Designing productively negative experiences with serious game mechanics

Qualitative analysis of game-play and game design in a randomized trial

Authors

First and corresponding author

Andrea Gauthier[1], BAA MScBMC
PhD Candidate
Institute of Medical Sciences
University of Toronto
andrea.gauthier@utoronto.ca

Second author

Jodie Jenkinson[2], PhD
Assistant Professor
Biomedical Communications, Biology
University of Toronto Mississauga
j.jenkenson@utoronto.ca

Mailing address

[1]Room 327, [2]Room 324
Terrence Donnelly Health Science Complex (HSC)
3359 Mississauga Road
Mississauga, Ontario, Canada
L5L 1C6

Abstract

Design, rather than medium, ultimately predicts learning outcomes, but the game-based learning literature has had difficulty successfully linking game design decisions to learning behaviours and outcomes. The current research investigates how explicit game design strategies can promote productive negativity (i.e. learning from failure), which has been identified as an important mechanism in both gaming and learning. We performed a randomized controlled trial with undergraduate biology students to investigate how game design might facilitate misconception resolution about random molecular behaviour through productive negativity. Students engaged with either a computer-based interactive simulation (n=20) or serious game (n=20) for 30 minutes, while their computer screens were recorded and click-stream data collected. We described in detail the theoretical framework underpinning our serious game and simulation using the Activity Theory Model of Serious Games (ATMSG); qualitatively coded and analysed video recordings of gameplay; and visually overlaid this data with the ATMSG models to draw conclusions about how game-design decisions influence learning-related behaviours. We found that the serious game resulted in significantly more productively negative experiences, while the interactive simulation allowed for greater exploratory or experimental behaviours. Based on our analyses of the qualitative gameplay data, we were able to recommend three game design strategies to enhance the occurrence of desired game-flow loops (e.g. productive negativity) with respect to an ATMSG framework: 1) including additional game mechanics on the primary game-flow axis of the ATMSG framework (i.e. mandatory interactions) limits the exploratory nature of the application; 2) integrating two or more primary-axis mechanics in a game-flow loop increases the frequency of interaction with this loop; and 3) gameplay loops that involve mechanics that fall off the primary-axis (i.e. non-mandatory mechanics) occur less frequently than those which involve primary-axis (i.e. mandatory) mechanics. This study is one of the first to successfully make direct comparisons between students' interactions in a game and a non-game application to provide concrete and actionable serious game design recommendations.

Keywords: serious game design; productive negativity; misconceptions; randomized controlled trial; Activity Theory Model of Serious Games (ATMSG)

1. Introduction

1.1. Background

There is a growing body of literature supporting the use of serious games to enhance learning and engagement in education, from kindergarten through to postsecondary levels (for recent meta-analyses and reviews, see Abdul Jabbar & Felicia, 2015; Clark et al., 2016; Boyle et al., 2016; Wouters et al., 2013). However, most studies investigate *whether or not* the serious game facilitates learning, rather than focusing on *how* the game design facilitates learning (Clark et al., 2016). Clark and colleagues (2016) conclude that the overall rigour in game-based learning research needs to improve; they encourage researchers to give more detailed accounts of both game and control interventions because their design, rather than their medium, ultimately predicts learning outcomes. Boyle et al. (2016) found that the great majority of games investigated from 2009-2014 did not integrate advanced gaming mechanics (such as those commonly found in adventure, role-playing, or strategy game genre) in a way that aligned with learning outcomes; they call for more research that investigates what gaming features are most effective at supporting learning and engagement. Another recent review of 165 papers, which specifically reported on how learning and gaming mechanics can be integrated into effective serious games, found that very few of these papers employed frameworks that succinctly linked learning and gaming elements with empirical evidence (Lameras et al., 2017). Overall, there is consensus that game-based learning literature needs to establish stronger theoretical links between pedagogical strategies and game design elements.

In response to this criticism, many serious game design models and frameworks have arisen in recent years to help game creators and evaluators describe the relationships between the serious game mechanics and instructional strategies more explicitly (for example: Amory, 2007; Arnab et al., 2015; Carvalho et al., 2015; De Freitas et al., 2010; Kelle et al., 2011; Kiili, 2005; Michie et al., 2008; Mislevy & Haertel, 2006; Starks, 2014). These frameworks each demonstrate unique benefits, though only a few allow the researcher or designers to make concrete associations between learning and gaming elements. In particular, Kelle and colleagues (2011) attempt this by identifying which learning functions (e.g. expectation, comparison, analysis) correspond to overarching game design patterns (e.g. goal-related patterns, information-related patterns, patterns for game mastery and balance). A limitation of this model is that it does not facilitate an understanding of how the patterns are integrated into the flow of the game. The Learning Mechanics-Game Mechanics (LM-GM) model is better at achieving this by providing a framework to visually map these patterns (Arnab et al., 2015). It recognizes that "serious game mechanics" are an interplay between fine-grained pedagogical and entertainment elements (i.e. *learning* mechanics and *game* mechanics) that results in higher order game design patterns, such as those described by Kelle et al. (2011). The LM-GM model provides a non-exhaustive list of both abstract and concrete learning mechanics (e.g. observation, experimentation, modelling) and game mechanics (e.g. role play, time pressure, rewards/penalties) derived from the literature on learning science and game design. It enables the designer to make quick pairings of learning strategies and gaming elements and visualize these in a graphical representation to give a succinct overview of the instructional intention of the game. However, Carvalho et al. (2015) suggest that the primary limitation of the LM-GM model is that "it does not expose the connection between concrete mechanics and the high-level educational objectives that the game is supposed to attain"; i.e. it is still too abstract.

The Activity Theory of Model of Serious Games (ATMSG) builds upon the LM-GM model and aims to rectify this by describing game design at an even more granular level (Carvalho et al., 2015). A generic ATMSG diagram is visualized in **Figure 1**. The ATMSG recognizes three main activities within serious game mechanics: (1) the gaming and (2) the learning activities, which centre around the actions and motives of the player, and (3) the instructional activity, which involves the actions and motives of the game designer (intrinsic) and/or of an instructor who might be facilitating gameplay in-person (extrinsic). Each of these activities can be further described by their more granular *actions*, which are mediated by concrete *tools* (i.e. visual game elements) that are implemented with specific *goals* in mind. The authors provide a template (**Figure 1**) in which designed game interactions can be represented visually in a detailed, linear game-flow map, while their corresponding gaming, learning, and instructional components (each with action, tool, and

goal sub-components) are charted in a table below the graphic. Mechanics that fall on the central, thicker arrow (referred to in this paper as the "primary game-flow axis") indicate interactions that are mandatory for successful level completion—these interactions *must* happen to complete a level (e.g. Mechanics A and D). Mechanics that fall off the primary game-flow axis are optional and it is either up to the player to decide if they should engage in the interaction or is dependent on computer logic/conditions in the game (e.g. Mechanics B and C). Overall, the ATMSG provides a detailed and precise depiction of how gaming and instructional elements are interwoven into the game-flow to facilitate learning (Carvalho et al., 2015).

[INSERT FIGURE 1]

Figure 1. Generic structure of an Activity Theory Model of Serious Game (ATMSG) diagram (based on Carvalho et al., 2015). Notably, mechanics that fall on the primary game-flow axis are considered mandatory for task completion, while those that fall off this axis are not mandatory. Diamonds represent conditions that affect the interaction flow in some way. Circles represent game states.

While the above models concern the details of the finalized game design, the Intervention Mapping approach to the design of behavioural change interventions provides a big-view framework for the *overall* design process, from problem conceptualisation to evaluation (Bartholomew et al., 1998; Michie et al., 2008). The Intervention Mapping approach involves characterizing the problem with a needs assessment of the target audience, generating objectives for solving the problem, identifying evidence-supported strategies to meet the objectives, designing and developing an intervention that applies these strategies, and, finally, implementing and evaluating the intervention. This approach has been used by some in the field to design serious games (Arnab et al., 2013; DeSmet et al., 2016); the current publication follows an overall Intervention Mapping approach, whilst—in the design and development step—using the ATMSG model to provide a robust description of how game elements map with learning elements in the game-flow. In the evaluation phase of Intervention Mapping, we describe the design of a highly-similar interactive simulation through a second ATMSG that can be directly compared to that of the game, to come to conclusions about the pedagogical impact of game design. This comparability is often lacking in game-based learning randomized trials where comparisons to non-equivalent media (e.g. lectures, texts, animations) are frequently made; such comparisons tell us more about the benefits of active learning and obscures the value-added effect of the game design itself (Clark et al., 2016; Prensky, 2011).

This paper also serves as one of the few examples investigating how specific game mechanics and design strategies can promote productive negativity or failure in STEM higher education. Failure is often perceived as a negative, undesirable event, that creates feelings of sadness, inadequacy, frustration, and confusion (D'Mello, 2013). However, the importance of introducing challenges and cognitive conflict to enhance learning is well recognized (D'Mello, 2013; Kapur, 2008, 2014; Kapur & Rummel, 2012). Specifically, Gadamer (1998) suggests that negativity—in the form of irritation, disillusionment, and failure—facilitates learning by exposing the learner to something new about a concept that highlights the limitations of their prior knowledge and of their own consciousness, allowing them to reach a new horizon of consciousness through the experience. When designed to be followed by a productive response (i.e. learning), failure is referred to as *productive failure* or *productive negativity* (Gadamer, 1998; Kapur, 2008, 2014; Kapur & Rummel, 2012). Productive negativity can play a particularly important role in misconception identification and resolution (further discussed in Section 2.3.1), as well as in gameplay (Section 2.3.2); our serious game and interactive simulation are designed to facilitate productive negativity to help students resolve misconceptions about the random nature of molecular environments.

1.2. Research questions, approach, and implications

We designed and developed an adventure-genre serious game, *MolWorlds*, to help students resolve misconceptions about the random nature of molecular environments and tested its value-added effect against a highly-similar interactive simulation (*MolSandbox*) in a randomized trial. The game differs from the simulation only in terms of added game mechanics, so that we might discern the direct impact that game design has on in-game behaviour and learning outcomes. The overarching hypothesis was that game design

would increase instances of productive negativity, which would result in better learning outcomes in the gaming group in comparison to the simulation group. The results of this evaluation have been described in Gauthier & Jenkinson (2017). A primary finding was that, while the simulation allowed greater freedom for the student to explore molecular factors, our gaming intervention increased instances of productive negativity — the quality of which held a trending relationship with a reduction in misconceptions.

To extend these findings, and in response to criticism of the literature described above, we endeavour to answer the follow questions in the current publication:

a) *How* does the presence of serious game design (specifically resource management, an immersed 3rd-person character, sequential level progression, and scoring) moderate the frequency and nature of how students demonstrate their conceptual knowledge within the interactive simulation?
b) *How* does the presence of serious game design fundamentally alter the frequency and nature of productive negativity experienced by the student within the interactive simulation?
c) Based on the findings to the two questions above, what are specific game design strategies to increase the occurrence of desired interactions and game-flow loops (e.g. productive negativity loops) that target learning outcomes?

To answer these questions, the current publication aims to:

a) Provide a detailed description of the theoretical framework underpinning the design of the serious game and simulation and how they are intended to facilitate misconception resolution;
b) Characterize how instances of productive negativity and demonstrations of correct conceptual knowledge elicited by the interventions differed between groups through a qualitative analysis of interaction data; and
c) Describe and visualize how these differences in interactions relate back to the interventions' interaction designs.

