Serious Game Facilitates Conceptual Change About Molecular Emergence Through Productive Negativity (RCT)

Andrea Gauthier, Jodie Jenkinson
University of Toronto, Toronto, Canada
andrea.gauthier@utoronto.ca
j.jenkinson@utoronto.ca

Keywords: Conceptual change, molecular biology, serious game, productive negativity, randomized controlled trial

Abstract: Throughout their undergraduate careers, biology students struggle to reconcile how randomness at the molecular level governs cellular systems, often misconceiving these emergent systems as mechanistic in nature. A serious game has potential to facilitate conceptual change by enabling instances of productive negativity—a player may attempt a challenge and fail under their current misconception, and then must re-evaluate their understanding in order to succeed. We designed a serious game, *MolWorlds*, under this premise and tested its efficiency against an interactive simulation that used the same graphics and simulation system as the game but lacked gaming elements such as score, sequential levelling structure, resource management, and a 3rd-person character immersed in the environment. 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.

We performed a 3x3 repeated measures linear mixed model 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. While educational level did not have an effect on misconceptions, the intervention type did (p<.001). *A priori* pairwise comparisons revealed that participants who were not exposed to any intervention retained significantly more misconceptions in comparison to those exposed to the interactive simulation (p=.007) as well as those exposed to the game (p<.001), while adjusting for educational level. A trending difference was found between the simulation group and the gaming group (p=.084), with gamers resolving more misconceptions. Analysis of gameplay data revealed that gamers 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).

1. Introduction

1.1 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 players (Momsen et al. 2010; Chi 2005; Chi et al. 2012; Garvin-doxas & Klymkowsky 2008; Chi & Roscoe 2002). It is important that students are able to 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). Game design mechanics and elements have potential to increase a student's willingness to participate in meaningful and intellectual play, thereby enhancing his or 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, especially in undergraduate STEM education, to which we hope to contribute with this publication.

1.2 Research objectives and hypotheses

In this study, we endeavoured to 1) facilitate conceptual change about molecular emergence through a serious game; and 2) characterize how game design influences this phenomenon by comparing the game to a similar interactive simulation without gaming elements.

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) that this conceptual change would be related to the quality of productively negative experience provoked by the game; and 3) achievement in the game (game score) would be predictive of the number of misconceptions held by the student.

2. Methods

2.1 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 students 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.

2.2 Materials

2.2.1 Stimuli: MolWorlds *(game) and* MolSandbox *(control)*

MolWorlds is a simulation-based, platform-genre, role-playing 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) in order 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. 2014). Specific game mechanics were implemented to directly encourage conceptual change:

1. **Resource management**: Players have to search for and collect the items in their inventory and can only carry five molecules of each type at a time. Having put effort into doing this, 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 has the ability to 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 from both stimuli to facilitate comparison. Though *MolSandbox* still has 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.

