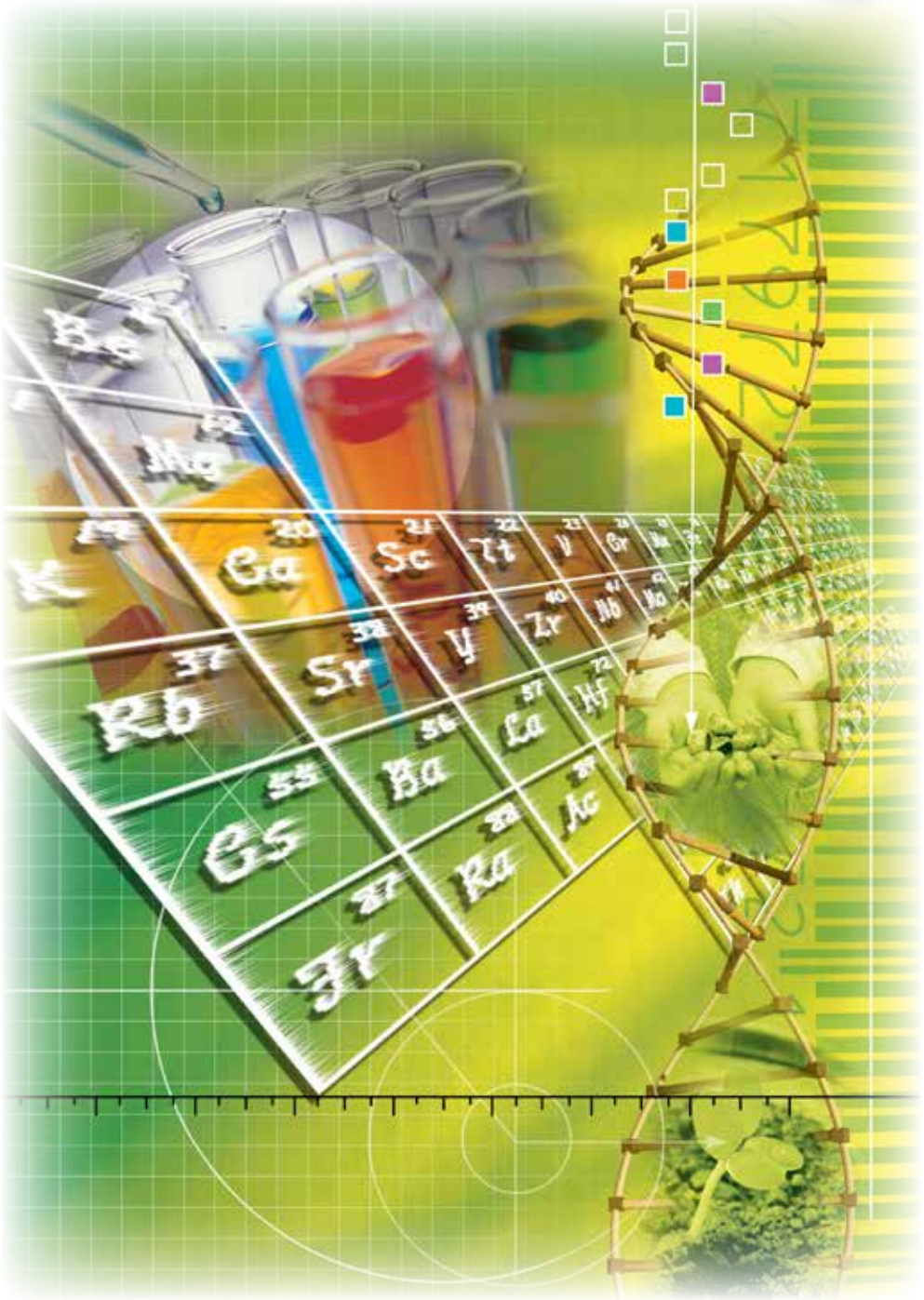
Alberta Science Education Journal

Vol 45, No 2
April 2018



a publication
of the
Science Council
of the
Alberta Teachers'
Association

Raging Skies: Development of a Digital Game-Based Science Assessment Using Evidence-Centred Game Design

Man-Wai Chu and Angie Chiang

Abstract

Digital game-based assessments have been gaining popularity; however, there is often an imbalance between entertainment and educational game elements, yielding barriers for both students and teachers. This paper examines the development processes of an interactive game-based assessment, Raging Skies, in which learning tasks are purposefully embedded and integrated into the game's design and framework so that specific knowledge and skill-based outcomes may be measured. This case study discusses some of the challenges and criticisms facing digital game-based assessments as outlined in the literature.

