Investigating the Feasibility of Using Assessment and Explanatory Feedback in Desktop Virtual Reality Simulations

Guido Makransky[1], Richard Mayer[2], Anne Nøremølle[3], Ainara Lopez Cordoba[4], Jakob Wandall[5], & Mads Bonde[4,6]

[1] Department of Psychology, University of Copenhagen, Copenhagen, Denmark

[2]Psychological and Brain Sciences, University of California Santa Barbara, CA

[3]Department of Cellular and Molecular Medicine, University of Copenhagen, Denmark

[4]Labster, Copenhagen, Denmark

[5]Department of Education, University of Aarhus, Aarhus, Denmark

[6] Department of Drug Design and Pharmacology, University of Copenhagen, Denmark

Corresponding Author: Guido Makransky, University of Copenhagen

Address: Øster Farimagsgade 2A, 1353 Copenhagen K, Denmark

Abstract

There is great potential in making assessment and learning complementary. In this study, we investigated the feasibility of developing a desktop virtual reality (VR) laboratory simulation on the topic of genetics, with integrated assessment using multiple choice questions based on Item Response Theory (IRT) and feedback based on the Cognitive Theory of Multimedia Learning. A pre-test post-test design was used to investigate three research questions related to: 1) students' perceptions of assessment in the form of MC questions within the VR genetics simulation; 2) the fit of the MC questions to the assumptions of the Partial Credit Model (PCM) within the framework of IRT; and 3) if there was a significant increase in intrinsic motivation, self-efficacy, and transfer from pre- to post-test after using the VR genetics simulation as a classroom learning activity. The sample consisted of 208 undergraduate students taking a medical genetics course. The results showed that assessment items in the form of gamified multiple-choice questions were perceived by 97% of the students to lead to higher levels of understanding, and only 8% thought that they made the simulation more boring. Items within a simulation were found to fit the PCM and the results showed that the sample had a small significant increase in intrinsic motivation and self-efficacy, and a large significant increase in transfer following the genetics simulation. It was possible to develop assessments for online educational material and retain the relevance and connectedness of informal assessment while simultaneously serving the communicative and credibility-based functions of formal assessment, which is a great challenge facing education today.

Key words: Simulations; desktop virtual reality, assessment; explanatory feedback; item response theory; cognitive theory of multimedia learning; retrieval practice

1. Introduction

When used correctly, accurate assessment and reliable feedback can motivate and encourage student learning, as well as create increased learning (Stiggins, Arter, Chappuis, & Chappuis, 2004). In many contexts however, assessment of student learning has primarily been conducted though large-scale standardized achievement testing since the 1960s (Shute, Leighton, Jang, & Chu, 2016). These assessments have the advantages of maintaining a high level of standardization, psychometric rigor, and legal defensibility; however they rarely lead to higher levels of motivational or increased learning, and have been criticized on a number of fronts (Sackett, Borneman, & Connelly, 2008). One is the concern that assessment can distract from the learning activity and disrupt the student's in the learning process (Shute & Ke, 2012). There is evidence that standardized assessments have a low level of ecological validity since assessment is typically conducted outside of the learning activity (Groth-Marnat, 2000). Furthermore, there is research showing that high-stakes standardized testing pressures teachers into teaching to the test, thus narrowing the curricula and undermining education (Au, 2007).

Informal formative assessments, on the other hand, provide a low-stakes yet ecologically strong method for collecting valuable information about students that can be used to provide targeted feedback or to guide adaptation of instruction to improve student learning (Black & Wiliam, 1998; Shavelson, et al., 2008). Furthermore, research shows that retrieval practice which occurs with informal, low-stakes classroom assessment can be a powerful tool for producing large gains in long-term retention as compared to repeated studying (Adescope, Trevisan & Sundarajan, 2017). Nonetheless, the low level of reliability and validity of this information limits its psychometric defensibility and use (Shute et al., 2016).

Advances in technology and the science of learning are causing considerable changes in the landscape and possibilities within education. These advances are forcing testing specialists to

consider significant changes in how large-scale and classroom assessments are conceptualized, designed, administered, and interpreted (Schraw et al., 2012; Shute & Becker, 2010). When implemented in simulations and games, innovative assessment methods have the potential of incorporating the psychometric rigor from large-scale standardized achievement tests with the ecological validity of the classroom settings (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015). A report by the National Research Council (NRC) concludes that simulations and games hold great promise for measuring important aspects of science learning that have been difficult to assess in formal large scale (as well as classroom) contexts (NRC, 2011). Informal formative assessments are crucial to individualized online learning because they can provide immediate feedback so students can self-assess and self-improve through retrieval practice activities (Tsai et al., 2015). Previous studies suggest that learning, motivation, and performance are improved in an online learning environment when online formative assessment is used (Gardner, Sheridan, & White, 2002; Henly, 2003; Khan, Davies, & Gupta, 2001). However, there are not many examples of successful applications of assessment that are effective without diminishing learning enjoyment (Shute et al., 2009; Wilson & Sloane, 2000). Furthermore, there is a gap in this research as few studies have shown that informal formative assessment could live up to the rigorous psychometric criteria that are requirements in formal assessment settings. This has caused leading experts to question if it is possible to maintain the relevance and connectedness that typically comes with informal assessment, while simultaneously maintaining the communicative and credibility-based functions of formal assessment (Mislevy, 2016).

The main objective of this study is to address this challenge by assessing the feasibility of developing a desktop virtual reality (VR) laboratory simulation in the field of genetics, with integrated formative assessment and feedback based on the Cognitive Theory of Multimedia Learning (CTML; Mayer, 2009). Using a pre-test, post-test design we will investigate whether it is

possible to use assessment to accurately measure student knowledge within a simulation using item response theory (IRT; Embretson & Reise, 2000), while simultaneously increasing students' intrinsic motivation, self-efficacy, and transfer, and without disturbing the learning process.

2. Literature review

2.1 Different types of assessment in education

Educational assessment can be divided into summative and formative assessment (Mayer, 2011; Pellegrino, Chudowsky, & Glaser, 2001). While summative assessments are conducted after a learning activity is complete, formative assessments are used during the learning process to provide feedback during the learning activity (Sadler, 1998). Furthermore, assessments can be formal or informal. Formal assessments are typically pre-planned data-based tests, whereas informal assessments are forms of assessment that can easily be incorporated into classroom activities by teachers or within games and simulations (Nitko, 1996).

2.2 The feedback principle from the cognitive theory of multimedia learning

The theoretical reference for the current study is the feedback principle from the CTML (Johnson & Priest, 2014). The feedback principle states that novice students learn better with explanatory feedback (providing students with principle-based explanations of why his or her answer was correct or incorrect) than with corrective feedback alone (simply informing the student whether their response was correct or not). In the learning process, feedback helps students' select relevant information, organize the information into a coherent mental representation, and integrate the information with existing knowledge stored in long-term memory. Feedback serves the purpose of modifying students' thinking or behavior to improve learning (Shute, 2008). Giving feedback to students has been demonstrated to be effective for the learning process (Hattie, 2009) and for correcting mistakes and misconceptions (Dantas & Kemm, 2008; Marcus, Ben-Naim, & Bain, 2011; Polly et al., 2014). As compared to corrective feedback, explanatory feedback guides the

student in selecting the appropriate information and consequently reduces the amount of extraneous processing necessary (Moreno, 2004). Several reviews and meta-analyses support this premise, and have found that explanatory feedback has much higher impact on achievement compared to corrective feedback alone (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kluger & DeNisi, 1996). Kluger and DeNisis (1996) also found that feedback had a negative impact on performance in one-third of the studies, supporting the notion that feedback should prompt essential processing in order to be effective (Johnson & Priest, 2014).

