**International Journal of Game-Based Learning**
Volume 7 • Issue 2 • April-June 2017

# Serious Game Leverages Productive Negativity to Facilitate Conceptual Change in Undergraduate Molecular Biology:
## A Mixed-Methods Randomized Controlled Trial

Andrea Gauthier, University of Toronto, Institute of Medical Sciences, Toronto, Canada
Jodie Jenkinson, University of Toronto Mississauga, Biomedical Communications, Mississauga, Canada

## ABSTRACT

We designed a serious game, MolWorlds, to facilitate conceptual change about molecular emergence by using game mechanics (resource management, immersed 3rd person character, sequential level progression, and 3-star scoring system) to encourage cycles of productive negativity. We tested the value-added effect of game design by comparing and correlating pre- and post-test misconceptions, interaction statistics, and engagement in the game with an interactive simulation that used the same graphics and simulation system but lacked gaming elements. We tested first-, second-, and third-year biology students' misconceptions at the beginning and end of the semester ($n = 526$), a subset of whom played either the game ($n = 20$) or control ($n = 20$) for 30 minutes prior to the post-test. A 3x3 mixed model ANOVA revealed that, while educational level (first-, second-, or third-year biology) did not influence misconceptions from pre-test to post-test, the intervention type (no intervention, simulation, or game) did ($p<.001$). Pairwise comparisons showed that participants exposed to the interactive simulation ($p = .007$), as well as those exposed to the game ($p<.001$), lost significantly more misconceptions in comparison to those who did not receive any intervention, while adjusting for educational level. A trending difference was found between the simulation group and the gaming group ($p = .084$), with the gaming group resolving more misconceptions. Quantitative analysis of click-stream data revealed the greater exploratory freedom of the control simulation, with greater accessibility to individuals who do not play games on a regular basis. However, qualitative analysis of gameplay data showed that MolWorlds-players experienced significantly more instances of productive negativity than control-users ($p<.001$) and that a trending relationship exists between the quality of productively negative events and lower post-test misconceptions ($p = .066$).

## KEYWORDS

Conceptual Change, Interactive Simulation, Molecular Biology, Productive Negativity, Randomized Controlled Trial, Serious Game

DOI: 10.4018/IJGBL.2017040102

## INTRODUCTION

### Background

In molecular biology, students have difficulty understanding how random, seemingly inefficient, mechanisms contribute to the functioning of complex, perceptually efficient, cellular systems and often compensate by attaching agency, or directedness, to molecular species (Momsen et al. 2010; Chi 2005; Chi et al. 2012; Garvin-doxas & Klymkowsky 2008; Chi & Roscoe 2002). It is important that students can reconcile randomness at the molecular level with the perceived efficiency of cellular systems as this lends meaning to more complex concepts, such as concentration gradients, protein specificity, or cell signalling cascades, and how these mechanisms may affect health and disease outcomes. However, these misconceptions are often robust and resistant to change; it requires that the student recognize that her understanding is incorrect, be provided with the tools to build a new mental model, and have the motivation in the first place to do so (Chi 2005; Modell et al. 2005).

Serious games are engaging spaces for active learning that may provide students with the motivation needed to trigger conceptual change. Cycles of productive negativity encourage schema building and are common in gaming environments—the player is challenged by a task and, under her current conception, she fails and must restructure her understanding in order to succeed (Mitgutsch & Alvarado 2012). This process corresponds with Chi (2005)'s conceptual change strategy involving misconception confrontation and schema building. Additionally, common game mechanics can be leveraged to promote productive negativity. For example, in their subversive game "*Afterland*", Mitgutsch and Alvarado (2012) employ inventory collection, health status, and enemy-evasion because they can predict how gamers will interact with and react to these mechanics. However, interaction with these leads to unexpected outcomes, resulting in productive negativity and a learning experience for the player. To generalize, serious games can apply commonly used mechanics in uncommon ways so that players behave predictably, increasing the chances for productive negativity—and conceptual change—to ensue.

Furthermore, game design mechanics and elements have potential to increase a student's willingness to participate in meaningful and intellectual play, thereby enhancing her understanding of target content and concepts (Squire 2011; Steinkuehler & Squire 2012; Gauthier et al. 2015). Much literature supports video games for learning (Gee 2007; Landers & Callan 2011; Squire 2006), but the empirical evidence can be contradictory. Recent meta-analyses reveal that serious games can increase learning, self-efficacy, and motivation in comparison to traditional learning or other non-gaming stimuli (Wouters et al. 2013; Sitzmann 2011; Clark et al. 2016) but fail to highlight the value-added effect of game design (Clark et al. 2016). The present publication strives to contribute to this area by investigating the value-added effect of a serious game in relation to a simulation application that employs similar interaction and visual design. This study is also, to the best of our knowledge, one of the first to implement conceptual change strategies—specifically, productive negativity—through game design to address misconceptions in undergraduate science.

### Research Objectives and Hypotheses

In this study, we compare and relate molecular misconceptions, game-play statistics, and player characteristics (e.g. level of education, gaming habits) among undergraduate biology students who engage with either a serious game or a 'control' interactive simulation. Specifically, we endeavoured to 1) facilitate conceptual change about molecular emergence through interactive media; 2) characterize how game design influences this phenomenon; and 3) explore how other factors such as perceived engagement with the app and player characteristics relate to interactions and, ultimately, learning.

We hypothesized that 1) serious game mechanics would help students achieve conceptual change about molecular emergence above and beyond standard education and an interactive simulation without gaming elements; 2) this conceptual change would be related to the quality of productively negative experience provoked by the game; 3) achievement in the game (game score) would be predictive of the number of misconceptions held by the student; 4) attitudes and engagement would be more positive for those exposed to the serious game over the control simulation; and 5) the serious game may be more effective for participants who play games on a regular basis.

