Game and simulation stimulate conceptual change about molecular emergence in different ways, with potential cultural implications

Supplementary Materials

Table of Contents

1. Materials

1.1. Demographics questionnaire

1.1.1 Pre-intervention demographics

Please remember that all data collected is completely anonymous and is not associated with your name or your student number in any way.

1. Age: ____
2. Gender identity:
    a. Female
    b. Male
    c. Other
    d. Prefer not to answer
3. How long have you lived in Canada?
    a. 0-1 year
    b. 2-5 years
    c. 5-10 years
    d. More than 10 years
    e. All my life
4. Is English your first language?
    a. No
    b. Yes
5. In which course are you currently enrolled?
    a. BIO152
    b. BIO206
    c. BIO372
    d. None of the above
6. What was your overall Grade Point Average (or equivalent) in your last year of study?
    a. 0.7 or less (0-52%)
    b. 1.0 (53-56%)
    c. 1.3 (57-59%)
    d. 1.7 (60-62%)
    e. 2.0 (63-66%)
    f. 2.3 (67-69%)
    g. 2.7 (70-72%)
    h. 3.0 (73-76%)
    i. 3.3 (77-79%)
    j. 3.7 (80-84%)
    k. 4.0 (85-100%)
    l. I prefer not to say
7. Do you own a smartphone?
    a. No
    b. Yes
8. On average, how often do you play digital games on your smartphone?
    a. Never
    b. Very seldom
    c. About once a month
    d. Several times a month
    e. About once a week
    f. Several times a week
    g. Everyday
9. On average, how often do you play digital games/video games on devices other than your phone or tablet? (Desktop/Laptop, Web, Xbox, PlayStation, Nintendo, etc.)
    a. Never
    b. Very seldom

c. About once a month
d. Several times a month
e. About once a week
f. Several times a week
g. Everyday

1.1.2 Post-intervention demographics

Please remember that all data collected is completely anonymous and is not associated with your name or your student number in any way.

1. What grade do you expect to achieve in this BIO course? (i.e. either BIO 152, 206, or 372)
    a. 0-50%
    b. 50-60%
    c. 60-70%
    d. 70-80%
    e. 80-90%
    f. 90-100%
    g. I prefer not to say
2. Thinking about this biology course (BIO152, BIO206, BIO372), please rate the following statements from "Strongly disagree" to "Strongly agree". (Questions are drawn from the Burch Engagement Survey for Students [1].)
    a. I am interested in the material I learn in this course. [emotional engagement]
        i. Strongly disagree
        ii. Disagree
        iii. Neither agree nor disagree
        iv. Agree
        v. Strongly agree
    b. I devote a lot of energy toward this course. [physical engagement]
        i. Strongly disagree
        ii. Disagree
        iii. Neither agree nor disagree
        iv. Agree
        v. Strongly agree
    c. When I am in the classroom for this course, my mind is focused on class discussion and activities. [cognitive engagement inside the class]
        i. Strongly disagree
        ii. Disagree
        iii. Neither agree nor disagree
        iv. Agree
        v. Strongly agree
    d. When I am reading or studying material related to this course, my mind is focused on class discussion and activities. [cognitive engagement outside the class]
        i. Strongly disagree
        ii. Disagree
        iii. Neither agree nor disagree
        iv. Agree
        v. Strongly agree

1.2. Engagement questionnaire based on the IMMS

The following questions were answered on a scale from 1 (strongly disagree) to 5 (strongly agree). They were modified from the Instructional Materials Motivations Survey (IMMS, Loorbach et al., 2014).

1. The material covered in the app was more difficult to understand than I would like for it to be.
2. The app had so much information that it was hard to pick out and remember the important points.
3. The app is so abstract that it was hard to keep my attention on it.
4. The app looks dry and unappealing.
5. The exercises in this app were too difficult.
6. The amount of repetition in this app caused me to get bored sometimes.
7. The app was not relevant to my needs because I already knew most of it.
8. The style of writing in the app is boring.
9. There are so many words in each exercise that it is irritating.
10. I could not really understand quite a bit of the material in this app.
11. Completing levels successfully was important to me.
12. I enjoyed the app so much that I would like to learn more molecular Biology concepts from it.
13. I can relate the content/concepts of this app to things I'm learning about in Biology.
14. It felt good to successfully complete levels in this app.
15. It was a pleasure to engage with this app and I would do so again if given the opportunity.

1.3. Telemetric events tracked during gameplay

Each interaction in *MolWorlds* and *MolSandbox* were recorded in the following format:

userID, level, x, y, event, detail1, detail2, detail3, timestamp

… where "userID" is the randomly-generated 10-digit alpha-numeric ID of the participant, "level" is the level in the game/simulation in which the event took place, "x" and "y" are the coordinates of the player (*MolWorlds*) or camera/viewport (*MolSandbox*), "event" is the type of interaction (see list directly below), "detailX" encompasses any additional information that may be pertinent, and "timestamp" is the time at which the user made the interaction event.

The following interactions and details were recorded:

1. Login
2. Level started
3. Level completed; details: time taken to completion, score, # of stars
4. Molecule info accessed; details: type of molecule
5. Collect molecule; detail: type of molecule
6. Release molecules; detail: type of molecule, number of molecules
7. Heat; detail: current temperature
8. Chill; detail: current temperature
9. Grow pinball/character
10. Shrink pinball/character
11. Collect power-up (game only)

2. Reducing dimensions

We hypothesized that students' personal attributes might influence their conceptual change and may have interaction effects with our intervention stimuli. We collected the following variables as possible confounders (in addition to gender and NES/NNES status already included in the model): GPA, expected course grade, emotional/physical/cognitive course engagement, mobile gaming habits, traditional gaming habits. However, this would be too many variables to be included in a single model with our sample size.

2.1. Categorical principal components analysis

We first reduced the number of ordinal variables (four Likert-scale course engagement questions, mobile gaming habits, and traditional gaming habits) using a categorical principal components analysis, using the 'grouping' discretization method and the variable principal normalization method, whilst specifying to output two components. The categorical principal components analysis resulted in two dimensions in 19 iterations: 1) "course engagement" with Cronbach's $\alpha = 0.62$, total Eigenvalue = 2.07, and 34.50% total variance accounted for, and 2) "gaming habits" with Cronbach's $\alpha = 0.45$, total Eigenvalue = 1.60, and 26.65% total variance accounted for (refer to Table 1 through Table 3). The object scores were saved as the new variable to be used for further analyses.

Table 1. Model summary. Total Cronbach's alpha based on the total Eigenvalue.

Dimension	Cronbach's Alpha	Total variance accounted for (Eigenvalue)	% of variance accounted for
1 – course engagement	0.620	2.070	34.499
2 – gaming habits	0.450	1.599	26.650
Total	0.873	3.669	61.149

Table 2. Iteration history

	Variance accounted for		Loss		
Iteration #	Total	Increase	Total	Centroid Coordinates	Restriction of centroid to vector coordinates
0 *	3.555	0.000003	8.445	8.226	0.218
19 **	3.669	0.000010	8.331	8.146	0.186

* Iteration 0 displays the statistics of the solution with all variables, except variables with optimal scaling level Multiple Nominal, treated as numerical.

** The iteration process stopped because the convergence test value was reached.

Table 3. Component loadings: Variable principal normalization.