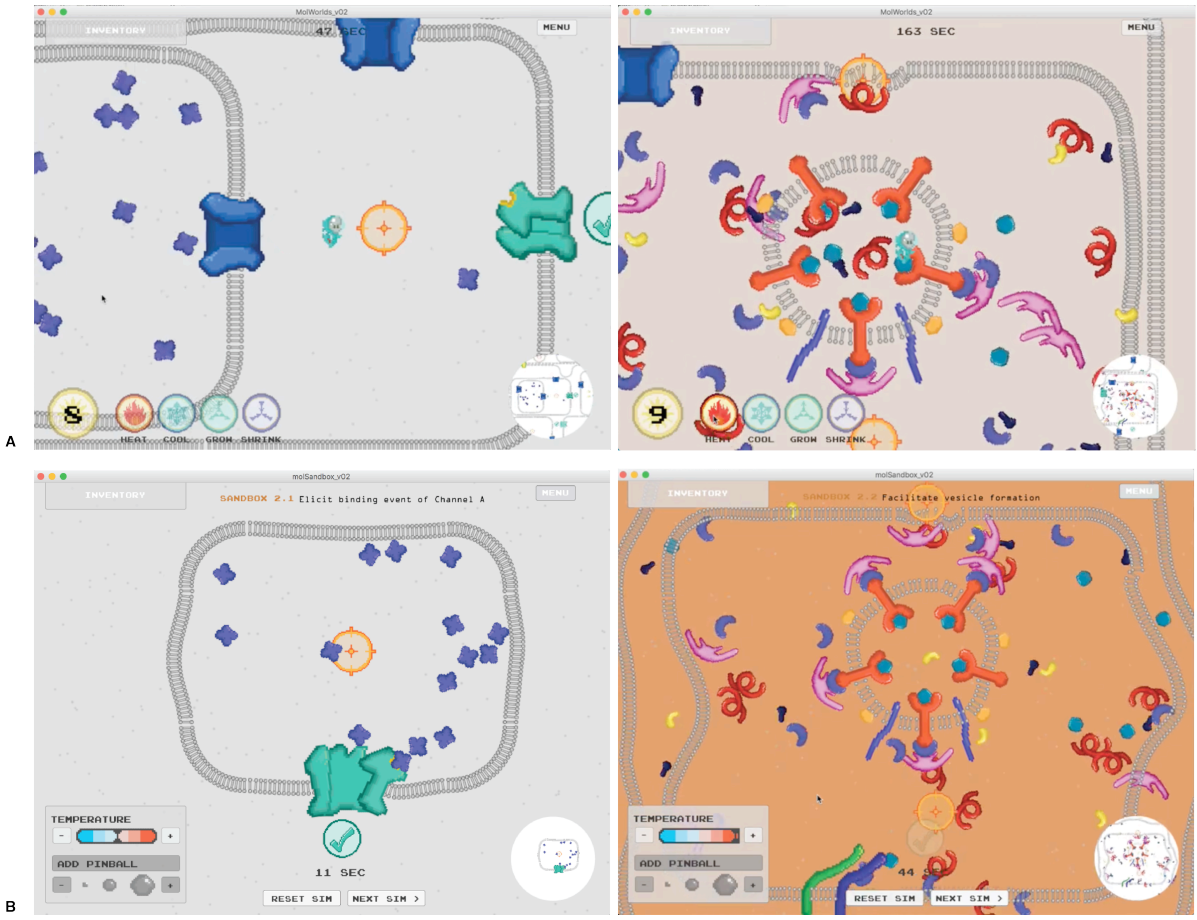

Figure 1: Screenshots of stimuli in 8-bit style. **A) *MolWorlds*** (game); Left: Level 6 in which the goal is to pass the character through the ligand-gated membrane channel to the checkpoint; Right: Level 7 which involves clathrin-coated vesicle formation to transport the character across the membrane.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

2.2.2 Demographics questionnaire

A web-based demographics questionnaire was administered at the pre-test (week 2) 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.

2.2.3 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 questions about what factors might affect the probability of binding events occurring. A maximum of 12 misconceptions are possible.

2.2.4 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".

2.3 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 in order to characterize the typical evolution of students' misconceptions over time. Those who completed the pre-test survey were later emailed and invited to register for the game randomized controlled trial, held in a computer lab on campus during week 11, prior to completing their post-survey. During this session, these 40 participants were randomized to engage either with the game, *MolWorlds*, or the interactive simulation, *MolSandbox*, for a period of 30 minutes. While they played, their cursor clicks and interactions within the applications were logged in a database using MySQL and their screens were recorded using QuickTime. After the intervention, they went on to complete the post-test survey and the engagement questionnaire. All data analyses were performed in SPSS Statistics v.23 (IBM Corporation 2013).

3. Data analysis and results

3.1 Group composition

In all, 526 students participated in this study. Of this, 486 completed both the pre- and post-test surveys and received no 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.

3.2 Molecular misconceptions

3.2.1 Change in misconceptions from pre-test to post-test

The *Molecular Concepts Adaptive Assessment* was marked for incorrect responses (i.e. misconceptions); therefore, 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.

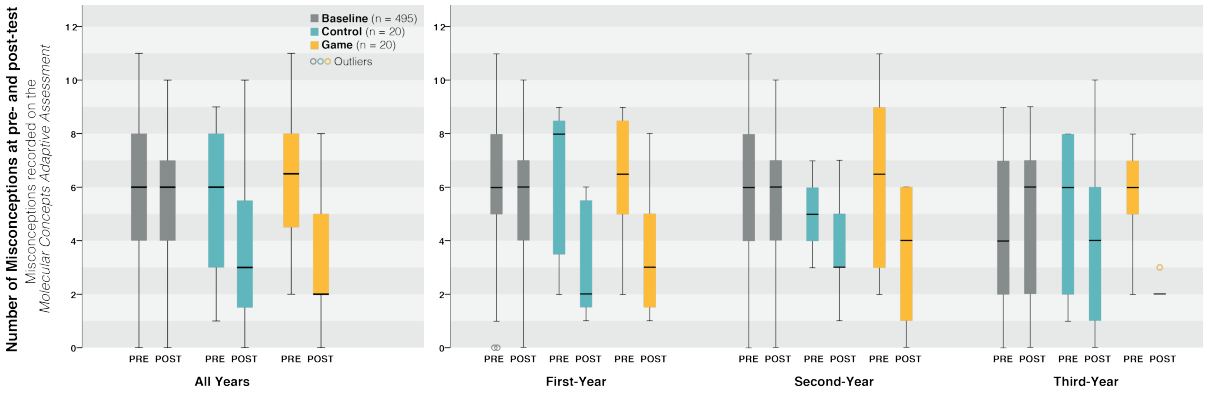

Figure 2: Misconceptions held at the beginning (pre-test) and end (post-test) of the semester as recorded on the *Molecular Concepts Adaptive Assessment*, across intervention groups (baseline, control stimulus, or game stimulus) and by educational level (first-, second, or third-year biology).

3.2.2 Affect 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 from different years but who were 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.