## METHODS

### Participants

Participants were undergraduate students enrolled in first- (n = 292), second- (n = 209), and third- (n = 34) year biology at the University of Toronto Mississauga. In the first-year course, molecular concepts are not specifically covered, so these students represent 'novice' learners with high school-level education. The second-year course is where students are first introduced to molecular phenomena (e.g. vesicle formation, RNA translation), many of which appear in our game and simulation, making these second-year 'initiate' learners a suitable target audience for the apps. The third-year students delve deeper into molecular biology concepts, representing an advanced learner group with a special interest in this subject matter.

### Materials

#### *Stimuli: MolWorlds (Game) and MolSandbox (Control)*

*MolWorlds* is a simulation-based, platform-genre, adventure game designed and developed by the Science Visualization lab at the University of Toronto Mississauga. In the game, players travel through a molecular realm and experience cellular processes (e.g. vesicle formation, RNA translation) while manipulating properties of the simulated emergent system (e.g. through temperature, macromolecular crowding, and concentration) to reach their destination. 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. The game has 13 levels in the current prototype; a 3-level version was piloted in 2015 and is described by Gauthier & Jenkinson (2015).

The overall design of the game was based on the concepts of evidence-centred design (Mislevy & Haertel 2006) and the learning mechanics-game mechanics model (Arnab et al. 2015). To summarize, the following game mechanics were implemented to directly encourage conceptual change:

1. **Resource Management:** Players must search for and collect the items in their inventory and can only carry five molecules of each type at a time. Having invested effort into increasing their inventory, the player is more likely to release only one molecule at a time, thereby decreasing the chance of a quick binding event and increasing the likelihood of eliciting productive negativity. This idea was inspired by the concept of subversive game design (Mitgutsch & Weise 2011). Further, temperature and crowding are controlled by power-ups found in the game world, so the player must pick the most opportune times in which to employ these;
2. **Immersed 3rd-Person Character:** The character, controlled by the player, is physically hindered from reaching a checkpoint by the emergent forces at hand, which he can modify (concentration, crowding, temperature). This is intended both to instil accountability in the player's actions and increase motivation by engaging the player in a narrative;
3. **Sequential Level Progression:** The player can only progress once a level is successfully completed, thus optimizing the probability that correct system-modifying mechanics will be used (of course, the molecular world is random, so some chance exists that they will progress without conceptual change occurring);

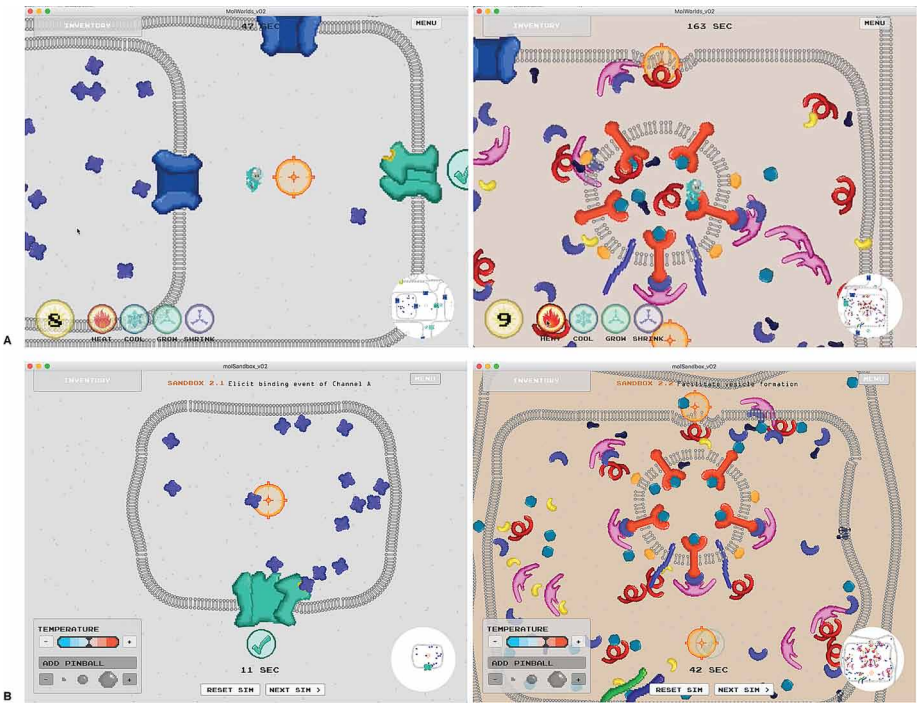
4. **Score and Feedback:** The more quickly a player moves through the level, the higher the score, which is reflected by a three-star system at the end of every level. This is intended to encourage repetition and, if level completion was due to random chance, another opportunity for conceptual change.

*MolSandbox* excludes these game mechanics. Figure 1 depicts screenshots from levels 6 and 7 of both stimuli to facilitate comparison. Although *MolSandbox* contains an inventory menu, items are automatically replenished for each simulation, thus removing the resource-management component. Additionally, temperature and crowding are adjusted with gauges without restrictions on usage (instead of power-ups contained in the game). The objective in each "sandbox" simulation mirrors the objective of the game, but without the immersed character. For example, in level 6 where a *MolWorlds*-player would have to move the character through a ligand-gated membrane channel to reach the checkpoint (Figure 1-A, left), a *MolSandbox*-user would simply have to elicit the same binding event by dropping items from their inventory (Figure 1-B, left). Users can progress through the app at will, skipping levels if they like. Lastly, while the time to objective completion is recorded, there is no associated score.