There are a few examples of studies which have underscored the importance of providing feedback in laboratory settings or with VR simulations. For instance, in a study conducted in a laboratory setting, Butler and Roediger (2008) found that receiving feedback provided a benefit over simply completing multiple choice (MC) questions on a final cued-recall test when students had to study historical passages. In a follow-up study, Butler, Karpicke, and Roediger (2008) found that feedback approximately doubled student performance over testing without feedback on a 5-minute delayed cued-recall test, when subject matter was general knowledge. Kruglikova and colleagues (2009) examined the impact of concurrent external feedback on learning curves and skill retention using a VR colonoscopy simulator. They found that a feedback group reached a level of proficiency on the simulator significantly faster and performed significantly better than the control group during a transfer test. With regard to using simulations for laboratory exercises, it was posited in Ronen and Eliahu (2000) that the main advantage of using simulations as a supplement to laboratory instruction is that simulations can provide immediate feedback to students about errors, so that less instructor intervention is required.

In a meta-analysis, Merchant, Goetz, Cifuentes, Keeney-Kennicutt, and Davis (2014) reported an insufficient number of studies involving the role of feedback in virtual simulations, although there is some evidence that instructor feedback improves learning of surgery skills in a VR

simulator (Strandbygaard, et al., 2013). Overall, this literature suggests the importance of providing students with elaborative feedback that includes explanations and examples.

2.3 Retrieval practice

In addition to explanatory feedback, retrieval practice--which is the act of having students recall information rather than providing it to them--has been shown to be a robust method for enhancing long-term memory performance (Dunlos, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Butler, 2011). This concept is also commonly referred to as testing effect, or retrieval-based learning. This research is relevant for the current study because the integrated assessment in the desktop VR genetics simulation requires students to retrieve information from long-term memory, which strengthen student's retention of this information (Roelle & Berthold, 2017). A recent meta-analysis concluded that there is an "overwhelming amount of evidence" supporting the learning benefits of retrieval practice; and that these benefits are maintained across a wide array of educational levels, settings, and retrieval procedures/testing formats (Adescope, Trevisan & Sundarajan, 2017). For instance, in a longitudinal study McDaniel et al. (2011) investigated whether low-stakes testing could be used to foster students' learning of course content, and they found that the act of being quizzed helped students learn better. Several other studies in the framework of retrieval practice have found that spending time giving students quizzes results in significant learning gains as compared to classic reviewing activities in classroom settings (Cranney et al., 2009; Larsen et al., 2009; McDermott et al., 2014). In a study to determine where to most effectively place retrieval practice with correct-answer feedback, Uner and Roediger (2017) found that it was the amount of retrieval practice and not the location that was most important. In their study, students in the intervention condition were given the task of reading a 40-page textbook chapter on biology, and were given practice tests after each section or after the whole chapter. They found that students recalled more information than two control groups who either read the chapter

once, or read and re-read the chapter. They also found that it did not matter if the questions occurred after each section or after the whole chapter. However, answering questions twice increased recall performance.

2.4 Desktop virtual reality laboratory simulations

VR can be defined as "a computer-mediated simulation that is three-dimensional, multisensory, and interactive, so that the user's experience is *as if* inhabiting and acting within an external environment" (Burbules, 2006, p. 37). Several different types of VR systems exist including cave automatic virtual environment (CAVE), head mounted displays (HMD), and desktop VR. In this study we use desktop VR, which enables the user to interact with a virtual environment displayed on a traditional computer monitor using a keyboard and mouse (Lee and Wong 2014; Lee et al. 2010). Cummings and Bailenson (2016) reason that the immersiveness of VR is dependent on system configurations or specifications. Accordingly, immersion can be regarded as an objective measure of the extent to which the VR system presents a vivid virtual environment. By this account, VR systems such as desktop VR, can be characterized as lower in immersion than CAVE or HMD environments.

Desktop VR laboratory simulations are designed to reproduce or amplify real world laboratory environments by allowing users to manipulate objects and parameters in a virtual lab. Simulations are becoming increasingly used educational tools because they provide a highly interactive environment in which learners become active participants in a computer-generated world (Jones, 2018; Bonde et al., 2014). Also, simulations enable learners to gain knowledge through experiencing an activity within a virtual environment that facilitates a high level of psychological presence (Ai-Lim Lee, Wong, & Fung, 2010; Makransky et al., 2017, 2018, 2019b). Meta-analyses have found that deliberate practice through simulations is an efficient way to acquire expertise (McGaghie, Issenberg, Cohen, Barsuk, & Wayne, 2011; McGaghie, Issenberg, Petrusa, & Scalese, 2010). Furthermore, there is evidence that simulations lead to enhanced learning (Bayraktar, 2000; Bonde et al., 2014; Polly, Marcus, Maguire, Belinson, & Velan, 2015; Rutten, Van Joolingen, &

Van Der Veen, 2012), and can lead to increased motivation and belief in one's ability to successfully complete a task (Makransky et al., 2019a, c; Makransky & Pedersen, 2019; Meyer et al., 2019).

2.5 Assessment of knowledge with item response theory

There are many formats which an assessment and feedback system can use within a learning simulation (Shute, 2011; Almond et al., 2015). In this study we used IRT for assessment based on MC questions that were integrated into the simulation. IRT has often been used for large-scale testing and has been suggested as a promising application for assessment of learning within simulations and games (NRC, 2011). The partial credit model (PCM; Masters, 1982) within the framework of IRT is appropriate for contexts where the objective is to measure a students' ability, and items are scored with successive integers. This model lends itself well to the current context because students get full credit when answering an item in the genetics simulation correctly on their first attempt (3 points), and one fewer point upon each extra attempt. The PCM describes the association between a student's level of an underlying trait (person parameter; in this case knowledge of medical genetics), and the probability of a specific item response in the simulation. The PCM provides the probability that a student with a given level of knowledge will be able to respond correctly to each item after the first, second, third or fourth or more attempts.

3. Research questions and background

There has been debate in educational literature regarding whether assessment and learning are contradictory or complementary (Boud, 1995). One criticism of assessment in learning environments is that it can distract from the learning activity and disrupt the student's in the learning process (Shute, 2011), which in turn can lead to lower levels of student motivation, self-efficacy, and ultimately lower learning and transfer. One solution to this problem has been to create embedded assessment methods that provide a measure of student progress and performance in the

instructional materials so they are virtually indistinguishable from the day-to-day classroom activities (Wilson & Sloane, 2000). For these solutions to be viable it is important to determine if: (1) They can be implemented without distracting students sense of presence in the learning process (Makransky & Lilleholt, 2018). (2) They have an acceptable level of reliability and validity to support their psychometric defensibility and use (Shute et al., 2016). (3) They can be implemented while simultaneously increasing students' intrinsic motivation, self-efficacy and learning (which are the ultimate goals of a learning intervention). The main objective of this paper is to assess the feasibility of developing a VR genetics simulation with integrated assessment using IRT and feedback based on the feedback principle from CTML; and to assess whether the implementation of the simulation in a classroom setting lives up to the three criteria outlined above. Therefore, we investigate three research questions based on each of these criteria and outline their relevance below.