In doing this, we hope to make our design decisions and the value-added effect of game mechanics transparent to other researchers, so that they may be replicated, critiqued, and improved upon to advance serious game design and the quality of research in this field. It has been long established that the engagement and learning capacity of games should be models for good instructional design (Dickey, 2005; Gee, 2005; Squire, 2013); we expect that our findings and approach have relevance beyond the realm of serious computer gaming, and could be applied to the design and evaluation of any digital interactive tool. While our interventions were designed to address problems in molecular biology at the undergraduate level, misconceptions are very common in all areas of science—from chemistry, to physics, math, evolution, and climate change—and at all age levels (Odom, 1995; Paz-Y-Miño-C & Espinosa, 2012; Robic, 2010; Rocklöv, 2016; Sanger & Iowa, 2000). The design strategies recommended to facilitate productive negativity might apply to interventions in other domains where misconceptions abound.

2. Design of the serious game and interactive simulation

We followed a basic intervention mapping approach to the design of *MolWorlds* (Michie et al., 2008). We characterized the conceptual problem (Section 2.1); defined learning objectives to address the problem (Section 2.2); determined which instructional strategies would achieve the learning objectives (Section 2.3); designed and developed the game and used the Activity-Theory Model of Serious Games to define the relationships between its gaming, learning, and instructional elements (Section 2.4); and, finally, prepared a comparative evaluation of the game (Section 2.5). These steps are described below in detail.

2.1. Characterizing the conceptual problem

Molecules, both large and small, move completely randomly. They rebound off each other in a random-walk pattern (referred to as "Brownian motion") until they collide with another molecule of a complementary chemistry (i.e. a binding partner) in the correct orientation. Cellular systems are composed of emergent molecular processes, where individual parts (molecules) move randomly, uniformly, independently,

simultaneously, and continuously (Chi, 2005, 2013; Chi et al., 2012). In essence, all cellular processes (and life itself!) rely completely on chance, moderated by the concentration of molecules and other cellular factors, like macromolecular crowding and temperature, amongst others. The concept of emergent processes is one of a number of so-called "threshold concepts" in life sciences that represent a transformation in one's understanding without which the learner cannot progress (J. H. F. Meyer & Land, 2005). In the context of biology, without an understanding of the emergent nature of molecular interactions, one cannot fully appreciate concentration gradients, signal transduction cascades, medication dosing effects, genetic mutations, or evolution (Garvin-Doxas & Klymkowsky, 2008). The concept of emergence also extends to other domains including chemistry (e.g. radioactive decay), physics (e.g. friction), economics (e.g. stock market), and earth sciences (e.g. weather patterns). In fact, Slotta and Chi (Slotta & Chi, 2006) showed that training students about emergent properties in one domain (e.g. molecular diffusion) could be transferred to other domains (e.g. electricity). Therefore, understanding how random mechanisms result in complex systems has application beyond the undergraduate molecular biology classroom.

Unfortunately, many students develop misconceptions about the random nature of molecular environments because the concept of randomness clashes with the way the world is perceived on an everyday basis (Coley & Tanner, 2012; Scholl & Tremoulet, 2000). We perceive the idea of randomness as being an inefficient mechanism and cellular processes as being highly efficient; two conditions which are contradictory. As such, the learner may try to assimilate the new concept (i.e. molecular emergence) into existing schemas (directed/sequential processes) by altering the details of that new concept, so that it works in relation to their pre-existing ideas (Piaget, 1974). Our recent research with undergraduate biology students suggests that students' understanding of random molecular behaviour is highly contextual. We developed the *Molecular Concepts Adaptive Assessment* (MCAA) and distributed the survey to biology students in three consecutive years (n = 1170) (Gauthier et al., 2019). The results showed that participants understood concepts of Brownian motion in isolation, but this comprehension fell apart when applying it to broader patterns observed in the cell. For instance, 75.3% of respondents believed that a molecule's path of motion is random but becomes more direct after being activated (by phosphorylation). Furthermore, about half of the sample thought large molecules (such as proteins) move more directly/purposefully than smaller molecules (like water or carbon dioxide), that diffusion ceases once equilibrium across a semipermeable membrane is reached, and that water does not diffuse in the same way as a solute does. In essence, they failed to understand the dimensions of emergent processes: individual parts (molecules) move randomly, uniformly, independently, simultaneously, and continuously (Chi, 2005, 2013; Chi et al., 2012). Rather, they identified molecules as causal agents that behave in a sequential manner, dependent on the interactions of other agents, and that behaviour terminates once the process is complete (i.e. the dimensions of a sequential process). Additionally, our analyses showed that a similar number of misconceptions were held across first-, second-, and third-year biology students, supporting other research that suggests that misconceptions about molecular emergence are resistant to change without targeted intervention (Chi, 2005; Garvin-Doxas & Klymkowsky, 2008; Tibell & Rundgren, 2010). The participants in the game-based randomized controlled trial reported in the current publication are drawn from the same population and were also assessed with the MCAA.

2.2. Defining the learning objectives of the intervention

After identifying the conceptual problem, the following learning objectives of the intervention were defined:

1. Molecular processes are not directed, under any circumstance; individual parts (e.g. proteins) do not have agency.
2. Molecular processes are emergent, meaning that the movements of individual parts are unconstrained or random, their behaviours are uniform, and the interactions between them are simultaneous, continuous, and independent (Chi, 2005).
3. The rate or probability of a molecular process can be affected by:
    a. The concentration of individual parts (i.e. how many molecules of one type there are)
    b. The macromolecular crowding in the environment (i.e. how many large molecules/proteins are in the space) and the size of the molecules of interest

c. Other environmental factors, such as temperature (additional examples might include pH or the availability of energy sources, however these have not been targeted in this version of the intervention).

We embedded our learning objectives within curriculum-related molecular biology content so that students might observe how these principles affect biological processes that they learn about in their lectures (e.g. ligand-gated membrane channel/transporter functionality, vesicular formation and docking, enzyme degradation and inhibition, regulatory mechanisms, mRNA translation).

2.3. Determining instructional strategies to achieve the learning objectives

2.3.1 Experiential learning and productive negativity

Our intervention attempts to help students resolve misconceptions about the random nature of molecular environments by providing a simulated, scaffolded, and active learning environment that leverages theories of experiential learning and of productive negativity.

The concept of experiential learning (Kolb, 1984) postulates that optimal learning takes place when learners engage in a concrete experiences, reflect upon their observations, and form hypotheses about the experience. They should then test their hypotheses through a process of experimentation and feedback. Chi (2005) suggests that scaffolded, interactive simulations could be used to help students develop a better understanding of emergent processes. Simulations model real-world or hypothesized processes and systems that may otherwise occur at sizes and timescales that we cannot naturally perceive, and can further abstract or simplify such processes to facilitate learning (Honey et al., 2011; Vogel et al., 2006). Other studies support this and have shown that simulations can help students overcome some (but not all) misconceptions surrounding similar processes (Meir et al., 2005; White & Bolker, 2008). By visualizing the random, uniform, simultaneous, continuous, and independent motion and interactions of molecules in the context of a cellular process, misconceptions can be naturally confronted. For example, the misconception that "large molecules have a more direct path of motion than small molecules" is contradicted when small and large molecules are simulated to both move randomly and uniformly; the misconception that "diffusion stops once equilibrium across a semi-permeable membrane is reached" is exposed when motion continues in an environment with such a membrane. Furthermore, by offering students opportunities to learn experientially by observing, interacting, hypothesizing, and experimenting, the student begins to understand how various factors (e.g. concentration, crowding, temperature) affect the system and develops a better appreciation about how randomness can be a mechanism for perceptually efficient cellular processes.

The concept of productive negativity, as previously introduced in Section 1.1, extends experiential learning. The efficacy of productive negativity or failure was demonstrated in ninth-grade students learning about the concept of variance (Kapur, 2014). One group of students received direct instruction on how to calculate variance, while another group was allowed to form their own hypotheses and attempt to calculate variance using their own formulas to test their naïve conceptions about variance. This second group failed in their attempts to calculate variance accurately and then received direct instruction on how to properly perform the calculation; while students in both groups were able to compute variance similarly well after the learning intervention, the group who first experienced failure outperformed the other in terms of their conceptual understanding of variance and their ability to transfer the concept to new scenarios. In a similar study comparing a productive failure group to a direct instruction group (this time with seventh-grade mathematics students solving problems calculating average speed), Kapur and Bielaczyc (2012) found that students who were allowed to fail outperformed those who received scaffolded instruction. However, it should be noted that failure in these cases may not have resulted because of robust misconceptions held by the student, but because of a general lack of knowledge about how variance and speed were calculated.

While these studies show promise, the literature base surrounding the effectiveness of failure and negativity on learning is still small; a recent meta-analysis by Darabi, Arrington, and Sayilir (2018) found only 62 articles within the past 10 years that empirically investigated productive failure and only 12 of these articles

were robust enough to include in their meta-analysis (i.e. many did not include a control condition or did not adequately report their data); additionally, 13 of the 23 included effect sizes in the synthesis were produced by the same author (Kapur), emphasizing the need for more wide-spread investigation in this field. Nevertheless, the authors found an overall significant medium-small effect of interventions that supported productively negative experiences over other control interventions.

2.3.2 Game design to enhance productive negativity

Section 2.3.1 described how misconceptions could be confronted through an interactive learning environment that allows users to test hypotheses, confront their expectations, and experience productive negativity. Here we would like to put forward that game mechanics might enhance the effectiveness of these strategies through their motivational properties and by promoting negativity.

Failure, frustration, and negativity can play a prominent role in gaming motivation and in-game learning. Failure is an integral part of gaming experiences, so much so that winning without prior failures actually creates a sense of dissatisfaction with the win (Hoffman & Nadelson, 2009; Juul, 2009). In general, games promote failure and negativity though challenging tasks and conflicts, thus ensuring the development of key skills or knowledge through repetition and experimentation in an engaging environment (Barab et al., 2010; Charsky, 2010; Gee, 2007; King et al., 2009; Salen & Zimmerman, 2003). While students often express a negative affect when faced with failure in an academic scenario (Clifford, 1988, 1991; D. K. Meyer & Turner, 2006), a serious game can, in contrast, leverage the fact that no real-world consequences result from failure in-game, thereby maximizing its beneficial effects (Gee, 2007). Effectively, the repetitive loops of trying, failing, readjusting, and re-trying in a game is *productive negativity*: a player attempts a challenge and, based on their current understanding, might fail (negative). They must then restructure their understanding (productive) if they want to succeed and progress in the game. The immersive challenges in games naturally encourage failure and the narratives, reward systems, and virtual identities might instigate the productive response (Gee, 2007; Mitgutsch & Weise, 2011a). For example, a recent study with middle school children playing the game *Virulent* (a virology game that teaches about cellular defense mechanisms) found that the number of level failures before first success significantly predicted post-test outcomes (Anderson et al., 2018), and that failure prompted productive discussion amongst classmates about the learning material in the game.

In addition to a game's challenges and conflicts, subversive game design might instigate cycles of productive negativity (Mitgutsch & Weise, 2011b, 2011a). In their research, Mitgutsch and Weise 'subvert' common game design elements—meaning they attach an unusual function or pattern to an element commonly used in games in other ways—to teach the player a lesson by leveraging their expectations in regard to the subverted element. For example, in their post-apocalyptic game *Afterland*, players start in an empty house and see an inventory checklist; they then proceed to explore the ravaged land, collecting all the elements on the checklist, filling their house with junk, while avoiding being seen by other people in the world. The subversion here is that the player only wins the game if they get rid of all the junk in their house—something that goes against the common gaming elements of collecting and managing an inventory—and make friends with the people who they originally considered to be enemies. This designed subversion forces the player into a predictable pattern of behaviour (collecting, hoarding, and avoiding enemies), which then makes the player reflect on the serious content of the game (social behaviours, mental health).