Introduction

Over the past decade, there has been a growing concern regarding the shortage of science, technology, engineering and mathematics (STEM) skilled workers to fill the many job vacancies in North America (US Department of Education 2015). A recent report has indicated that in Canada, the supply-and-demand ratio for STEM workers has improved, but concerns are now being raised regarding the quality and level of their skills (Council of Canadian Academies [CCA] 2015; National Science Foundation [NSF] 2015). To address this concern, the CCA and NSF have indicated a need to develop STEM-proficient students through high-quality programs from preprimary education through to secondary school. Their hope is that initial investments in building fundamental STEM skills at a young age will develop higher-quality STEM students for the workforce. However, the types of educational programming needed to develop a high-quality STEM-literate population warrant investigation. Before designing new educational programs to improve the quality of the STEM skills students acquire, it is important to investigate gaps in the current methods of teaching and assessing science knowledge and skills.

Developing a strong foundation of science content knowledge is important for success in the field, but equally important is an understanding of scientific inquiry, which explains the process of how scientists came to form these theories (National Research Council [NRC] 2006, 2014). While there are many tools to assess students' conceptual understanding of content knowledge, there are very few tools to assess the process of science inquiry, particularly in a standardized way. Hence, there is a need for assessment tools that can capture evidence of science inquiry skills, which requires an investigation of the process students use to complete a task. In order to capture this evidence, we need new assessment formats that break the mould of traditional paper-and-pencil tests.

Assessments are currently facing a turning point at which the impact of technological advances, coupled with a wave of innovation in learning sciences, has opened the doors for new possibilities. These advances and innovations have created an environment that is ripe for investigation, because formats and capabilities of assessment have been revolutionized as a result (Shute et al 2016). These improvements allow for the development of high-quality, authentic digital tasks, resulting in the measurement of both content knowledge and process skills (Shute and Ventura 2013). Assessments that take the format of digital tasks (eg, technology-rich environments [TRE], search and simulation [Sim] scenarios) are being developed and used by large testing agencies such as the National Assessment of Educational Progress (Bennett et al 2007). Many of these digital task assessments use simulations to guide

students through a learning environment such as a nature conservatory, like Taiga Park (Barab, Gresalfi and Ingram-Goble 2010), or a science laboratory, like TRESim (Bennett et al 2007). These digitally simulated assessments allow students to interact with a dynamic environment that is responsive to their actions and performances. The computer logs that capture students' actions throughout the simulation are analyzed for evidence of content knowledge and process skills (Shute and Ventura 2013). Although these assessments allow for interactive components, they still mimic and use traditional assessment formats such as multiple-choice items (Bennett et al 2007). Although the interactive learning environment often increases students' engagement, the embedded tasks often emulate that of traditional assessments, which may continue to elicit test anxiety-related performance (Chu 2017).

In order to combat the reliance on traditional test formats, such as multiple-choice items, as a measure of performance, some researchers have started to capitalize on digital activities for assessments insofar as they are embedded in actual digital gameplay (Mislevy et al 2014). Digital game-based assessments (eg, Physics Playground) have started to gain popularity in recent years (Shute and Ventura 2013). Some of these assessments use existing commercial games (eg, Portal 2 and Lumosity), which are primarily designed to provide entertainment, to measure skill-based outcomes such as problem solving, spatial skill and persistence (Shute, Ventura and Ke 2015). Critics of these game-based assessments have indicated that the problem with retrofitting commercial games for use in education settings is that the observable evidence needed to support the resulting inferences made on a specific skill may not be built into the game (Mislevy et al 2014). For example, commercial games may be used by researchers to measure persistence, but if the game was not originally designed to assess this skill then the results may not be valid. Therefore, making conclusions regarding students' skill levels based on the data collected from these games may lead to weak and inaccurate inferences.

Conversely, there are educational games that focus more exclusively on teaching and assessing specific educational knowledge and skills than their commercial counterparts (Shute and Ventura 2013). However, these games have been criticized for essentially being a high-tech worksheet instead of using the evidence collected during the interactive portions of the task in a more purposeful way (Mislevy et al 2014). Additionally, these games do not develop good game mechanics (eg, in-game rewards, such as points or trophies, for high performance), which leads to lower student engagement levels when interacting with these assessments (Shute and Ventura 2013). Hence, there appears to be a need to develop digital game-based tools that more equally balance entertainment and education so that new assessments may be developed that capitalize on the benefits of each (Mislevy et al 2014). Specifically, these assessments should incorporate the development of data capture methods that more purposefully demonstrate the acquisition of specific content knowledge and process skill outcomes from a program of study.

This paper describes the development of a digital game-based assessment that used a framework called *evidence-centred game design* (ECgD), which was designed to balance the entertainment and education elements during the development stages. This dynamic, digital game-based assessment is called Raging Skies (a more thorough description follows), and aims to measure both science content knowledge and process skills. Raging Skies directly and purposefully embeds various outcome-informed tasks into gameplay. This assessment tool was specifically designed to measure a set of content knowledge and process skill outcomes related to an elementary school science program of study. This investigation of the developmental stages of the game seeks to resolve the imbalance that much of the literature on game-based assessments identifies and to offer solutions to the competing priorities of entertainment and education games.