Variable	Dimension 1 Course engagement	Dimension 2 Gaming habits
Emotional engagement	0.702	0.121
Physical engagement	0.648	-0.408
Cognitive engagement (in class)	0.770	0.014
Cognitive engagement (out of class)	0.743	0.149
Mobile gaming habits	-0.033	0.839
Traditional/platform gaming habits	0.103	0.832

2.2. Factor analysis

Secondly, we reduced the number of continuous variables to be included our model (GPA, expected grade) by performing a factor analysis, with heterogeneous (two-step) correlations and matrices, principal axis extraction type, and varimax rotation, based on Eigen values greater than 1. When a participant was missing data (e.g. a participant could choose not to disclose their GPA), the means were computed for that participant. Factor scores were saved using the regression method. The factor analysis resulted in one additional variable that we term as "academic achievement", where GPA and expected grade loaded together at 0.62, with a total Eigen value of 1.38 and accounted for 37.96% of the extracted sums of squared loading variance. Further details of these transformations are displayed in Table 4 and Table 5.

Please note that, since only one factor was extracted, varimax factor rotation was not performed.

Table 4. Communalities and factor loadings

	Communalities		Factor Matrix
Variable	**Initial**	**Extraction**	**Factor 1***
GPA	0.145	0.380	0.616
Expected grade	0.145	0.380	0.616

* Extraction method: Principal Axis Factoring; 1 factor extracted, 8 iterations required.

Table 5. Total variance explained

	Initial Eigenvalues			Extraction sums of squared loadings		
Factor	**Total**	**% of variance**	**Cumulative %**	**Total**	**% of variance**	**Cumulative %**
1	1.381	69.028	69.028	0.759	37.963	37.963
2	0.619	30.972	100.000			

* Extraction method: Principal Axis Factoring

3. Participant composition

3.1. Descriptive statistics

Table 6. Descriptive statistics of demographic characteristics between stimulus groups, for continuous and ordinal variables.

	Baseline			*MolSandbox*			*MolWorlds*		
Variable	**Mean**	**SD**	**Min, max**	**Mean**	**SD**	**Min, max**	**Mean**	**SD**	**Min, max**
Course engagement	-0.075	0.995	-2.83, 2.08	0.051	1.192	-4.48, 1.99	0.197	0.791	-1.91, 1.68
— Emotional engagement [a]	4	1	1, 5	4	1	1, 5	4	1	1, 5
— Physical engagement [a]	4	1	1, 5	4	2	1, 5	4	1	2, 5
— Cognitive engagement (in class) [a]	3	2	1, 5	4	2	1, 5	4	1	1, 5
— Cognitive engagement (out of class) [a]	4	1	1, 5	4	1	1, 5	4	1	2, 5
Gaming habits	-0.038	1.059	-2.07, 2.74	0.146	0.901	-1.62, 2.15	-0.021	0.910	-1.53, 2.30
— Mobile gaming hab. [b]	1	3	0, 6	1	3	0, 6	1	2	0, 6
— Traditional gaming hab.[b]	1	3	0, 6	1	1	0, 5	1	2	0, 6
Student Achievement	-0.006	0.721	-1.54, 1.67	0.050	0.685	-1.48, 1.19	-0.030	0.862	-1.33, 1.67
— GPA [c]	7.885	2.094	3, 11	7.707	2.148	2, 11	7.425	2.581	2, 11
— Expected grade (%) [d]	3.943	0.912	2, 6	4.122	0.843	2, 6	4.056	0.959	2, 6
Science literacy scores	8.86	1.375	3, 10	8.98	0.924	7, 10	8.64	1.165	6, 10
Bioliteracy scores	5.18	1.722	1, 9	5.12	1.941	1, 9	5.57	1.810	0, 9
Pre-intervention molecular misconceptions	5.59	2.256	0, 10	5.67	2.386	0, 10	6.12	2.340	1, 10
Ave. confidence in answer misconception questions	58.867	18.256	1.44, 94.89	53.024	18.864	11.45, 95.38	59.059	20.315	0.00, 100.00

a. Ordinal variable on Likert scale from 1-5 (strongly disagree to strongly agree). Medians and interquartile range replace means and standard deviations.

b. Ordinal variable on scale from 0-6 (0 – Never, 1 – Very seldom, 2 – About once a month, 3 – Several times a month, 4 – About once a week, 5 – Several times a week, 6 - Everyday). Medians and interquartile range replace means and standard deviations.

c. Batched scale variable (max GPA = 4.0): 1 = 0.7 or less (0-52%), 2 = 1.0 (53-56%), 3 = 1.3 (57-59%), 4 = 1.7 (60-62%), 5 = 2.0 (63-66%), 6 = 2.3 (67-69%), 7 = 2.7 (70-72%), 8 = 3.0 (73-76%), 9 = 3.3 (77-79%), 10 = 3.7 (80-84%), 11 = 4.0 (85-100%)

d. Batched scale variable: 1 = 0-50%, 2 = 50-60%, 3 = 60-70%, 4 = 70-80%, 5 = 80-90%, 6 = 90-100%

Table 7. Descriptive statistics of demographic characteristics between stimulus groups, for categorical variables.

	Gender		**English as a first language**		**Canada-born**	
Stimulus group	**Male**	**Female**	**No**	**Yes**	**No**	**Yes**
Baseline	44	93	51	87	61	77
MolSandbox (Interactive sim.)	10	32	18	24	24	18
MolWorlds (Serious game)	10	32	13	29	21	21

Our sample of second-year students proved to be homogenous across our three intervention groups. To assess the homogeneity of group composition prior to intervention exposure, we performed a MANOVA using intervention/stimulus type as our independent variable (baseline group, SIM group, and GBL group) and

multiple continuous dependent variables including: gaming habits, academic achievement, course engagement, bioliteracy pre-test scores, molecular misconceptions pre-test scores, as well as average confidence scores. Science literacy scores were not included since nearly everyone scored highly on this test (M = 8.84/10, SD = 1.26), resulting in highly skewed data (Shapiro-Wilk = .807, $p < .001$). Regardless of unequal sample sizes between experimental groups, Box's test for equality of covariance matrices was not significant for the MANOVA ($M = 52.34$, $F(42, 40181.46) = 1.17$, $p = .201$), nor were Levene's tests for equality of error variances in any dependent variable (Table 8). There was no relationship between stimulus group (baseline, interactive simulation, or serious game) and course engagement ($F(2, 219) = 1.26$, $R^2_{adjust} = .002$, $p = .212$, *partial* $\eta^2 = 0.01$), gaming habits ($F(2, 219) = 0.55$, $R^2_{adjust} = -.004$, $p = .578$, *partial* $\eta^2 < 0.01$), academic achievement ($F(2, 219) = 0.13$, $R^2_{adjust} = -.008$, $p = .875$, *partial* $\eta^2 < 0.01$), bio-literacy ($F(2, 219) = 0.90$, $R^2_{adjust} > -.001$, $p = .409$, *partial* $\eta^2 < 0.01$), pre-test misconceptions ($F(2, 219) = 0.85$, $R^2_{adjust} > -.001$, $p = .429$, *partial* $\eta^2 < 0.01$), or average confidence in answering the pre-test misconceptions survey ($F(2, 219) = 1.68$, $R^2_{adjust} = .006$, $p = .189$, *partial* $\eta^2 = 0.02$).