### *Demographics Questionnaire*

A web-based demographics questionnaire was administered at the pre-test (week 2 of the semester) that collected data on age, gender, biology courses completed, self-reported grade-point average (GPA), mobile gaming habits as well as gaming habits on other platforms (7-point scale ranging from never to everyday). At the post-test (week 11), students were additionally asked what grade they expected to achieve in the course and how engaged they were in their biology course.

**Figure 1. Screenshots of stimuli in 8-bit style. A) MolWorlds (game); Left: Level 6 in which the goal is to pass the character through the ligand-gated membrane channel to the checkpoint; Right: Level 7 which involves clathrin-coated vesicle formation to transport the character across the membrane. B) MolSandbox (control) representing the same level as those in MolWorlds.**



### *Molecular Concepts Adaptive Assessment*

In collaboration with Harvard Medical School, Center for Molecular and Cellular Dynamics, our lab developed and evaluated a *Molecular Concepts Adaptive Assessment*. The survey is a web-based, adaptive, multiple-choice test that assesses understanding of complex molecular motion, interactions, and systems. For example, one identified misconception was that molecules (e.g. extracellular ligands) have some sort of agency and objectives in that they actively seek out complementary receptors. The first survey question asks "True or False: An extracellular molecule tries to move toward a complementary receptor." If the student answers "True" the survey follows up with the question "Based on your previous answer and assuming there are several of the complementary receptors present, an extracellular molecule tries to move toward… [options A through D]" in order to gain a more nuanced understanding of their misconception. Proceeding true and false questions delve deeper into this concept of intent and directedness; for example, "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". Further, if the student is able to correctly identify random collisions as the mechanism of molecular motion, the survey includes questions geared toward identifying factors might affect the probability of binding events occurring. A maximum of 12 misconceptions are possible.

### *Attitudes and Engagement Questionnaire*

In order to gauge participants' perceived engagement with the stimuli, a subset of 10 statements from the *Instructional Materials Motivation Survey* (Loorbach et al. 2014) were selected and refined to apply to our interventions. Statements were rated on a 5-point Likert scale from strongly disagree to strongly agree. A few example statements include "the material in this app was more difficult to understand than I would like for it to be", "there was so much information that it was difficult to pick out important points", "the app looked dry and unappealing", "the app was not relevant to my needs because I knew it all already", and "the amount of repetition in this app caused me to get bored sometimes".

## Procedure

First- and second-year students participated during the Fall 2015 semester, whereas third-years participated in the Winter 2016 semester. The *Molecular Concepts Adaptive Assessment* survey and the demographics questionnaire were administered online near the beginning (week 2) and end (week 11) of the semester to characterize the typical evolution of students' misconceptions over time. Those who completed the pre-test survey were later contacted via email and invited to register for the game randomized controlled trial, held on campus in a 25-seat lab outfitted with iMacs during week 11, prior to completing their post-survey. Several sessions were conducted during this week and participants registered for the session that best fit with their schedule, leading to group sizes of 3-10 people per session (40 participants total). Participants were individually seated at computers spread throughout the lab to discourage communication during the study. Each computer displayed the study-hub webpage within a browser window. Upon providing informed consent, participants logged into this webpage using their student ID and password, at which point they were randomized to either the gaming group (exposed to *MolWorlds*) or the control interactive simulation group (exposed to *MolSandbox*). They first completed a few of demographic questions, after which the invigilators opened their assigned application. Participants used their app for a period of 30 minutes, while their cursor clicks and interactions within the applications were logged in a database using MySQL and their screens were recorded using QuickTime. Following the intervention, they returned to the study-hub webpage where they completed the post-test survey and the engagement questionnaire. All data analyses were performed in SPSS Statistics v.23 (IBM Corporation 2013).

## DATA ANALYSIS AND RESULTS

### Group Composition

In all, 526 students participated in this study. Of this, 486 completed both the pre- and post-test surveys without participating in the game intervention; this group—our "baseline" group—consisted of 277 first-, 196 second-, and 22 third-year students, with 357 females, 132 males, and 2 individuals with undisclosed gender. The final control-stimulus group (n = 20) consisted of 7 first-years, 7 second-years, and 6 third-years, with an average age of 18.85 years. The gaming stimulus group (n = 20) consisted of 8 first-years, 6 second-years, and 6 third-years, with a mean age of 19.40 years, which is not statistically different from that of the control, $t(38) = -1.39$, $p = .172$. The control group was comprised of 13 females and 7 males, while the game group consisted of 15 females and 5 males.

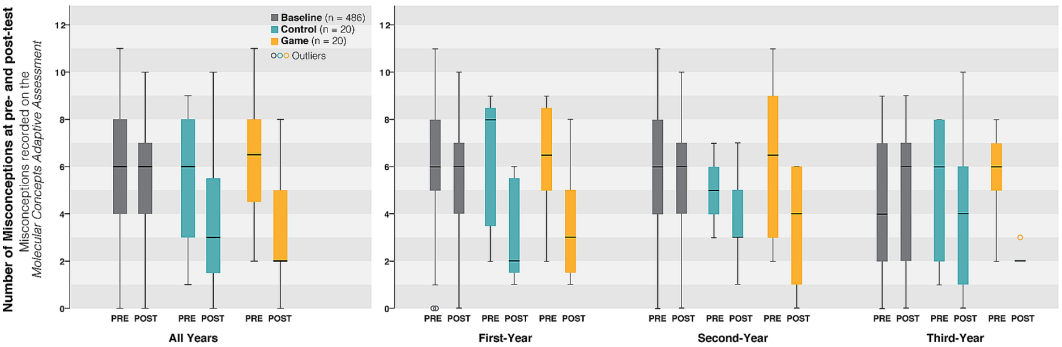
Several demographic characteristics were compared between gaming and control groups to ensure that any observed differences in learning could be attributed to the stimuli themselves. Mann-Whitney U tests comparing gaming habits revealed no significant difference in either mobile gaming habits (U = 186.50, Z = -0.35, p = .729) or platform/desktop gaming habits (U = 198.00, Z = -0.056, p = .955). Further, t-tests were used to compare continuous variables of self-reported GPA (t(1,33) = -1.12, p = .271) and expected grade (t(1,38) = 0.97, p = .339), also revealing no significant differences between groups. Therefore, we can suggest that our intervention groups had similar compositions.