Our discussion is structured into three sections: describing the ECgD framework in more detail, a description of the developmental process of Raging Skies and a survey of the challenges faced during the development of the game.

Evidence-Centred Game Design (ECgD)

The analysis of the production framework is guided by ECgD, in which digital games function as both assessments and learning tools to measure content knowledge and skill-based competencies (Mislevy et al 2014). One aim of ECgD is to synthesize two design development frameworks—game and assessment—into one unified process, as shown in Figure 1.

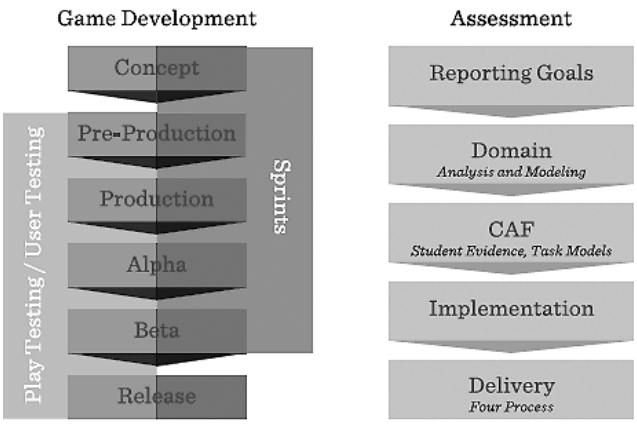


Figure 1. *Design frameworks for games and assessments that are integrated using ECgD. Adapted from Mislevy et al 2014, 135. Reprinted with permission.*

On the right side of Figure 1 is the evidence-centred design (ECD) framework that is often used to develop assessments based on evidentiary reasoning so that judgments of students' level of knowledge and skills may be made (for more details please see Mislevy, Almond and Lukas 2003). The ECD framework guides educators to articulate the observable evidence needed to support the inferences they wish to make regarding students' achievement of specific knowledge and skills (Behrens et al 2010). The left side of Figure 1 shows the design process typically used to guide the development of recreational digital games, emphasizing repetitive implementation, testing and enhancing of the product during what is called the "sprint" period. The majority of the development process is done after alpha- and beta-user testing phases when feedback is provided to the development team outlining usability, requirements and constraints (Mislevy et al 2014).

By unifying both of these frameworks, ECgD attempts to reflect a meaningful integration of both game and assessment, as shown in Figure 2. It illustrates the importance of developing an assessment product that has a meaningful context for students to learn and educators to measure specific content knowledge and skill-based competencies. Once this meaning or macro-level defining stage is complete, micro-level designs follow to address the types of actions students need to perform during an activity. These actions indicate whether or not students have provided sufficient evidence of mastering a construct. Considering the constellation of perspectives outlined in Figure 2—meaning, construct, knowledge, actions, evidence and activities—it is important to develop a product that adequately represents each domain (ie, games, learning and assessment) and evokes evidence of players' capabilities (Mislevy et al 2014).

The integration of games and assessment results in an ECgD framework that follows four phases (Mislevy et al 2014, 136):

1. Definition of competencies from a non-game realm
2. A strategy for integrating externally defined competency with gameplay competency
3. A system for creating formative feedback that is integral with the game experience
4. A method for iteration of the game design for fun, engagement and deep learning, simultaneous with iteration of the assessment model for meaning and accuracy

It is important to note that ECgD is not a retrospective process, instead designing the game's mechanics to suit the assessment and learning needs of interest during the initial planning stages. It is, therefore, important to consider the goals of games, assessment and learning early during the development process.

ECgD builds upon the principles of technology-rich and simulated learning environments that situate assessment tasks within a digital game environment. ECgD often seamlessly embeds assessments into the learning environment so that students are not pulled away from an engaging flow of tasks with an explicit test (see Shute and Ventura 2013). This seamless integration allows the digital game environment to be

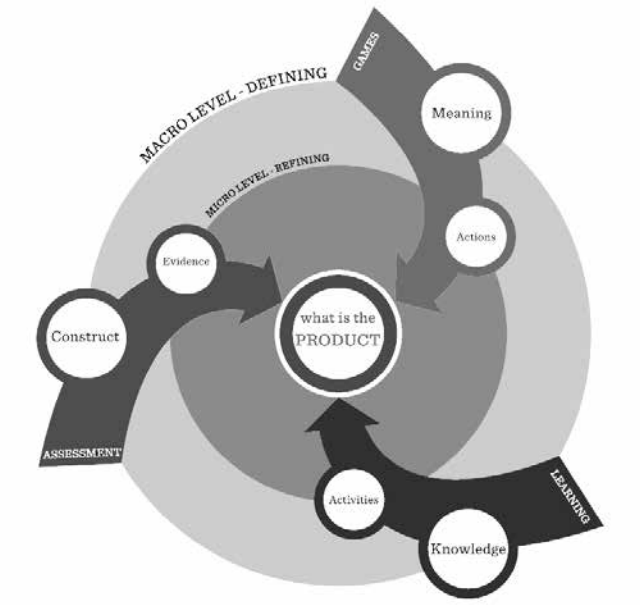


Figure 2. *Model of unifying frameworks from the disciplines of games, assessment, and learning. Adapted from Mislevy et al 2014, 136. Reprinted with permission.*

