Technology, Knowledge and Learning (2023) 28:1825–1839
https://doi.org/10.1007/s10758-022-09617-7

ORIGINAL RESEARCH

If We Build It, Will They Learn? An Analysis of Students' Understanding in an Interactive Game During and After a Research Project

Paul Horwitz[1] · Frieda Reichsman[1] · Trudi Lord[1] · Chad Dorsey[1] · Eric Wiebe[2] · James Lester[2]

Accepted: 21 July 2022 / Published online: 5 August 2022

Abstract

Studies of educational games often treat them as "black boxes" (Black and Wiliam in Phi Delta Kappan 80: 139–48, 1998; Buckley et al. in Int J LearnTechnol 5:166–190, 2010; Buckley et al. in J Sci Educ Technol 13: 23–41, 2010) and measure their effectiveness by exposing a treatment group of students to the game and comparing their performance on an external assessment to that of a control group taught the same material by some other method. This precludes the possibility of monitoring, evaluating, and reacting to the actions of individual students as they progress through the game. To do that, however, one must know what to look for because superficial measures of success are unlikely to identify unproductive behaviors such as "gaming the system." (Baker in Philipp Comput J, 2011; Downs et al. in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, USA, 2010) The research reported here advances the ultimate goal of creating educational games that can provide real time, meaningful feedback on the progress of their users, enabling teachers or the game itself to intervene in a timely manner. We present the results of an in-depth analysis of students' actions in *Geniventure*, an interactive digital game designed to teach genetics to middle and high school students. *Geniventure* offers a sequence of challenges of increasing difficulty and records students' actions as they progress. We analyzed the resulting log files, taking into account not only whether a student achieved a certain goal, but also the quality of the student's performance on each attempt. Using this information, we quantified students' performance and correlated it to their learning gain as estimated by scores on identical multiple-choice tests administered before and after exposure to *Geniventure*. This analysis was performed in classes taught by teachers who had participated in professional development as part of a research project. A two-tailed paired-sample t-test of mean pre-test and post-test scores in these classes indicates a significant positive difference with a large effect size. Multivariate regression analysis of log data finds no correlation between students' post-test scores and their performance on "practice" challenges that invite experimentation, but a highly significant positive correlation with performance on "assessment" challenges, presented immediately following the practice challenges, that required students to invoke relevant mental models. We repeated this analysis with similar results using a second group of classes led by teachers

Extended author information available on the last page of the article who implemented *Geniventure* on their own after the conclusion of, and with no support from, the research project.

Keywords Modeling · Science education · Logging · Assessment

1 Introduction and Theoretical Framework

A voluminous body of science education research has demonstrated the importance of supporting students in the formation of mental models of target concepts (Bransford et al., 2000; Fishwick et al., 2014; Harrison & Treagust, 1996; Hestenes, 2015; White, 1993; Wright, 2012). Simulations of relevant phenomena and processes have proven effective in achieving this goal (Bossel, 2018; Fishwick et al., 2014; Horwitz, 1995; Perkins, 2020), particularly when coupled to appropriately scaffolded sequences of challenges (Franklin et al., 2009; Mayer, 2016; White, 1993) in which case they are often referred to as "serious games." (de Freitas, 2018; Noemí & Máximo, 2014; Riopel et al., 2019). Properly designed, such games can engage students in authentic STEM practices such as problem solving and model formation. Moreover, by logging and analyzing students' actions we can adapt games for "stealth" assessment (Gobert et al., 2013; Pellegrino & Quellmalz, 2010; Shute, 2011; Shute & Ventura, 2013, 2015) of students' learning.

However, merely analyzing a student's success or failure within the structure of a game intended to teach certain concepts cannot necessarily be considered a valid measure of understanding those concepts. Students can succeed at a game that is intended to teach a certain set of concepts without actually learning them (Annetta et al., 2009; Baker et al., 2009, 2013; Horwitz & Christie, 2000), a process sometimes referred to as "gaming the system." The term implies purposeful actions by a student seemingly designed to avoid learning the target content, but the behavior needn't be intentional. There is evidence that students can succeed at a game without really understanding why they have succeeded (Aleven et al., 2010; Horwitz & Christie, 2000). When this happens the knowledge they acquire is superficial and narrowly contextualized within the parameters of the game. Thus, it fails to transfer and does not contribute to success on other forms of assessment of content knowledge.

There is a pressing need, therefore, to distinguish between in-game performance that implies content mastery and that which is attributable to the acquisition of contextualized skills that convey success in achieving the objectives of the game but do not correlate with conceptual understanding of the STEM concepts underlying the game. In this paper we demonstrate that this goal can be achieved by designing game challenges wherein success requires that students form mental models of the target concepts. We compare two types of superficially similar challenges. The *practice challenges* enable students to become familiar with the user interface and to improve their in-game skills. These challenges offer students the opportunity to develop a mental model of an underlying STEM concept, but do not require such a step for success. Immediately following each practice challenge, however, we present a corresponding *assessment challenge* designed to test students' access to just such a mental model.