2.3.3 Summary

In summary, we determined that our intervention would promote productive negativity in a task-based, interactive, simulated environment to expose misconceptions to the learner and facilitate a better understanding of emergent molecular systems. Ultimately, both gaming and simulation conditions in our research would adopt this strategy, but the game would also strategically integrate gaming mechanics and patterns to promote productive negativity (Section 2.4)

2.4. Designing and developing the serious game intervention

2.4.1 Basic game concept

Briefly, *MolWorlds* (prototype v.2.0) is a 2D, 13-level, simulation-based, platform-genre, adventure game developed for desktop using the Unity Game Engine. The narrative involves a scientist, Dr. Goodcell, who, having been shrunk down to the size of a protein by his evil academic colleague and subsequently trapped in a molecular world, is trying to find a way home (sample frames from the narrative are visualized **Figure 2**-A). Players travel through the molecular realm and experience cellular processes (e.g. vesicle formation, RNA translation), while manipulating properties of the simulated emergent system through temperature, macromolecular crowding, and concentration—factors that affect the rate of molecular processes governed by random motion, as identified by our learning objectives (Section 2.2)—to help them on their journey. Each level is completed by reaching a checkpoint. The challenge in each level requires the player to facilitate one or more molecular processes by collecting and releasing molecules at appropriate locations and passing the character over the checkpoint; in a simple level (e.g. level 2), this might be opening a ligand-gated membrane channel to pass the character through the channel to the next level. In a more advanced level (e.g. level 10), this might involve translating an mRNA strand to form a protein used in vesicle formation, then situating the character inside the vesicle to be transported across the membrane to the checkpoint. In any situation, the player can use power-ups (when available) to modify the temperature of the environment or their character size to play with environmental crowding. Manipulating size and temperature will affect the character's ability to navigate and collect resources, as well as influence the overall simulation. **Figure 2**-B and C illustrates the basic user-interface and world elements. The molecular environment is simulated interpretively, with water represented by a particle generator on a soft grey background (refer to screenshot in **Figure 2**-B), and the Brownian (i.e. random) motion of the other proteins approximated to make it appear as though they were interacting in water. A brief video of gameplay can be viewed on the Science Vis Lab website: www.sciencevis.org.

[INSERT FIGURE 2]

Figure 2. A) Selected frames from *MolWorlds'* introductory narrative depicting Dr. Goodcell shrinking down to protein-size and entering the molecular world. B) Level 6 and C) level 7 of *MolWorlds*, highlighting interface elements in orange numbers: 1) Inventory menu – stores collected molecules; 2) Timer – restarts at the beginning of each level; 3) Main menu – keeps track of performance on each level; previously completed levels can be reloaded here; 4) Dropzone – collected molecules can only be released on these targets; 5) Checkpoint – must be reached for level completion; 6) Powerup count; 7) Powerup buttons – clicking one of the powerups (heat, chill, grow, and shrink) will have an effect for 10 seconds; 8) Map – zoomed-out view of the level; 9) Inventory menu pulled down – displays all collected molecules with sliders to select concentrations of each; and 10) Molecule info pop-up – by clicking on objects in the environment or in the inventory, a brief description of the molecule's function appears.

We can summarize *MolWorlds'* serious game design into four broad game design patterns: 1) resource management, 2) an immersed 3rd-person character, 3) sequential level progression, and 4) scoring/feedback. Each of these patterns is meant to enhance productive negativity. Firstly, by integrating resource management as a primary pattern, the player is encouraged to be retentive with their resources, making it likely that they will only release a single molecule to accomplish a task, as depicted in **Figure 2**-B; this concept is akin to the subversive design (hoarding resources) described by Mitgutsch & Weise (2011). A negative experience is likely to ensue when the player realizes that the released molecule does not move directly to its target, and a productive response (e.g. increase concentration) might follow. Secondly, an immersed 3rd-person character will elicit productive negativity by allowing the player to physically experience the effect of emergent forces on the character as they travel through the crowded environment. The player is often bumped around, and his progress hindered by other fast-moving molecules (negative); by making modifications to the character size, crowding of the environment, or concentrations of certain molecules (positive), their passage through the molecular world to the checkpoint is made easier. Thirdly, if the character does not reach the checkpoint (regardless of whether they successfully facilitated the correct molecular interactions), they must replay the level, or portions of the level, before they can progress in the game (negative). In subsequent attempts, they should reflect on the reasons why they did not reach the checkpoint and make appropriate actions to ensure

their success (positive). Lastly, scoring is displayed upon reaching a checkpoint. The score is calculated based on the time-to-level-completion and is displayed in seconds as well as stars (1 to 3); this 3-star system is intended to promote productive negativity by encouraging the player to replay the level and achieve a full, three-star status, which would require them to optimize their interactions in the level.

Section 2.4.2 describes how specific mechanics are implemented in these four broader game design patterns in more detail using the Activity-Theory Model of Serious Games (ATMSG; Carvalho et al., 2015).

2.4.2 Defining serious game mechanics: Activity-Theory Based Model of Serious Games

Here, we use the ATMSG (refer to Section 1.1) to describe the game-flow and implementation of instructional and learning mechanics in the game. The ATMSG model for *MolWorlds* is visualized in **Figure 3** and is elaborated upon in **Table 1**. **Figure 3** represents the sequence of game mechanics diagrammatically; the actions, tools, and goals for the gaming, learning, and/or instructional components of each serious game mechanic are indicated in the layered table beneath the diagram. These components were chosen directly from Carvalho et. al.'s (2015) taxonomy for serious game components, apart from those related to productive negativity as an instructional strategy (something that is not addressed in the ATMSG's current taxonomy). The implementation of the serious game mechanics from **Figure 3** is described more thoroughly in **Table 1**, which should be read in conjunction with the figure. For a more thorough description of how the ATMSG works, we encourage the reader to access Carvalho et. al.'s (2015) article.

By implementing this framework, we can specify the granular mechanics and elements that make up *MolWorlds'* four broad serious game design patterns: 1) resource management — identifying, selecting, collecting, and releasing molecules and using powerups; 2) an immersed character — exploring, navigating the environment, experiencing; 3) sequential level progression — reaching the checkpoints, behavioural momentum, repetition; and 4) scoring and feedback — a 3-star system for quick feedback that encourages reflection and repetition.

[INSERT FIGURE 3 – Full page, rotated 90º for vertical orientation]

Figure 3. ATMSG model for *MolWorlds*, describing serious game mechanics (numbered i-x) from the overall game flow (top) and from the core gameplay (bottom). Each mechanic is further described by gaming, learning, and/or intrinsic instructional mechanics with action, tool, and goal components and is elaborated upon in Table 1.

Table 1. Description of the ATMSG implementation in *MolWorlds* (refer to Figure 3). (PN = productive negativity)

[INSERT TABLE 1 - Full page]

2.5. Implementing and evaluating the intervention

The final step in the game design process was to implement the game with end-users and evaluate its efficacy in promoting productive negativity and in helping students resolve misconceptions related to molecular randomness. This first involved piloting an early prototype of the game (v.1.0) (Gauthier & Jenkinson, 2015) and making changes based on these results and user-feedback—these changes are reflected in the current description of the game. Secondly, we designed and programmed a non-game, interactive simulation (Section 2.5.1) to be used in a randomized trial to investigate how the serious game mechanics affected interactions and subsequent learning outcomes, beyond the underlying interactive simulation (Sections 3-5).

2.5.1 MolSandbox: designing a non-game intervention for comparison

Simulations and games share many similarities and may exist on a broad spectrum of gamefulness (Whitton, 2010, p. 22). Both involve simulated, real-world or hypothesized phenomena or environments that allow the user a certain degree of control (Honey et al., 2011), but a game may become distinct from a non-game, interactive simulation by introducing aspects of play, such as competition, challenge, exploration, fantasy, goals and rules (Whitton, 2010, p. 23). Specifically, while a non-game stimulation may be playful, in that it allows the user to explore the parameters of the simulated environment (Whitton, 2010, p. 28-29), it does not

incorporate gameplay, defined as a formalization of "interaction that occurs when players follow the rules of a game and experience its system through play" (Salen & Zimmerman, 2003).

A primary aim of our research is to characterize the specific influence of game design on productively negative experiences and learning outcomes, beyond the interactive, experiential learning that can be delivered through an interactive simulation. As such, we developed a mirror intervention, *MolSandbox*, which excludes most game design features but retains the underlying simulated environment and basic system-modifying mechanics. In *MolSandbox*, students are presented with the same 13 levels as *MolWorlds* in the same graphic style. **Figure 4** depicts screenshots from levels 7 and 11 from both interventions to facilitate comparison. While overarching goals still remain (the goal in each "sandbox" simulation parallels the goal in each game level, e.g. facilitate vesicle formation and docking in level 7), the gamefulness is reduced by removing the immersed 3rd-person character and the rules, restrictions, and penalties that go along with it. For example, in level 7 where a *MolWorlds*-player would have to situate the character inside the vesicle to be transported to the other side of the membrane and reach the checkpoint (**Figure 4**-A, left), a *MolSandbox*-user would simply have to elicit the formation and docking of the vesicle by interacting with the interface using the mouse/cursor (**Figure 4**-B, left). *MolSandbox*-users are not required to successfully complete a level before proceeding to the next level and can access levels in any order after completing the introductory tutorial. Additionally, molecules in the *MolSandbox* inventory are automatically replenished for each simulation (concentration is increased in the environment through releasing molecules on "dropzone" targets, like in the game, and decreased by clicking and scrubbing the cursor over molecules to collect them) and temperature and crowding are adjusted with gauges that have no usage restrictions (in contrast to power-ups contained in the game), thus completely removing the resource-management pattern. Furthermore, the score is not calculated at the end of each level, though users are shown their time to goal-completion, as well as an indication of whether each level was successfully completed in the level menu.

A final differentiating feature is the inclusion of the "add/remove pinball" function in *MolSandbox*. In *MolWorlds*, the player can increase or decrease the size of the character to experiment with environmental crowding. The character is a "foreign body" in this molecular world and may be flagged for degradation (death) by ubiquitination enzymes (**Figure 4**-A, right). A similar mechanism was required for the interactive simulation to ensure that students in either condition would be exposed to the same educational content. In *MolSandbox*, the user can insert a "foreign body" in the form of a pinball whose size may be increased or decreased proportionally, much like the character in the game. Ubiquitination processes behave in the same way with this pinball as they do with the game character and users can experiment with how quickly the pinball degrades depending on factors of temperature and size—however, pinball degradation has no influence on the user's success in the environment, whereas the degradation of Dr. Goodcell requires the player to restart the level. In other levels, the pinball may be inserted and enlarged to investigate concepts associated with crowding.

[INSERT FIGURE 4]

Figure 4. Screenshots of levels 7 (left) and 11 (right) from A) *MolWorlds* (game) and B) *MolSandbox* (simulation). Background colours represent the temperature modifications made by the user. The orange numbers highlight differences in *MolSandbox* in comparison with *MolWorlds*. 1) Inventory menu – the inventory is replenished for every level; 2) Activity – the goal or intended outcome of each level is stated at the top of the screen; 3) Main menu – levels are marked only as complete or incomplete and any level may be accessed here; 4) Modify temperature and pinball size – powerups are not necessary to make adjustments to the system; 5) Reset simulation button; 6) Next simulation button – does not require completion of the current level to proceed to the next; 7) Depiction of pinball (flagged by ubiquitin (smaller pink balls)); and 8) Level complete checkmark – analogous to the checkpoint in MolWorlds, but the checkmark appears as an element on the user interface instead of embedded in the environment. The checkmark changes from translucent to opaque once the level's activity has been completed.