highly immersive and engaging, thus helping to reduce test or evaluation anxiety (Shute and Ventura 2013). Part of this engagement is due to the real-time interactions between the user and the digital game, which is often viewed as feedback. This real-time feedback is made possible by using computers as a method for administering ECgD assessments. Although many, if not most, of the ECgD assessments are administered using computers (Rowe, Asbell-Clarke and Baker 2015; Rupp et al 2010), the framework itself does not mandate the use of digital technology.

Raging Skies

Using the ECgD framework, a team of researchers and digital-game developers created a computer game-based assessment entitled Raging Skies. This role-playing game transports students into the world of storm chasers, asking them to use various resources (eg, weather balloon and thermometer) located on their vehicle to collect information regarding the weather phenomena (ie, wind speed and temperature). The game uses real-time footage of storms across North America as players are asked to collect data, identify the type of storm and report on it. The footage is overlaid with animated elements to mimic a first-person experience. Figure 3 shows a diagram of the vehicle dashboard that players will use throughout the game to activate each of the tools. This game-based assessment was developed to capitalize on its format so that both content knowledge and process skills may be measured. The development of the game was guided by the four steps of the ECgD model, which are presented in the following sections.

Definition of Competencies

Competencies are the knowledge and/or skills that the game-based assessment intends to measure. The competencies that were assessed in Raging Skies are outlined by the specific learner outcomes listed in Alberta's Grade 5 science program of studies under the Weather Watch unit (Alberta Education 1996; Leighton and Gierl 2007). Two types of learner outcomes in the program were of particular interest—content knowledge and science inquiry skills. These two types of outcome were specifically selected because they support one another during the learning process. For example, prerequisite content knowledge is needed so that it can be applied during the process of science inquiry. On the other hand, science inquiry is defined as the process of acquiring new knowledge. Hence, these two types of learner outcome form a mutualistic relationship in which both knowledge and skills benefit when addressed together. The specific learner outcomes that were used to guide the development of Raging Skies are as follows:

Knowledge Outcomes

5.8.2 Describe patterns of air movement, in indoor and outdoor environments, that result when one area is warm and another area is cool.

5.8.3 Describe and demonstrate methods for measuring wind speed and for finding wind direction.

5.8.5 Describe and measure different forms of precipitation, in particular, rain, hail, sleet, snow.

5.8.8 Identify some common types of clouds, and relate them to weather patterns.

Figure 3. *Screen capture of the vehicle dashboard from the proof-of-concept prototype from the digital-game-based assessment Raging Skies. Students may click on the icons, highlighted by the boxes, on the dashboard to activate the different tools used to collect data regarding the weather outside of the vehicle. Copyright 2016 by MindFuel. Reprinted with permission.*

5.8.10 Recognize that weather systems are generated because different surfaces on the face of Earth retain and release heat at different rates. (Alberta Education 1996, 27)

Skill-Based Outcomes

5.1.2 Identify one or more possible answers to questions by stating a prediction or a hypothesis.

5.2.3 Record observations and measurements accurately, using a chart format where appropriate. Computer resources may be used for record keeping and for display and interpretation of data.

5.2.4 Reflect and interpret: state an inference, based on results. The inference will identify a cause and effect relationship that is supported by observations. (Alberta Education 1996, 24)

Instead of selecting all 28 Weather Watch learner outcomes, only these 8 were specifically targeted for Raging Skies. Focusing on a few selected outcomes allows for more evidence for each outcome to be collected, thereby improving the reliabilities of the claims made from the assessment (American Educational Research Association [AERA], American Psychological Association [APA] and National Council on Measurement in Education [NCME] 2014).

Just as the two types of learner outcomes are interconnected, the eight specific outcomes that represent the two types are also connected. To represent the connection between the eight specific learner outcomes, a competency model was developed. Competency models are often developed using extensive literature reviews of the constructs being measured, such as those undertaken by the curriculum specialists when developing the learner outcomes (Shute 2011; Shute and Ventura 2013).