Table 8. Levene's test of equality of error variances

	F	df1	df2	*p*
Course engagement	1.812	2	219	.166
Gaming habits	1.553	2	219	.214
Student Achievement	1.376	2	219	.255
Bioliteracy scores	0.162	2	219	.850
Pre-intervention molecular misconceptions	0.079	2	219	.924
Ave. confidence in answer misconception questions	0.075	2	219	.927

Secondly, we performed multinomial logistic regression analysis using our binary variables of gender and native language to predict stimulus group, to ensure that these categories were evenly distributed between groups. The test revealed no effect of gender (-2 Log likelihood = 32.63, $\chi^2(2) = 1.78$, $p = .410$) or of native language (-2 Log likelihood = 32.14, $\chi^2(2) = 1.29$, $p = .524$), meaning that males and females, native and non-native English-speakers were evenly distributed between stimulus groups (final model: -2 Log likelihood = 401.98, $\chi^2(430) = 1.78$, $p = .830$) (further details in Table 9).

Table 9. Parameter estimates for multinomial regression

Stimulus group [a]		B	SE	Wald	df	*p*	Exp(B)	95% CI Low	95% CI High
Baseline	Intercept	1.396	.374	13.933	1	.000			
	Gender = female	-.407	.406	1.002	1	.317	.666	.300	1.476
	Gender = male	0[b]	.	.	0	.	.	.	.
	Native language = not English	.236	.379	.388	1	.533	1.266	.603	2.662
	Native language = English	0[b]	.	.	0	.	.	.	.
Interactive simulation	Intercept	-.203	.483	.177	1	.674			
	Gender = female	.018	.514	.001	1	.972	1.018	.372	2.788
	Gender = male	0[b]	.	.	0	.	.	.	.
	Native language = not English	.515	.457	1.271	1	.260	1.674	.684	4.099
	Native language = English	0[b]	.	.	0	.	.	.	.

a. The reference category is: Game stimulus
b. This parameter is set to zero because it is redundant

4. Conceptual change ANCOVA (baseline, control, game)

4.1. Post-hoc details

Table 10. Post-hoc pairwise comparisons between stimulus-native-language subgroups using estimated marginal means. Lower scores represent more favourable outcomes.

					95% confidence intervals	
Group (I)	Group (J)	Mean diff (I-J)	SE	p*	Low	High
Baseline-NNES	Baseline-NES	-0.401	0.493	1.000	-1.863	1.060
	Control-NNES	2.116	0.802	.127	-0.261	4.493
	Control-NES	0.008	0.784	1.000	-2.315	2.332
	Game-NNES	0.530	0.885	1.000	-2.091	3.150
	Game-NES	1.956	0.760	.150	-0.295	4.207
Baseline-NES	Control-NNES	2.518*	0.755	.015	0.281	4.755
	Control-NES	0.410	0.738	1.000	-1.778	2.598
	Game-NNES	0.931	0.840	.991	-1.557	3.419
	Game-NES	2.357*	0.711	.016	0.250	4.464
Control-NNES	Control-NES	-2.108	0.944	.333	-4.904	0.688
	Game-NNES	-1.587	1.041	.874	-4.671	1.498
	Game-NES	-0.160	0.944	1.000	-2.957	2.636
Control-NES	Game-NNES	0.521	1.035	1.000	-2.544	3.586
	Game-NES	1.948	0.932	.440	-0.815	4.710
Game-NNES	Game-NES	1.426	1.016	.929	-1.583	4.436

Covariates appearing in the model are evaluated at the following values: Course Engagement = -.0024, Gaming habits = .0056, Academic achievement = -.0075727.

*Sidak adjustments made for multiple comparisons

4.2. Effects of attrition at the delayed follow-up

Table 11. Repeated measured ANOVA investigating the effects of stimulus intervention (baseline, interactive simulation, or serious game), English as a native language, and whether or not the participant was retained at the delayed follow-up time point.

	df	***F***	***p***	**partial η^2**	**Obs. power**
Change	1	19.674	< .001 *	.086	.993
Change * stimulus	2	4.083	.018 *	.037	.720
Change * language	1	0.290	.591	.001	.083
Change * retained	**1**	**0.206**	**.650**	**.001**	**.074**
Change * stimulus * language	2	4.002	.020 *	.037	.711
Change * stimulus * retained	**2**	**0.890**	**.412**	**.008**	**.202**
Change * language * retained	**1**	**0.438**	**.509**	**.002**	**.101**
Change * stimulus * language * retained	**2**	**0.275**	**.760**	**.003**	**.093**

Type III sum of squares was used. Significant factors are highlighted (*).

5. Quantitative interaction data: correlations with gaming habits

Table 12. Spearman correlations between gaming habits (reduced variable combining mobile and traditional gaming habits) and various interaction statistics in both the gaming group and the interactive simulation group.

	***MolSandbox* (Control)**		***MolWorlds* (Game)**	
Interaction	***r***	***p***	***r***	***p***
Levels started	.045	.777	.437	.004 *
Levels completed	.435	.004 *	.382	.012 *
Unique levels completed	.414	.006 *	.268	.086
Collect molecules	-.045	.777	.403	.008 *
Release molecules	.143	.368	.326	.035 *
Temperature mods.	.091	.567	.334	.031 *
Crowding mods.	.112	.481	.209	.184
Item info accessed	-.195	.215	.347	.024 *
Powerups used (temp. + crowd.)	-	-	.420	.006 *
Total game score	-	-	.404	.008 *

6. Qualitative gameplay coding process and coding scheme

Game and interactive simulation videos were coded in an alternating fashion to achieve consistency across conditions. Codes were produced both deductively and inductively. Based on our findings from our previous trial [3], three types of demonstrations of correct conceptual knowledge and five types of productive negativity were defined prior analysing this dataset. However, we allowed for the flexibility to identify new codes; the coding scheme was finalized after an analysis of 20 videos (10 game and 10 control) and this first set of videos was coded a second time to ensure that the analysis conformed to the final scheme. The coding scheme can be viewed in the subsections below. All videos were coded by a primary coder who was blinded to the demographic characteristics and pre/post-test results of the participants. To ensure reliability of the primary coder's analysis—and due to the time-consuming nature of this task—a secondary coder coded 25% of the videos, whilst blinded to the assessment of the primary coder [4]. The secondary coder was given the coding scheme, as well as written examples of what constituted each code, and was trained on four randomly selected videos (2 game and 2 control). The secondary coder then proceeded to code a random selection of 22 remaining videos (11 game and 11 control; random selections were made by using the online random number generator at stattrek.com). Codes were summarized by total counts of each code type per participant; interrater reliability was assessed on these statistics using the intra-class correlation coefficient (ICC), where a value of 0.75 or greater would be deemed acceptable (results provided in body of article).

To better understand the coding scheme described below, we recommend our publication on the design of *MolWorlds* and *MolSandbox* [3], which summarizes the game flow and mechanics using the Activity Theory Model of Serious Games [5].

6.1. Demonstration of correct conceptual knowledge

A demonstration of correct conceptual knowledge (DCCK) is identified as a series of actions wherein the user made appropriate adjustments to the interactive simulation (i.e. in concentration, temperature, or crowding) to aid them in completing the objective at hand.

For all sub-categories, **do not code if the user is prompted by a tutorial**. If the user returns to the same area and performs the proper interactions again, in this case you may code it as a DCCK the 2nd, 3rd, etc. time around.

6.1.1 Concentration

Applicable to both game and interactive simulation conditions.

Concentration can be adjusted by releasing molecules from the inventory (increase concentration) and collecting molecules from the environment (decrease concentration).

To earn a DCCK by increasing concentration, the user should release **many** molecules of that type from their inventory (i.e. more than needed for the desired event to occur).