Given that one criteria for successfully integrating assessment into a VR learning activity is that students do not perceive that it disrupts them in the learning process (Shute, 2011), it is relevant to investigate students' perceptions of the assessment. In this study, we investigate if the assessment in the simulation increased students' motivation, and if it negatively influenced their perceived understanding of the material**.** The integrated assessment using IRT proposed in this study uses MC questions with explanatory feedback within the VR simulation in order to stimulate retrieval practice. Therefore, it is important to investigate if the introduction of MC questions results in negative perceived outcomes such as lack of motivation or a lower level of understanding.

Research question 1: What are students' perceptions of the assessment in the form of MC questions within the VR genetics simulation with respect to understanding (1a) and motivation (1b)?

The second important criteria when implementing MC questions to assess knowledge within a learning simulation is whether these questions lead to accurate assessment of student knowledge. Accurate assessment requires drawing inferences in real time from diverse behaviors and performances. One challenge is that most conventional psychometric methods are not well suited for such modeling and interpretation (NRC, 2011). IRT has been suggested to overcome these limitations (NRC, 2011); however, the validity of using IRT in a complex simulation environment must be tested prior to its implementation. Although there is abundant evidence for the value of IRT within educational measurement, we have not found any literature investigating the feasibility of this methodology for assessment within VR science simulations. An important difference between the genetics simulation in this study and a traditional assessment setting is that assessment questions in the genetics simulation are an integral part of the larger learning environment and do not exist in isolation. This important difference could potentially have a negative impact on the fundamental assumptions of IRT, including the assumption of local independence (items are assumed to be conditionally independent given an individual's standing on the latent construct), and unidimensionality (the items measure only one latent construct). Thus, a fundamental research question that has not been satisfactorily answered within the field of VR simulations is whether IRT is suitable for assessment within virtual simulations. This can be tested by assessing the internal validity of the items by testing them against the assumptions of the model. The PCM, which is one model within the framework of IRT, is used in this study. This model is appropriate because students get multiple chances to respond to each question and gain a lower score with each additional attempt. This leads to the second research question.

Research question 2: Do the MC items inserted in the VR genetics simulation fit the assumptions of the Partial Credit Model (PCM) within the framework of IRT?

Since the goal of any educational activity is to increase students' knowledge and skills (Pellegrino & Hilton, 2012; Perkins, 1994) in addition to non-cognitive outcomes such as intrinsic motivation and self-efficacy (Kraiger, Ford, & Salas, 1993), the ultimate test of the activity should be to investigate if it has the expected positive educational impact on students. Intrinsic motivation and self-efficacy have commonly been investigated in educational research and there is abundant evidence that they play an essential role in academic success (Ryan & Deci, 2000, 2016; Schunk & DiBenedetto, 2016; Zimmermann, 2000). *Self-efficacy is defined as one's perceived capabilities for learning or performing actions at designated levels* (Schunk & DiBenedetto, 2016 p. 34). Richardson et al. (2012) reported a correlation of 0.31 between grade point average (GPA) and academic self-efficacy (general perceptions of academic capability), and a correlation of 0.59 between GPA and performance self-efficacy (perceptions of academic performance capability) in their systematic review and meta-analysis of the psychological correlates of university students' academic performance. Intrinsic motivation is defined as interest or enjoyment in the task itself that comes from within the individual, rather than being influenced through factors outside of the individual (Ryan & Deci, 2000). Intrinsic motivation is relevant because students who are intrinsically motivated are more likely to engage in an educational task willingly and work to improve their skills, which increases their capabilities (Wigfield et al., 2004). Richardson et al. (2012) found a correlation of 0.17 between academic intrinsic motivation (i.e. self-motivation for and enjoyment of academic learning and tasks) and GPA.

Therefore, as addressed in research question 3, one way to test if the design of the VR genetics simulation is successful is to administer it as an intervention and assess if students increased from pre- to post-test on three outcome variables: intrinsic motivation to study medical genetics; self-efficacy for learning medical genetics; and transfer (ability to apply knowledge in a new situation; Mayer, 2009).

Research question 3: Is there a significant increase in intrinsic motivation, self-efficacy, and transfer from pre- to post-test after using the VR genetics simulation as a classroom learning activity?

4. Research design

4.1 Participants

The sample consisted of 208 first year medical students (65% female) who were participating in a medical genetics course within the undergraduate medical program at the University of Copenhagen. The students were given the VR genetics simulation toward the end of the medical genetics course as an integrated part of the education. The study was conducted according to the Helsinki Declaration (Williams, 2008) for safeguarding the ethical treatment of human subjects. In accord with the Helsinki Declaration, participation in the experiment was voluntary. and all participants were given written and oral information describing the research aims and the experiment before participation. Data was treated anonymously. The study protocol was submitted to the Regional Committees on Health Research Ethics for Southern Denmark, who indicated that no written consent was required by Danish law. A total of 300 students used the simulation as part of the course activities; however, we only used data from the 208 students who took part in the pre-test and post-test and gave their written consent to participate in the study.

4.2 Materials

4.2.1 Scales used in the study

Intrinsic motivation in genetics was assessed with 5 questions adapted from the Interest/Enjoyment scale from the Intrinsic Motivation Inventory (Deci, Eghrari, Patrick, & Leone, 1994; e.g., “I enjoy working with medical genetics very much”). Self-efficacy for learning medical genetics was measured with 5 questions adapted from the Motivated Strategies for Learning Questionnaire (Pintrich, Smith, Garcia, & McKeachie, 1991; e.g., “I am confident that I can

understand the basic concepts of medical genetics"). Students indicated their responses to the intrinsic motivation and self-efficacy items using a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree). These scales are two of the most widely used measures of intrinsic motivation and self-efficacy, respectively, and they have been used in their current form in educational research (Makransky et al., 2019c; Thisgaard & Makransky, 2017). Transfer was assessed using 20 MC items intended to assess general competencies within medical genetics (e.g., Individuals with three copies of autosomal chromosomes normally do not survive (except trisomies of e.g. 13, 18 and 21), whereas individuals with an extra X chromosome have relatively mild phenotypes. Why? A) The extra X chromosomes will be inactivated; B) The X chromosome carries fewer genes than the autosomes; C) The X chromosome only determines the sex of the individual; D) All of the autosomes are acrocentric). These items were different from the MC items that were integrated in the simulation. They were intended to measure the degree to which students had gained a deep understanding of the topic and could use the knowledge they had gained to solve more general problems. All items are listed in the Appendix.

4.2.2 Desktop VR genetics simulation based with integrated assessment and explanatory feedback

In this study we address the feasibility of including assessment and explanatory feedback in simulations by developing a simulation for learning medical genetics (genetics simulation). The genetics simulation was developed by the educational technology company Labster in cooperation with the department of Cellular and Molecular Medicine at the University of Copenhagen. The genetics simulation was designed to facilitate learning within the field of genetics at a university level by allowing the user to virtually work through the relevant procedures in a lab using and interacting with state-of-the art lab equipment, through an inquiry-based learning approach. In this simulation students are introduced to a young pregnant couple whose fetus may suffer from a

syndrome caused by a chromosomal abnormality. The students are asked to make a genome-wide molecular cytogenetics analysis of the fetal and parental DNA and karyotypes in the virtual laboratory. Then they have to practice communicating their conclusions to the couple. The simulation consists of three main parts: 1) An introduction of the case story, where students are presented with a case-based challenge aimed at engaging them emotionally and providing them with essential knowledge and concepts of the material (a screenshot of this part of the simulation is shown in the top panel of Figure 1). 2) A series of laboratory experiments that the students have to complete in order to respond to research question associated with key learning goals (see bottom panel of Figure 1). 3) A final task where the students have to interpret their results and give genetic counseling. This section is designed to stimulate reflective thinking and deep learning (for a short video that gives an overview of the simulation see Labster, 2019).