While *MolSandbox* lacks many components of a serious game, it is still an experiential, playful, application that incorporates learning strategies into its interaction design. As such, we can also describe its design with an ATMSG model, allowing us to make direct comparisons with *MolWorlds'* framework. **Figure 5** and

Table 2 summarize the ATMSG for *MolSandbox*. One major differentiating feature is the number of mechanics that fall on the primary game-flow axis (the central line with broader arrows). In *MolSandbox*, only one mechanic (iii. Release molecules) falls on this axis, identifying it as the only necessary interaction for level completion. In *MolWorlds* (**Figure 3**), six mechanics fall on the primary axis, requiring a great deal more interaction and chances for productive negativity to ensue before successfully completing a level.

[INSERT FIGURE 5]

Figure 5. ATMSG model for *MolSandbox*, describing the interaction mechanics (numbered i-v). Interaction mechanics in the teal-coloured boxes (iii-v) represent the core simulation modifying mechanics present in both *MolSandbox* and *MolWorlds*, while the solid grey mechanics are generic interactions present in both digital applications. Each mechanic is further described by gaming, learning, and/or intrinsic instructional mechanics with action, tool, and goal components. Mechanics that are required for level completion appear on the primary axis, indicated by thicker grey arrows.

Table 2. Description of the ATMSG implementation in *MolSandbox*, interactive computer simulation (refer to Figure 5). (PN = productive negativity)

[INSERT TABLE 2]

3. Evaluation methodology

3.1. Summary of methods

The purpose of the evaluation was 1) to test the efficacy of our interactive simulation and serious game in facilitating a better understanding about molecular randomness through productive negativity; and 2) to compare how patterns of productive negativity experienced by participants differed between intervention groups in relation to the presence/absence of game design. To do this, we invited undergraduate first-, second-, and third-year biology students at the University of Toronto Mississauga to complete the *Molecular Concepts Adaptive Assessment* (Gauthier et al., 2019), an online questionnaire that tests misconceptions about the emergent nature of molecular environments, at the beginning and end of the Fall 2015 (first- and second-year students) and Winter 2016 (third-year students) semesters (Gauthier & Jenkinson, 2017). In all, the survey addressed 11 misconceptions in a total of 13 possible questions. Examples of MCAA items include: A) True or False: An extracellular molecule tries to move toward a complementary receptor; E) What is the mechanism of an extracellular molecule's movement toward a complementary receptor? (options: the extracellular molecule propels itself; the extracellular molecule is released with the correct initial trajectory; the extracellular molecule uses other "helper" molecules to carry it closer; the extracellular molecule collides randomly with other molecules); I) True or False: A molecule's path of motion is more direct when it has been activated (e.g. by phosphorylation), whereas its path of motion is more random when it is inactive; and L) True or False: In the case of simple diffusion across a semi-permeable membrane, once solute molecules reach equilibrium, they cease to cross the membrane (Gauthier et al., 2019). Items on the test assess students' nuanced and contextual understanding of different aspects of emergent molecular systems, such as randomness (A, E), independence (E), uniformity (I), and continuity (L).

Those who completed the surveys were rewarded with a 0.5% bonus mark on their final course grade. Before completing the post-assessment, a subset of these students volunteered to participate in a randomized controlled trial to evaluate the efficacy of game design on their understanding of molecular biology concepts. To blind participants from knowing to which group they were assigned, they were told that they would be randomized to one of two "different types of gaming conditions that apply game design to various extents" and that we would be investigating the differences in interactions, learning, and engagement evoked by the differences in design. In this way, simulation-group participants were unaware that they were not given the full game experience. Students used their assigned application for 30 minutes before completing the post-test and engagement questionnaire. During play-time, their computer screens were recorded using QuickTime and their click-stream data (button clicks and other interactions in the application) were transcribed into an online

MySQL database. These students were remunerated with a $20 gift-card to the University's Bookstore, along with the original 0.5% bonus mark for pre-post-test completion. The full protocol, including a more detailed description of the participants, recruitment, materials, and procedure, is available in Gauthier & Jenkinson (2017).

3.2. Coding schema for screencasts

All 40 30-minute screencasts were reviewed and qualitatively coded for demonstrations of correct conceptual knowledge (defined below) and instances of productive negativity, using a deductive approach to thematic analysis. Thematic analysis is "a method for identifying, analysing, and reporting patterns (themes) within data" (Braun & Clarke, 2006). Our coding scheme and full details about our qualitative approach are provided as a supplementary material to this publication. Our process was deductive in that we hypothesized productive negativity would be instigated by specific game design patterns (resource management, an immersed 3rd-person character, sequential level progression, and scoring), along with negativity caused the simulation alone, so the coders were instructed to look for negativity emerging around these themes. Furthermore, coders were instructed to identify demonstrations of correct conceptual knowledge when the participant properly used one of the three primary system-modifying mechanics (i.e. concentration, temperature, or crowding).

As identified in our learning objectives (Section 2.2)—and implemented with interactions in the interventions—the rate of a random molecular process can be affected by the concentration of molecules involved in interaction, the crowding of the environment, and other environmental factors, such as temperature. Therefore, a demonstration of correct conceptual knowledge was identified as an action or a series of actions wherein the user made adjustments to the simulation—a change in either 1) concentration, 2) temperature, or 3) crowding—that directly benefitted goal achievement, i.e. facilitated the desired molecular process. For example, if the goal or activity in the level was to stop the degradation of a foreign body (the pinball in the interactive simulation, or the character in the game) by ubiquitination enzymes and proteasomes (**Figure 4**, right), the player might chill the environment to slow the rate of molecular collisions; this would count as one demonstration of correct conceptual knowledge. They could also reduce the size of the character/pinball, further reducing the likelihood of collisions, as well as increase the concentration of de-ubiquitination enzymes (these proteins will "un-flag" the character/pinball for degradation by removing ubiquitin) in the environment; these actions would count as two additional demonstrations of correct conceptual knowledge. As such, demonstrations of correct conceptual knowledge were coded into three categories: concentration, temperature, and crowding.

An instance of productive negativity was identified as a series of actions that results in some sort of negativity (frustration, delay of progress, level failure) but which was then followed by an action that was indicative of a correct conceptual understanding of molecular emergence (i.e. a demonstration of correct conceptual knowledge). This would suggest that the student re-evaluated their understanding of the system to progress. The most typical example of this would occur if, under a misconception of directed motion, the user releases only one ligand expecting it to bind directly to a ligand-gated membrane channel (example depicted in **Figure 2**-B). When this does not happen (a negative event), the user might increase the concentration of the ligand and increase the temperature to progress more quickly (two demonstrations of correct conceptual knowledge). Ultimately, five sources (or themes) of productive negativity related to these mechanics were agreed upon: 1) resource retentiveness, 2) difficult resource collection, 3) resources lost due to overheating, 4) navigation and reaching checkpoints, and 5) simulation-based negativity (details in Section 4.2.2). Several examples of interaction patterns that constitute each code are included in the supplementary materials.

4. Data analysis & results

Gauthier & Jenkinson (2017) provides a detailed account of this study's sample composition and the effect of the interventions on misconception resolution, as well as the overall frequency of productive negativity and demonstrations of correct conceptual knowledge (summary presented in Section 4.1). The focus of the

current paper is to provide a more granular, qualitative analysis of how the nature of productively negative experiences and demonstrations of correct conceptual knowledge (defined in Section 3.2) differed between *MolWorlds* and *MolSandbox*, and how these differences connect to their ATMSG design models.

4.1. Summary of results reported in Gauthier & Jenkinson (2017)

A total of 40 undergraduate students from first-year (n = 15), second-year (n = 13), and third-year (n = 12) biology participated in the randomized controlled trial, with 20 in the *MolWorlds* gaming group and 20 in the *MolSandbox* simulation group. Groups did not differ in gaming habits, scholarly achievement, or biology course engagement levels. A total of 486 students from the same population completed the pre- and post-assessments but were not exposed to any intervention, serving as a non-randomized baseline comparison.

Through a repeated-measures mixed model, we found that both interactive simulation ($p = .007$) and gaming ($p < .001$) conditions successfully facilitated a decrease in misconceptions beyond classroom instruction (measured by the baseline sample), with no differences between level of enrolment (i.e. first-, second-, or third-year students) (Gauthier & Jenkinson, 2017). Game-players improved marginally more than simulation-users ($p = .084$).

As reported in Gauthier & Jenkinson (2017), simulation-group participants exhibited a far greater number of demonstrations of correct conceptual knowledge than did the gaming participants. We suggested that this was due to the lack of rules and restrictions in the non-game condition, which allowed students to experiment freely with different manipulable factors and resulted in a high number of demonstrations of correct conceptual knowledge ($p = .003$). For example, simulation participants modified the temperature an average of 93.40 (SD=49.90) times, while the game group modified temperature an average of 14.50 (SD=9.93) times; this difference can largely be attributed to the fact that *MolWorlds* requires the player to use powerups to modify temperature and it is impossible to find and use ~93 powerups within 30 minutes. Resultantly, their demonstrations of correct conceptual knowledge (i.e. beneficial modifications) are also lower than the simulation group. On the other hand, serious gaming participants experienced significantly higher numbers of productively negative events than simulation participants ($p < .001$), likely due, as hypothesized, to the natural gameplay loops described in Section 2.4.2; these assumptions are supported by several new qualitative analyses described in Sections 4.2.1 and 4.2.2. Due to this ratio of high demonstrations of knowledge and low instances of productive negativity, simulation participants exhibited higher "quality" of productive negativity, i.e. how many demonstrations of correct conceptual knowledge resulted from each negative experience. However, no relationship was seen between this quality and post-test misconceptions amongst the simulation group ($p = .442$), whereas a trending relationship was observed in the gaming group ($p = .066$). Therefore, we surmised that the quality of productively negative experiences may be indicative of conceptual understanding in the gaming group only.

Below, we give a new, more granular qualitative analysis of the demonstrations of correct conceptual knowledge (Section 4.2.1) and instances of productive negativity (Section 4.2.2) experienced in our interactive simulations and relate this back to their ATMSG designs.

4.2. Granular qualitative analysis of productive negativity and demonstrations of correct conceptual knowledge

4.2.1 How game design influenced demonstrations of correct conceptual knowledge

We coded demonstrations of correct conceptual knowledge into three categories: concentration, temperature, and crowding. In *MolWorlds*' ATMSG (**Figure 3**), these mechanics are indicated by items viii-a (followed by vi and/or vii) and viii-b; in *MolSandbox*' (**Figure 5**), these mechanics are indicated by items iv-a (followed by iii and/or v) and iv-b. **Table 3** summarizes this data in both *total of each type* of demonstration per individual and by *each type as a percentage* of that individual's total interactions with those mechanics. To clarify with an example, a user might have made 10 demonstrations of correct conceptual knowledge in the temperature category but had modified the temperature a total of 20 times; therefore, they made productive

modifications to temperature 50% of the time (i.e. productive modifications to temperature divided by all modifications to temperature x 100%). A detailed list of interaction data (i.e. total temperature modifications, molecules collected, released, etc.) can be referred to in Gauthier & Jenkinson (2017), Table 1. To examine the differences in demonstrations of correct conceptual knowledge between groups, Mann-Whitney U tests were performed on each demonstration subcategory and on the overall total. Test results for each comparison are presented in **Table 3**.