Table 13. Examples where **releasing** molecules (i.e. increasing concentration) would and would not be coded as a demonstration of correct conceptual knowledge in "concentration".

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
There is a ligand-gated membrane channel that needs to be opened. Only one ligand is required to bind for this to happen.	Both apps	Releasing 2+ ligands Releasing a vesicle containing transporters that pump the ligands into the appropriate area; even though it is a single vesicle, the presence of the vesicle itself increases the concentration of other ligands	Releasing only 1 ligand; it is the exact number needed for the event to occur
There is a ligand-gated channel that requires 2 ligands.	Both apps	Releasing 3+ ligands If there are already 2 ligands in the area, releasing any number of more ligands (e.g. 1 or 2) because now local concentration is greater than the exact number required	Releasing 1 or 2 ligands when none others of the same type are present in the area
There are 5 empty cargo receptors, each requiring a single cargo molecule.	Both apps	Releasing 5/5 cargo if 4 or fewer receptors are empty (maybe some bound earlier) Releasing 5/5 cargo but there are already loose/un-bound cargo in the area; concentration of cargo in the area is now greater than 5	Releasing 5/5 cargo when none other are in area; 5 are needed total so concentration is exact. If there are loose cargo in the area that are then (re-)collected and (re-)released on the dropzone to put in closer proximity to the receptors; the concentration has not changed
Level 11 (Box 2.6), the user is tasked to keep the pinball alive for as long as possible	Simulation only	They release 10/10 de-ubiquitination enzymes before or after the pinball is inserted	They release 10/10 de-ubiquitination enzymes and either 1) do not release the pinball at all, or 2) releases them after the pinball is degraded and do not re-add a pinball

Note that in the gaming condition, the player can only carry a maximum of 5 molecules of any type at a time until level 10 (where they can carry 10 at a time after collecting the inventory expansion pack); the player carries the inventory from level to level. In the simulation condition, there is no limit and the inventory is emptied and refreshed on every level.

A player can reduce molecular concentrations by collecting molecules in the environment. In the game, the player uses the character to collect molecules by pressing space bar while coming in contact with the molecule(s); in the control, the user collects molecules by scrubbing the mouse over them. We only code collection events when they have a direct effect on the targeted cellular process. In certain cases, we need to judge whether the user is intentionally collecting the items.

Table 14. Examples where **collecting** molecules (i.e. decreasing concentration) would and would not be coded as a demonstration of correct conceptual knowledge in “concentration”.

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
There is a ligand-gated membrane channel that needs to be opened with ligand A; in another region, there are a bunch of ligand A bustling around	Game only	-	The player collects ligand A thereby reducing the concentration in that area. This would NOT be coded because the action of decreasing concentration here is not relevant to the opening of the channel… rather, once they release the ligands, that will constitute “increase concentration” instead
There is a ligand-gated membrane channel that stays closed when one or two inhibitors are bound to it	Both apps	The user collects the inhibitors in the region The user releases the transporter vesicle from their inventory which pumps the inhibitor across the membrane, away from the channel	-
In the 9th level (W2Z4), the objective is to open a ligand-gated membrane channel (Channel C). In this area, there are also enzymes that degrade Ligand C	Both apps	If the player had either read about the enzyme or if they had seen the effects first-hand, and then collected some enzymes	Before reading about the enzyme (e.g. the user can click for info) or before seeing the actual effects of the enzyme from experience (e.g. they released ligand C and it was degraded), collecting some of these enzymes would NOT be coded since it is assumed that the player is collecting just for the sake of collecting, without an appreciation of its effects

6.1.2 Crowding

Applicable to both game and interactive simulation conditions.

Crowding pertains to changes in size to the character with power-ups (game condition) or the pinball with a gauge (simulation condition).

The user uses the Grow function to increase the size of either the character (game condition) or the pinball (simulation condition) to directly benefit their goal completion by increasing crowding.

Table 15. Examples where **increasing** the **character/pinball size** (i.e. increase crowding) would and would not be coded as a demonstration of correct conceptual knowledge in “crowding”.

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
To help collect molecules	Game only	To catch a certain molecule more easily, the player might enlarge the character size Player can collect multiple different molecules quickly just by standing still in a crowded environment, holding space, and using grow	The player is not collecting molecules while grow is in use.
	Simulation only	Inserting a large pinball into a tiny area may crowd the space enough to make it much easier to catch smaller molecules (does not happen frequently, though)	The player is not collecting molecules.
To make area more crowded	Game only	With a large character size, the player can position the character to block molecules into an area, which may help increase binding chances	The intention of blocking is not obvious (player can position character)
	Simulation only	A control-user might insert a large pinball to block other molecules into a particular area. Sometimes, we see the user toggling the pinball size as desired molecules move in and out of the area	The pinball is enlarged in an area where crowding is not beneficial.

The user uses the Shrink function to decrease the size of either the character (game condition) or the pinball (simulation condition) to directly benefit their goal completion by decreasing crowding.

Table 16. Examples where **decreasing** the **character/pinball size** (i.e. decrease crowding) would and would not be coded as a demonstration of correct conceptual knowledge in "crowding".

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
To make navigation or movement easier	Game only	In a very crowded area, the player shrinks the character size so that he can slip more easily between molecules	If they use shrink to pass between membranes where there is a small space, this does NOT count, since this is never required in the levels and it's not deep thinking/reflection on molecular mechanisms… i.e. "There's a small space, therefore I'll make myself small"
To make area less crowded	Simulation only	Since toggling between large and tiny pinballs is quick and easy, a tiny pinball may be switched to allow other molecules to pass easily around it	The pinball is shrunk in an area where decreased crowding is not beneficial.
To avoid degradation	Both apps	In level 11 (a.k.a. W2Z6 and box 2.6), there are ubiquitination enzymes and proteasomes. Shrinking the character or the pinball reduces collisions, resulting in a longer chance of survival	-

6.1.3 Temperature

Applicable to both game and interactive simulation conditions.

Temperature can be increased and decreased through power-ups (game) or a gauge (control). Temperature can be normal, 3 increasingly cold temperatures, or 3 increasingly hot temperatures. Power-ups run out after 10 seconds, whilst the gauges in *MolSandbox* must be manually changed.

Only distinct changes in temperature are coded separately, as exemplified in Table 17.

Table 17. Examples of "distinct" temperature changes.

Example context	Applies to	How many DCCKs?
The player chills/heats once, power-up runs out, then player chills/heats again	Game only	If these actions were productive (examples below), then this would be coded as 2 DCCKs for chill/heat.
The player chills/heats 3x rapidly in a row, making the environment very cold/hot	Game only	Even though this is 3 power-ups, it only counts as 1 DCCK for either chill or heat if it is productive.
The player cranks the temperature to max. heat or max. chill all at once	Simulation only	Similarly, it only counts as 1 DCCK if productive
Heat/Chill is on for medium strength for a while… it is then increased to max strength	Simulation only	This would be coded as 2 DCCKs, if they are both productive. If only one of these temperatures are productive (e.g. they only release the proper molecules whilst on temp #2), then it is only 1 DCCK

Chilling (Table 18)and heating (Table 19) of the environment must aid in task completion. See examples below.

Table 18. Examples where **decreasing the temperature** would and would not be coded as a demonstration of correct conceptual knowledge in "temperature".