### Molecular Misconceptions

#### *Change in Misconceptions from Pre-Test to Post-Test*

The *Molecular Concepts Adaptive Assessment* was marked for incorrect responses (i.e. misconceptions); as such, higher scores indicate negative outcomes. On the pre-test, the baseline group recorded a mean 5.87 misconceptions ($SD = 2.32$), the control group an average of 5.45 ($SD = 2.50$), and the game group 6.15 ($SD = 2.62$). On the post-test, the baseline group averaged 5.55 misconceptions ($SD = 2.33$), the control scored a mean of 3.75 misconceptions ($SD = 2.55$), and the game group averaged 3.10 ($SD = 2.17$). Therefore, the baseline group lost an average of 0.34 ($SD = 2.55$) misconceptions over the course of the semester, while the control and gaming interventions generated an average loss of 1.70 ($SD = 2.72$) and 3.05 ($SD = 3.20$) misconceptions respectively. Figure 2 illustrates pre- and post- misconceptions across stimuli groups and educational levels.

**Figure 2. Misconceptions held at the beginning (pre-test) and end (post-test) of the semester as recorded on the Molecular Concepts Adaptive Assessment, across intervention groups (baseline, control stimulus, or game stimulus) and by educational level (first-, second, or third-year biology)**



### *Effect of Stimulus and Educational Level on Misconceptions*

We performed a 3x3 repeated measures mixed model analysis (using the "unstructured: correlation metric" repeated covariance type to compensate for unequal sample sizes) to determine how educational level (first-, second-, or third-year biology) and intervention type (no intervention, simulation, or game) affected students' molecular misconceptions from pre-test to post-test. There was an overall significant effect of testing time on misconceptions ($F(1, 526) = 32.65$, $p<.001$) and, while educational level did not have an effect on the change in misconceptions from pre-test to post-test ($F(4, 526) = 0.95$, $p = .435$), the intervention type did ($F(4, 526) = 8.94$, $p<.001$). Further, there was no significant interaction effect between the testing time, stimulus, or educational level ($F(8, 526) = 0.43$, $p = .903$), meaning that individuals in different years of study, exposed to the same stimulus, changed in similar ways. Figure 3 depicts the estimated marginal means from the model across stimuli groups and educational levels.

*A priori* pairwise comparisons revealed that participants who were not exposed to any intervention (baseline group) retained significantly more misconceptions in comparison to those exposed to the control interactive simulation ($p = .007$, 95% CI[0.45, 2.82]) as well as those exposed to the serious game ($p<.001$, 95% CI[1.85, 4.23]), while adjusting for educational level. A trending difference was found between the simulation group and the gaming group ($p = .084$, 95% CI[-0.19, 2.99]), with gamers resolving a greater number of misconceptions.

## Gameplay/Usage Statistics (Game and Control Groups Only)

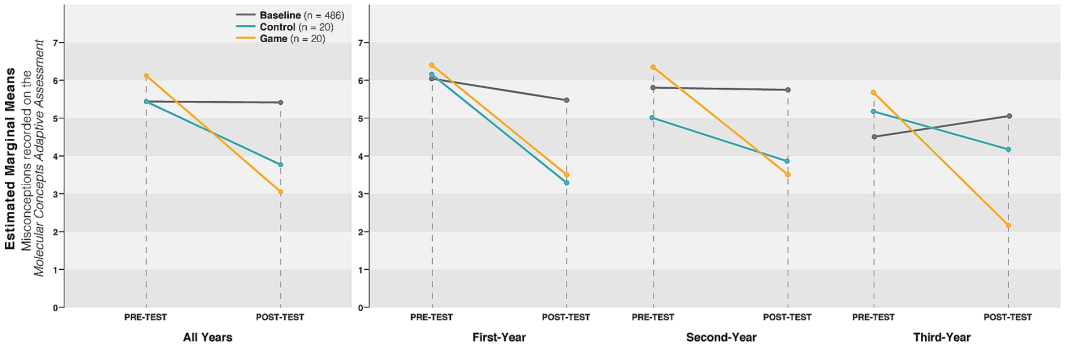
To gain a more nuanced understanding of the findings above, it is important to investigate how the presence of game design affected interactions in, and perceptions of, the app. Here, we analyse differences in click-stream data, qualitative game-flow events, and perceived engagement.

### *App Completion: Use of System-Modifying Mechanics*

Table 1 summarizes the raw gameplay statistics coalesced from the click-stream data. We employed *t*-tests to compare app-use statistics between the control and gaming group, using the Welch-Satterthwaite correction where Levene's test for equality of variance was significant.

During the 30 minutes that each intervention group was exposed to their respective stimulus, the control group was able to attempt ($t(24.40) = 8.47$, $p<.001$, 95% CI[16.15, 26.55]) and complete ($t(22.13) = 6.05$, $p<.001$, 95% CI[6.21, 12.69]) significantly more levels than the game group. As such, the game-players spent significantly more time on each attempted level than did simulation-users ($t(28.48) = 9.58$, $p<.001$, 95% CI[1.48, 2.25]). Further, control participants were able to progress

**Figure 3. Estimated marginal mean misconceptions (recorded on the Molecular Concepts Adaptive Assessment) from pre-test to post-test across intervention groups (baseline, control stimulus, or game stimulus) and by educational level (first-, second, or third-year biology), outputted from the mixed model**



to—and complete—significantly more unique levels (out of a total 13) than the gaming participants ($t(32.26) = 7.98$, $p<.001$, 95% CI[2.72, 4.58]).