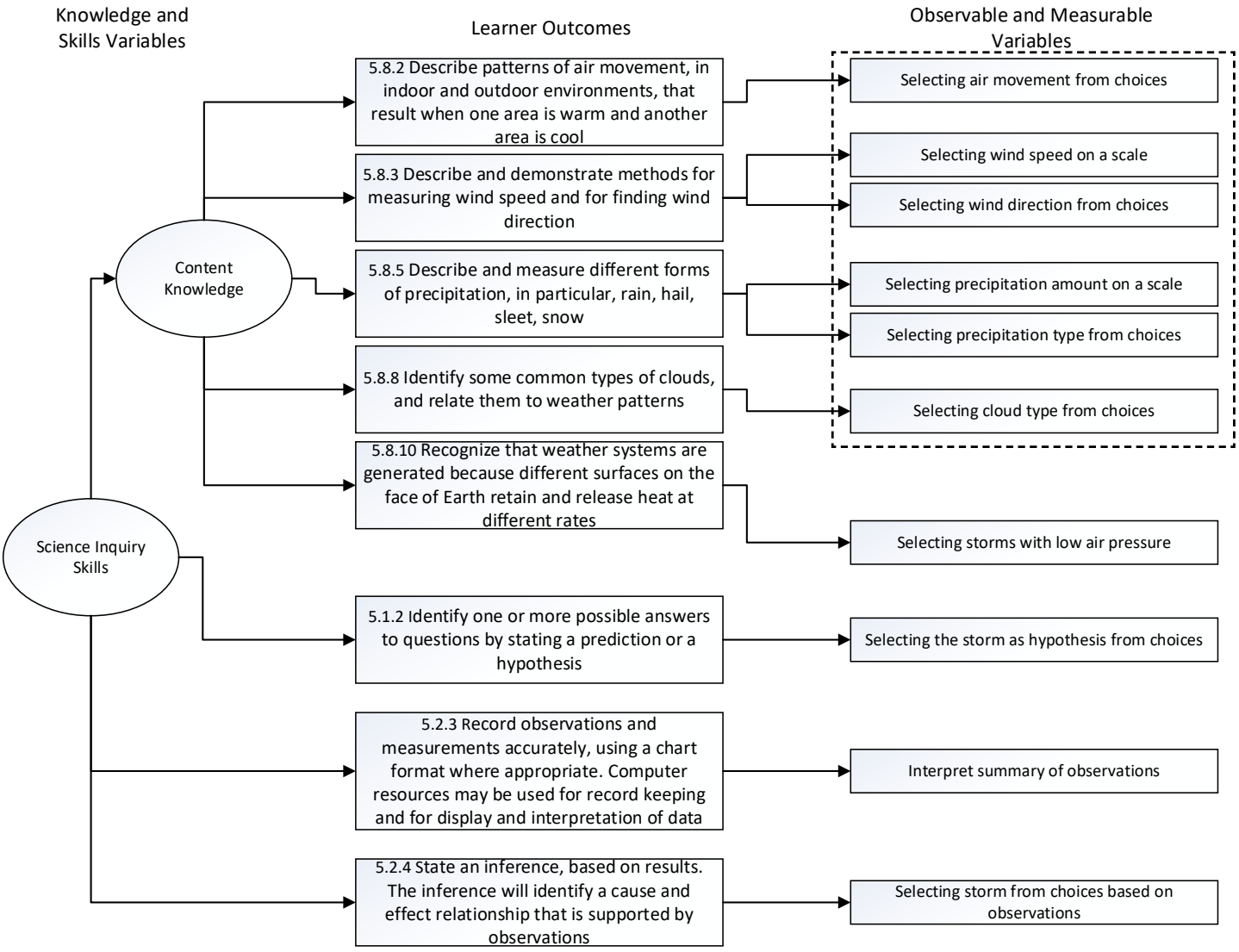
Figure 4 (page 42) shows the competency model used to guide the development of Raging Skies, which illustrates how the learner outcomes are connected to each other. The competency model also identifies the observable and measurable variables that are used as evidence to support the corresponding learner outcomes. Using the model as a whole, the evidence collected from the observable and measurable variables is then used to support claims of proficiency of the learner outcomes and corresponding knowledge and skills variable. For example, during the assessment, students are asked to collect information on six observable and measurable variables that represent different aspects of the storm; they are highlighted using a dash-lined box in Figure 4. Students' performance on these six variables is used to indicate their proficiency on the corresponding four learner outcomes. In order for the assessment to provide the necessary opportunities to collect evidence of students' performances for each learner outcome, the team of researchers and digital-game developers worked collaboratively to integrate the competencies with the gameplay. This process is discussed in the next section.

Integrating Competencies with Gameplay

The production team started to write a story that would be realistic for students while ensuring that the game mechanics could properly capture evidence for each learner outcome. The introduction of the game was designed to be highly captivating and realistic to students so that they would become immersed into the game-based assessment's story line. This immersion into the assessment's game-based environment is referred to by many gaming communities as *flow* (Shute et al 2009). Research in game design has suggested that optimal flow is achieved when the intrinsically motivating environment has elements of challenge, control and fantasy to keep the players engaged in the game such that they lose their self-consciousness and sense of time (Gee 2007; Rieber 1996). Raging Skies is designed to maximize students' flow so that they do not view the game as an assessment, which often elicits anxieties related to testing (Shute and Ventura 2013).