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
To help collect molecules	Both apps	Chilling slows down the rate of molecular movement and the player collects some items.	The player does not actually collect anything (or visibly try to collect anything) while chilled.
To make navigation of character easier	Game only	In the midst of navigating (or just before starting to move) the player chills the environment and passes through.	The character does not attempt to move to another location.
To slow molecular processes	Both apps	Inhibitors are keeping channel D closed; when there are only a few molecules remaining, chilling may prevent them from binding long enough for it to stay open a while. Chilling may slow down collisions with enzymes, thus preventing key ligands from degrading. Chilling will prolong the life of the pinball and character in the ubiquitination level because collisions are reduced.	There is no benefit to chilling (collisions are not harmful).

Table 19. Examples where **increasing the temperature** would and would not be coded as a demonstration of correct conceptual knowledge in "temperature".

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
To help collect molecules	Game only	It may be counter-intuitive but if the player stays still and holds down the space bar and cranks the heat, it's an efficient way of collecting because the molecules in the area will collide with the character quickly, especially if they use a Grow power-up with it.	It is not evident that the player is holding space bar to collect.
	Simulation only	-	Heating doesn't make as much sense for the control since it just makes it more difficult for the user to collect with the cursor
Speeding up molecular processes	Both apps	This is the most common use of heating. If the user has placed molecules in the correct areas, then heating (more often than not) will speed up the interactions. Note that it may also lead to productive negativity... see Section 6.2.3	Heating does not noticeably aid in achieving a goal (whether it be navigation, collection, slowing inhibition, etc.). You will often see, in both conditions, the user crank up the heat just to see if things happen without having made the correct molecular transpositions. If stuff does happen then it might be coded, but often it doesn't if other actions weren't taken first.

6.2. Productive negativity

An instance of productive negativity (PN) is identified as a series of actions not indicative of a correct conception and that does not result in immediate success; i.e. some sort of failure, frustration, confusion or other negativity happens. However, this negativity then prompts a demonstration of correct conceptual knowledge, or a complete do-over of the level in which they perform differently.

Note: We do not code for "just negativity" without a production response, since it is only with a productive response that we can assume that the user perceived the event in a negative way, or that they noticed that something was wrong.

6.2.1 Resource retentiveness

Applicable to both game and interactive simulation conditions.

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 (due to random motion), and followed this with a productive modification (Table 20).

Table 20. Examples of **resource retentiveness** as a productively negative experience.

Example of negativity	Applies to	WOULD be coded if…	WOULD NOT be coded if…
There are 3 inactive cargo receptors each, requiring 1 cargo molecule before vesicle formation can initiate	Both apps	Player releases only 3 cargo molecules from the inventory (instead of several more to increase the probability of a binding event occurring); the participant waits around a while, then follows with a DCCK, such as collecting and releasing more cargo, or heating, or crowding the area.	They release 3/3 cargo, they wait (negative), they follow with a DCCK… this is instead "Simulation only", since they are not technically being retentive (it is all they have in their inventory).
There are 5 empty cargo receptors, each requiring 1 cargo molecule before vesicle formation can initiate.	Both apps	Player releases 5/10 cargo (or any number where total inventory held is greater than 5 and yet they release only 5), waits around (negative), and follows up with a DCCK.	Player releases 5/5 cargo, waits (negative), then does a DCCK. Not coded because 5 is all that they carry. Would be a PN by "simulation-only" instead.
In level 10 (W2Z5, box 2.5), 3 tRNA are needed to produce one cargo. Five cargo are needed to form a vesicle. Therefore 15 tRNA are needed total.	Game app	After level 9, players can now carry 10 of each molecule at a time. Therefore, they would have to make two trips for tRNA. So, if they release 10/10 tRNA, went immediately to get more, then release 5/10, this would be coded as "exact" ... a DCCK must follow.	The total number of tRNA is less than 15.
	Simulation app	Releasing 15/30 tRNA, waits, releases more.	

6.2.2 Difficult resource collection

Applicable to both game and interactive simulation conditions.

This is coded 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. 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. Unlike using the character to collect molecules in *MolWorlds*, collecting molecules in *MolSandbox* involves clicking and scrubbing the cursor over the molecules.

Table 21. Examples of **difficult resource collection** as a productively negative experience.

Example of negativity	Applies to	WOULD be coded if…	WOULD NOT be coded if…
In areas in which there are low concentrations of desired molecules, the player pursues a single, randomly-moving molecule with some difficulty (either with cursor or character)	Game only	The player then increases the size of the character using power-ups. They may also increase the heat and run into a crowded space to collect upon collision.	-
	Both apps	The player chills the environment in response to difficult collection.	It does not appear that the user was having trouble collecting molecules, which then leads to them reducing the temperature. We would only code for a "chill" DCCK

6.2.3 Resources lost due to overheating

Applicable to both game and interactive simulation conditions.

If the user keeps the heat high for a long time, then the membrane becomes loose and molecules will have enough energy to escape.

Table 22. Examples losing **resources due to overheating** as a productively negative experience.

Example of negativity	Applies to	WOULD be coded if…	WOULD NOT be coded if…
The user keeps the heat high for a long time; many molecules escape the membrane	Both apps	The user increases concentration again in response. The user restarts the level.	Chilling or decreasing the temperature; because chilling does not help get to the goal… it just prevents the same negativity from happening twice.

This negativity source often occurs alongside "Detrimental release or collection" (Section 6.2.6) since collecting a bunch of resources into the inventory can drastically reduce local concentrations and enhance the

effect when more resources are lost due to overheating... we must choose one or the other. Choose the one closest to the productive response.

6.2.4 Navigation and reaching checkpoint

Applicable to **GAME** ONLY.

In the game, the player must navigate through crowded environments and physically reach a checkpoint at the end of each level. This may lead to some negativity through difficult navigation, reaching the checkpoint (even if they successfully initiated a cellular process, they still may not successfully reach the checkpoint), or a combination of difficult navigation and reaching the checkpoint.

Table 23. Examples of **navigation and reaching checkpoints** as a productively negative experience.

Example of negativity	Applies to	WOULD be coded if…	WOULD NOT be coded if…
In crowded areas with a lot of molecules buzzing about the player may have trouble getting to where they need to go (negative)	Game only	They chill the environment to make passage through the crowded area easier (they don't get bumped so much). They make the character smaller, also decreasing collisions. They collect molecules as move along to get them out of the way (decrease concentration).	A DCCK is made when no visible negativity has taken place, do not code as PN, just a DCCK. A DCCK in decrease concentration should be distinguishable from "difficult resource collection" (Section 6.2.2)… make a judgement about the player's intentions here.
Ligand A binds to its channel and the channel opens with the check point on the other side. The player moves to pass through but misses and the ligand dissociates.	Game only	The player increases concentrations of Ligand A. The player increases the temperature. The player decreases the character size to fit more easily through the channel when it opens again	
The player successfully forms a vesicle but doesn't get inside of it to be transported across a membrane (maybe they were hindered by the crowded environment or they just didn't realize they need to be in it)	Game only	They were successful in the cellular process but didn't reach the checkpoint (negative) forcing them to restart (positive)	

6.2.5 Scoring less than 3 Stars

Applicable to **GAME** ONLY.

Productive negativity due to feedback in the game. The player is shown 1 or 2 out of 3 stars. They restart the level and to try to do better to achieve a full 3 Stars. It is NOT coded for the very first level of the game when the player simply has to move the character in a straight line through the channel (impossible for a DCCK to occur in this level).

6.2.6 Incorrect actions

Applicable to both game and interactive simulation conditions.

