the same instructions, and students received printed instructions, a breakdown happened in implementing the module consistently. This breakdown, resulting in a failure to record a number of the groups, is our greatest limitation in the present study. On the other hand, these results are valuable in suggesting the importance of facilitation. Although we aimed to reduce the role of the facilitator and the results may stem from a bias between classrooms, future research should consider facilitator interaction more carefully, to ensure consistency is achieved. Additionally, this limitation emphasizes a key area in which scaffolding would be necessary within the game. It would be useful to provide early, temporary checklists for players to confirm that necessary steps in the instructions of any task have been followed.

After the initial instructions were provided, 3.79% of the coded conversations focused on additional interaction with the scaffolding (facilitator interactions and comparing their work to the example visual narrative and questions). This suggests that players were able to create their experiments

with limited scaffolding and instruction. One benefit that allowed this was likely that players collaborated with peers to design their experiments. Indeed, 23.69% of the coded conversations focused on collaborating (i.e., teaching each other) and by-products of that collaboration (i.e., compliments and laughter). We believe there are benefits to allowing collaboration between players. However, further research will need to compare scenarios created by individuals, collaborative groups of players, and players collaborating with in-game AI agents. We hope to use our findings from the collaborative efforts of these groups to inform collaborative AI support in the game, including the need for teaching, complimenting, recommending alternative actions, and joking. However, through additional testing, decisions about whether and when players collaborate with other players or AI support will need to be made to achieve the goal of teaching scientific thinking to non-scientist populations.

Players were able to discuss their experimental design using appropriate concepts, which may have been a by-product of the non-scientific language used to focus players on answering questions about their experimental design and the several examples provided. Players discussed the experimental design (IV, DV, and hypothesis) 13.98% of the time. These conversations largely revolved around operationalizing variables and identifying and justifying the variables of interest for their studies. Additionally, several of the conversations focused on controlling for confounding variables although players were not instructed to do so. This suggests that players participated in additional aspects of scientific thinking facilitated by the task in the game.

Confusion (expressed by the player or observed by the coder) and frustration in the task was found in only 6.12% of the total coded conversations. This is promising because it suggests that what participants were asked to do, using the tools and structure provided in *Mad Science*, was for the most part easy to follow. However, it is important to note that the confusion consistently contained issues with understanding how to manipulate the IV and how to differentiate dialogue, choices, and actions in the game. These were also the larger issues when players taught each other and when players had to seek input from the facilitator. This provides insight for the need to provide further scaffolding and training for understanding how to use the *Mad Science* scripter (dialogue, actions, and choices) and how to manipulate independent variables, thus increasing players ability to create playable scenarios and true experiments.

Groups created experiments that varied in the theories being tested. This means that the tools that were shown in *Mad Science* and the structure around which participants were encouraged to shape their scenarios allowed participants to see many possibilities for creating experiments. However, there were a large number of students who created experiments that tested relationship theories, which we were not surprised to find. Because the class focuses on this topic, participants were likely adhering to a priming effect. The idea that participants were primed was further supported by the larger number of experiments focused on testing the effect of levels of disclosure on liking, which was the theory from the class used as an example for the homework assignment. Therefore, we acknowledge the need to avoid examples that are too specific when creating examples moving forward because it may impact the creativity of the players. Instead, it will be important for AI support to ask players about their interests and provide examples for the experimental design process that build on the players stated interests rather than using a standard example.

Additionally, participants tended to create ethnocentric and stereotypical examples within the game. For example, most of the relationship studies focused on heterosexual relationships. And, even in the alien restaurant scenario, the wait staff were female. Lawyers and cops were male and a male was the protection provider for the female zombie apocalypse survivor. These examples go beyond inclusiveness in games and playing into stereotypes, they violate an important aspect of the experimental design process. By not utilizing representative samples, the results of the studies are not able to be generalized back to a diverse population. For example, players may respond in a particular way to females in need of protection. It would then be incorrect to state that the results generalize to how people respond to all people in need of protection. This may not be important to the playful scenarios, but they do reinforce an error in experimental designs and proper AI support would be necessary to caution players or educate players about external validity in research. *Mad Science* is not meant to teach people to create realistic cover stories, but to teach players how to conduct scientific research. To teach players the correct way to conduct good scientific research, scaffolding would need to be provided to players to warn about common errors.

Social science experiments often require the use of deception to test their hypotheses. This usually involves the creation of elaborate cover stories. The creativity of the cover story can often engage participants while keeping the true nature of the experiment a mystery. Gameful scenarios are perfect examples of well told cover stories. Participants in our study found this to be the most important aspect, as supported by the amount of time spent on establishing creative cover stories. In fact, 41.36% of the coded conversations focused on creating their cover story, including identifying characters (confederates), narration and sequencing (script), and the setting in which the experiment takes place. Interestingly, players created settings and characters that extended beyond reality (e.g., space dogs and alien wait staff), showing that *Mad Science* allows players to use their own creativity in exploring the experimental design process. Our expectation is that by allowing playful design, players will feel more comfortable exploring scientific thinking; however, further studies are needed to explore issues of

engagement and learning in our game.