We measure students' learning gain by comparing their scores on identical assessments of STEM content delivered before and after exposure to the game (Pellegrino, 2014). This assessment parallels the *Geniventure* curriculum but introduces novel species and traits. It contains some items that relate to the practice and assessment challenges as well as others taught by the game but only indirectly related to those particular challenges. Notwithstanding the presence of these far transfer items we hypothesized that students' performance on the assessment challenges should correlate with their learning gain because success on those challenges is an indication of model building, which has been shown to correlate with content mastery. In contrast, we expected to find a significantly weaker correlation, or none at all, with performance on the practice challenges, which prioritize highly contextualized in-game skills.

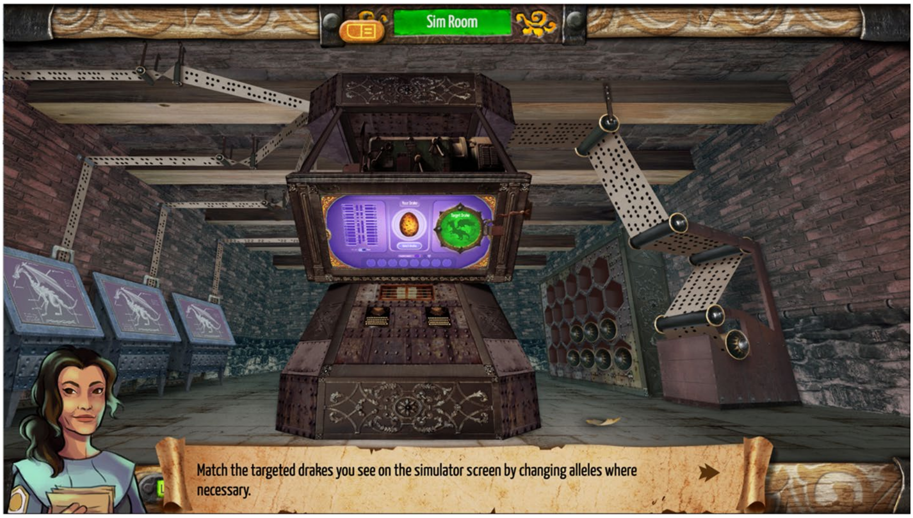
Our target domain is genetics, specifically *transmission genetics*: the study of how a sexually reproducing organism's physical traits are related to the traits of its parent organisms, and the mechanism by which those traits are inherited. *Geniventure* addresses this subject matter in general. In contrast, the key construct addressed in this paper is the mapping between an organism's genes—its genotype—and the collection of its observable traits—its phenotype. This mapping is not one-to-one in that pairs of genes act in concert to affect phenotype, nor is it unique: different traits have different mappings. *Geniventure* addresses this issue by providing students with a sequence of similar challenges, described below, involving traits with increasingly complex modes of inheritance. Performance on this set of challenges, and ultimately their performance on a post-test, forms the focus of the research presented here.

2 Description of the Intervention

Geniventure introduces students to genetics principles and requires problem solving in concert with a growing understanding of genetics (McElroy-Brown & Reichsman, 2019; Mutch-Jones et al., 2021; Rachmatullah et al., 2021). The game involves a narrative about dragons and their model species, drakes, in which a war has broken out between kingdoms, endangering the dragon population. The goal is for each player to breed drakes to learn about the genetics of certain dragon traits that would be useful during the war. A diverse cast of characters in a scientific Guild present a total of 65 challenges organized into missions of three to eight related challenges (Fig. 1).

As students progress through the game, each of their actions generates an array of parameters that are saved in a log file. For example, an action that alters a gene will report the specific gene that was changed and the initial and final alleles (variants) of that gene. Since all the actions are time-stamped, the information contained in each log file is sufficient to enable the complete reconstruction of the session.

We employ a correlational research design (Asamoah, 2014; Steinkamp & Maehr, 1983; Thompson et al., 2005) based on two delayed cohorts both of which used identical versions of the game and assessments. The research cohort (Rachmatullah et al., 2021) consists of six high school teachers and their students who implemented the intervention during the spring of 2019. The teachers had all participated in the National Science Foundation-funded project that created *Geniventure* and had had experience with similar earlier versions of the genetics game. All of them had attended a three-day workshop held during the summer of 2018 that included sessions on genetics, guidance on implementation strategies, and an introduction to supplementary materials such as worksheets and student handouts. The teachers in this group were observed by research staff on multiple occasions during the implementation (Mutch-Jones et al., 2021; Wu et al., 2019) and were asked to complete an online survey after each implementation day. They could submit help requests as needed and these were addressed by project staff, generally within 24 h.



Match the targeted drakes you see on the simulator screen by changing alleles where necessary.

Fig. 1 The Geniventure narrative unfolds as the students level up through the challenges. A cast of scientists guide students through the game with instructions and hints

In addition to the teachers who participated in the research and were supported as described, an extended cohort of teachers used *Geniventure* on their own during the 2019–2020 school year, after the research phase of the project had ended. These teachers received no professional development or other services from us, and we had no communication with them other than answering occasional requests for technical assistance. They had access to the same worksheets and discussion guides that had been provided to the research teachers, and also to an online course created late in the project cycle and thus not available to teachers in the research cohort. The teachers in this extended cohort represented over 400 schools and taught almost 20,000 students (though, as described below, we limited our data analysis to just 433 of them). These teachers were not requested to provide information concerning the nature of the school they taught in or the level of their classes. Although extended cohort teachers were by definition self-selected, their experience with the intervention and that of their students offers the best evidence we have regarding the long-range impact of the research project.