The simulation was developed using inspiration from the NRC' (2011) guidelines and had the following general goals: 1) Increase student motivation by having them experience the excitement and increase their interest to learn about genetics phenomena with a realistic case. 2) To increase conceptual understanding through retrieval practice of key terms with the MC questions. 3) To increase science processing skills by asking students to observe, manipulate, explore, test, and make sense of relevant material in a virtual environment. 4) To understand the nature of science by having students be a virtual doctor who has to diagnose and communicate the lab results to a patient and their family. 5) To give students the opportunity to experience what it is like to use genetics in a realistic setting to solve a realistic problem.

Student learning is prompted through their interaction with the gamified laboratory environment, and by obtaining explanatory feedback to MC questions in which the accuracy of their results provides a running score (see the top panel in Figure 2 for a screen shot of an integrated MC

question in the simulation). In this study the desktop VR simulation was administered on standard computers in a computer lab containing 25 computers.

4.2.3 Explanatory feedback in the genetics simulation

In the genetics simulation, students receive immediate explanatory feedback following their response to the MC questions in which four response options appear randomly (see the bottom panel of Figure 2 for a screenshot of the explanatory feedback in the simulation). If the response is correct, an explanation is provided and the student gets full credit. If the response is incorrect, participants are provided with the opportunity to access the background theory related to the question to ensure that students can always read relevant information about the question they are trying to answer, thereby gaining a deeper understanding of the content (Mayer, 2008). Then the response options are randomly shuffled, and the student is asked to try again. The logic for this design is based on findings that feedback should support intentional and effortful processing of information in order to have an impact on student understanding (Moreno & Valdez, 2005). Therefore, mixing the response options requires students to cognitively engage in the content, thereby increasing generative processing. The lack of randomization could encourage students to use a "trial and error" strategy rather than engaging in more effortful processing in order to actively make sense of the material. This process continues until the correct response is given. The genetics simulation included 41 MC questions with four response options for each question. The number of questions was chosen by trying to reach a compromise between the desire to have many questions to obtain an accurate assessment of student knowledge, while maintaining the value of giving students an authentic virtual laboratory learning experience. To maintain that balance, a total of 41 MC questions over the course of a two-hour learning session allowed for students to encounter a question approximately every three minutes.

4.3. Data collection and analysis

First, students were given a general introduction to the desktop VR simulation including important information about how to use the simulation, the learning objectives, as well as other relevant instructions. This was followed by the pre-test post-test design consisting of a 20-minute pre-test to determine students' baseline on scales that measured intrinsic motivation, self-efficacy, and transfer in the field of genetics (see Appendix for the source and list of items); a 2-hour session where the students used the VR genetics simulation; and a 30-minute post-test to reassess student's motivation, self-efficacy, and transfer. The pre-test also included questions about demographic characteristics, and a question asking students if they accepted that their data would be used anonymously in a research study. Three teachers were available to assist the students during this session, and students were allowed to work and talk together (except during the pre- and post-tests). In general, students did not have prior experience with this type of simulation or desktop VR.

Research question 1 was investigated by having students respond to two questions regarding the influence of the MC questions for their perceived learning and engagement following their participation in the simulation on a four point Likert scale ranging from strongly disagree to strongly agree. The questions were: 1) the multiple choice items in the simulation improved my understanding of the material; and 2) the multiple choice items in the simulation made the simulation boring.

Research question 2 was investigated by assessing the fit of 41 MC items within the genetics simulation to the PCM using the RUMM 2030 program (Andrich, Sheridan, & Luo, 2010). RUMM 2030 is a commonly used statistical program for conducting PCM analyses. Fit to the PCM supports the construct validity of the scale, and is an indication that the items are functioning optimally within the simulation context. A variety of fit statistics that are typically used to assess the fit of a scale to the PCM were used. These included assessing: item fit, the unidimensionality of the scale, local dependence between all possible pairs of items, and the internal reliability in the form of Cronbach's alpha and the person separation index (PSI). A thorough description of these fit indices, and the

criteria that are used are only described briefly below and can be found in more detail in (Makransky et al., 2014; Pallant & Tennant, 2007; Tennant & Conaghan, 2007).

The fit of each item according to the expectations of the model was evaluated based on fit residuals larger than ± 2.5 (Pallant & Tennant, 2007). Unidimensionality was evaluated according to a formal test proposed by Smith (2002). This test uses the first residual factor in a principal components analysis (of residuals) to determine two groups of items: those with positive and those with negative residuals. Each set of items is then used to calculate an independent trait estimate for each person in the sample. If items form a unidimensional scale, it is expected that the person estimates from the two item subsets should be similar. An independent samples t-test is used to determine whether there is a significant difference between the two person estimates. This is conducted for each person with the expectation that the percentage of tests lying outside the range of -1.96 to 1.96 should not exceed 5%. Local dependence (which can indicate redundancy among items in a scale) was assessed by investigating if residual correlations between the items in the scale were larger than a critical value using the criteria outlined in (Makransky et al., 2017).

Research question 3 was investigated by conducting paired samples t-tests to assess the impact of the genetics simulation on intrinsic motivation to study genetics, self-efficacy in genetics, and transfer in genetics.

5. Results

5.1 Impact of assessment and explanatory feedback using the multiple-choice questions on perceived learning and engagement

Research question 1 was investigated by asking the students to report their level of perceived learning and engagement on a four-point scale. The responses to the two questions are presented in Figure 3. In general, the students were positive about the use of the MC questions in the genetics simulation. The results show that 97% of the students either agreed or strongly agreed with the

statement “the multiple choice items in the simulation improved my understanding of the material”. Furthermore, only 8% of the students agreed with the statement “the multiple choice items in the simulation made the simulation more boring”.

5.2 Construct validity of implementing assessment in the genetics simulation

Research question 2, the construct validity of implementing assessment based on the MC questions, was assessed by investigating the fit of the items to the PCM. The MC questions in the genetics simulation had a good general fit to the PCM, χ^2 *(123)* = 127.54, p = 0.37. There were also no items with significant chi-squared statistics, indicating that all of the items measured the latent trait of knowledge in medical genetics in a consistent way. The scale was also found to be unidimensional according to a formal test proposed by Smith (2002), meaning that the questions all measured the same latent trait. There were five item pairs that had positive residuals over the critical value, indicating redundancy or local dependence. Finally, the internal consistency as measured by the PSI was 0.64, and Cronbach’s alpha was 0.77.

5.3 Was there an increase in intrinsic motivation, self-efficacy, and transfer?

Table 1 provides reliability values for the measures in the study. Intrinsic motivation had Cronbach’s alpha reliability values of 0.86, and 0.78 in the pre- and post-test respectively. Self-efficacy had values of 0.85 for the pre- and post-test. Finally, the transfer test had Cronbach’s alpha values of 0.74 and 0.64 for the pre- and post-test respectively. Therefore, all scales used in the pre/post-test had acceptable reliability for the sample, with the exception of the transfer test in the post-test session which was slightly lower than an acceptable value of 0.7 (DeVellis, 1991). Table 1 also shows the results of the paired samples t-tests used to investigate research question 3: was there a significant increase in intrinsic motivation, self-efficacy, and transfer from the pre- to the post-test?