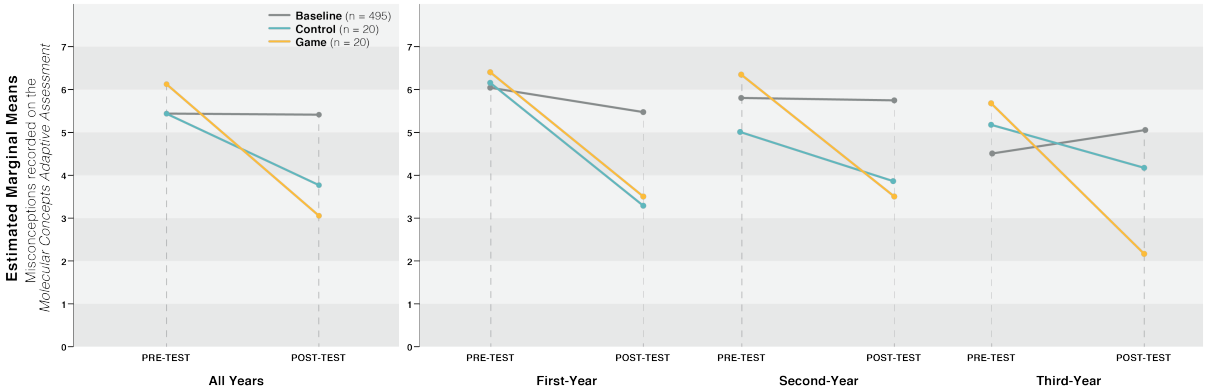

Figure 3: Estimated marginal mean misconceptions (recorded on the *Molecular Concepts Adaptive Assessment*) from pre-test to post-test across intervention groups (baseline, control stimulus, or game stimulus) and by educational level (first-, second, or third-year biology), outputted from the mixed model.

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 more misconceptions.

3.3 Gameplay/usage statistics (game and control groups only)

3.3.1 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 the intervention groups were exposed to their 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 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 ($t(38)=1.34$, $p=.187$), while 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.

3.3.2 Instances of productive negativity and demonstrations of correct conceptual knowledge

All 40 screencasts were watched and hand-coded for instance of productive negativity and demonstrations of correct conceptual knowledge (Table 1) by a coder who was blinded to the participants' demographic information and responses to *the Molecular Concepts Adaptive Assessment*. 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 9th 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 6th 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 reveal 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]).

In order 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 section 3.4.1.

3.3.3 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, with the exception of 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 ($U=125.50$, $Z=-2.11$, $p=.035$).

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)

3.4 Bivariate relationships

3.4.1 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 usage statistics listed in Table 1. For 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).

For the control group, no correlation existed between post-test misconceptions and the quality of productively negative experiences (r=-0.18, p=.442) but, in the game group, we found a trending negative correlation (r=-0.42, p=.066). In other words, as productively negative events resulted in more demonstrations of correct conceptual knowledge, misconceptions went down for game-players. Game score also 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).

3.4.2 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).

4. Discussion

4.1 Major findings

This research empirically shows that a serious game successfully facilitated conceptual change about the emergent nature of molecular environments in undergraduate students, beyond standard education. In addition, the RCT attempted to highlight 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 by an amount 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 lower misconceptions in the gaming group may be due to the quality of productively negative experiences elicited in *MolWorlds*. As intended by the design (refer to section 2.2.1), 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 is hindered; and 2) forced to re-evaluate their mental model if they want 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, furthering this point.

The lack of game mechanics 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 due to the fact that, 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 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.

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.

4.2 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 subjective to the coder.

5. 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 that compelled the player to re-evaluate their understanding and make appropriate adjustments to the game world in order to progress.

6. Acknowledgements

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.

7. References

Arnab, S. et al., 2014. Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology*, 46(2), pp.391–411.

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

Chi, M.T.H. et al., 2012. Misconceived causal explanations for emergent processes. *Cognitive science*, 36(1), pp.1–61.

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, 3-27*. Netherlands: Kluwer Academic Publishers.

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, pp.227–233.

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), pp.24–34.

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*. Steinkjer, Norway: Academic Conferences and Publishing International Limited, pp. 656–663.

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

Halverson, R. & Owen, E., 2014. Game-Based Assessment: An integrated model for capturing evidence of learning in play. *[Special Issue: Game-Based Learning] International Journal of Learning Technology*, 9(2), pp.111–138.

IBM Corporation, 2013. SPSS Statistics.

Landers, R.N. & Callan, R.C., 2011. Serious Games and Edutainment Applications M. Ma, A. Oikonomou, & L. C. Jain, eds. , pp.399–423.

Loorbach, N. et al., 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), pp.204–218.

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

Mitgutsch, K. & Alvarado, N., 2012. Purposeful by Design ? A Serious Game Design Assessment Framework. In *FDG 2012, FDG '12 Proceedings of the International Conference on the Foundations of Digital Games*. 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), pp.20–26.

Momsen, J.L. et al., 2010. Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE life sciences education*, 9(4), pp.435–40.

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

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.