All of the teachers in the research cohort were required to administer identical assessments to their students before and after their engagement with *Geniventure*. The extended cohort teachers were under no such constraint but twenty-two of them chose to do so anyway. We limited our analysis to those teachers. The 2019–2020 school year was cut short by the appearance of the COVID-19 virus, which necessitated the shutting down of schools across the country. Consequently, we limited our analysis of the extended cohort data to students who completed the intervention prior to April 2020. For both cohorts we excluded from our analysis any students who had not completed all the relevant challenges in the game or had not answered at least 95% of the items on both the pre-test and the post-test. Consequently, the length of time that students spent with the game was comparable across cohorts. After filtering with these criteria, we ended up with 338 students from the research cohort and 433 from the extended cohort.

3 Research Design

3.1 Research Questions

By nalysing log files and pre- and post-test scores, we sought answers to the following research questions:

1. Is there evidence that both the research cohort and the extended cohort learned the target content?
2. Is there a significant difference between the practice challenges and the assessment challenges with respect to how well they predict the post-test score?
3. Are there significant differences between the research cohort and the extended cohort with respect to how well the target match challenges predict the post-test score?

3.2 Methodology

Geniventure presents students with 65 challenges, each requiring multiple steps to solve. At any given time the student is faced with a choice of actions to initiate. Some are productive in that they move the student closer to the goal, some are counterproductive, and some are simply redundant. Students' ability to discriminate between these possibilities, informed by their understanding of the underlying genetics, can be estimated by abelled their actions. We applied this strategy to a subset of challenge types: the *target match* challenges. Both types focus on a specific aspect of genetics: the mapping between genotype and phenotype, which is the target construct for this research.

The 22 target match challenges share a common goal: to change a given drake's genes from one allele to another in order to make it look like a target drake, randomly generated at the start of the challenge. When the student thinks she has achieved the goal, she submits her drake for inspection. The genotypes of the two drakes need not match – indeed, the target drake's genotype is not revealed to the student. All that is required is that the phenotypes and sexes of the submitted and target drakes be identical.

The challenges also set a subsidiary goal—to achieve a match with as few actions, called "moves" in the game, as possible, where a move is defined as either an allele change or a sex chromosome change (e.g. from two X chromosomes to an X and a Y). The minimum number of moves required to change the given drake to match the target drake is calculated by the game, based on the phenotype of the target drake and the initial genotype of the given drake. If the student produces the target drake using no more than two moves over the theoretical minimum, she is rewarded with a crystal, the color of which denotes how successful she was. Achieving a match with the minimum number of moves results in a blue crystal—the most prized of all. One excess move results in a yellow crystal, two excess moves and the crystal is red. Matching the target drake with more than two moves over the minimum is counted as a completion but produces no crystal at all. If the drakes do not match, students continue the challenge with the same target drake; if they do match, the students are offered an opportunity to try to achieve a better crystal color by revisiting the challenge with a different target drake.

The target match challenges occur in four clusters throughout the game, separated by challenges of other types. Each cluster starts with the given drake continuously visible to the student, changing its appearance (or not) after a move, in accordance with a scientifically realistic inheritance pattern that is not explicitly presented to the student. In this practice "visible drake" condition students can make as many moves as they like, observing the effect and only submitting their drake when it looks exactly like the target drake. After two or more challenges with a visible drake, the immediately following challenges change to a "hidden drake" condition: the given drake is now contained inside an egg and only becomes visible when the drake is submitted and the egg hatches