No differences were observed in molecule-collection events ($t(38) = 1.34$, $p = .187$), whereas control participants partook in many more molecule-releasing events than the gaming participants ($t(38) = 4.93$, $p<.001$, 95% CI[11.73, 28.07]). They also engaged in more temperature- ($t(20.50) = 6.94$, $p<.001$, 95% CI[55.20, 102.60]) but slightly fewer crowding- ($t(22.28) = -1.96$, $p = .062$, 95% CI[-3.19, 0.09]) modifying events than gamers.

### *Instances of Productive Negativity and Demonstrations of Correct Conceptual Knowledge*

The 40 30-minute screencasts were watched and coded for instance of productive negativity and demonstrations of correct conceptual knowledge (Table 1). One half of the screencasts (10 control and 10 game) were coded by two raters in a fully crossed design. Only one of the raters (the "primary rater") coded the remaining twenty videos, after Hallgren (2012) in consideration of the time-consuming nature of the task. Raters were blinded to the participants' demographic information and learning outcomes on *the Molecular Concepts Adaptive Assessment*. To determine the trustworthiness of data from the primary rater, inter-rater reliability was calculated using the Intraclass Correlation Coefficient (ICC) with a two-way mixed model of absolute agreement type, where a value of .700 would indicate an acceptable level of agreement. An exceptional level was achieved for demonstrations of correct conceptual knowledge (ICC = .982, 95% CI [.942, .994]) and good levels were achieved for instances of productive negativity (ICC = .841, 95% CI [.603, .937]), confirming the reliability of the primary rater whose data were used in our analyses.

A demonstration of correct conceptual knowledge was identified as a series of actions wherein the user made appropriate adjustments to the simulation (i.e. in concentration, temperature, or crowding) in order to complete the objective at hand. For example, in the 9$^{th}$ level, the objective is to open a ligand-gated membrane channel (and, in *MolWorlds*, get the character to the other side); in this area, there also exists an enzyme that will degrade the ligand once released; to achieve the goal more efficiently, the user could balance reducing the concentration of the enzyme, increasing the concentration of the ligand (and possibly an inhibitor) and increasing the temperature. The preceding example would have been coded as three demonstrations of correct conceptual knowledge (two concentration, one temperature).

An instance of productive negativity was identified as a series of actions not indicative of a correct conception and that does not result in immediate success, but which then prompts a demonstration of correct conceptual knowledge. For example, the 6$^{th}$ level (Figure 1, left) also requires the opening of a ligand-gated channel but without the presence of other obstacles; under a misconception of molecular agency or directed motion, the user might release a single ligand, expecting it to bind immediately; when they see that it does not, they may then increase the concentration of the ligand in the environment to heighten the probability of a binding event. This example would have been coded as one instance of productive negativity, followed by one demonstration of correct conceptual knowledge once the concentration is increased.

*T*-tests revealed that the game elicited more instances of productive negativity than did the interactive simulation ($t(26.67) = 5.00$, $p<.001$, 95% CI[1.47, 3.53]), but the interactive simulation elicited more demonstrations of correct conceptual knowledge than the game ($t(38) = 3.17$, $p = .003$, 95% CI[2.53, 11.47]).

To test our second hypothesis, we calculated a productive negativity impact rate for each participant by dividing the number of demonstrations of correct conceptual knowledge by the number of productively negative events—in essence, the quality of the productively negative events. In the gaming group, each productively negative event was associated with a mean 2.47 ($SD = 1.16$) demonstrations of correct knowledge whereas the control app was associated with 9.85 ($SD = 8.27$), a significant difference ($t(19.75) = 3.96$, $p = .001$, 95% CI[3.49, 11.28]). The linear relationship of this rate to assessment outcomes is recorded in the section below on bivariate relationships.

## Attitudes and Engagement Questionnaire

The ten *IMMS* statements were negatively phrased; therefore, lower scores (toward the "disagree" end of the 5-point Likert scale) represent more positive attitudes. All items scored a median of 2 to 2.5 for both stimuli resulting in no difference between groups, except for the statement "the amount of repetition in this app caused me to get bored sometimes". For this statement, gamers scored a median of 2 (disagree), whereas simulation-users rated a median 3.5 (between neutral and agree), resulting in a significant difference when tested with a Mann-Whitney U test ($U = 125.50$, $Z = -2.11$, $p = .035$).

## Bivariate Relationships

### *Relationship between Gameplay/Usage Statistics and Misconceptions*

We used two-tailed Pearson correlations to determine if a relationship existed between post-test misconceptions and the quantitative and qualitative usage statistics listed in Table 1.

In the control group, we found a negative correlation between post-test misconceptions and breadth of app completion ($r = -0.66$, $p = .003$). That is, as the number of unique completed levels increased, misconceptions decreased. In the gaming group, post-test misconceptions held negative trending correlations with attempted levels ($r = -0.40$, $p = .084$), completed levels ($r = -0.42$, $p = .065$), and a significant negative correlation with breadth of completion ($r = -0.45$, $p = .049$).