While simulation-users outperform game-players in raw counts of overall demonstrations (any type), as well as demonstrations of concentration and of temperature, a different picture emerges when we measure their "productive" actions as a percentage of their total actions. Upon doing this, we see that the percentage of productive interactions regarding concentration, crowding, and overall demonstrations are similar across groups, while the game group made a higher proportion of productive modifications in temperature.

It is worth noting that the percentage of productive modifications to concentration seems very low; this is because, telemetrically, total molecules collected were recorded at the level of a single molecule (e.g. 15 enzymes were collected), since each molecule was collected individually, while a demonstration of correct conceptual knowledge was coded at the event level (e.g. the player correctly collected many enzymes to reduce concentration), resulting in a finding skewed towards a low percentage. Hence, this number should not be taken as a literal representation of the proportion of collection/release events that were productive but does offer a comparison between our two stimuli groups that compensates for the greater interactive freedom—and thus, greater number of modifications made—in the interactive simulation.

Table 3. Types of correct conceptual knowledge (as raw counts and as a percentage of total environmental modifications of that type) demonstrated by simulation (*MolSandbox*) and gaming (*MolWorlds*) participants during the 30-minute exposure time, compared using a Mann-Whitney U test.

[INSERT TABLE 3]

4.2.2 How game design influenced instances of productive negativity

While *MolWorlds*-players exhibited fewer demonstrations of correct conceptual knowledge, they experienced a significantly greater number of productively negative events than did *MolSandbox*-users (**Table 4**). In our video-coding process, we interpreted the source of each productively negative event and established five categories that relate back to the interventions' ATMSGs: resource retentiveness, difficult resource collection, resources lost due to overheating, navigation and reaching the checkpoint (game condition only), and simulation-based negativity. **Figure 6** depicts each of these negativity sources and their resultant productive loops using wide, semi-transparent arrows overlaid on the ATMSGs of both stimuli. To examine the differences in productive negativity between groups, Mann-Whitney U tests were performed on each negativity source subcategory and on the overall total. Test results for each comparison are presented in **Table 4**. We describe each source in more detail below; additional examples of gameplay interactions for each type of productive negativity can be found in the supplementary materials document.

Table 4. Sources of productive negativity experienced by simulation (*MolSandbox*) and gaming (*MolWorlds*) participants during the 30-minute exposure time, compared using a Mann-Whitney U test.

[INSERT TABLE 4]

A productively negative event from resource retentiveness was identified when the participant released the exact number of molecules needed for a molecular process to ensue, leading to negativity when binding was not immediate, and followed this with a productive modification. For example, if there are three inactive cargo receptors each requiring a cargo molecule before vesicle formation can initiate, releasing only three cargo molecules from the inventory (instead of several more to increase the probability of a binding event occurring) would constitute "resource retentiveness". This sequence of interactions is visualized in **Figure 6-A**. Seven out of 20 simulation participants and 14 of 20 gaming participants exhibited this behaviour in significantly different proportions when analyzed by a Chi-square test of independence ($\chi^2(1) = 4.91$, $p =$

.027, $\phi = 0.35$). Furthermore, game players recorded a significantly greater number of productively negative events from this source than simulation-users, as shown in **Table 4**.

The source of a productively negative event from "difficult resource collection" was identified when the negativity was initiated by the apparent chasing of molecules in the environment, either with the cursor in the simulation stimulus, or with the character in the gaming stimulus. Five simulation-users and eight game-players encountered productive negativity from this source ($\chi^2(1) = 1.03$, $p = .311$), with a significantly greater number exhibited by the game-users on average (**Table 4**). **Figure 6-B** illustrates this sequence of interactions.

The ability to interactively increase the temperature in the simulation led to another source of negativity. If a user increased the environmental temperature too much for a prolonged time, the membrane became loose and molecules required for certain processes often escaped the area, resulting in low concentrations and, therefore, low rates of desired molecular interactions (**Figure 6-C**). This occurred with four game- and six simulation-users ($\chi^2(1) = 0.53$, $p = .465$), showing relatively equal distribution of this source of productive negativity across stimuli.

A large proportion of *MolWorlds* players were exposed to a source of productive negativity unique to the game, involving navigating and reaching the level's checkpoint (**Figure 6-D**). Even if environmental conditions (concentration, temperature, etc.) were optimized and the intended cellular events achieved, it is not certain that the player would progress in the game. For example, in level 7, many players were impeded by the crowdedness of the environment and did not situate themselves in the vesicle in time to get transported across the membrane to the checkpoint and had to repeat the level. Navigation by itself also caused negativity; while travelling through crowded environments, the player was often bumped and jostled by other molecules, which prompted the player to shrink the character size or make the environment cooler and, thus, easing navigation. These types of negativity occurred at least once to 15 of 20 *MolWorlds* participants, a significant proportion when evaluated with a goodness of fit test ($\chi^2(1) = 5.00$, $p = .025$).

Lastly, productive negativity in either condition could be purely simulation-driven (**Figure 6-E**). The source was identified as purely simulation-driven if the user made one or more correct adjustments to the environment (i.e. 1+ demonstration of correct conceptual knowledge) but the programmed behaviour of the molecules led to a negative experience, regardless. For example, in level 9, cytosolic enzymes may degrade all the ligands necessary to open a ligand-gated membrane channel due to random chance, even if the student increased the concentration of the ligand and an appropriate molecular inhibitor. This may prompt the user to reflect on other factors that affect an emergent system. Six simulation participants and six gaming participants experienced productive negativity in this manner, an equal distribution across groups ($\chi^2(1) = 0.00$, $p = 1.000$).

Productive negativity was *not* observed through the scoring and 3-star reward system at the end of each level. *MolWorlds* players did not immediately repeat levels when presented with 1- or 2-star feedback as was expected and outlined by the game's ATMSG (**Figure 3**, **Table 1**: mechanic #v).

[INSERT FIGURE 6 – Full page]

Figure 6. Productive negativity interaction loops transposed over the *Activity Theory Model for Serious Games* (ATMSG) for both simulation and gaming stimuli. Negativity sources, labelled A-E, are represented by wide, dark arrows, while the productive response is represented by a lighter grey arrow. Refer to Figure 3 and Table 1 for the more detailed ATMSG for *MolWorlds*; for *MolSandbox*, refer to Figure 5 and Table 2.

We were also interested in investigating which source or combination of mechanics in each intervention was responsible for the most productively negative events. To do this, we performed a Friedman test for each intervention comparing mean source type, while excluding the navigation source for the simulation comparison. There was a significant difference in productively negative events produced by source types for both *MolWorlds* ($\chi^2(4) = 23.26$, $p < .001$) and *MolSandbox* ($\chi^2(3) = 9.83$, $p = .020$). We conducted post hoc analyses with Wilcoxon signed-rank tests and a Bonferroni correction to compensate for multiple comparisons, resulting in a significance level of .005 for the gaming group (10 comparisons) and .008 for the

simulation group (6 comparisons). For *MolWorlds*-players, both resource retentiveness (Z = -3.02, p = .003) and navigation (Z = -3.11, p = .002) were significantly more frequent sources of productive negativity than losing resources to overheating. All other comparisons in the gaming group were not significant. In *MolSandbox*, resource retentiveness and simulation-only sources were more probable than a difficult resource collection source (both Z = -2.71, p = .007), while all other comparisons were statistically similar.

Lastly, we explored what types of correct conceptual knowledge were demonstrated immediately following negative events from different sources; this analysis reflects the probability of the different productive-flow arrows in **Figure 6**, illustrated by the wide, light grey arrows. For example, in **Figure 6**-A, we observe that participants responded with either a modification to concentration (mechanic viii-a in **Figure 3**, or iv-a in **Figure 5**) or a modification to temperature/crowding (mechanic viii-b in **Figure 3**, or iv-b in **Figure 5**) following negativity from resource retentiveness. **Figure 7** provides a graphical summary of the data, splitting demonstrations of correct conceptual knowledge by negativity source. Using our raw dataset of coded instances of productive negativity and demonstrations of correct conceptual knowledge (i.e. not summarized by participant), we performed a Chi-square goodness of fit test on each negativity source. Many instances of negativity elicited more than one type of productivity in sequence (e.g. a modification in both temperature and concentration), in which case that instance is represented more than once in the dataset for this analysis only. Also, note that the productivity following a negative event was sometimes classified as "other" when one of our three categories of demonstrations did not immediately follow but the participant reloaded the level and approached the challenge in a different, more successful way, thus qualifying the negative event as productive.

[INSERT FIGURE 7]

Figure 7. Raw counts of different types of demonstrations of correct conceptual knowledge elicited immediately following a negative experience.

Amongst non-gaming participants, simulation-driven productive negativity was equally likely to be followed by demonstrations about concentration and temperature ($n_{negEvent}$ = 14, $\chi^2(1) = 1.14$, p = .285), with no occurrences of crowding or other types. Low event counts elicited by the simulation group prevented us from performing valid chi-square analyses on all other negativity source types, although the distribution can be observed visually in **Figure 7**. Amongst gaming participants, the goodness of fit test was significant for resource retentiveness ($n_{negEvent}$ = 33, $\chi^2(2) = 11.63$, p = .003), navigation ($n_{negEvent}$ = 39, $\chi^2(3) = 29.82$, $p <$ .001), and simulation-only ($n_{negEvent}$ = 17, $\chi^2(2) = 8.94$, p = .011) sources, while difficult resource collection and overheating sources failed the assumptions of the Chi-square, similar to the simulation group. We performed an examination of the adjusted standardized residuals to identify which types of interactions were most likely, where values greater than 2 or less than -2 would be considered significant; this method of one-way Chi-square post-hoc analysis is supported in the literature (Agresti, 2013; Delucchi, 1993; Sharpe, 2015). The analysis suggests that proper modifications in concentration will occur more frequently than chance following negativity caused by resource retentiveness (adj. res. = 3.7), navigation (adj. res. = 5.27), and by simulated molecular behaviour (adj. res. = 3.73). Productive temperature modification is more likely to follow negativity from navigation (adj. res. = 2.11, though less likely than concentration here), while modifications to crowding is not likely to result from navigation (adj. res. = -2.85) and was not elicited at all from resource retention and simulation-driven sources. Other productivity (e.g. restarting with a different strategy) is also less likely to be elicited by retentiveness (adj. res. = -2.52), navigation (adj. res. = -2.56), and simulation-based (adj. res. = -2.03) negativity.

5. Discussion

5.1. Overall findings

Both *MolSandbox* and *MolWorlds* present an interactive, simulated molecular environment where students can learn experientially through digital experimentation of cellular factors. As such, both applications afford students with the opportunity to confront their expectations related to molecular behaviour and alter their

conceptions. Both stimuli proved to be efficacious at facilitating a better understanding of the role of randomness at the molecular level but did so by instigating very different patterns of interactions that link back to their mechanic designs, which we believe is of interest to discuss. Below we answer our three research questions about how game design alters the nature and frequency of students' demonstrations of conceptual knowledge (Section 5.1.1) and their experience of productive negativity (Section 5.1.2) in an interactive simulation, then glean insights from our findings about how we can apply game design to achieve desired game-flow interactions (Section 5.1.3).

5.1.1 How does the presence of serious game design moderate the frequency and nature of students' demonstrations of conceptual knowledge within the interactive simulation?