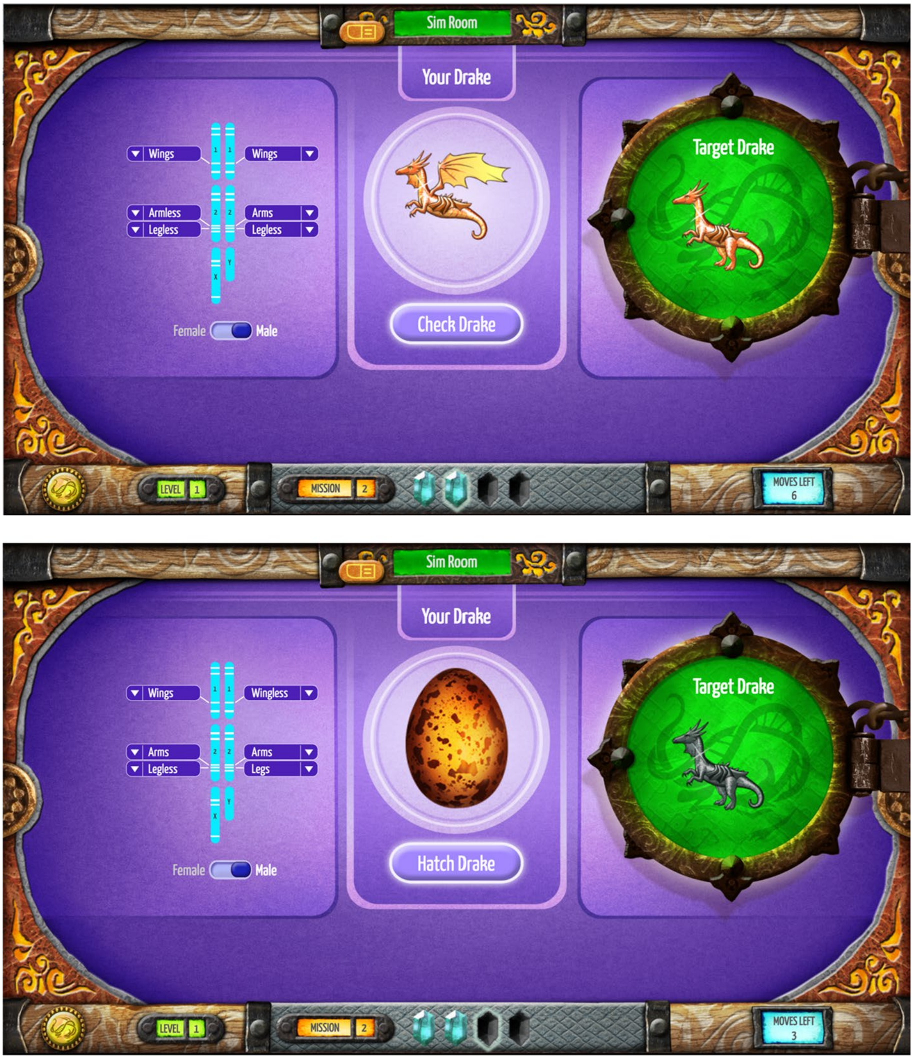


Fig. 2 Top: visible drake target match challenge. "Your Drake" is the given drake with alleles that can be changed by the student. Students see the effect, if any, of an allele change (called a "move" in the game) immediately, and can visually compare the given drake to the target drake before submitting it as a match ("Check Drake" button). Bottom: hidden drake target match. The given drake is encased in its egg and students must apply a correct mental model of the genotype-to-phenotype mapping in order to avoid making redundant or unproductive moves. Students can view the outcome only after "hatching" the drake, which submits the alleles they have chosen. Student performance is scored using colored "crystals"

(see Fig. 2). Thus, in the assessment condition students cannot observe the effect of their actions but must apply their knowledge of the mapping between genotype and phenotype. The transition from the practice (visible) to the assessment (hidden) condition repeats in each of the four clusters of challenges throughout the game, each recurrence involving a different set of traits and a more complex mapping from genotype to phenotype.

The outcomes of the target match challenges are not limited to "correct" or "incorrect." On each correct submission, the color of a crystal received is an indication of *how well* the student succeeded, that is, how efficiently they made their moves to match the target. The crystal color, therefor, is not an arbitrary measure; it is an indicator of the student's understanding of the genetics involved in each challenge.

We illustrate this point by a simple example. The presence or absence of wings is a trait controlled by a single gene that occurs in two different forms, or alleles, abelled, respectively, "Wings" and "Wingless." But wings are a dominant trait, meaning that it only takes one "Wings" allele to give the drake wings. So if the target drake has wings and the manipulable drake has two "Wingless" alleles, the student can match the target by altering one or both genes, but *only needs to alter one.* Altering both genes, which entails an unnecessary move, is evidence of an incomplete understanding of the genotype-to-phenotype mapping for the wings gene, and will result in an inferior crystal. In the alternative situation, if the target drake is wingless and the manipulable drake starts with two "Wings" alleles, the student must change *both* genes to "Wingless" in order to achieve a match. Although students can of course make lucky guesses or careless mistakes, the minimum move requirement for multiple challenges serves as a measure of their grasp of this classic Mendelian inheritance pattern.

3.3 Data Analysis

To support the analysis of students' actions, we developed a rubric that assigns a single numerical score, ranging from zero to five, to each target match challenge. The rubric takes into account the best crystal color awarded to the student and the number of attempts the student made to receive that crystal. (In the rare event that a student attempted the challenge again after receiving a blue crystal, the "extra" attempts were not counted in computing the score.) The scoring rubric is summarized in Table 1.