For this negativity source, the negativity that occurs (and its productive response) is directly linked to a blatantly ***incorrect*** action from the user. For other sources, the negativity is a side effect of engaging in the interactive simulation (e.g. losing resources due to overheating in Section 6.2.2, navigating through crowded environment in Section 6.2.4). Even for resource retentiveness (Section 6.2.1), the user is still releasing the *correct* molecule. So here, we look for incorrect action in our three main interactions: 1) collecting/releasing molecules (i.e. concentration modification); 2) Heating/Chilling the environment (temperature modification); 3) Growing/shrinking the character/pinball (crowding modification). The action is detrimental to goal completion, leading to a productive response.

Incorrect crowding: The user uses the Grow or Shrink function on the pinball/character incorrectly, hindering their progress in some way. A demonstration of correct conceptual knowledge follows. More often than not, it will be Grow that is detrimental, rather than shrink.

Table 24. Examples of an incorrect use of crowding as a productively negative experience.

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
In level 11 (W2Z6 or Box 2.6), the player must keep the character/pinball 'alive' for as long as possible in the presence of degradation enzymes.	Both apps	The player uses Grow, which makes the pinball/character degrade quickly because collision frequency is increased with higher crowding. They restart the level over insert a new pinball and try again.	If the user does not try again after degradation, or there is no productive response.
To decrease crowding	Both apps	Using Grow will make the pinball/character large, which may block molecules from getting to where the user wants them. They might use shrink to counteract, or (in the control) remove the pinball.	-

Incorrect concentration: Incorrect collection/release (i.e. concentration) comes in two flavours: A) incorrect item is released or the correct item is released in the wrong location); B) detrimental collection or release.

A) The user might release an incorrect molecule to achieve a task, or they might release the correct molecule in the wrong place. This may indicate incorrect *factual* knowledge. This action must be followed by a productive response.
B) Detrimental collect/release may be harder to identify. Whereas the type above was about incorrect *factual* knowledge, this is more closely related to incorrect *conceptual* knowledge. A DCCK must follow, of course.
    - **Detrimental collection (decrease concentration):** It is very easy to collect molecules that you don't need in areas where there are lots of different types. However, retaining these molecules may result in low concentrations where they are needed, thus impeding the cellular process under investigation. This type of productive negativity often happens at the same time as "resources lost to overheating" …molecules escape the area if the heat is too high and if the user has 3+ clathrin/adaptin in their inventory, there may not be enough for the vesicle to form. We must choose one or the other. Choose the one closest to the productive response.
    - **Detrimental release (increase concentration):** The molecules that are released are harmful to the processes being investigated.

Table 25. Examples of an incorrect use of concentration as a productively negative experience.

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
A) In level 7 (W2Z2 or Box 2.2), the user needs to release cargo molecules in the area above the cargo receptors (at the start, this area has none)	Both apps	The user releases the cargo in the area below the receptors (instead of above), but then either a) restarts, or b) collects them again and releases above The user releases other molecules above this area, such as clathrin, adaptin, HSC70, auxillin, or ligand A/B, and then either a) collects them and transposes them back across, or b) restarts.	The user releases dynamin above this area, that is OK because it will pinch off the vesicle from this side too.
B - decrease) In level 7, clathrin and adaptin often get collected inadvertently and then, later, the vesicle wont form because there are little/none left in the area.	Both apps	The player can release the items from their inventory, or heat to speed reactions (if all are not gone).	If the user only has 1 or 2 of the molecule in their inventory and there are plentiful molecules in the area (losing a few won't make a difference). Use best judgement. It should slow down the cellular process and result in clear negativity.
B - increase) The user re-releases degrading enzymes, proteasomes, ubiquitination enzymes into the same area as they were collected in.	Both apps	This leads to the degradation of key molecules needed (negative). The player might remove these again, increase concentrations of target molecules (if available), chill the environment to decrease collisions, or restart.	-

Incorrect temperature: Incorrect use of chill/heat that hinders the progress in the game/app, with a productive response.

Table 26. Examples of an incorrect use of temperature as a productively negative experience.

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
Raising the heat in level 11 (ubiquitin level), causing the proteasomes and ubiquitination enzymes to degrade the pinball/character quickly.	Both apps	User restarts level or adds in new pinball.	-
Environment is chilled when user is waiting for ligands to bind, resulting in very long time-to-reaction.	Both apps	They respond by either changing the temp to be warmer, increasing concentrations, increase crowding.	-
Cranks heat when waiting for ligands to bind	Both apps	-	The molecules begin to disappear because they escape the confines of the membrane. This is instead “resources lost due to overheating” (Section 6.2.2) because the action itself was helpful initially.

6.2.7 Simulation-only

Applicable to both game and interactive simulation conditions.

When all other sources do not fit but there is still negativity, it is probably due to the simulation (i.e. the programed random behaviour of the molecules). The negativity is not a direct cause of interactions made by the user; the user has performed one or more DCCK and there is still negativity, leading to more DCCK(s).

Table 27. Examples of productively negative experiences induced by the underlying simulation.

Example context	Applies to	WOULD be coded if…	WOULD NOT be coded if…
The user releases the more (correct) molecules than are needed for a process to take place.	Both apps	There is a still delay. They respond with a DCCK (temp/concentration/crowding increase)	The negativity fits into one of the sources described above.
In level 9 (W2Z4, box 2.4) they release all the ligand C (with or without inhibitor) that they can.	Both apps	The enzymes still degrade all the ligand C, requiring them to collect more inhibitor/ligands or a restart, etc.	
Player has increased concentrations (in a way that does not result in negativity in and of itself)	Both apps	The process is still taking a while regardless of high concentrations, leading to further modifications.	
When stuff is just taking a while (even though they’ve performed a DCCK)	Both apps	they decide to restart and try again using a different strategy	

7. Qualitative gameplay: detailed results

7.1. Demonstrations of correct conceptual knowledge

Table 28. Demonstrations of correct conceptual knowledge (DCCKs) produced by *MolWorlds* and *MolSandbox* players, as well as the percentage of total modifications that were classified as DCCKs. Total DCCKs and subcategories are presented. Total modifications can be viewed in Table 6.

	***MolWorlds* (serious game)**			***MolSandbox* (interactive sim.)**			**Difference**			
Interaction	**Mean**	**SD**	**Min, max**	**Mean**	**SD**	**Min, max**	**U**	***Z***	***p***	**+**
Concentration total	9.90	6.96	1, 28	17.02	6.33	5, 34	386.50	-4.33	< .001	S
Temperature total	6.61	6.77	0, 29	10.00	4.98	0, 20	495.00	-3.47	.001	S
Crowding total	2.02	2.25	0, 11	1.05	1.48	0, 6	577.00	-2.69	.007	G
Total demonstrations	19.24	12.69	4, 58	25.17	7.18	8, 44	532.50	-3.00	.003	S
% Concentration mods.	6.84	2.73	1.0, 11.4	12.67	7.72	2.6, 44.2	352.50	-4.62	<.001	S
% Temperature mods.	33.69	19.60	0.0, 83.3	12.76	5.95	0.0, 31.0	273.50	-5.35	<.001	G
% Crowding mods.	30.53	21.19	0.0, 72.7	15.83	18.85	0.0, 57.1	530.00	-3.09	.002	G
% Total modifications	11.38	3.68	4.6, 21.8	11.19	4.45	3.3, 25.5	804.50	-0.52	.607	-

+ = indicates whether greater interactions are seen in the interactive simulation (S) or serious game (G) group for significant ($p < .050$) results, for ease of viewing.
SD = standard deviation

7.2. Instances of productive negativity: detailed results

Table 29. Instances of productive negativity (PN).