7. Conclusion

With limited prior study and limited scaffolding, we observed that non-experts were able to make social experiments in a short amount of time, with the important warning that the topic was familiar and that we were dealing with University students. Additionally, players tended to apply scientific principles beyond what was required for the study. The tools in *Mad Science* translated well to the design of experiments and, despite needing additional scaffolding in a few key areas, appear to support players in the design process. Players were creative in their approach to their designs and scenarios and spent a large portion of their time creating the cover story for the experiments. Finally, and very importantly, participants had fun throughout the process as evidenced by the overwhelming amount of laughter throughout the process. Learning scientific thinking is essential for achieving crowdsourced citizen-led science, but it is through the enjoyment of the experience that players will return to interact with scientific concepts at deeper levels. Our findings are encouraging for the development of citizen science games that aim to educate their users complex skills. Our methodology for rigorously analyzing paper prototypes may provide an exemplar for others to engage with the complex design process that is inherent with the development of such games.

References

[1] R. Bonney, C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk, "Citizen science: a developing tool for expanding science knowledge and scientific literacy," *BioScience*, vol. 59, no. 11, pp. 977–984, 2009.

[2] J. Silvertown, "A new dawn for citizen science," *Trends in ecology & evolution*, vol. 24, no. 9, pp. 467–471, 2009.

[3] A. Wiggins and K. Crowston, "From conservation to crowdsourcing: A typology of citizen science," in *44th Hawaii International Conference on System Sciences*. IEEE, 2011, pp. 1–10.

[4] J. P. Cohn, "Citizen science: Can volunteers do real research?" *BioScience*, vol. 58, no. 3, pp. 192–197, 2008.

[5] C. Cooper, "Pearls across the Zooniverse: When Crowdsourcing Becomes Citizen Science," Feb. 2013. [Online]. Available: http://goo.gl/Oi3ifC

[6] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovi, and F. Players, "Predicting protein structures with a multiplayer online game," *Nature*, vol. 446, pp. 756–760, 2010.

[7] N. Lazzaro, "The four fun keys," in *Game Usability: Advancing the Player Experience*, K. Isbister and N. Schaffer, Eds. Burlington, MA: Elsevier, 2008, pp. 315–344.

[8] J. R. Carbonell, "Mixed-Initiative Man-Computer Instructional Dialogues. Final Report." 1970.

[9] G. Smith, J. Whitehead, and M. Mateas, "Tanagra: Reactive Planning and Constraint Solving for Mixed-Initiative Level Design," *IEEE Transactions on Computational Intelligence and AI in Games (TCIAIG)*, vol. 3, no. 3, Sep. 2011.

[10] J. Nielsen, *Usability engineering*. Elsevier, 1994.

[11] J. Gee, *What Video Games Have To Teach Us about Learning and Literacy*. New York, NY: Palgrave Macmillan, 2003.

[12] C. Harteveld, *Triadic game design: Balancing reality, meaning and play*. London, UK: Springer, 2011.

[13] R. Garris, R. Ahlers, and J. E. Driskell, "Games, motivation, and learning: A research and practice model," *Simulation & Gaming*, vol. 33, no. 4, pp. 441–467, 2002.

[14] M. J. Raddick, G. Bracey, K. Carney, G. Gyuk, K. Borne, J. Wallin, and S. Jacoby, "Citizen science: status and research directions for the coming decade," *AGB Stars and Related Phenomenastro 2010: The Astronomy and Astrophysics Decadal Survey*, p. 46P, 2009.

[15] N. R. Prestopnik and K. Crowston, "Gaming for (citizen) science: exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system," in *e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on*. IEEE, 2011, pp. 28–33.

[16] D. J. Ketelhut, B. C. Nelson, J. Clarke, and C. Dede, "A multiuser virtual environment for building and assessing higher order inquiry skills in science," *British Journal of Educational Technology*, vol. 41, no. 1, pp. 56–68, 2010.

[17] K. D. Squire and M. Jan, "Mad City Mystery: Developing scientific argumentation skills with a place-based augmented reality game on handheld computers," *Journal of Science Education and Technology*, vol. 16, no. 1, pp. 5–29, 2007.

[18] C. Harteveld and S. Sutherland, "The goal of scoring: Exploring the role of game performance." in *Proceedings of the 2015 ACM Conference on Computer-Human Interaction*, Apr. 2015.

[19] T. Fullerton, C. Swain, and S. S. Hoffman, *Game design workshop: a playcentric approach to creating innovative games*, 2nd ed. Burlington, MA: Morgan Kaufmann Publishers, 2008.

[20] K. Salen and E. Zimmerman, *Rules of Play: game design fundamentals*. Cambridge, MA: MIT Press, 2004.

[21] K. Isbister and N. Schaffer, Eds., *Game usability: Advice from the experts for advancing the player experience*. Burlington, MA: Morgan Kaufmann Publishers, 2008.

[22] M. Seif El-Nasr, A. Drachen, and A. Canossa, Eds., *Game analytics: Maximizing the Value of Player Data*. London, UK: Springer, 2013.

[23] S. C. Sutherland, C. Harteveld, G. M. Smith, J. Schwartz, and C. Talgar, "Exploring Digital Games as a Research and Educational Platform for Replicating Experiments," in *NEDSI Conference*, Boston, MA, 2015.

[24] D. J. Trumbull, R. Bonney, D. Bascom, and A. Cabral, "Thinking scientifically during participation in a citizen-science project," *Science education*, vol. 84, no. 2, pp. 265–275, 2000.

[25] I. L. Janis, *Groupthink: Psychological studies of policy decisions and fiascoes*. Houghton Mifflin Boston, 1982.

Acknowledgment

We thank our students Nolan Manning, Yuyang Zhao, and Huichen Guan for realizing our vision, Dr. Joseph Schwartz for allowing to run this study in his class, and Dr. Brooke Foucault Welles for assisting in the facilitation. We further thank the College of Arts, Media & Design and Northeastern University for funding this project.