There was a significant small effect size increase in intrinsic motivation from the pre-test (M = 3.88, SD = 0.57) to the post-test (M = 3.96, SD = 0.52) $t(207) = 2.50$, $p = .013$; $d = 0.15$. There was also a significant small effect size improvement in self-efficacy from the pre-test (M = 3.28, SD = 0.64) to the post-test (M = 3.48, SD = 0.61) $t(207) = 6.08$, $p < .001$; $d = 0.32$. Finally, there was a significant large improvement in transfer from the pre-test (M = 9.25, SD = 2.72) to the post-test (M = 14.78, SD = 2.91) $t(207) = 30.19$, $p < .001$; $d = 1.97$.

6. Discussion

6.1 Empirical contribution

The results suggest that the use of MC questions and measurement using IRT combined with explanatory feedback can be successfully implemented in a simulation. Most of the students in the study (97%) agreed with the statement that the MC questions with explanatory feedback used in the simulation increased their understanding of the material. Furthermore, the majority (92%) disagreed with the statement that the MC questions with explanatory feedback made the simulation more boring. Therefore, the results indicated that the MC questions with explanatory feedback did not seem to disrupt the majority of the students' in the learning process; rather they suggest that they helped facilitate the learning process. It should be noted, however, that a minority of students reported being negative about the use of the MC questions. The results also provide evidence of the construct validity of the items within the genetics simulation based a good fit to the PCM within the framework of IRT (Masters, 1989). The combined findings suggest that assessment based on gamified MC items could introduce the ecological validity typically associated with informal classroom assessments within a simulation context, while maintaining a high level of reliability and validity typically associated with formal assessment. Furthermore, significant increases in intrinsic motivation for genetics, self-efficacy in genetics, and deep learning measured through a scale measuring transfer, were observed from prior to after using the simulation.

6.2 Theoretical implications

One of the main assumptions underlying innovative assessment methods such as stealth assessment (Shute, 2011) or Bayesian nets (Almond et al., 2015), is that assessment should be seamlessly introduced in educational material without interrupting students' in the learning process. Inserted MC questions are often seen as disruptive for learning because they disengage students from the material. In contrast, the results suggest that the MC questions with explanatory feedback enhance the learning process and sustain students' motivation. This is in line with the abundant research supporting the learning impact of using quiz questions as retrieval practice (Brown, Roediger, & McDaniel, 2014; Roediger & Butler, 2011). It is, however, important to note that the results would probably not generalize to all learning environments. Settings that provide little guidance (De Jong & Van Jooligen, 1998) or where students learn through exploration or through trial and error (Kapur, 2008) may benefit more from the use of less intrusive assessment methodologies like stealth assessment or Bayesian nets, as these methodologies provide the opportunity to assess students in a more natural way.

6.3 Practical implications

The results regarding how the MC questions impacted students' perceived learning and their engagement in the genetics simulation showed that students were in general very positive about the use of the gamified MC questions. This finding is particularly relevant because it is a great deal easier to validly assess constructs such as knowledge by posing concrete MC questions, rather than by assessing students' behavior in the e-learning environment. Furthermore, it is relatively inexpensive to develop and score MC questions as compared to open-ended questions (NRC, 2011). The use of MC questions based on IRT requires students to actively respond to the questions, thereby potentially disrupting the sense of presence in the learning process. In the genetics simulation questions were carefully developed as an integrated part of the simulation,

where students learn through answering the questions. Students are rewarded based on the number of times it takes them to answer correctly; and all students are given an explanation of the correct response before moving on in the simulation, thereby ensuring that students gain fundamental knowledge of the topic. The results support the large body of literature that has found formative assessment and immediate feedback to improve motivation and learning. Here an important consideration is that the assessment items were implemented in a way that increases students' engagement in the simulation.

The results also provided support for the fit of the items within the genetics simulation to the PCM, which is a prerequisite for developing accurate assessment, and makes it possible to develop an adaptive simulation which adjusts to different levels of student knowledge. A significant challenge in developing assessment within simulations is that the content is strongly tied to a case story. This means that the content that is presented is often situation-specific, which can create dependencies that do not match the fundamental assumptions of IRT – specifically, local independence and unidimensionality. The finding that the MC questions in the genetics simulation fit the PCM is important because it provides evidence that it is possible to meet these fundamental assumptions.

6.4 Future research and limitations

The results of this study show the viability of using IRT and the feedback principle in developing VR simulations with integrated accurate assessment. The use of IRT also would allow instructional designers to develop adaptive simulations that adjust to different student's knowledge levels. Future research should investigate whether an adaptive version of a simulation (where the difficulty of items is adjusted adaptively based on students' responses) results in better psychological and educational outcomes. Another line of research is to investigate adaptability based on constructs other than knowledge, or in combination with knowledge.

Although the majority of the students were positive about the use of the MC questions, there was a minority who did not agree that the questions increased their understanding (3%). Furthermore, some students agreed with the statement that the MC questions made the simulation more boring (8%). A limitation in this study was that we were not able to follow-up and investigate why these students were not as positive about the MC questions as the majority. Therefore, more research is needed to investigate what might cause these differences between students. Future research could also investigate the underlying factors that lead to these results.

As described above, previous research has found that more retrieval practice is better (Uner & Roediger, 2017). However, it is likely that there is a limit to the number of MC questions that should be used to optimize motivation and transfer. The simulation used in this study had 41 MC questions over the course of a two-hour learning session, which means that students encountered a question approximately every three minutes. Future research should investigate if and how the rate of the questions influences students' learning and motivational outcomes. Furthermore, gamified MC questions that provided a running score throughout the simulation were used in this study. Future research should investigate the consequences of the gamification factor in the assessment by comparing a format both with and without gamification. Another potential limitation in this study is the finding that the transfer test had a reliability value that was below standard in the post-test. Although this is a limitation in this study, the goal of the transfer test was to assess a wide range of information in order to obtain a general measure of deep learning. By measuring a broad construct in this study, we prioritized content validity rather than internal consistency reliability.

Another limitation was that it was not possible to develop an experimental design in the given setting because the teachers did not want to disadvantage any students, inasmuch as the genetics simulation was a mandatory part of the curriculum. However, the strength of the design was that data was collected in a real classroom environment, which made it possible to assess the impact of

the genetics simulation in a realistic setting. Testing the value of educational technology in applied settings is recognized as an important supplement to classic lab experiments. For instance, Gerjets and Kirschner (2009) suggest that more research is necessary to test learning principles in natural learning settings such as in classrooms or with self-directed learning, rather than experimental laboratory conditions where there is system-controlled pacing of materials. It is possible that some of the principles are less valid or may even be reversed under more natural conditions because other variables such as adaptive strategies or collaboration become more important (Rummer, Schweppe, Scheiter, & Gerjets, 2008). Therefore, future research should use an experimental design to investigate the consequences of developing different versions of VR simulations based on different forms of feedback with the intention of testing whether there are significant educational differences between using the different versions. One example could be to develop a simulation with corrective feedback, and then to compare it to the version described in this study in order to investigate whether the positive results from this study were specifically due to the explanatory feedback alone. Future research should also investigate whether the results from this study can be generalized to other types of e-learning material, other subjects, other samples of students, and other settings.