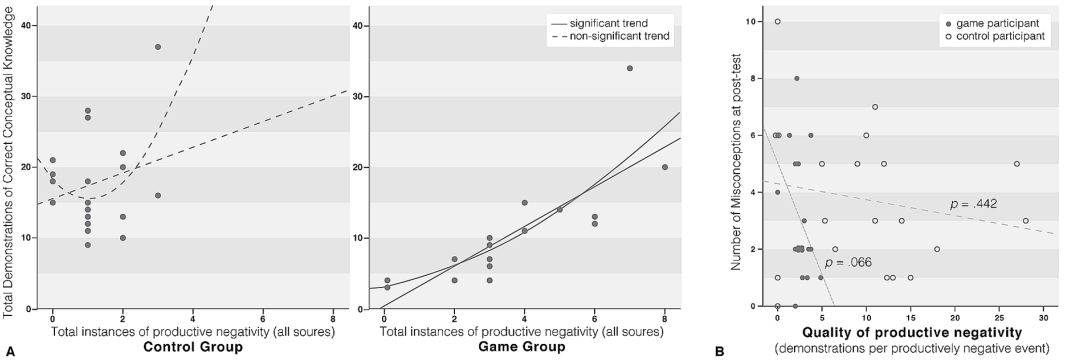
With respect to the control group, no correlation existed between post-test misconceptions and the quality of productively negative experiences ($r = -0.18$, $p = .442$). Conversely, in the game group, we found a trending negative correlation ($r = -0.42$, $p = .066$) (Figure 4-B). In other words, as productively negative events resulted in more demonstrations of correct conceptual knowledge, misconceptions decreased for game-players. This suggests that as more negativity is experienced during gameplay, the stronger the student grasps the concepts and applies their new knowledge to new challenges in increasing amounts. As such, we should see a curvilinear relationship between total instances of productive negativity and demonstrations of correct conceptual knowledge (Figure 4-A, Game Group).

To analyse this, we performed a quadratic regression analyses using productive negativity to predict demonstrations of correct conceptual knowledge in each stimulus group. We found no curvilinear relationship in the control group ($R^2 = 0.177$, $F(2, 19) = 1.83$, $p = .190$) nor did we find a

**Table 1. Gameplay/usage statistics over a 30-minute period**

| | *MolSandbox* (Control) | | | *MolWorlds* (Game) | | |
|---|---|---|---|---|---|---|
| | **Min** | **Max** | **Mean (SD)** | **Min** | **Max** | **Mean (SD)** |
| Levels attempted | 13 | 53 | 32.60 (10.34) | 7 | 22 | 11.25 (4.01) |
| Levels completed | 8 | 29 | 17.95 (6.71) | 6 | 14 | 8.50 (1.93) |
| Minutes per attempted level | 0.57 | 2.31 | 1.03 (0.40) | 1.36 | 4.29 | 2.91 (0.78) |
| Unique completed (out of 13) | 8 | 13 | 11.15 (1.72) | 6 | 9 | 7.50 (1.10) |
| Molecule collection events | 38 | 237 | 122.20 (59.85) | 36 | 224 | 99.10 (48.33) |
| Molecule release events | 10 | 73 | 40.70 (15.04) | 8 | 41 | 20.80 (9.97) |
| Temperature modification | 22 | 217 | 93.40 (49.90) | 3 | 38 | 14.50 (9.93) |
| Crowding modification | 3 | 7 | 3.45 (0.99) | 2 | 15 | 5.00 (3.38) |
| Productive negativity | 0 | 3 | 1.15 (0.93) | 0 | 8 | 3.65 (2.03) |
| Correct conceptual knowledge | 9 | 37 | 17.65 (6.87) | 3 | 34 | 10.65 (7.10) |
| Quality of productive negativity | 0 | 28 | 9.86 (8.27) | 0 | 4.86 | 2.47 (1.16) |

**Figure 4. A) Bivariate linear and quadratic relationships between total demonstrations of correct conceptual knowledge for the gaming group (n = 20) and the control group (n = 20). B) Linear relationship between number of post-test misconceptions held by students and the quality of their productively negative experiences.**




linear relationship ($R^2 = 0.061$, $F(1, 19) = 1.16$, $p = .294$). In the gaming group, we found significant quadratic ($R^2 = 0.673$, $F(2, 19) = 17.52$, $p < .001$) and linear ($R^2 = 0.646$, $F(2, 19) = 32.85$, $p < .001$) relationships. These relationships are illustrated in Figure 4-A. However, a closer inspection of the difference between linear and quadratic models reveals that the quadratic adjustment did not have a significant impact on the overall fit of the data above and beyond the linear model ($\Delta R^2 = 0.027$, $\Delta F(1, 17) = 1.42$, $p = .249$). Regardless, this slight curve in the line is reflected in the trending relationship seen between the quality of productive negativity and post-test misconceptions in the gaming group (Figure 4-B), supporting, in part, our second hypothesis.

In and of themselves, demonstrations of correct conceptual knowledge (control: $r = 0.02$, $p = .792$; game: $r = 0.12$, $p = .456$) and instances of productive negativity (control: $r = 0.02$, $p = .792$; game: $r = 0.12$, $p = .456$) held no relationship with misconceptions. Furthermore, game score did not correlate with misconceptions ($r = -0.16$, $p = .499$) but held a strong relationship with the quality of productive negativity ($r = 0.60$, $p = .005$).

### *Relationship between Self-Reported Engagement and Misconceptions*

Spearman correlations were performed with our ordinal engagement items to test for a relationship between self-reported engagement and misconceptions. In the control group, a positive correlation existed between perceived difficulty of the material and post-test misconceptions ($r = 0.48$, $p = .034$). In the game group, there was a positive trending correlation between misconceptions and a difficulty picking out important details due to too much information ($r = 0.43$, $p = .061$).

### *Relationship between Gaming Habits, Usage Statistics, and Misconceptions*

Misconception improvement from pre- to post-test did not correlate with mobile gaming habits (control: $r = -0.01$, $p = .988$; game: $r = -0.34$, $p = .146$) or with "traditional" gaming habits (i.e. on platforms other than mobile) (control: $r = -0.02$, $p = .939$; game: $r = -0.08$, $p = .724$) in for either stimulus group.