While both the serious game and interactive simulation offered the same environmental modifications—changes to molecular concentrations, crowding, and temperature—the presence of game mechanics in *MolWorlds* fundamentally changed the frequency and nature of these interactions over a fixed 30-minute period. Game participants exhibited significantly fewer overall interactions (i.e. productive and/or unproductive modifications), as well as fewer demonstrations of correct conceptual knowledge (i.e. only productive modifications) due to the rules and constraints present in the game (e.g. requiring the user to find and collect power-ups). Meanwhile, simulation-users were free to make as many modifications as they pleased. For example, *MolSandbox* participants made adjustments to the temperature about six times more frequently than *MolWorlds* participants. This interactive freedom may appear to be a distinct benefit of the simulation over the game. However, by looking more holistically at total modifications made, total demonstrations of correct conceptual knowledge (i.e. helpful modifications), and the proportion of these two statistics (demonstrations of correct conceptual knowledge divided by total modifications), we can interpret how well the player understood the effect that these modifications would have on the simulated environment. For instance, a higher proportion of correct modifications to temperature would indicate a better understanding of the consequences of temperature change on the rates of molecular interactions. With respect to temperature modifications, the simulation group showed significantly less productivity than the game group. We often observed *MolSandbox*-users increasing the temperature upon beginning a level, before investigating what proteins were present, to see if anything in the environment would interact immediately, resulting in high counts of temperature interactions that were not productive. This type of behaviour was observed much less frequently in *MolWorlds*, perhaps because the players wanted to use their power-ups strategically. In **Figure 3**, this rule can be identified by the diamond shape leading to viii-b and is notably missing in **Figure 5**, leading to this difference in interaction. If we look at demonstrations of correct conceptual knowledge as a proportion of entire modifications, we see similar productivity between groups with respect to molecular concentration and crowding.

Overall, this data suggests that students engage with the interactive simulation in a much more exploratory manner because of the lack of rules and restrictions that would otherwise be present in a game. Similar findings are reported in Gauthier & Jenkinson (2015) where a serious game-study aid was compared with a similar non-gaming study aid in the study of human anatomy.

5.1.2 How does the presence of serious game design fundamentally alter the frequency and nature of productive negativity experienced within the interactive simulation?

As a game, *MolWorlds* integrates more mechanics in its primary game-flow axis (**Figure 3**) than does *MolSandbox* (**Figure 5**) and contains nearly double the instances where the user must make a critical decision about which actions to undertake next (represented by the diamond structures in the diagrams). The presence of these mechanic patterns in *MolWorlds*—specifically resource management (collecting molecules and power-ups), an immersed character (navigation/exploration), and sequential level progression (reaching checkpoints)—increased the probability for productive negativity to occur. The five types of productive negativity observed can be broadly categorized as (a) "mechanic-based" sources (namely resource retentiveness, difficulty collecting resources, resources lost due to overheating, and difficulty navigating and reaching checkpoints) because the negativity was due to specific user interactions; and as (b) "simulation-based" sources, which are due to the random nature of the simulation and not primarily due to users'

interactions. In *MolWorlds*, mechanic-based sources of negativity made up the bulk of their productively negative experiences, we believe due to their prominence on the main game-flow axis. For the *MolSandbox* group, a simulation-based source of negativity was the leading source of negativity, contributing to a mean 43% of productively negative experiences, compared to 15% in the gaming group. Overall, participants in both groups experienced equal amounts of negativity from simulation-based sources. Accordingly, the presence of game design does not dampen the underlying simulation's innate effectiveness at promoting productive negativity.

In contrast, most mechanic-based sources of negativity did differ between groups, most notably resource retentiveness. Since users had direct control over the concentration of molecules in the environment, they had the opportunity to confront their expectation of directed molecular motion by releasing one (or very few molecules) into the environment, potentially leading to a productively negative experience when binding was not immediate. We see a higher proportion of *MolWorlds*-players performing these actions because, as suggested in Section 2.4.2, the game mechanics in *MolWorlds* leverage a player's natural tendency to conserve collected resources (similar to Mitgutsch and Weise [2011a, 2011b]'s findings with *Afterland*) and the idea of lock-and-key mechanisms in molecular biology. Whether or not the player has a persistent misconception of directed molecular motion, by leveraging these three things, the player is most likely to release only one ligand (key) to elicit a conformational change (or unlock) the ligand-gated membrane channel (door) and continue his quest. In relation to the ATMSG models, resource retentiveness in *MolWorlds*, as depicted in **Figure 6**-A, involves the interplay between three mechanics (navigation/exploration, collecting molecules, and releasing molecules) on the primary game-flow axis, whilst in *MolSandbox* it depends only on a single mechanic; this quantitative change in the design of the applications made a significant difference in the frequency of this type of productive negativity. By adding navigation and molecule collection on the primary game-flow axis—that is, making it a required interaction—it drastically increased resource retentiveness in game-players and an appropriate response (i.e. modifications to concentration, temperature, or both) often followed (**Figure 7**).

Differences in "difficult resource collection" as a negativity source can also be explained by game design. In our interventions, the molecules are programmed to move in random directions to approximate Brownian motion. In *MolWorlds*, the player collects molecules by moving the character around with the ASWD/arrow keys and holding down the spacebar, which collects molecules that collide with the character; this is an interplay between navigation and collection, both of which appear on the primary game-flow axis. In areas in which there are low concentrations of desired molecules, we sometimes observed the player pursuing a single, randomly-moving molecule for some time before realizing that, if they decreased the temperature and increased the size of the character using power-ups (as shown by the distribution of demonstrations of correct conceptual knowledge following "difficult collection" in **Figure 7)**, catching the molecule becomes much easier. Unlike using the character to collect molecules in *MolWorlds*, collecting molecules in *MolSandbox* involves clicking and scrubbing the cursor over the molecules. This scrubbing mechanism was intended to be less game-like, while producing the same telemetric data as *MolWorlds* for data comparison. However, under high temperature conditions (which, in *MolSandbox*, was often) the molecules move very quickly, and collection becomes more difficult. We observed this type of productively negative experience in one simulation-user; we can suggest that she reflected on the effects of temperature because she then reduced the temperature gauge to near freezing to collect the desired molecules more easily. However, this type of negativity was only observed in a single user, likely because the collection mechanic in *MolSandbox* falls off the primary game-flow axis, making it a non-required mechanic to complete a level successfully. Contrastingly, in *MolWorlds*, the collection mechanic is central to gameplay and is also accompanied by the navigation mechanic, furthering the potential for challenge. Therefore, we can suggest that game design is directly responsible for the increase in productive negativity caused by difficult resource collection.

Navigating a 3rd-person character in the molecular environment proved to be a leading source of productive negativity in *MolWorlds* and was unique to the serious game. The player was physically impeded by the presence of macromolecules; the difficulty of navigating through them was directly related to the molecules' size, quantity, and their movement, governed by temperature. As such, we observed this source of negativity

often being followed by a cooling of temperature that prevented the character from being bumped around too much; decreases in character size that allowed Goodcell to slip past molecules more easily; and decreases in the concentration of surrounding molecules, also facilitating easy passage. Furthermore, requiring the character to physically reach a checkpoint in order to progress to the next level afforded another opportunity for productive negativity related to navigation, not present in the interactive simulation. For example, in level 7, if the player does not position the character inside of the vesicular bud, the vesicle will form without Goodcell inside of it, stranding him and requiring the player to redo the level. We observed many players restart this level and then make careful adjustments to temperature and concentrations to form the vesicle while properly positioning the character. This was not the case for *MolSandbox*-users who, if they successfully initiated vesicle formation and docking, would receive "completed" status on that level, demonstrating that navigation with a 3rd-person character as a game mechanic is an efficient source of negativity in our interactive molecular environment.

Lastly, productive negativity instigated by losing resources due to overheating did not differ between groups, even though the rules governing temperature interactions were distinctly different. In *MolWorlds*, temperature was adjusted using power-ups (which had to be found and collected), the effects of which ran out after 10 seconds, whereas in *MolSandbox*, temperature could be adjusted freely and would remain at the modified temperature until changed again. To elicit quicker motion and interactions of molecules, a common strategy among all participants was to increase the heat to the highest capacity, sometimes resulting in molecules escaping the confines of the membrane. This caused productive negativity when the remaining low concentrations of the molecules in the environment resulted in reduced interactions of interest. One might assume that, having no restrictions on temperature modification, the simulation-users would have experienced a greater number of productively negative experiences from this source. However, this source was observed equally between groups. This may be explained because this source involved the same number of mechanics in both stimuli (i.e. only one: temperature adjustment). Furthermore, the temperature adjustment mechanic falls off the primary game-flow axis in both stimuli, which may be responsible for the relatively low frequency of this negativity source overall.

A final point of discussion surrounds scoring and the 3-star feedback system as a negativity source, represented by mechanic #x in **Figure 3**. This mechanic was intended to promote level repetition (productive response) when three stars were not achieved (negativity source). We attribute this to the 30-minute limit that was placed on participants who, we believe, moved on to complete new levels instead of spending time trying to achieve higher scores in lower levels. This is an important finding about how research protocol design can alter certain interactions in a gaming environment; the players may have felt more rushed, so did not engage with the scoring mechanic as described by the ATMSG. This finding will be re-investigated in our ongoing research which allows students to play *MolWorlds* for a voluntary length of time.

5.1.3 What are specific game design strategies to increase the occurrence of desired interactions and game-flow loops (e.g. productive negativity loops)?

By thoroughly describing our game and simulation intervention designs using the Activity Theory Model of Serious Games, we have communicated how we intended gaming, learning, and instructional mechanics to elicit desired game-flow loops that we hypothesized would impact our learning objectives. Furthermore, by performing a rigorous qualitative analysis of screencasts of game/simulation use, we have provided empirical evidence that these mechanics did elicit these desired interactions. By comparing the ATMSG models of a serious game and similar non-gaming intervention, and overlaying these with our qualitative data, we can make concrete design suggestions about the integration of game mechanics into interactive learning tools. Ultimately, the frequency of engaging in such game-flow loops is dependent on students' understanding of how the interactions affect the simulated system, but these design suggestions may influence the probability of initiating interaction in the loop or pattern. Our suggestions are as follows:

1. Including additional game mechanics on the primary game-flow axis may limit the exploratory nature of the application but does not impede overall productive interactions from occurring.

In other words, making a mechanic mandatory before another mechanic can be accessed reduces the frequency of the later interaction, in addition to changing the students' approach to its use. Take temperature change as a concrete example from this study. In *MolWorlds*, the player is required to navigate, find, collect, and conserve power-ups in order to modify temperature at a later time, whilst temperature can be modified freely in *MolSandbox*. This forces *MolWorlds*-players to use power-ups strategically, resulting in a high frequency of effective temperature modifications, whereas the more exploratory nature of *MolSandbox* provoked a higher number of overall temperature modifications, though a smaller proportion of these were productive.

2. Integrating two or more primary-axis mechanics in a game-flow loop will increase the frequency of interaction with this loop.

An ideal example of this is resource retention as a productive negativity source. *MolWorlds* involves three primary axis mechanics in this negativity loop that are intrinsically linked (navigate/explore, collect, release), whilst *MolSandbox* involves its one-and-only primary-axis mechanic: release molecules. Resultantly, resource retentiveness was significantly more frequent in *MolWorlds* than in the interactive simulation.

3. Gameplay loops involving mechanics that fall off the primary axis (i.e. non-mandatory mechanics) may occur less frequently than those which involve primary axis (i.e. mandatory) mechanics.

To illustrate this concept using *MolWorlds* data as an example, sources of negativity, such as the retention of resources or navigation to a checkpoint (which both required interaction with two or more mandatory mechanics), occurred significantly more frequently than negativity incurred by the loss of resources due to overheating (which involved only non-mandatory mechanics). On the other hand, resource retentiveness loops occurred in equal amounts to navigation/checkpoint loops, even though resource retentiveness involved more mandatory mechanics; this could be due to the integral nature of navigation to the entire game. Further research is needed to explore whether the relationship between loop frequency and its number of mechanics is linear.