Table 1 Challenge scores as a function of attempts and best outcome

Outcome[a]	1 attempt	2 attempts	3 attempts	> 3 attempts
Blue	5	3	1	0
Yellow	4	2	0	0
Red	3	1	0	0
None	2	0	0	0
Incorrect	0	0	0	0

[a] Outcomes key. Blue crystal: minimum number of moves. Yellow crystal: one excess move. Red crystal: two excess moves. None: more than two moves over the minimum results in no crystal. Incorrect: no correct drake was submitted and no crystal was received

Table 2 Results of a t-test of the pre- and post-test means of the research cohort

	Mean	Standard deviation
Pre-test	10.5	0.067
Post-test	14.9	0.076
n	338	
Pearson correlation	0.619	
t statistic	19.1	
t critical two-tailed	1.97	
P (two-tailed)	< .001	
Cohen's d	0.832	

Table 3 Results of a t-test of the pre- and post-test means of the extended cohort

	Mean	Standard deviation
Pre-test	12.3	0.051
Post-test	14.9	0.045
n	433	
Pearson correlation	0.663	
t statistic	14.3	
T critical two-tailed	1.97	
P (two-tailed)	< .001	
Cohen's d	0.574	

3.4 Research Question 1: Is There Evidence that both the Research and the Extended Cohort Learned the Target Content?

To answer our first research question we performed a two-tailed, paired-sample t-test on the means of the pre- and post-test, which consisted of 27 identical items. The results of this test are summarized in Table 2.

There is a significant difference between the means of the pre-test and post-test scores, and thus evidence of learning, in the research cohort. The Cohen's d statistic of 0.832 is generally considered to represent a large effect size (Thalheimer & Cook, 2002).

An identical analysis of the extended cohort data yields similar results, as reported in Table 3.

There is a significant difference between the means of the pre-test and post-test scores, and thus evidence of learning, in the extended cohort. The Cohen's d statistic of 0.574 is generally considered to represent a medium effect size (Thalheimer & Cook, 2002).

Table 4 Results of a multiple regression of data from the research cohort using post-test score as the dependent variable

Independent variable	Regression coefficient	Standard error	t statistic	P value
Pre-test score	0.523	0.0483	10.8	<0.001
Mean practice challenge score	0.336	0.506	0.662	0.508
Mean assessment challenge score	1.73	0.334	5.19	<0.001

Table 5 Results of a multiple regression of data from the extended cohort using post-test score as the dependent variable

Independent variable	Regression coefficient	Standard error	t statistic	P value
Pre-test score	0.495	0.0333	14.9	<0.001
Mean practice challenge score	0.529	0.313	1.69	0.092
Mean assessment challenge score	1.37	0.203	6.77	<0.001

3.5 Research Question 2: Is There a Significant Difference Between the Practice Challenges and the Assessment Challenges with Respect to How Well They Predict the Post-test Score?

We evaluated the predictive power of the target match challenge scores in both practice (visible drake) and assessment (hidden drake) conditions. Having established the absence of collinearity as well as the normality of the residuals, the criteria for the validity of ordinary least squares estimation (Hutcheson, 2011), we performed a multiple regression with the students' post-test scores as the dependent variable and their pretest scores and their mean challenge scores (across the four challenges) in the two conditions as independent variables. These results are reported in Table 4 for the research cohort (n = 338) and Table 5 (n = 433) for the extended cohort.

As seen in Tables 4 and 5, the post-test score is significantly correlated with the pre-test score (as expected) in both cohorts. In addition, the post-test score for both cohorts is also significantly correlated with the mean assessment (hidden drake) challenge score but not with the practice (visible drake) challenge score. F tests for both cohorts were highly significant (F = 89.8, p < 0.001 for the research cohort, F = 165, p < 0.001 for the extended cohort), indicating that the model with these three independent variables fits the data significantly better than the mean alone. Moreover, for both cohorts the regression coefficient of the mean assessment challenge score is highly significant, whereas the correlation coefficient for the mean of the practice challenge scores is not. In other words, controlling for pretest score, the assessment challenge score predicts the post-test score but the practice score does not.

3.6 Research question 3: Are There Significant Differences Between the Research Cohort and the Extended Cohort with Respect to How Well the Target Match Challenges Predict the Post-test Score?

To answer this question we conducted an analysis of variance (ANOVA), including the cohort as a fourth independent variable. The results of that analysis are reported in Table 6.

As expected, the regression coefficients of the pre-test score and mean assessment challenge score differ significantly from zero and that of the mean practice challenge score does not. Moreover, we found no significant difference between the cohorts with respect to the prediction of post-test score. The coefficient of the cohort variable is slightly negative, meaning that the extended cohort scored a little lower, on average, on the post-test than the research cohort, but the value of that coefficient is not statistically different from zero. As measured by their performance on the post-test, the two cohorts are statistically indistinguishable.

4 Discussion

The first two research questions are answered affirmatively, the third negatively. Specifically, our analysis indicates that:

> Both cohorts learned the target concepts during the intervention, as indicated by the differences in their mean scores on identical pre- and post-tests. The effect size was large for the research cohort and medium for the extended cohort.
> In both cohorts, students' performance on the target match challenges was significantly and positively predictive of their post-test scores in the assessment condition but not in the practice condition.
> There is no significant difference in post-test scores between the two cohorts.