7. Conclusions

The purpose of this study was to investigate the feasibility of developing a desktop virtual reality (VR) laboratory simulation in the field of genetics (genetics simulation) with integrated assessment and feedback. Using a pre-test, post-test design we investigated whether it was possible to use assessment to accurately measure student knowledge within a simulation using IRT, while simultaneously increasing students' intrinsic motivation, self-efficacy, and transfer, and without disturbing them in the learning process. The results showed that assessment items in the form of gamified MC questions were perceived by most students to lead to higher levels of understanding and did not lead to lack of motivation. Items within a simulation were found to fit the Partial Credit

Model (PCM), illustrating that IRT can be a viable methodology for accurate assessment with a desktop VR simulated environment. Finally, results showed that the sample had a small significant increase in intrinsic motivation and self-efficacy, and a large significant increase in transfer, following the genetics simulation. These results suggest that it is possible to develop online educational material that retains the relevance and connectedness traditionally associated with informal assessment, while simultaneously serving the communicative and credibility-based functions traditionally associated with formal assessment.

8. References

Ai-Lim Lee, E., Wong, K. W., & Fung, C. C. (2010). How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Computers & Education*, *55*(4), 1424–1442. http://doi.org/10.1016/j.compedu.2010.06.006

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian Networks in Educational Assessment*. New York: Springer.

Andrich, D., Sheridan, B., & Luo, G. (2010). *Rasch models for measurement: RUMM2030*. Perth, Australia: RUMM Laboratory.

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*(5), 258–267. http://doi.org/10.3102/0013189X07306523

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, *61*(2), 213–238. http://doi.org/10.3102/00346543061002213

Bayraktar, S. (2000). A meta-analysis on the effectiveness of computer-assisted instruction in science education. *Journal of Research on Technology in Education*, *34*(2), 173–189. http://doi.org/10.1080/15391523.2001.10782344

Black, P. J. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1): 7–73.

Bonde, M. T., Makransky, G., Wandall, J., Larsen. M. V., Morsing M., Jarmer H., Sommer M. O. (2014). Improving Biotechnology Education through Simulations and Games. *Nature Biotechnology*. 32: 7, 694-697. doi:10.1038/nbt.2955

Boud, D. (1995). Assessment and learning : contradictory or complementary ? *Assessment for Learning in Higher Education*, 35–48.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful*

learning. Cambridge, MA: Harvard Universoty Press.

Burbules, N. C. (2006). Rethinking the Virtual. In J. Weiss, J. Nolan, J. Hunsinger & P. Trifonas (Eds.), *The international handbook of virtual learning environments* (pp. 37-58). Dordrecht: Springer. doi:10.1007/978-1-4020-3803-7

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918.

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, *19*(2), 272–309. https://doi.org/10.1080/15213269.2015.1015740.

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*, 919-940.

Dantas, A. M., & Kemm, R. E. (2008). A blended approach to active learning in a physiology laboratory-based subject facilitated by an e-learning component. *Advances in Physiology Education*, *32*, 65–75. http://doi.org/10.1152/advan.00006.2007

Deci, E., Eghrari, H., Patrick, B., & Leone, D. (1994). Facilitating internalization: the self-determination theory perspective. *Journal of Personality*, *62*(1), 119–42.

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of educational research*, *68*(2), 179-201.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Thousand Oaks, CA, USA: Sage Publication, Inc.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving

students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4-58.

Embretson, S., & Reise, S. P. (2000). Item response theory for psychologists. *Books.google.com*. Retrieved from http://books.google.com/books?hl=en&lr=&ie=UTF-8&id=rYU7rsi53gQC&oi=fnd&pg=PP11&dq=ORDERED+LATENT+CLASS+MODELS+IN+NONPARAMETRIC+ITEM+RESPONSE+THEORY&ots=ZAESC95fcK&sig=Rvmsiq0-E7GYnGqw9ejqivzaKd4

Gardner, L., Sheridan, D., & White, D. (2002). A web-based learning and assessment system to support flexible education. Journal of Computer Assisted Learning, 18, 125-136.

Gerjets, P., & Kirschner, P. (2009). Learning from multimedia and hypermedia. *Technology-Enhanced Learning,* 251-272.

Groth-Marnat, G. (2000). Visions of clinical assessment: Then, now, and a brief history of the future. *Journal of Clinical Psychology, 56*(3), 349–365. http://doi.org/10.1002/(SICI)1097-4679(200003)56:3<349::AID-JCLP11>3.0.CO;2-T

Hattie J. (2009). Visible learning: A synthesis of 800+ meta-analyses on achievement. Routhedge

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. *Contemporary Sociology, 21*. http://doi.org/10.2307/2075521

Johnson, C. I., & Priest, H. A. (2014). The Feedback Principle in Multimedia Learning. In *The Cambridge handbook of multimedia learning*, 449-463.

Jones N. (2018). The Virtual Lab: Can a simulated laboratory experience provide the same benefits for students as access to a real-world lab? *Nature*. 562:S5-S7.

Kapur, M. (2008). Productive failure. *Cognition and instruction*, *26*(3), 379-424.

Khan, K. S., Davies, D. A., & Gupta, J. K. (2001). Formative self-assessment using multiple true-false questions on the Internet: feedback according to confidence about correct knowledge.

Medical Teacher, 23, 158e163.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. http://doi.org/10.1037/0033-2909.119.2.254

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*(2), 311–328. http://doi.org/10.1037/0021-9010.78.2.311

Kruglikova, I., Grantcharov, T. P., Drewes, A. M., & Funch-Jensen, P. (2010). The impact of constructive feedback on training in gastrointestinal endoscopy using high-fidelity virtual-reality simulation: a randomised controlled trial. *Gut*, *59*(2), 181-185.

Labster. (2019). Labster - Cytogenetics lab. Retrieved January 16th 2019 from https://www.youtube.com/watch?v=_VsabhW1LkA.

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomized controlled trial. *Medical Education*, *43*, 1174-1181.

Lee, E. A.-L., & Wong, K. W. (2014). Learning with desktop virtual reality: Low spatial ability learners are more positively affected. *Computers & Education*, *79*, 49–58. https://doi.org/10.1016/j.compedu.2014.07.010.

Lee, E. A.-L., Wong, K. W., & Fung, C. C. (2010). How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Computers & Education*, *55*(4), 1424–1442. https://doi.org/10.1016/j.compedu.2010.06.006.

Makransky, G., Borre-Gude S., Mayer, R. E. (2019a). Motivational and Cognitive Benefits of Training in Immersive Virtual Reality Based on Multiple Assessments. *Journal of Computer Assisted Learning*. DOI: 10.1111/jcal.12375

Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development, 66*, 1141–1164.

Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior, 72*, 276-285. http://doi.org/10.1016/j.chb.2017.02.066

Makransky, G., Petersen G. B. (2019). Investigating the process of learning with desktop virtual reality: A structural equation modeling approach. *Computers & Education*. https://doi.org/10.1016/j.compedu.2019.02.002

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019b). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, *60*, 225-236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Makransky G., Mayer R. E., Veitch N, Hood M., Christensen K. B., Gadegaard H. (2019c) Equivalence of using a desktop virtual reality science simulation at home and in class. *PLoS ONE* 14(4): e0214944. https://doi.org/10.1371/journal.pone.0214944

Makransky, G., Schnohr, C., Torsheim, T., Currie, C. (2014). Equating the HBSC Family Affluence Scale across survey years: A method to account for item parameter drift using the Rasch model. *Quality of Life Research*. 23: 10, 2899-2907. DOI: 10.1007/s11136-014-0728-2

Makransky, G., Wismer, P. Mayer, R. E. (2018). A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12335

Marcus, N., Ben-Naim, D., & Bain, M. (2011). Instructional support for teachers and guided feedback for students in an adaptive elearning environment. In *Information Technology: New*

Generations (ITNG), 2011 Eighth International Conference on (pp. 626-631). IEEE.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. http://doi.org/10.1007/BF02296272

Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Mayer, R. E. (2009). Multimedia learning (2nd ed.). New York: Cambridge University Press.