Whilst these design strategies may seem obvious to game designers, linking empirical data to visual design models lends greater credibility to designers' intuitions and provides non-designers with a more design-oriented approach to evaluating serious games. These same strategies can be applied beyond serious games to the design of other digital interactive environments where different learning strategies need to be leveraged.

5.2. Limitations and future directions

5.2.1 Limitations and future directions of the game/simulation design

Figure 7 shows a visual distribution of the kinds of correct interactions that directly followed our five different types of productive negativity. It is notable that productive changes in crowding (i.e. a change to character size or pinball size) were uncommon in both computer applications, which may suggest that the students did not grasp the role that crowding plays in molecular systems. An appreciation of the relationship between the size of an entity and its movement at the molecular level is an important concept; more scaffolding around the use of this interaction should be integrated into both the game and the interactive simulation. Furthermore, the inherent 'usefulness' of character-size change in *MolWorlds* is greater than pinball-size change in *MolSandbox*. Since the player controls the character, an increase in character size facilitates resource collection because of increased collisions (we see the most instances of crowding made after negativity from resource collection), while a decrease in character size facilitates navigation because of decreased collisions. The benefits to changing the size of the pinball in *MolSandbox* are not as obvious, apart from level 11 (**Figure 4**-B) where the educational task is to prevent the pinball from being flagged for degradation by ubiquitination (correct action: reduce pinball size). In other levels, increasing the pinball size could facilitate resource collection by placing it in a tight environment, thus crowding the other molecules together and making them easier to collect with the cursor; however, this is may not be an obvious strategy compared to increasing character size in *MolWorlds* to enhance collisions with—and, thus, collection of—

molecules. Future iterations of both interventions should consider strategies for making the consequences of adjustments to crowding equivalent between groups and to enhance its use.

A second modification would involve the 3-star feedback screen at the end of each level in *MolWorlds*. We did not observe feedback as a productive negativity source. To encourage repetition, we could implement motivational statements (e.g. "finish X seconds faster to earn 3 stars!"). However, another probable explanation for lack of level repetition may have to do with our protocol design, discussed below in Section 5.2.2.

5.2.2 Limitations and future directions of the research

This intervention had high internal validity as it took place in a controlled academic setting, with participants computationally randomized and blinded to their intervention groups. However, with high internal validity comes low ecological validity. Individuals assigned to the serious game completed between 6 and 9 out of a total 13 levels of the game, rarely repeating the same level twice. Participants were told that they had a predetermined 30 minutes of playtime, so it is possible that they felt pressured to try to complete as much of the game as they could within the given timeframe, which resulted in lack of level repetition that is not necessarily representative of gameplay in an authentic context. We recently completed a second randomized controlled trial with these same two interventions that employed a variable, voluntary intervention time. Participants were required to use their computer application for a minimum of 20 minutes but could choose to continue its use; preliminary analysis of this data reveals increased level repetition with this approach (data unpublished). Similarly, in their study of voluntary versus compulsory play of a serious game, Rodríguez-Aflecht and colleagues (2017) found that voluntary play led to an increase in mathematical skills learning beyond the compulsory-play group, suggesting that learning may be enhanced in a voluntary play context. Additionally, a longer exposure time would have generated more opportunities for students to experience negativity and master productive manipulations of the environment, which may have led to overall greater learning outcomes. Furthermore, if students had access to the full breadth of the game, they would have been required to transfer their conceptual knowledge to more complex cellular processes and scenarios, which could have further consolidated their understanding.

A second limitation was that participants were aware that we were doing research on educational games, so they were primed to expect to engage in a gaming environment. To blind participants to their assigned group, we told them that they would be assigned to one of two "different types of games that apply game design to various extents" and that we would be investigating differences in interactions, learning, and engagement invoked by the differences in design. Specifically, while *MolWorlds* is clearly a fully-fledged game, we felt that this wording would help blind the simulation condition, so that users would not lose interest in participating when they realized that they were not assigned to the gaming condition. However, this blinding may have had other unintended effects; research shows that framing an activity as a game can hold as much psychological power as implementing actual game mechanics (Lieberoth, 2015). Since we have documented significant differences in interactions between the game and simulation in this paper, it is unlikely that our blinding had influenced the participants in any meaningful way, but this may be responsible for similarities in self-reported engagement discussed in our previous reports (Gauthier & Jenkinson, 2017).

By using two raters to independently code our gameplay screencasts and achieving good inter-rater agreement, we can be moderately confident in our qualitative data. However, other methods, such as participant checking (e.g. asking participants if they agreed that a productively negative event took place at a certain timeframe in the recording), or method triangulation (e.g. conducting a think-aloud play session with a few participants for comparison), would have enhanced the reliability of the data and should be considered in future studies (Twining et al., 2017).

Beyond other limitations of the study (e.g. small sample size) (Gauthier & Jenkinson, 2017), this paper is limited by being one of the first to apply the Activity Theory Model of Serious Games to empirically evaluate the value-added effect of serious game design in a randomized controlled trial and, as such, we were unable to find current literature to support our findings. Recent studies have applied similar frameworks, such as the

Learning Mechanics and Game Mechanics model, to support the integration of pedagogy within the game (Arnab et al., 2013; Callaghan et al., 2016; Koivisto et al., 2016; Proulx et al., 2017)—an important step for the game-based learning literature—but should extend their research by connecting measured interaction data to the serious game design to able to make concrete conclusions about the value-added effect of game design. Our ongoing research will continue to employ similar analytical approaches as described in this paper and we encourage other researchers to try similar strategies to determine the validity of the design considerations described above (Section 5.1.3).

6. Conclusions

Digital game-based learning is an emerging field of research often criticized for lack of rigour. Among the most common criticisms is that research on educational games often lacks a detailed description of how learning and instructional strategies are incorporated with game mechanics (Clark et al., 2016). Without this information, it is difficult for the community to appreciate how the design affects both the educational and entertainment value of the game at hand. Furthermore, many studies draw comparisons between a serious game and "standard education" or other passive learning interventions; this approach may determine whether the game is effective but does not inform our understanding of how the game "works".

The current paper takes a step towards filling these gaps in the serious gaming literature. We described and compared in detail, using the Activity Theory Model of Serious Games, the design of our game and its interactive simulation counterpart, and specified how each mechanic is meant to further the learning objective by facilitating productive negativity. By qualitatively analysing how negativity and its productive response was generated differentially in two stimuli, we found that adding game design limits the exploratory nature of the environment, while making way for the player to demonstrate their conceptual knowledge in a strategic manner. Secondly, we found that the increase of productively negative experiences in game-players (vs. simulation-users) was due to the implementation of more mandatory mechanics, and that negativity sources involving non-mandatory mechanics occurred much less frequently. As such, we are able to recommend the following general game design strategies in relation to the Activity Theory Model of Serious Games: 1) including additional game mechanics on the primary game-flow axis may limit the exploratory nature of the application but does not impede overall productive interactions from occurring; 2) integrating two or more primary-axis mechanics in a game-flow loop will increase the frequency of interaction with this loop; and 3) gameplay loops that involve mechanics that fall off the primary-axis (i.e. non-mandatory mechanics) may occur less frequently than those which involve primary-axis (i.e. mandatory) mechanics. We also conclude that protocol design—such as the amount of time allotted for play and the potential priming effect of group assignment—in a randomized controlled trial should be carefully considered as these decisions may affect how participants approach and respond to feedback in a gaming environment.

The main findings described in Gauthier & Jenkinson (2017) showed significant learning outcomes about the random nature of molecular environments in both interactive applications, with a trend toward increased performance in the gaming group. However, only the game group showed a relationship between misconceptions and the quality of productive negativity experienced during the intervention. This suggests that the more negativity experienced in the game, the more nuanced understanding they developed about the emergent nature of the system, which was reflected in their post-test scores; this highlights the importance of defining a relationship between game design decisions and pedagogical strategies so that learning outcomes can be interpreted. Overall, the results of this study emphasize the relevance of robustly describing game designs and making value-added comparisons with similar interactive stimuli. By understanding how game design alters interactions within a digital application, we have learned how we can leverage game-flow loops (e.g. productive negativity) with gaming, learning, and instructional elements, to further learning objectives in the design of future serious games and other computer applications.

References

Abdul Jabbar, A. I., & Felicia, P. (2015). Gameplay Engagement and Learning in Game-Based Learning: A Systematic Review. *Review of Educational Research*, *85*(4), 1–40. doi:10.3102/0034654315577210

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken NJ: Wiley.

Amory, A. (2007). Game object model version II: A theoretical framework for educational game development. *Educational Technology Research and Development*, *55*(1), 51–77. doi:10.1007/s11423-006-9001-x

Anderson, C. G., Dalsen, J., Kumar, V., Berland, M., & Steinkuehler, C. (2018). Failing up: How failure in a game environment promotes learning through discourse. *Thinking Skills and Creativity*, (March), 1–10. doi:10.1016/j.tsc.2018.03.002

Arnab, S., Brown, K., Clarke, S., Dunwell, I., Lim, T., Suttie, N., … De Freitas, S. (2013). The development approach of a pedagogically-driven serious game to support Relationship and Sex Education (RSE) within a classroom setting. *Computers and Education*, *69*, 15–30. doi:10.1016/j.compedu.2013.06.013

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., de Freitas, S., Louchart, S., … De Gloria, A. (2015). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology*, *46*(2), 391–411. doi:10.1111/bjet.12113

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational Play: Using Games to Position Person, Content, and Context. *Educational Researcher*, *39*(7), 525–536. doi:10.3102/0013189X10386593

Bartholomew, L. K., Parcel, G. S., & Kok, G. (1998). Intervention Mapping: A Process for Developing Theory and Evidence-Based Health Education Programs. *Health Education & Behavior*, *25*(5), 545–563. doi:10.1177/109019819802500502

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., … Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers and Education*, *94*, 178–192. doi:10.1016/j.compedu.2015.11.003

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. doi:10.1191/1478088706qp063oa

Callaghan, M. J., Mcshane, N., Eguíluz, A. G., Teillès, T., & Raspail, P. (2016). Practical Application of the Learning Mechanics – Game Mechanics (LM-GM) framework for Serious Games Analysis in Engineering Education. In *13th International Conference on Remote Engineering and Virtual Instrumentation* (pp. 382–386). doi:10.1109/REV.2016.7444510

Carvalho, M. B., Bellotti, F., Berta, R., De Gloria, A., Sedano, C. I., Hauge, J. B., … Rauterberg, M. (2015). An activity theory-based model for serious games analysis and conceptual design. *Computers and Education*, *87*, 166–181. doi:10.1016/j.compedu.2015.03.023

Charsky, D. (2010). From Edutainment to Serious Games: A Change in the Use of Game Characteristics. *Games and Culture*, *5*(2), 177–198. doi:10.1177/1555412009354727

Chi, M. T. H. (2005). Commonsense Conceptions of Emergent Processes: Why Some Misconceptions Are Robust. *Journal of the Learning Sciences*, *14*(2), 161–199. doi:10.1207/s15327809jls1402_1

Chi, M. T. H. (2013). Two kinds and four sub-types of misconceived knowledge, ways to change it, and the learning outcomes. In *International Handbook of Research on Conceptual Change* (pp. 49–70).

Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive science*, *36*(1), 1–61. doi:10.1111/j.1551-6709.2011.01207.x

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis. *Review of Educational Research*, *86*(1), 79–122. doi:10.3102/0034654315582065

Clifford, M. M. (1988). Failure tolerance and academic risk-taking in ten- to twelve- year-old students. *British Journal of Educational Psychology*, *58*(1), 15–27.