4.1 Implications for Game Design

Educational games aimed at the sciences are often designed to induce students to form mental models involving unseen entities in order to explain observable phenomena (Mayer, 2016; Buckley et al., 2010; Pedro et al., 2014; Schwarz et al., 2017). A common strategy for accomplishing this goal is to create simulations in which the relevant components of the target model are visible and/or manipulable, whether or not they would be in the real world (Horwitz, 1995; McElroy-Brown & Reichsman, 2019; White, 1993). Another important

Table 6 Results of a comparison of data from both cohorts using post-test score as the dependent variable and including the cohort as a categorical independent variable

Independent variable	Regression coefficient	Standard error	t statistic	P value
Pre-test score	0.510	0.0280	18.2	< 0.001
Mean practice challenge score	0.433	0.276	1.57	0.117
Mean assessment challenge score	1.51	0.189	8.36	< 0.001
Cohort	− 0.362	0.261	− 1.39	0.165

goal for the designers of educational games is to distinguish between the students whose improved performance is an indication that they are learning the target content and those who are simply getting "good at the game" (Baker et al., 2009, 2013; Downs et al., 2010).

The choice of what information to make visible and what to hide from students suggests a simple strategy for addressing both goals. The two different types of *Geniventure* target drake challenges are an example of such a strategy. By alternating between the visible and the hidden drake conditions, *Geniventure* presents two modes of game play, roughly equivalent to a practice phase followed by an assessment phase. In both, the goal is to match a manipulable drake to a fixed target. However, success at the practice challenges can be achieved simply by observing one's incremental progress toward the goal. Students who succeed at this task are evincing evidence of learning how to play the game, but are not necessarily learning the target construct (the mapping from genotype to phenotype). It is striking that students' within-game scores on these challenges correlate with post-test scores of content knowledge *only in the assessment condition.* This strongly suggests that *Geniventure* is able to distinguish between achieving superficial facility at a game (i.e., recognizing when two drakes look the same) and learning its underlying content (i.e., understanding how a drake's genotype determines its phenotype). This feature of *Geniventure* is an essential first step toward providing valuable feedback in real time to students and teachers.

4.2 Replicability in the Absence of Project Support

We find that the effectiveness of our intervention was not significantly impacted by the absence of teacher support typical of a research project. The results indicate that the *Geniventure* materials that were made available online (and continue to be available) after the project ended are sufficient to yield learning of transmission genetics commensurate with that produced in classrooms where teachers had received additional support from the research team. This bears on the broader impact of educational research projects beyond their termination date.

Scalability and sustainability are important goals for education research projects (Roesken-Winter et al., 2015; Blumenfeld et al., 2000) but are difficult to achieve because the constraints imposed for carrying out reliable research often conflict with those presented by the real world (Buzhardt et al., 2006; Carpenter et al., 2004; Dede, 2006). Consequently, findings from educational research projects may fail to offer a direct benefit to teachers and students who did not participate in the original research (Looi & Teh, 2015; Ross & Morrison, 2021). Interventions that develop interactive educational software in the service of research have the potential to address this problem (Morrison et al., 2009). If the technology and supporting materials can be made available beyond the termination of the project that produced them, they may be adopted by teachers who did not participate in their creation. If the software is instrumented and the necessary backend technology exists, then log files produced by those students can be analyzed and the educational effect of the software evaluated.

Geniventure fits into that category. During the school year subsequent to the end of the funding period, 19,524 students used the game, forming a potential comparison group, the "extended cohort." The analysis of data obtained from the research cohort compared students' actions in *Geniventure* to their scores on a pre- and post-test external to the game. Therefore, our analysis of the extended cohort was limited to those students who had completed both tests. As one might expect, a small fraction—just 2.6%—of the students who were not part of the research cohort took both tests, but that was a large enough group to demonstrate that the students' performance in *Geniventure* was predictive of their scores on those summative tests. Once such a correlation has been established with a subset of students, the in-game scores themselves can be validated and used as reliable predictors of learning outcomes. Our study suggests that such serendipitous, uncontrolled interventions can be effective in achieving educational goals.

4.3 Limitations of the Study

The fact that the two cohorts in our study performed so similarly suggests that the combination of online teacher support materials and the *Geniventure* software itself enabled at least the students of the subset of teachers who administered the pre- and post-tests to achieve results similar to those obtained by students of teachers who participated in the research project. That subset, however, is small and arguably not representative of the target population. This is a limitation of the present study and begs the question: what happened in those other classes? This is a pressing subject for further research.

Since our research design did not include a control group, we cannot argue that learning genetics through a game is superior to any other mechanism, nor do we make such a claim. Rather, our purpose is to demonstrate that a fine-grained analysis of log data acquired through students' use of a game can distinguish between behavior likely to reflect learning of target content and that which merely correlates with getting better at achieving in-game objectives.

5 Directions for Future Research

Educational games can easily detect the students who struggle to achieve within-game goals; it is more challenging to identify those who are succeeding at the game without learning what the game is intended to teach. All too often, the existence of such students becomes apparent only when their superficial and contextualized knowledge fails to transfer to performance on an assessment that, for example, introduces a new species with novel traits but genetics identical to that exposed in the game. The challenge, then, is to instrument educational games with appropriate and validated diagnostic challenges, and to use these to inform teachers in a timely manner.

We expect that the research we have reported on here will be a useful milestone in the quest to equip teachers with the information they need to assist students who appear to be winning the game without learning the content.