	***MolWorlds* (serious game)**			***MolSandbox* (interactive sim.)**			**Difference**			
PN type	**Mean**	**SD**	**Min, max**	**Mean**	**SD**	**Min, max**	**U**	***Z***	***p***	**+**
Resource retentiveness	0.78	0.82	0, 4	0.21	0.52	0, 2	486.00	-3.97	< .001*	G
Difficult resource collection	0.29	0.60	0, 2	0.07	0.26	0, 1	729.00	-1.97	.049*	G
Resources lost to overheating	0.17	0.38	0, 1	0.43	0.81	0, 4	731.00	-1.59	.112	-
Navigation, reaching checkpoint	1.37	1.07	0, 5	-	-	-	-	-	-	-
NEW: Scoring less than 3 stars	0.17	0.67	0, 4	-	-	-	-	-	-	-
NEW: Incorrect actions	0.68	0.76	0, 3	1.62	1.43	0, 5	520.00	-3.26	.001*	S
Simulation only	1.34	1.97	0, 11	1.45	1.23	0, 2	718.00	-1.36	.174	-
Total instances of PN	4.59	3.01	1, 15	3.29	1.83	1, 8	638.50	-2.06	.040*	G

+ = indicates whether greater interactions are seen in the interactive simulation (S) or serious game (G) group for trending ($p < .100$) and significant (*, $p < .050$) results for ease of viewing.
SD = standard deviation.

Table 30. Summary of types of productive negativity (PN) sources with examples.

PN source	Description of negative event	Example of negativity source	Example of productive response
Resource retentiveness	**Both apps**. The player releases the exact number of molecules needed for a molecular event to occur; the random motion of the molecule(s) results in delayed binding.	The player must open Channel A, requiring 1 Ligand A. Although they have 5 ligands in their inventory, they release only 1/5, which moves off in a random direction once released.	In response to the delay, the player A) releases more ligands, B) increases the temperature, or C) increases crowding (or in combination), all of which increases the probability of Ligand A binding.
Difficult resource collection	**Both apps**. Molecules can be difficult to collect using the character (game) or cursor (control), due to the chaotic random motion of the molecules.	There is apparent "chasing" of molecules by the character (game) or cursor (control).	The player chills the environment to slow the movement of molecules, or increases the crowding to limit molecules' range of motion to make collecting resources easier.
Resources lost to overheating	**Both apps**. High heat can be helpful to speed rates of interactions. However, heating for too long results in a loose membrane where molecules can escape, subsequently resulting in slowed or halted interactions, depending on how many are lost.	In level 7, the player releases 5 cargo molecules to bind to the 5 empty cargo receptors and initiate vesicle formation. They crank the heat up to high repeatedly. While 2 cargo bind, 3 others escape the confines of the membrane and the vesicle cannot form.	If available, the player collects and releases more cargo molecules in the correct location. If less than 3 cargo remain, the player restarts the level to try again, since the vesicle cannot form without 5 total cargo.
Navigation, reaching checkpoint	**Game only**. When navigating through the molecular world, the player gets bumped around by molecules, making it difficult to reach the checkpoint. This may result in them missing their chance to reach the checkpoint altogether.	In level 7, the vesicle is about to bud off. The player is outside of it because they were increasing concentrations of molecules on the exterior of the vesicle. They race the character back to the vesicle but is bumped along the way by other molecules and ends up missing their ride to the checkpoint.	The level must be restarted in this scenario. Next time, the player shrinks the character to decrease local crowding and chills the environment to make navigation easier. They could also modify concentrations of molecules on the exterior of the vesicle prior to releasing cargo.
NEW: Scoring less than 3 stars	**Game only**. Game score is based on time to level completion. Based on this score, 1-3 are shown after level completion to encourage the player to try again.	The player gets 2/3 stars after completing a level.	To achieve 3 stars, the player restarts the level and produces more DCCKs (e.g. increase concentration of key molecule, which will lead to faster interactions) to finish more quickly.
NEW: Incorrect actions	**Both apps**. The player does something blatantly incorrect, such as heating the environment when chilling would be more appropriate, increasing concentrations when they should be decreased, or releasing molecules in incorrect locations.	In level 11, the goal is to prevent the character/pinball from being degraded by enzymes. The player increases the temperature and the size of the character/pinball, resulting in almost immediate degradation.	On their subsequent attempt, the player chills the environment, shrinks the pinball/character, and collects enzymes to reduce their concentration, leading to the longer survival of the character/pinball.
Simulation only	**Both apps**. The player performs one or more DCCKs, but a negative event still occurs.	The player must open Channel A, requiring 1 Ligand A. They collect 5 ligands (max) and then release all 5/5. The random motion of the ligand results in delayed binding, regardless.	The player finds and releases more Ligand A, increases temperature, and/or increases crowding, resulting in a higher probability of binding.

DCCK = demonstration of correct conceptual knowledge

8. Intervention engagement survey results

8.1. IMMS statements: descriptive results and comparisons by stimulus type

Table 31. Descriptive statistics and Mann-Whitney U comparison of Stimulus groups for modified IMMS statements [2] to evaluate intervention (app) engagement.

	MolSandbox			*MolWorlds*			Comparison		
Modified IMMS statement	**Median**	**IQR**	**Min, max**	**Median**	**IQR**	**Min, max**	**U**	***Z***	***p***
1. The material covered in the app was more difficult to understand than I would like for it to be.	3	2	1, 5	2	1	1, 5	576.00	-2.899	.004 *
2. The app had so much information that it was hard to pick out and remember the important points.	2	2	1, 5	2	1	1, 5	781.50	-0.952	.341
3. The app is so abstract that it was hard to keep my attention on it.	2	1	1, 5	2	1	1, 5	819.00	-0.608	.543
4. The app looks dry and unappealing.	2	0	1, 4	2	1	1, 4	709.00	-1.718	.086
5. The exercises in this app were too difficult.	2	1	1, 5	3	1	1, 4	639.00	-2.278	.023 *
6. The amount of repetition in this app caused me to get bored sometimes.	3	2	1, 5	3	2	1, 5	818.00	-0.599	.549
7. The app was not relevant to my needs because I already knew most of it.	2	1	1, 4	2	1	1, 4	798.00	-0.800	.424
8. The style of writing in the app is boring.	3	1	1, 4	2	1	1, 4	699.50	-1.734	0.83
9. There are so many words in each exercise that it is irritating.	2	1	1, 5	2	1	1, 4	820.00	-0.596	.551
10. I could not really understand quite a bit of the material in this app.	3	2	1, 5	2	1	1, 5	528.00	-3.237	.001 *
11. Completing levels successfully was important to me.	3	1	1, 5	3	1	1, 5	882.00	0.000	1.000
12. I enjoyed the app so much that I would like to learn more molecular Biology concepts from it.	4	1	2, 5	4	1	2, 5	792.00	-0.915	.360
13. I can relate the content/concepts of this app to things I'm learning about in Biology.	4	1	2, 5	4	2	2, 5	796.50	-0.817	.414
14. It felt good to successfully complete levels in this app.	4	1	3, 5	4	0	2, 5	796.50	-0.895	.371
15. It was a pleasure to engage with this app and I would do so again if given the opportunity.	4	1	1, 5	4	0	2, 5	784.50	-0.942	.346

Scale: 1 = Strongly disagree; 2 = disagree; 3 = Neither agree nor disagree; 4 = Agree; 5 = Strongly agree

IQR = inter-quartile range

8.2. IMMS statements: descriptive results and comparisons by stimulus-native-language subgroup

Table 32. Descriptive statistics by Stimulus-native-language subgroups (control-NNES, control-NES, game-NNES, game-NES) for modified IMMS statements [2] to evaluate intervention (app) engagement.