Mobile gaming habits did not correlate with any usage statistics in either intervention group. In the control group, "traditional" gaming habits (i.e. on platforms other than mobile) correlated positively with the breadth of app completion ($r = 0.47$, $p = .035$) and demonstrations of correct conceptual knowledge ($r = 0.45$, $p = .046$). In the gaming group, traditional gaming habits correlated positively with game score ($r = 0.60$, $p = .006$), levels completed ($r = 0.47$, $p = .039$), molecule collection ($r = 0.72$, $p<.001$) and release ($r = 0.59$, $p = .007$) events, demonstrations of correct conceptual knowledge ($r = 0.61$, $p = .005$), instances of productive negativity ($r = 0.58$, $p = .007$), and the

quality of productive negativity ($r = 0.48$, $p = .031$). It should be noted that mobile gaming habits and traditional gaming habits did not, between themselves, correlate in either group.

## DISCUSSION

### Major Findings

This research empirically demonstrates that a serious game successfully facilitated conceptual change in undergraduate students understanding of the emergent nature of molecular environments, beyond traditional education methods alone. In addition, the aim of the RCT was to characterize the specific contribution of game mechanics (namely resource management, an immersed 3rd-person character, score, and sequential level progression) by comparing it to an interactive simulation without these elements. The game mechanics increased the effectiveness of the simulation to a degree that trended toward significance. Educational level (first-, second-, third-year biology) did not influence the effectiveness of either stimulus.

Analyses of gameplay videos and interaction data suggest that the reduced misconceptions in the gaming group may be attributed to the quality of productively negative experiences elicited in *MolWorlds*. As intended, the game encouraged greater numbers of productively negative experiences through to its mechanics. Therefore, players were 1) more likely to exhibit behaviour reflective of their misconceptions (e.g. releasing a single molecule under the conception of directed motion), thus confronting their misconception when their progress was hindered; and 2) forced to re-evaluate their mental model if they wanted to progress, both physically in the level and through the game. Thus, as the quality of productively negative events (i.e. the number of demonstrations of correct conceptual generated by each instance of productive negativity) increased, the number of post-test misconceptions decreased—though only trendingly—suggesting that the revised actions of the gamers may be representative of their actual understanding. We also observed that as level attempts, successful completions, and overall game progression increased, misconception on the post-test decreased, further supporting this argument.

The lack of game mechanics in *MolSandbox* allowed control users greater room for experimentation (as evidenced by a larger number of attempted levels, completed levels, breadth of completion, temperature modifications, and demonstration of correct conceptual knowledge), but a relationship with misconceptions only exists with the number of unique completed levels (breadth of completion). This may be explained by examining time constraints, since, in the allotted 30 minutes, control-users had plenty of time to review the entire app and complete the simple levels (e.g. a ligand-gated channel binding event) multiple times, leading to the high attempt and completion counts; this may also be why control participants were more likely to rate their app as being boring due to repetition. However, more advanced and complex simulations were often passed over when the outcome (e.g. vesicle formation or translation) was not immediately achieved; those participants who made the effort to work through and experiment in these more complex "sandboxes"—thus achieving greater breadth of app completion—met with better outcomes on the post-test. On the other hand, game players were less likely to revisit very simple, early levels as more effort was placed in progressing through the game, leading to stronger correlations between level completions and a reduction in misconceptions. Further, no relationship is seen between misconceptions and the quality of productively negative events in the control group, suggesting that the demonstrations of correct conceptual knowledge resulting from these events are not necessarily reflective of learning outcomes and may be due to random experimentation.

We see no relationship between misconceptions and raw counts of productively negative experience or demonstrations of correct conceptual knowledge in either group. A relationship to misconceptions was not expected with raw instances of productive negativity, since the presence of a negative event may or may not reflect a lack of understanding. However, one might expect to see a

negative correlation between demonstrations of correct conceptual knowledge and misconceptions; we do not, and yet, we see a trending relationship between quality of productive negativity (demonstrations per negative event) and misconceptions. This could suggest that negativity is a critical ingredient for meaningful interactions to take place.

Game score did not reflect learning outcomes. This was likely due to the 30-minute timeframe allotted for gameplay, which did not allow enough time for game completion; in fact, the highest level completed by any gamer was 9 out of 13. We can surmise that players were focused on trying to complete the game rather than re-attempting levels to achieve three stars. Future research should extend the timeframe to allow gamers to 1) finish the game and 2) repeat levels to achieve a higher score, thus producing scores that reflect their conceptual understanding more accurately.

An individual's gaming habits did not, ultimately, mitigate the effectiveness of either stimulus to facilitate conceptual change, at least in the short time-frame allotted for play in this study. However, it is worth noting that gaming habits (mobile gaming excluded) had a significant impact on what participants could accomplish in the game, but to a lesser extent in the control simulation. Though gaming habits did not correlate with learning outcomes themselves, they should be considered as a possible confounder in future research that employs more complex models with larger sample sizes. Further, it would be interesting to investigate whether mobile gaming habits would relate to gameplay data should the game be placed on a mobile platform.

### Limitations

First and foremost, the relatively small sample sizes of our two intervention groups may have been responsible for several of our trending results, prohibiting us from performing analytical models with more than a couple independent variables of interest, and limiting us to exploring bivariate relationships. Secondly, as mentioned above, the timeframe was insufficient for those assigned to the gaming condition to finish the game and reattempt levels, thus obscuring a potential relationship between in-game performance (i.e. game score) and misconceptions. Lastly, though we could qualitatively sense by observation in the lab that the game generated a higher level of engagement than the simulation, our survey failed to reflect this. Only one of 10 items proved significantly better for the evaluation of *MolWorlds*. In future work, we should consider the use of the full 32-item *IMMS* survey as well as an analysis of facial expressions, which could prove a better measure of engagement as well as support recorded instances of productive negativity, which, at this point, are subject to the coders interpretation.