Acknowledgements We would like to thank the National Science Foundation for their funding of the project (DRL-1503311) that created *Geniventure* and gathered the data reported on here. We would also like to thank project evaluators Karen Mutch-Jones and Santiago Gasca at TERC; our research and development partners Danielle Boulden and Robert Taylor at North Carolina State University; and all the teachers and students who participated in our research.

Data Availability The data is available to qualified researchers by request to The Concord Consortium.

Declarations

Conflict of interest The authors have not disclosed any conflict of interest.

References

Vincent, A., Myers, E., Easterday, M., and Ogan, A. (2010). Toward a Framework for the Analysis and Design of Educational Games. In *2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*, pp. 69–76. ieeexplore.ieee.org.

Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M.-T. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers & Education, 53*(1), 74–85.

Asamoah, M. K. 2014. Re-Examination of the Limitations Associated with Correlational Research. *Journal of Educational Research and*. http://sciencewebpublishing.net/jerr/archive/2014/July/pdf/Asamoah.pdf.

Baker, R. S. j. d., De Carvalho, A., Raspat, J., Aleven, V., Corbett, A. T., and Koedinger, K. R. (2009). Educational Software Features That Encourage and Discourage 'gaming the System.' In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, pp. 475–482. books.google.com.

Baker, R. S. J. d. (2011). Gaming the system: A retrospective look. *Philippine Computing Journal*. http://radix.www.upenn.edu/learninganalytics/ryanbaker/PSCS-gaming-v6.pdf.

Baker, R. S. J. D., Corbett, A. T., Roll, I., Koedinger, K. R., Aleven, V., Cocea, M., Hershkovitz, A., de Caravalho, A. M. J. B., Mitrovic, A., & Mathews, M. (2013). Modeling and studying gaming the system with educational data mining. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 97–115). Springer New York.

Black, P., & Wiliam, D. (1998). Inside the Black Box : Raising Standards Through Classroom Assessment. *Phi Delta Kappan, 80*, 139–148.

Blumenfeld, P., Fishman, B. J., Krajcik, J., Marx, R. W., & Soloway, E. (2000). Creating usable innovations in systemic reform: Scaling up technology-embedded project-based science in Urban Schools. *Educational Psychologist, 35*(3), 149–164.

Bossel, Hartmut. 2018. *Modeling and Simulation*. AK Peters/CRC Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn* (Vol. 11). National academy press.

Buckley, B. C., Gobert, J. D., Horwitz, P., & O'Dwyer, L. M. (2010). Looking inside the black box: Assessing model-based learning and inquiry in BioLogica$^{\mathrm{TM}}$. *International Journal of Learning Technology*. https://doi.org/10.1504/IJLT.2010.034548

Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlits, B., Wilensky, U., Dede, C., & Willett, J. (2004). Model-based teaching and learning with BioLogica™: What do they learn? How do they learn? How do we know? *Journal of Science Education and Technology, 13*(1), 23–41.

Buzhardt, J., Greenwood, C. R., Abbott, M., & Tapia, Y. (2006). Research on scaling up evidence-based instructional practice: Developing a sensitive measure of the rate of implementation. *Educational Technology Research and Development: ETR & D, 54*(5), 467–492.

Carpenter, T. P., Blanton, M. L., Cobb, P. A., Franke, M. L., Kaput, J., and McClain, K. (2004). Scaling up Innovative Practices in Mathematics and Science. *Education*, no. February: 1–16.

de Freitas, S. (2018). Are games effective learning tools? A review of educational games. *Journal of Educational Technology & Society, 21*(2), 74–84.

Dede, C. (2006). Scaling up: Evolving innovations beyond ideal settings to challenging contexts of practice. collegechangeseverything.org. 2006. https://www.collegechangeseverything.org/dotAsset/d352af01-fb00-43c3-a956-6a2f092a7c67.pdf.

Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. (2010). Are your participants gaming the system? Screening mechanical Turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2399–2402. CHI '10. New York: Association for Computing Machinery.

Fishwick, P., Brailsford, S., Taylor, S. J. E., Tolk, A., and Uhrmacher, A.. (2014). Modeling for everyone: Emphasizing the role of modeling in stem education. In *Proceedings of the Winter Simulation Conference 2014*, 2786–96. ieeexplore.ieee.org.

Franklin, T., Morge, S., Narayan, S., Tagliarini, G., Knezek, G., Christensen. R., Tyler-Wood. T., Liu. C., and Chelberg. D. (2009). STEM Learning in Middle School with Games and Simulations. In *Society for Information Technology & Teacher Education International Conference*, pp.1445–1449. Association for the Advancement of Computing in Education (AACE).

Gobert, J. D., Pedro, M. S., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521–563.

Harrison, A. G., & Treagust, D. F. (1996). Secondary students' mental models of atoms and molecules: Implications for teaching chemistry. *Science Education*. https://doi.org/10.1002/(SICI)1098-237X(199609)80:5%3c509::AID-SCE2%3e3.0.CO;2-F

Hestenes, D. (2015). Modeling theory and modeling instruction for STEM education. In *S. Chandrasekhara (Chair), epiSTEME 6 International Conference to Review Research on Science, Technology and Mathematics Education. Symposium Conducted at the Meeting of epiSTEME*. Vol. 6. secure.hbcse.tifr.res.in. https://secure.hbcse.tifr.res.in/epi6/papers/Review-talks/epiSTEME6_ReviewTalk_David%20Hestenes.pdf.