Mayer, R. E. (2011). *Applying the science of learning.* Boston: Pearson.

Meyer O.A., Omdahl M.K. & Makransky G., (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. *Computers & Education*. doi: https://doi.org/10.1016/j.compedu.2019.103603.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399-414.

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20,* 3-21.

McGaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H., & Wayne, D. B. (2011). Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Academic Medicine : *Journal of the Association of American Medical Colleges, 86*(6), 706–711. http://doi.org/10.1097/ACM.0b013e318217e119

McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003-2009. *Medical Education, 44*(1), 50–63. http://doi.org/10.1111/j.1365-2923.2009.03547.x

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers and Education, 70,* 29–40. http://doi.org/10.1016/j.compedu.2013.07.033

Moreno, R. (2004). Decreasing Cognitive Load for Novice Students: Effects of Explanatory versus Corrective Feedback in Discovery-Based Multimedia. *Instructional Science, 32*(1/2), 99–113. http://doi.org/10.1023/B:TRUC.0000021811.66966.1d

Moreno, R., & Valdez, A. (2005). Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development, 53*(3), 35–45. http://doi.org/10.1007/BF02504796

Mislevy, R., J. (2016). Postmodern Test Theory. The Gordon Comission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/rsc/pdf/mislevy_postmodern_test_theory.pdf

National Research Council. (2011). Learning Science Through Computer Games and Simulations.

Nitko, A. J. (1996). *Educational assessment of students*. Prentice-Hall Order Processing Center, PO Box 11071, Des Moines, IA 50336-1071.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1–18. http://doi.org/10.1348/014466506X96931

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* Washington, DC: National Academies Press.

Perkins, D. (1994). Do students understand understanding? *Education Digest, 59*(5), 21.

Pintrich, P. R. R., Smith, D., Garcia, T., & McKeachie, W. (1991). A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). Ann Arbor. Michigan, 48109, 1259.

Polly, P., Marcus, N., Maguire, D., Belinson, Z., & Velan, G. M. (2014). Evaluation of an adaptive virtual laboratory environment using Western Blotting for diagnosis of disease. *BMC medical education, 14*(1), 222.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353–387. https://doi.org/10.1037/a0026838

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences, 15*(1), 20-27.

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction*, *49*, 142-156.

Ronen, M., & Eliahu, M. (2000). Simulation—a bridge between theory and reality: the case of electric circuits. *Journal of computer assisted learning*, *16*(1), 14-26.

Rummer, R., Schweppe, J., Scheiter, K., & Gerjets, P. (2008). Lernen mit Multimedia: die kognitiven Grundlagen des Modalitätseffekts. *Psychologische Rundschau*, *59*(2), 98-107.

Rutten, N., Van Joolingen, W. R., & Van Der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers and Education, 58*(1), 136–153. http://doi.org/10.1016/j.compedu.2011.07.017

Ryan, R. M.; Deci, E. L. (2000). "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being". *American Psychologist*. 55 (1): 68–78. CiteSeerX 10.1.1.529.4370. doi:10.1037/0003-066X.55.1.68.

Ryan, R. M., & Deci, E. L. (2016). Facilitating and hindering motivation, learning, and well-being

in schools: Research and observations from self-determination theory. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of Motivation at School* (2nd ed; pp. 96-119). New York: Rout;edge.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-Stakes Testing in Higher Education and Employment. *American Psychologist, 64*, 215–227. http://doi.org/10.1037/0003-066X.63.4.215

Sadler, D. R. (1998). Formative assessment: revisiting the territory. Assessment in Education, 5(1), 77e84

Schraw, G., Mayrath, M. C., ClarkeMidura, J., & Robinson, D. H. (Eds.). (2012). *Technology Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. IAP.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., ... & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied measurement in education*, *21*(4), 295-314.

Schunk, D. H., & DiBenedetto, M. K. (2016). Self-Efficacy Theory in Education. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of Motivation at School* (pp. 34–54). New York, NY: Routledge.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research, 78*(1), 153–189. http://doi.org/10.3102/0034654307313795

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

Shute, V. J., & Ke, F. (2012). Assessment in Game-Based Learning. In Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives, (pp. 43–58). http://doi.org/10.1007/978-

1-4614-3546-4

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the Science of Assessment. *Educational Assessment, 21*(1), 34–59. http://doi.org/10.1080/10627197.2015.1127752

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. *Serious games: Mechanisms and effects, 2,* 295-321.

Smith Jr., E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*(2), 205–231. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12011501

Strandbygaard, J., Bjerrum, F., Maagaard, M., Winkel, P., Larsen, C. R., Ringsted, C., ... & Sorensen, J. L. (2013). Instructor feedback versus no instructor feedback on performance in a laparoscopic virtual reality simulator: a randomized trial. *Annals of surgery*, *257*(5), 839-844.

Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). Classroom assessment for student learning: doing it right--using it well. Assessment Training Institute.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care and Research, 57(8), 1358–1362. http://doi.org/10.1002/art.23108

Thisgaard M., Makransky, G. (2017). Virtual Learning Simulations in High School: Effects on Cognitive and Non-Cognitive Outcomes and Implications on the Development of STEM Academic and Career Choice. *Frontiers in Psychology.* https://doi.org/10.3389/fpsyg.2017.00805

Tsai, F.-H., Tsai, C.-C., & Lin, K.-Y. (2015). *The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment.*

*Computers & Education, 81, 259–269.*doi:10.1016/j.compedu.2014.10.013

Uner, Oyku, and Henry L. Roediger III. "The Effect of Question Placement on Learning from Textbook Chapters." *Journal of applied research in memory and cognition* 7, no. 1 (2018): 116-122.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). Testlet response theory and its applications. New York: Cambridge University Press.

Waldrop, M. (2013). The virtual lab. *Nature, 499*, 268–270.

Wigfield, A.; Guthrie, J. T.; Tonks, S.; Perencevich, K. C. (2004). "Children's motivation for reading: Domain specificity and instructional influences". *Journal of Educational Research. 97* (6): 299–309. doi:10.3200/joer.97.6.299-310.

Williams, J. R. (2008). The Declaration of Helsinki and public health. Bulletin of the World Health Organization. http://doi.org/10.2471/BLT.08.050955

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208.

Zacharia, Z. C., Olympiou, G., & Papaevripidou, M. (2008). Effects of experimenting with physical *and virtual manipulatives on students' conceptual understanding in heat and temperature. Journal of Research in Science Teaching*, *45*(9), 1021–1035. http://doi.org/10.1002/tea.20260

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology*, 25(1), 82-91.

Table 1: Reliability coefficients, means (standard deviations), and p-values for the dependent samples t-tests investigating the differences between the pre- and post-test scores.