	MolSandbox						***MolWorlds***					
	NNES			**NES**			**NNES**			**NES**		
Modified IMMS statement	**Med.**	**IQR**	**Min, max**	**Med.**	**IQR**	**Min, max**	**Med.**	**IQR**	**Min, max**	**Med.**	**IQR**	**Min, max**
1. The material covered in the app was more difficult to understand than I would like for it to be.	2.5	2	1, 4	3	2	2, 5	2	2	1, 4	2	2	1, 5
2. The app had so much information that it was hard to pick out and remember the important points.	2	1	1, 4	2	2	1, 5	2	2	2, 4	2	2	1, 5
3. The app is so abstract that it was hard to keep my attention on it.	2	1	1, 3	2.5	2	1, 5	2	2	1, 4	2	0	1, 5
4. The app looks dry and unappealing.	2	1	1, 4	2	0	1, 4	2	1	1, 3	2	1	1, 4
5. The exercises in this app were too difficult.	2	1	1, 3	2	1	1, 5	3	2	1, 4	3	1	1, 4
6. The amount of repetition in this app caused me to get bored sometimes.	3	2	1, 4	3	2	2, 5	3	2	1, 4	3	2	1, 4
7. The app was not relevant to my needs because I already knew most of it.	2	1	1, 4	2	1	1, 4	2	1	1, 4	2	2	1, 4
8. The style of writing in the app is boring.	3	1	1, 4	3	2	1, 4	2	2	1, 4	2	1	1, 4
9. There are so many words in each exercise that it is irritating.	2	2	1, 4	2	1	1, 5	2	2	1, 4	2	1	1, 4
10. I could not really understand quite a bit of the material in this app.	3	2	1, 5	3	2	2, 5	2	3	1, 4	2	1	1, 5
11. Completing levels successfully was important to me.	3.5	1	2, 4	3	1	1, 5	3	1	2, 4	3	1	1, 5
12. I enjoyed the app so much that I would like to learn more molecular Biology concepts from it.	4	1	2, 5	4	1	3, 5	4	1	2, 5	4	1	2, 5
13. I can relate the content/concepts of this app to things I'm learning about in Biology.	4	1	2, 5	4	1	2, 5	4	1	2, 5	4	2	2, 5
14. It felt good to successfully complete levels in this app.	4	1	3, 5	4	1	3, 5	4	1	3, 5	4	1	2, 5
15. It was a pleasure to engage with this app and I would do so again if given the opportunity.	4	1	2, 5	4	1	1, 5	4	1	2, 5	4	1	2, 5

Scale: 1 = Strongly disagree; 2 = disagree; 3 = Neither agree nor disagree; 4 = Agree; 5 = Strongly agree

IQR = inter-quartile range; Med. = Median

Table 33. Kruskal-Wallis comparisons by Stimulus-native-language subgroups (control-NNES, control-NES, game-NNES, game-NES) for each modified IMMS statements [2] to evaluate app engagement.

Modified IMMS statement	χ^2	df	*p*
1. The material covered in the app was more difficult to understand than I would like for it to be.	10.710	3	.013 *
2. The app had so much information that it was hard to pick out and remember the important points.	3.616	3	.306
3. The app is so abstract that it was hard to keep my attention on it.	14.530	3	.002 *
4. The app looks dry and unappealing.	4.051	3	.256
5. The exercises in this app were too difficult.	5.617	3	.132
6. The amount of repetition in this app caused me to get bored sometimes.	2.397	3	.494
7. The app was not relevant to my needs because I already knew most of it.	3.679	3	.298
8. The style of writing in the app is boring.	6.455	3	.091
9. There are so many words in each exercise that it is irritating.	1.803	3	.614
10. I could not really understand quite a bit of the material in this app.	11.361	3	.010 *
11. Completing levels successfully was important to me.	0.036	3	.998
12. I enjoyed the app so much that I would like to learn more molecular Biology concepts from it.	0.935	3	.817
13. I can relate the content/concepts of this app to things I'm learning about in Biology.	5.013	3	.171
14. It felt good to successfully complete levels in this app.	4.394	3	.222
15. It was a pleasure to engage with this app and I would do so again if given the opportunity.	2.613	3	.455

* = significant at .050); NES = native-English speakers; NNES = non-native-English speakers

Table 34. Post-hoc Mann-Whitney U pairwise comparisons by Stimulus-native-language subgroups for significant Kruskal-Wallis tests analysing modified IMMS statements [2] (Table 33).

Modified IMMS statement	Comparison groups		U	Z	*p*
1. The material covered in the app was more difficult to understand than I would like for it to be.	Control-NNES	Control-NES	181.00	-0.939	.348
		Game-NNES	104.00	-0.560	.575
		Game-NES	174.00	-2.035	.043
	Control-NES	Game-NNES	113.50	-1.433	.152
		Game-NES	184.50	-3.072	.002 *
	Game-NNES	Game-NES	143.00	-1.333	.182
3. The app is so abstract that it was hard to keep my attention on it.	Control-NNES	Control-NES	91.00	-3.389	.001 *
		Game-NNES	70.00	-2.020	.043
		Game-NES	197.00	-1.600	.110
	Control-NES	Game-NNES	119.50	-1.207	.227
		Game-NES	210.50	-2.673	.008 *
	Game-NNES	Game-NES	150.00	-1.145	.252
10. I could not really understand quite a bit of the material in this app.	Control-NNES	Control-NES	184.50	-0.834	.404
		Game-NNES	90.00	-1.122	.262
		Game-NES	167.00	-2.161	.031
	Control-NES	Game-NNES	98.00	-1.962	.050
		Game-NES	183.00	-3.139	.002 *
	Game-NNES	Game-NES	173.50	-0.441	.659

* = significant at Sidak adjusted value of .0085 (for 6 comparisons); NES = native-English speakers; NNES = non-native-English speakers

9. References

[1] G. F. Burch, N. a. Heller, J. J. Burch, R. Freed, and S. a. Steed, “Student Engagement: Developing a Conceptual Framework and Survey Instrument,” *J. Educ. Bus.*, vol. 90, pp. 224–229, 2015, doi: 10.1080/08832323.2015.1019821.

[2] N. Loorbach, O. Peters, J. Karreman, and M. Steehouder, “Validation of the Instructional Materials Motivation Survey (IMMS) in a self-directed instructional setting aimed at working with technology,” *Br. J. Educ. Technol.*, vol. 46, no. 1, pp. 204–218, 2015, doi: 10.1111/bjet.12138.

[3] A. Gauthier and J. Jenkinson, “Designing productively negative experiences with serious game mechanics: Qualitative analysis of game-play and game design in a randomized trial,” *Comput. Educ.*, vol. 127, no. 2018, pp. 66–89, 2018, doi: 10.1016/j.compedu.2018.08.017.

[4] K. A. Hallgren, “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial,” *Tutor. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, 2012, doi: 10.1016/j.biotechadv.2011.08.021.Secreted.

[5] M. B. Carvalho *et al.*, “An activity theory-based model for serious games analysis and conceptual design,” *Comput. Educ.*, vol. 87, pp. 166–181, 2015, doi: 10.1016/j.compedu.2015.03.023.