Horwitz, P. (1995). Linking models to data: Hypermodels for science education. *The High School Journal, 79*(2), 148–156.

Horwitz, P., & Christie, M. A. (2000). Computer-based manipulatives for teaching scientifc reasoning: An example. In M. J. Jacobson & R. B. Kosma (Eds.), *Innovations in science and mathematics education*. Routledge.

Hutcheson, G. D. (2011). Ordinary least-squares regression. *L. Moutinho and GD Hutcheson, The SAGE Dictionary of Quantitative Management Research*, 224–228.

Looi, C. K., & Teh, L. W. (Eds.). (2015). *Scaling educational innovations*. Springer.

Mayer, R. E. (2016). The role of metacognition in STEM games and simulations. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment* (pp. 207–229). Routledge.

McElroy-Brown, K., & Reichsman, F. (2019). Genetics with dragons: Using an online learning environment to help students achieve a multilevel understanding of genetics. *Science Scope, 42*(8), 62–69.

Morrison, G. R., Ross, S. M., & Lowther, D. L. (2009). Technology as a change agent in the classroom. In L. Moller, J. B. Huett, & D. M. Harvey (Eds.), *Learning and instructional technologies for the 21st Century: Visions of the future* (pp. 1–23). Springer US.

Mutch-Jones, K., Boulden, D. C., Gasca, S., Lord, T., Wiebe, E., & Reichsman, F. (2021). Co-teaching with an immersive digital game: Supporting teacher-game instructional partnerships. *Educational Technology Research and Development: ETR & D, 69*(3), 1453–1475.

Noemí, P.-M., & Máximo, S. H. (2014). Educational games for learning. *Universal Journal of Educational Research, 2*(3), 230–238.

Pedro, M. A. S., Gobert, J. D., & Betts, C. G. (2014). Towards scalable assessment of performance-based skills: Generalizing a detector of systematic science Inquiry to a simulation with a complex structure. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent tutoring systems* (pp. 591–600). Springer International Publishing.

Pellegrino, J. W. (2014). A learning sciences perspective on the design and use of assessment in education. *The Cambridge Handbook of the Learning Sciences, 2*, 233–252.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*, 119.

Perkins, K. (2020). Transforming STEM learning at scale: PhET interactive simulations. *Childhood Education, 96*(4), 42–49.

Rachmatullah, A., Reichsman, F., Lord, T., Dorsey, C., Mott, B., Lester, J., & Wiebe, E. (2021). Modeling secondary students' genetics learning in a game-based environment: Integrating the expectancy-value theory of achievement motivation and flow theory. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09896-8

Riopel, M., Nenciovici, L., Potvin, P., Chastenay, P., Charland, P., Sarrasin, J. B., & Masson, S. (2019). Impact of serious games on science learning achievement compared with more conventional instruction: An overview and a meta-analysis. *Studies in Science Education, 55*(2), 169–214.

Roesken-Winter, B., Hoyles, C., & Blömeke, S. (2015). Evidence-based CPD: Scaling up sustainable interventions. *ZDM: the International Journal on Mathematics Education, 47*(1), 1–12.

Ross, S. M., & Morrison, J. R. (2021). Achieving better educational practices through research evidence: A critical analysis and case illustration of benefits and challenges. *ECNU Review of Education, 4*(1), 108–127.

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). *Helping students make sense of the world using next generation science and engineering practices*. NSTA Press.

Shute, V. J., and M. Ventura. (2015). Stealth Assessment. *The SAGE Encyclopedia of Educational*. http://myweb.fsu.edu/vshute/pdf/sa_handbook.pdf.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction, 55*(2), 503–524.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. MIT Press.

Steinkamp, M. W., & Maehr, M. L. (1983). Affect, ability, and science achievement: A quantitative synthesis of correlational research. *Review of Educational Research, 53*(3), 369–396.

Thalheimer, W., & Cook, S. (2002). How to calculate effect sizes from published research: A simplified methodology. *Work-Learning Research, 1*, 1–9.

Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71*(2), 181–194.

White, B. Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition and Instruction, 10*(1), 1–100.

Wright, T. L. (2012). The Effects of Modeling Instruction on High School Physics Academic Achievement. In *ProQuest LLC*. ProQuest LLC. http://www.proquest.com/en-US/products/dissertations/individuals.shtml.

Wu, Z., Mott, B. W., Min, W., Taylor, R., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E. N., and Lester, J. C. (2019). Predicting challenge outcomes for students in a digital game for learning genetics. In *EDM (Workshops)*, (pp. 51–59). ceur-ws.org.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Paul Horwitz[1] · Frieda Reichsman[1] · Trudi Lord[1] · Chad Dorsey[1] · Eric Wiebe[2] · James Lester[2]

✉ Paul Horwitz
phorwitz@concord.org

[1] The Concord Consortium, Concord, MA, USA

[2] North Carolina State University, Raleigh, NC, USA