	Pre-test		**Post-test**		
	Cronbach's alpha	Mean	Cronbach's alpha	Mean	p-value
Motivation	0.86	3.88 (0.57)	0.78	3.96 (0.52)	p=0.013*
Self-efficacy	0.85	3.28 (0.64)	0.85	3.48 (0.61)	p<0.001**
Transfer	0.74	9.25 (2.72)	0.64	14.78 (2.91)	p<0.001**

Note: ** indicates significance at alpha of 0.01, * indicates significant at the alpha 0.5 level.

Figure 1: Screen shots from the genetics simulation. The top panel shows the introduction of the case story. The bottom panel is an example of one of a series of laboratory experiments that the students have to carry out in order to answer the research question associated with key learning goals.

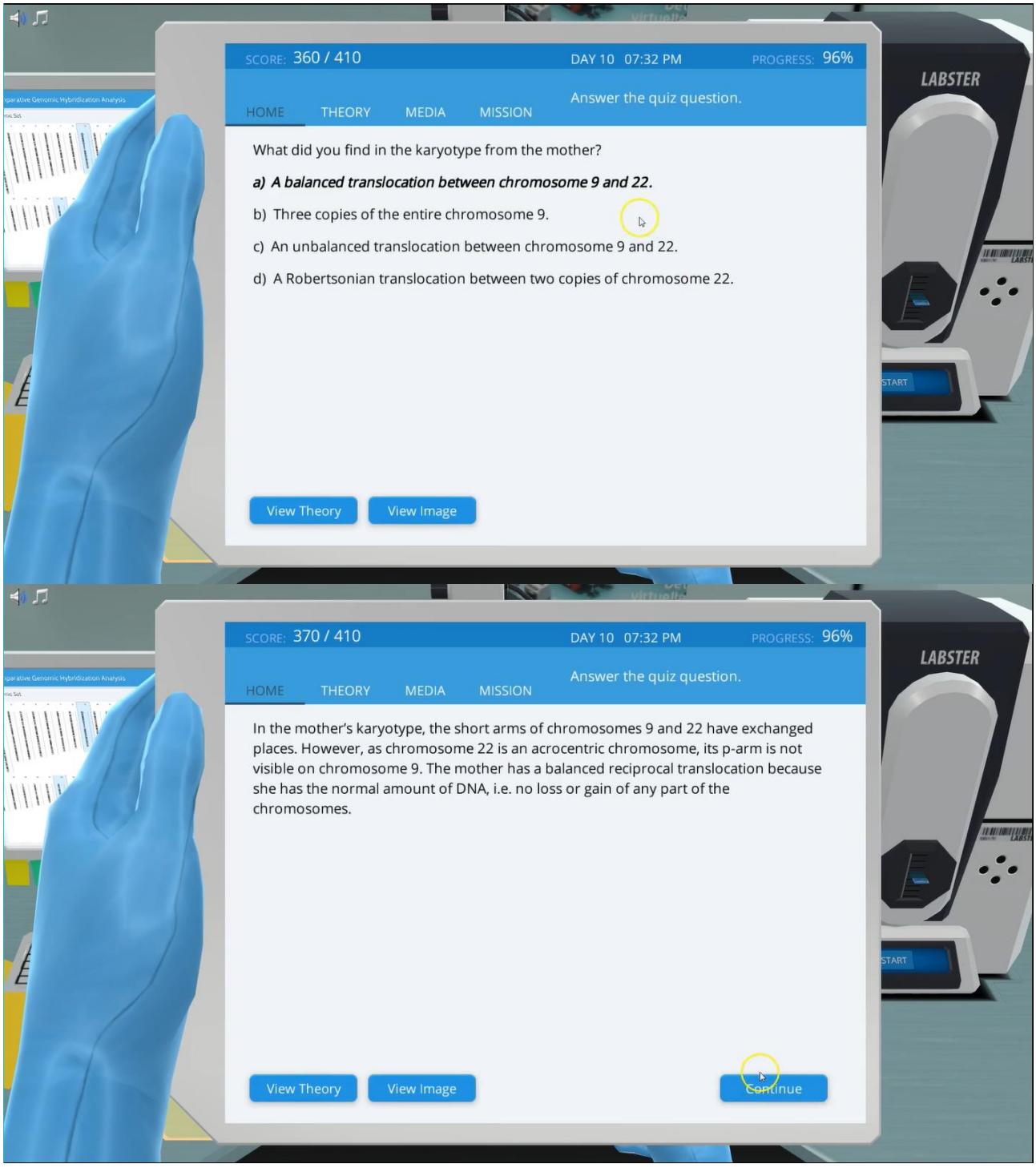

Figure 2: Screen shots from the genetics simulation. The top panel shows an example of a MC question, and the bottom panel shows an example of the explanatory feedback in the simulation.

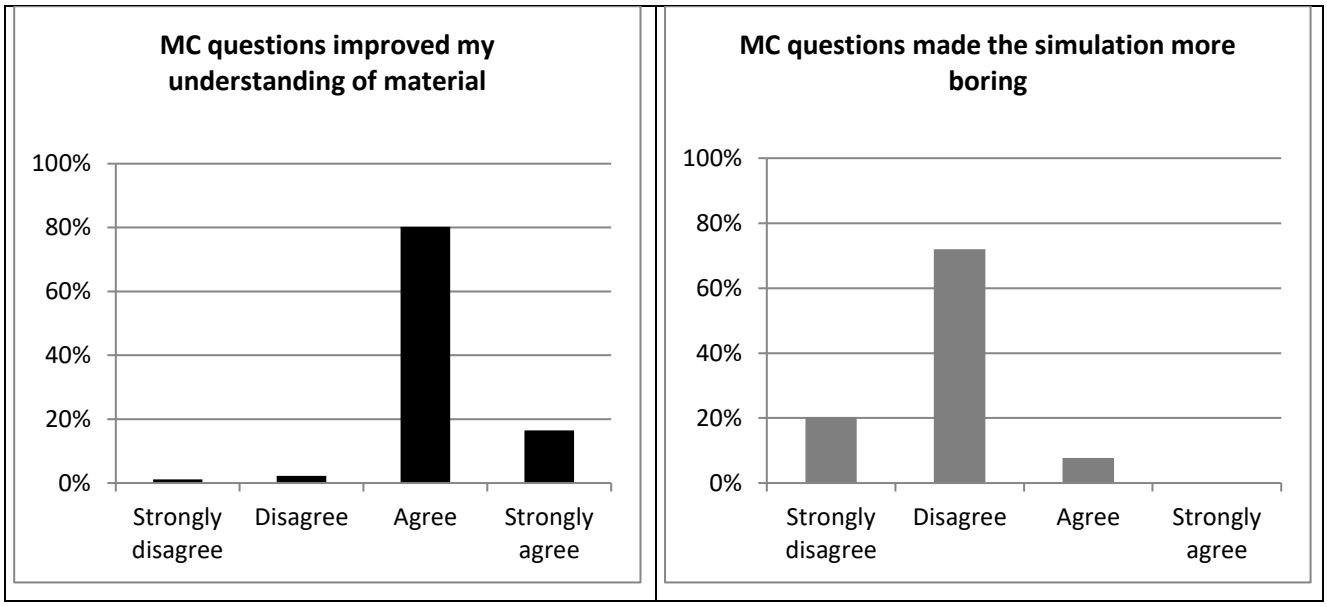

Figure 3: Student evaluation of the multiple choice (MC) questions in the Genetics simulation.