Clifford, M. M. (1991). Risk Taking : Theoretical , Empirical , and Educational Considerations Risk Taking : Theoretical , Empirical , and Educational Considerations. *Educational Psychologist*, *26*(3 & 4), 37–41.

Coley, J. D., & Tanner, K. D. (2012). Common origins of diverse misconceptions: cognitive principles and the development of biology thinking. *CBE life sciences education*, *11*(3), 209–15. doi:10.1187/cbe.12-06-0074

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with

technology. *Journal of Educational Psychology*, *105*(4), 1082–1099. doi:10.1037/a0032674

Darabi, A., Arrington, T. L., & Sayilir, E. (2018). Learning from failure: a meta-analysis of the empirical studies. *Educational Technology Research and Development*, 1–18. doi:10.1007/s11423-018-9579-9

De Freitas, S., Rebolledo-Mendez, G., Liarokapis, F., Magoulas, G., & Poulovassilis, A. (2010). Learning as immersive experiences: Using the four-dimensional framework for designing and evaluating immersive learning experiences in a virtual world. *British Journal of Educational Technology*, *41*(1), 69–85. doi:10.1111/j.1467-8535.2009.01024.x

Delucchi, K. L. (1993). On the use and misuse of chi-square. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 294–319). Hillsdale, NJ: Lawrence Erlbaum Associates.

DeSmet, A., Van Cleemput, K., Bastiaensens, S., Poels, K., Vandebosch, H., Malliet, S., … De Bourdeaudhuij, I. (2016). Bridging behavior science and gaming theory: Using the Intervention Mapping Protocol to design a serious game against cyberbullying. *Computers in Human Behavior*, *56*, 337–351. doi:10.1016/j.chb.2015.11.039

Dickey, M. D. (2005). Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development*, *53*(2), 67–83. doi:10.1007/BF02504866

Gadamer, H. G. (1998). *Truth and Method*. New York: Continuum.

Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding Randomness and its Impact on Student Learning : Lessons Learned from Building the Biology Concept Inventory (BCI). *CBE–Life Sciences Education*, *7*, 227–233. doi:10.1187/cbe.07

Gauthier, A., Jantzen, S., McGill, G., & Jenkinson, J. (2019). Molecular Concepts Adaptive Assessment (MCAA) characterizes undergraduate misconceptions about molecular emergence. *CBE–Life Sciences Education*, *18*(ar4), 1–17. doi:DOI:10.1187/cbe.17-12-0267

Gauthier, A., & Jenkinson, J. (2015). Game Design for Transforming and Assessing Undergraduates' Understanding of Molecular Emergence (Pilot). In R. Munkvold & L. Kolås (Eds.), *Proceedings of the 9th European Conference on Games Based Learning* (pp. 656–663). Steinkjer, Norway: Academic Conferences and Publishing International Limited.

Gauthier, A., & Jenkinson, J. (2017). Serious Game Leverages Productive Negativity to Facilitate Conceptual Change in Undergraduate Molecular Biology: A Mixed-Methods Randomized Controlled Trial. *International Journal of Game-Based Learning*, *7*(2), 20–34. doi:10.4018/IJGBL.2017040102

Gee, J. P. (2005). Good Video Games and Good Learning. *Phi Kappa Phi Forum*, *85*(2), 33–37.

Gee, J. P. (2007). *What Video Games Have To Teach Us About Learning And Literacy* (2nd ed.). New York, New York, USA: Palgrave MacMillan.

Hoffman, B., & Nadelson, L. (2009). Motivational engagement and video gaming: a mixed methods study. *Educational Technology Research and Development*, *58*(3), 245–270. doi:10.1007/s11423-009-9134-9

Honey, M. A., Hilton, M., & National Research Council's Committee on Science Learning. (2011). *Learning Science Through Computer Games and Simulations*. (M. A. Honey & M. Hilton, Eds.). Washington, DC: National Academies Press.

Juul, J. (2009). Fear of failing? the many meanings of difficulty in video games. In M. J. P. Wolf & B. Perron (Eds.), *The video game theory reader 2* (pp. 237–252). New York: Routledge. doi:10.1017/CBO9781107415324.004

Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, *26*(3), 379–424. doi:10.1080/07370000802212669

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022. doi:10.1111/cogs.12107

Kapur, M., & Bielaczyc, K. (2012). Designing for Productive Failure. *Journal of the Learning Sciences*, *21*(1), 45–83. doi:10.1080/10508406.2011.591717

Kapur, M., & Rummel, N. (2012). Productive failure in learning from generation and invention activities. *Instructional Science*, *40*(4), 645–650. doi:10.1007/s11251-012-9235-4

Kelle, S., Klemke, R., & Specht, M. (2011). Design patterns for learning games. *Int. J. Technology Enhanced Learning*, *3*(6), 555–569. doi:10.1504/IJTEL.2011.045452

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *Internet and Higher Education*, *8*(1), 13–24. doi:10.1016/j.iheduc.2004.12.001

King, D., Delfabbro, P., & Griffiths, M. (2009). Video Game Structural Characteristics: A New Psychological Taxonomy.

International Journal of Mental Health and Addiction, *8*(1), 90–106. doi:10.1007/s11469-009-9206-4

Koivisto, J.-M., Haavisto, E., Niemi, H., Katajisto, J., & Multisilta, J. (2016). Elements Explaining Learning Clinical Reasoning Using Simulation Games. *International Journal of Serious Games*, *3*(4), 2384–8766. doi:10.17083/ijsg.v3i4.136

Kolb, D. A. (1984). *Experiential learning : experience as the source of learning and development*. New Jersey: Prentice Hall.

Lameras, P., Arnab, S., Dunwell, I., Stewart, C., Clarke, S., & Petridis, P. (2017). Essential features of serious games design in higher education: Linking learning attributes to game mechanics. *British Journal of Educational Technology*, *48*(4), 972–994. doi:10.1111/bjet.12467

Lieberoth, A. (2015). Shallow Gamification Testing Psychological Effects of Framing an Activity as a Game. *Games and Culture*, *10*(3), 229–248. doi:10.1177/1555412014559978

Meir, E., Perry, J., Stal, D., Maruca, S., & Klopfer, E. (2005). How effective are simulated molecular-level experiments for teaching diffusion and osmosis? *Cell biology education*, *4*(3), 235–48. doi:10.1187/cbe.04-09-0049

Meyer, D. K., & Turner, J. C. (2006). Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review*, *18*(4), 377–390. doi:10.1007/s10648-006-9032-1

Meyer, J. H. F., & Land, R. (2005). Threshold concepts and troublesome knowledge (2): Epistemological considerations and a conceptual framework for teaching and learning. *Higher Education*, *49*(3), 373–388. doi:10.1007/s10734-004-6779-5

Michie, S., Johnston, M., Francis, J., Hardeman, W., & Eccles, M. (2008). From Theory to Intervention: Mapping Theoretically Derived Behavioural Determinants to Behaviour Change Techniques. *Applied Psychology*, *57*(4), 660–680. doi:10.1111/j.1464-0597.2008.00341.x

Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, (Winter), 6–20.

Mitgutsch, K., & Weise, M. (2011a). Subversive Game Design for Recursive Learning. In *DiGRA 2011 Conference: Think Design Play* (pp. 1–16).

Mitgutsch, K., & Weise, M. (2011b). Afterland – From well theorized to well learned? In D. Davidson (Ed.), *Well Played* (1.0., Vol. 1, pp. 33–48). Pittsburgh, Pennsylvania: ETC Press. doi:10.1017/CBO9781107415324.004

Odom, A. L. (1995). Secondary & College Biology Students ' Osmosis Misconceptions About Diffusion and Osmosis. *The American Biology Teacher*, *57*(7), 409–415.

Paz-Y-Miño-C, G., & Espinosa, A. (2012). Introduction: why people do not accept evolution: using protistan diversity to promote evolution literacy. *The Journal of eukaryotic microbiology*, *59*(2), 101–4. doi:10.1111/j.1550-7408.2011.00604.x

Piaget, J. (1974). *Understanding causality*. (W. W. Norton, Ed.). New York, NY.

Prensky, M. (2011). Comments on research comparing computer games to other instructional methods. In S. Tobias & J. D. Fletcher (Eds.), *Computer Games and Instruction* (pp. 251–278). Charlotte, NC: Information Age Publishing.

Proulx, J.-N., Romero, M., & Arnab, S. (2017). Learning Mechanics and Game Mechanics Under the Perspective of Self-Determination Theory to Foster Motivation in Digital Game Based Learning. *Simulation & Gaming*, *48*(1), 81–97. doi:10.1177/1046878116674399

Robic, S. (2010). Mathematics , Thermodynamics , and Modeling to Address Ten Common Misconceptions about Protein Structure , Folding , and Stability. *CBE—Life Sciences Education*, *9*, 189–195. doi:10.1187/cbe.10

Rocklöv, J. (2016). Climate science: Misconceptions of global catastrophe. *Nature*, *532*, 317–318.

Rodríguez-Aflecht, G., Hannula-Sormunen, M., McMullen, J., Jaakkola, T., & Lehtinen, E. (2017). Voluntary vs Compulsory Playing Contexts. *Simulation & Gaming*, *48*(1), 36–55. doi:10.1177/1046878116673679

Salen, K., & Zimmerman, E. (2003). *Rules of Play: Game Design Fundamentals*. Cambridge, MA: MIT Press.

Sanger, M. J., & Iowa, N. (2000). Addressing student misconceptions concerning electron flow in aqueous solutions with instruction including computer animations and conceptual. *International Journal of Science Education*, *22*, 521–537.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309. doi:10.1016/S1364-6613(00)01506-0

Sharpe, D. (2015). Your Chi-Square Test is Statistically Significant: Now What? *Practical Assessment, Research &*

Evaluation, *20*(8), 1–10.

Slotta, J. D., & Chi, M. T. H. (2006). Helping Students Understand Challenging Topics in Science Through Ontology Training. *Cognition and Instruction*, *24*(2), 261–289. doi:10.1207/s1532690xci2402_3

Squire, K. (2013). Video Game – Based Learning : An Emerging Paradigm for Instruction. *Performance Improvement Quarterly*, *21*(2), 7–36. doi:10.1002/piq

Starks, K. (2014). Cognitive behavioral game design: A unified model for designing serious games. *Frontiers in Psychology*, *5*(FEB), 1–10. doi:10.3389/fpsyg.2014.00028

Tibell, L. A. E., & Rundgren, C.-J. (2010). Educational challenges of molecular life science: Characteristics and implications for education and research. *CBE life sciences education*, *9*(1), 25–33. doi:10.1187/cbe.08-09-0055

Twining, P., Heller, R. S., Nussbaum, M., & Tsai, C.-C. (2017). Some guidance on conducting and reporting qualitative studies. *Computers & Education*, *106*, A1–A9. doi:10.1016/j.compedu.2016.12.002

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer Gaming and Interactive Simulations for Learning: A Meta-Analysis. *Journal of Educational Computing Research*, *34*(3), 229–243. doi:10.2190/FLHV-K4WA-WPVQ-H0YM

White, B. T., & Bolker, E. D. (2008). Interactive computer simulations of genetics, biochemistry, and molecular biology. *Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology*, *36*(1), 77–84. doi:10.1002/bmb.20152

Whitton, N. J. (2010). *Learning with digital games: A practical guide to engaging students in higher education*. (F. Lockwood, A. W. (Tony) Bates, & S. Naidu, Eds.). New York, NY: Routledge.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*(2), 249–265. doi:10.1037/a0031311