One way that Raging Skies keep players engaged is the use of a reward system, such as in-game money to add a competitive element and increase replayability. Players are provided with an account to track the amount of in-game money accumulated and used; this gives students the ability to purchase upgrades to and customizations of their equipment (eg, change the colour of their dashboard) as well as gas for driving to future storms. Players are able to access their account frequently to determine how much more they need to reach their next level of upgrade or customization. To get players motivated to start the game, Raging Skies uses a guided tutorial during the first administered task so that enough in-game money can be earned to keep the player engaged.

Figure 4. *Competency model of learner outcomes from Alberta's Grade 5 program of study's Weather Watch unit, designed specifically for the game-based assessment Raging Skies. The dotted line outlines six observable and measurable variables that represent different aspects of the storm.*



In order to further increase student engagement, a competitive element of racing against computer-generated storm chasers was added. The first student to identify the storm (when playing against the computer) earns extra in-game money, which allows them to purchase gas and upgrade/customize their equipment. The amount of in-game money rewarded to students is proportional to their performance during the storm task. As such, the amount of in-game money received by the student is a form of formative feedback regarding their performance during the storm task.

Formative Feedback During Game

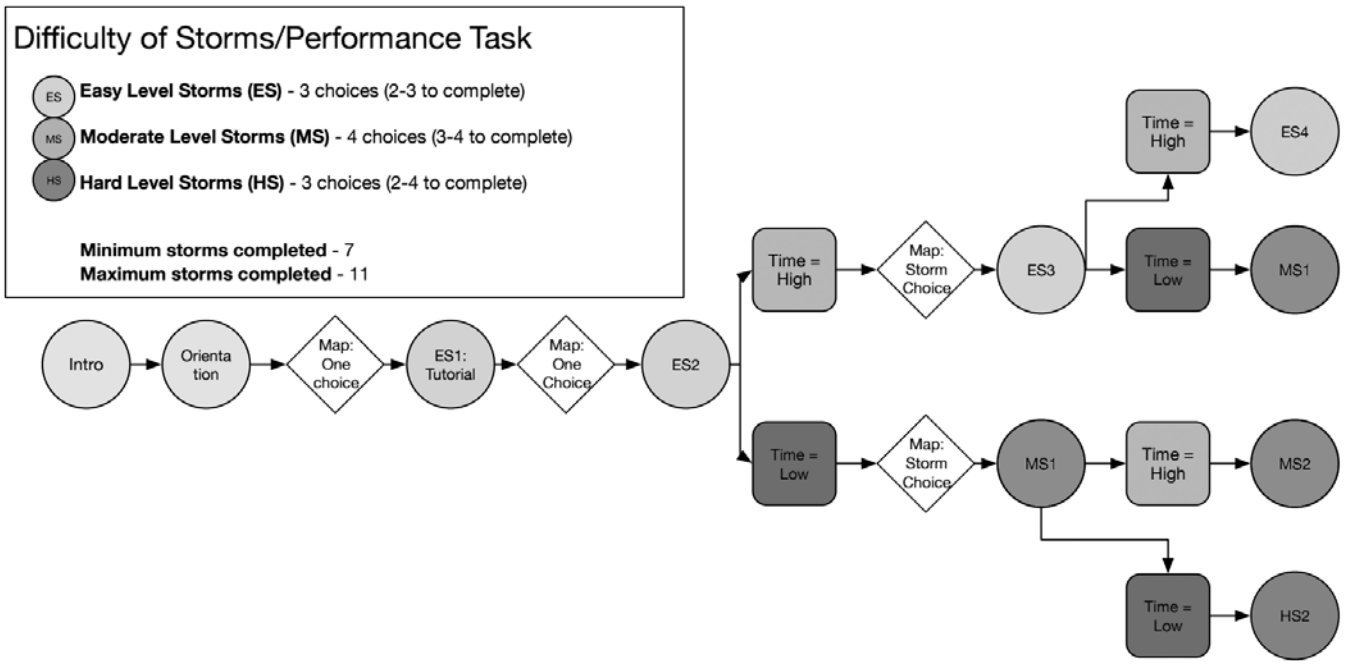
After students identify the storm type, they are given formative feedback regarding their performance in measuring the six different elements of the storm and identifying the storm type. An example of a student feedback report is provided in Figure 5. The feedback report indicates students' level of performance using in-game money as their reward system. The better the student performs during the storm task, the more in-game money they will receive.

Figure 5. *Screenshot of the formative feedback report that students receive after a storm task. Students are rewarded with in-game money so that they may purchase upgrades, customize their equipment (eg, change the color of their dashboard) and purchase gas to reach the next location.*

After students are shown their formative feedback report, they are able to click on the portions for which they did not receive the full amount of in-game money (eg, variables in which the student received only $50) to review their answers and reselect their choice. Students are given one opportunity to reselect their choice for each variable for which they did not receive the full amount of in-game money. However, a correct selection during the second attempt does not result in additional in-game money being given; instead, this opportunity is designed to provide students with formative feedback so that they may improve their performance on later storms.