### Practical Implications

Based on these primary findings and limitations, we can suggest the following practical implications for future serious game design and assessment: 1) game mechanics can be implemented to facilitate productively negative game-flow loops, encouraging the player to demonstrate their knowledge in increasing amounts through interaction in the game; 2) adding game design may limit accessibility of content, especially to people who do not play games on a regular basis; and 3) a predetermined, limited play duration may obstruct relationships between game score and learning outcomes, so a longer play-time should be considered if assessing learning through game performance.

## CONCLUSION

Most undergraduate biology students fail to comprehend how random mechanisms at the molecular level might lead to perceptually efficient cellular processes, misconceiving these events as directed in nature. This randomized control trial documents conceptual change via a serious game and interactive simulation in a population whose misconceptions otherwise remain robust to change. We observed that game mechanics, such as resource management, an immersed character, and sequential level

progression, helped to elicit conceptual change beyond the interactive simulation by encouraging a greater number of productively negative events, compelling the player to re-evaluate their understanding and make appropriate adjustments to the game world in order to progress, regardless of level of education and gaming habits.

## ACKNOWLEDGMENT

This research was supported by the Social Sciences and Humanities Research Council of Canada and by the University of Toronto's Information Technology Innovation Fund. We would also like to thank Drs. Espie, Arts, and Anderson from the Department of Biology, University of Toronto Mississauga, for allowing us access to their students and for facilitating our research.

## REFERENCES

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

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

Chi, M. T. H., & Roscoe, R. D. 2002. The Processes and Challenges of Conceptual Change. In M. Limon & L. Mason (Eds.), Reconsidering Conceptual Change. Issues in theory and Practice (pp. 3-27). Netherlands: Kluwer Academic Publishers. doi:10.1007/0-306-47637-1_1

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. Available at http://www.ncbi.nlm.nih.gov/pubmed/22050726 doi:10.1111/j.1551-6709.2011.01207.x PMID:22050726

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. Retrieved from http://rer.sagepub.com/cgi/doi/10.3102/0034654315582065 doi:10.3102/0034654315582065 PMID:26937054

IBM Corporation, 2013. SPSS Statistics.

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*(2), 227–233. doi:10.1187/cbe.07-08-0063 PMID:18519614

Gauthier, A., Corrin, M., & Jenkinson, J. (2015). Exploring the influence of game design on learning and voluntary use in an online vascular anatomy study aid. *Computers & Education*, *87*(September), 24–34. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0360131515000950 doi:10.1016/j.compedu.2015.03.017

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.

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

Hallgren, K. a. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23–34. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/\n

Landers, R. N., & Callan, R. C. (2011). In A. Oikonomou & L. C. Jain (Eds.), (pp. 399–423). Ma: Serious Games and Edutainment Applications. Retrieved from http://www.springerlink.com/index/10.1007/978-1-4471-2161-9

Loorbach, N., Peters, O., Karreman, J., & Steehouder, M. (2014). Validation of the Instructional Materials Motivation Survey (IMMS) in a self-directed instructional setting aimed at working with technology. *British Journal of Educational Technology*, *46*(1), 204–218. doi:10.1111/bjet.12138

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., & Alvarado, N. 2012. Purposeful by Design? A Serious Game Design Assessment Framework. *Proceedings of the International Conference on the Foundations of Digital Games FDG '12*, New York, New York, USA.

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

Modell, H., Michael, J., & Wenderoth, M. P. (2005). Helping the Learner To Learn: The Role of Uncovering Misconceptions. *The American Biology Teacher*, *67*(1), 20–26. doi:10.1662/0002-7685(2005)067[0020:HTLTLT]2.0.CO;2

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE Life Sciences Education*, *9*(4), 435–440. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995761&tool=pmcentrez&rendertype=abstract doi:10.1187/cbe.10-01-0001 PMID:21123690

Sitzmann, T. (2011). A Meta-Analytic Examination of the Instructional Effectiveness of Computer-Based Simulation Games. *Personnel Psychology*, *64*(2), 489–528. Retrieved from http://doi.wiley.com/10.1111/j.1744-6570.2011.01190.x doi:10.1111/j.1744-6570.2011.01190.x

Squire, K. (2006). From Content to Context : Videogames as Designed Experience. *Educational Researcher*, *35*(8), 19–29. doi:10.3102/0013189X035008019

Squire, K. (2011). *Video games and Learning: Teaching and Participatory Culture in the Digital Age*. New York, New York, USA: Teachers College Press.

Steinkuehler, C., & Squire, K. (2012). Videogames and Learning. In K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences*. New York: Cambridge University Press.

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. Retrieved from http://doi.apa.org/getdoi.cfm?doi==10.1037/a0031311 doi:10.1037/a0031311

*Andrea Gauthier is a biomedical communicator from Toronto, Canada, and is a PhD candidate at the Institute of Medical Sciences (University of Toronto) in the Science Visualization Lab (www.sciencevis.ca). Her doctoral work investigates how digital serious games can influence students' understanding of mechanistic and emergent systems in undergraduate health and life sciences education, with a focus on the specific role of design over medium.*

*Jodie Jenkinson is an Assistant Professor of Biomedical Communications at University of Toronto (Canada) and principle investigator of the Science Visualization Lab (www.sciencevis.ca). She holds a PhD in Education (University of Toronto), specializing in cognition and learning, with a focus on technology in education. Her research focuses on the role that visual representations play in learning. This includes investigation along various lines of inquiry including the efficacy of visual media within different learning contexts, the design of visual representations for optimal impact, and the development of standards of visual communication in the scientific visualization community.*