The design of Raging Skies is based on both game design (eg, presenting easy tasks first and then increasing the difficulty) and assessment principles (eg, computer adaptive testing) (Weiss 1982). These principles indicate the importance of administering different storm tasks to students based on their performance on previous storm tasks. Figure 6 presents a diagram of the adaptive process based on students' performance on previous storm tasks. The first storm task administered to students after their tutorial storm is rated to be at moderate difficulty level. If students perform well on this storm task, they are provided with in-game money

Figure 6. *Adaptive process based on students' performance*



so that they may purchase gas to travel to their next location. Alternatively, a weak performance on this first storm task would result in a smaller amount of in-game money for gas. Less in-game money for gas will result in the student only being able to reach closer storms, which are rated to be a lower difficulty level.

Students are incentivized to reach storm tasks with a higher difficulty rating because they will be able to receive more in-game money during those tasks. For example, students may receive a maximum of $1,250 during a moderate-difficulty storm task, but they may receive up to $1,875 during a high-difficulty storm task. This adaptive process continues throughout the assessment so that each student will have a customized experience that matches their performance. Throughout the assessment, students are presented with multiple storm tasks (ie, 7 to 11 storm tasks) so that enough evidence may be collected to ensure that reliable claims may be made regarding their performance of each learner outcome.

The previous sections discussed the first three phases of the ECgD framework in terms of game development (ie, concept and preproduction) and assessment (ie, reporting goals, domain and conceptual assessment framework [CAF]). Raging Skies is currently in the production stage, in which the developers have taken the designs from the previously discussed three phases to write the computer codes needed for this assessment. The next section of this paper will discuss the next steps for this project, which will involve the fourth, and final, step of the ECgD framework.

Iterations of Game Design and Assessment Model

ECgD stresses the importance of the iterative process involved when developing a game-based assessment. Although there are four steps to consider when developing such an assessment, it is imperative that the development is informed by both game-design and assessment principles. Raging Skies is currently being validated through a process in which information about whether or not this game-based assessment is in fact measuring the intended learner outcomes is being collected (Kane 2013). Validation is critical to good assessment development and is important for ensuring that irrelevant concepts are not interfering with the measurement of the intended purposes and goals (AERA, APA and NCME 2014). The validation results allow for enhancements to be made to the assessment by both the digital game developers and educational assessment researchers.

Throughout the development of this game-based assessment, the ECgD framework guided this project. Previous researchers who developed game-based assessments using this framework identified some of the challenges they encountered (Mislevy et al 2014). The Raging Skies game developers and educational researchers were able to avoid some of these challenges. However, the team still encountered additional challenges during the development process. A discussion of some of the challenges faced during this development process is discussed in the next section.

Challenges During the Development of Raging Skies

Although the ECgD framework does a good job of integrating the entertainment and educational elements of game-based assessment, some challenges presented themselves during the development process. The challenges that the development team faced occurred during Phase 1 of the ECgD framework—defining the constructs. Specifically, the development team had difficulties in identifying proper methods to collect the evidence needed to support the observable and measurable variables and in representing the open-ended nature of science inquiry skills.

The development of Raging Skies used the lessons learned from existing game-based assessment development literature to prevent recurrence of previously encountered issues. For example, studies that investigated the difficulties associated with developing game-based assessments indicated that the multidimensionality of the constructs (eg, creativity) were problematic when trying to define the construct and when identifying observable and measurable evidence (Kim and Shute 2015). The development of Raging Skies avoided this issue of needing to define multidimensional constructs by using learner outcomes from the program of study. The program of study defined the constructs of interest (ie, content knowledge and science inquiry skills) for the Weather Watch unit by identifying the outcomes that are associated with each construct (Alberta Education 1996). Although the outcomes identified by the program of study may not fully encompass all aspects of the construct for the unit, the curriculum team who wrote the program identified the

main elements that are important for Grade 5 students to know. Developing a game-based assessment that is guided by specific learner outcomes allows teachers who use the Alberta Grade 5 science program of study to use Raging Skies as a classroom resource.

Although the development team was able to avoid the challenges associated with defining constructs by using learner outcomes, one difficulty that the team encountered was the issue of using a creative format to collect evidence of student performance. The team identified observable and measurable variables for each learner outcome, as shown in Figure 4, so that appropriate evidence could be collected. However, when operationalizing the observable and measurable variables, the team faced the challenge of designing a collection format that would mimic how real-life storm chasers measure and document their findings. One of the main challenges was to develop a collection method that did not emulate multiple-choice items because the development team did not want this game-based assessment to be viewed as a fancy digital worksheet. However, for many of the observable and measurable variables, a multiple-choice item format was implemented to collect the necessary information. For example, when students are asked to identify the wind direction using the weather balloon launched from their vehicle, they are provided with only three choices: straight, clockwise and counter-clockwise. The three choices made the recording of the measurement seem like a multiple-choice item.

Of course, this item format was avoided when possible. For example, when students were asked to record the wind speed, they were given a full scale, such as the one shown in Figure 7, to record their findings. This format of recording wind speed emulates the real world in terms of providing students with a scale so that they can focus on the accuracy of their measurement. However, it could be argued that this scale still emulates the multiple-choice item format because it provides students with 20 possible choices to select. The research team was unable to develop a better method of recording students' measurements during the storm task; this is an area that will, hopefully, be enhanced with future iterations of this assessment.

Another challenge faced by the development team was ensuring that the storm tasks allowed for the open-ended solutions needed to assess science inquiry skills. Science inquiry focuses on how a scientist would acquire knowledge and skills, a process that is relatively open ended. However, the learner outcomes selected from the program of studies represented only a small subset of this construct. Additionally, the selected subset of outcomes were typically quite closed ended. For example, learner outcome 5.1.2 indicates that students should "identify one or more possible answers to questions by stating a prediction or a hypothesis" (Alberta Education 1996, 24). The learner outcome indicates the need to focus on the open-ended nature of answering a problem because there could be multiple correct answers. However, in the context of Raging Skies, the objective of the game was for students to identify the storm type; hence, only one correct answer is present. This created a closed-ended problem for each of the storm tasks administered, and also prevented the open-ended nature of science inquiry to be assessed.

Figure 7. *Wind speed scale*

During the validation process, these two challenges will be shared with students and educators in hopes that possible solutions will be presented to the development team so that future iterations of Raging Skies will be enhanced. Future research needs also to focus on using more real-life formats to document the measurements taken during a storm task and providing more open-ended storm tasks that allow for multiple processes and solutions to be accepted. Possible solutions to these challenges will greatly enhance science assessments that aim to be authentic and measure science inquiry skills.

Classroom Relevance

A preliminary version of Raging Skies is currently available on the MindFuel's Wonderville.org website (https://wonderville.org/asset/stormchasers) for use by students and educators. Although Raging Skies is still being validated, the game-based assessment provides students and educators with an opportunity to approach the Weather Watch learner outcomes with a realistic simulation. This new approach of addressing the learner outcomes may provide additional insight in terms of guiding students' learning and informing teachers' practices. Once this assessment has been validated, it will also provide students with feedback regarding their performance on skill-based outcomes.

Digital game-based assessments, although starting to become more popular, are still in their infancy (Shute and Ventura 2013). In order to address some of the criticisms of entertainment and education game-based assessments, it is important that new games use proper game design and rigorous assessment properties (Mislevy et al 2014). The development process of Raging Skies led the development team to spend a substantial amount of time researching the learning objectives and mapping them to the different levels of student performance. This ensured that the assessment developed would be aligned with learner outcomes and adaptive to reflect student performance. By introducing game-based science assessments that are well aligned with learner outcomes from a program of study, this educational tool may be used in the classroom to provide evidence that students are learning specific content knowledge and process skills. Formative feedback provided to students and educators will target specific areas of weaknesses so that instruction may be adapted. With more of these educational tools being developed to enhance students' content knowledge and process skills, the vision of a high-quality STEM-literate population is possible.

References

Alberta Education. 1996. *Science (Elementary)*. Edmonton, Alta: Alberta Education. Available at www.education.alberta.ca/media/654825/elemsci.pdf (accessed January 18, 2018).

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME). 2014. *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Barab, S A, M S Gresalfi and A Ingram-Goble. 2010. "Transformational Play: Using Games to Position Person, Content, and Context." *Educational Researcher* 39, no 7: 525–36. Available at http://ase.tufts.edu/DevTech/courses/readings/Barab_Transformational_Play_2010.pdf (accessed January 18, 2018).

Behrens, J T, R J Mislevy, K E DiCerbo and R Levy. 2010. *An Evidence Centered Design for Learning and Assessment in the Digital World* (CRESST Report 778). Available at http://files.eric.ed.gov/fulltext/ED520431.pdf (accessed January 18, 2018).

Bennett, R E, H Persky, A R Weiss and F Jenkins. 2007. *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project*. Washington, DC: National Center for Education Statistics. Available at http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp (accessed January 18, 2018).

Chu, M-W. 2017. "Using Computer Simulated Science Laboratories: A Test of Pre-Laboratory Activities with the Learning Error and Formative Feedback Model." Doctoral dissertation, University of Alberta.

Council of Canadian Academies (CCA). 2015. *Some Assembly Required: STEM Skills and Canada's Economic Productivity.* Ottawa, Ont: CCA. Available at http://scienceadvice.ca/uploads/ENG/AssessmentsPublicationsNewsReleases/STEM/STEMFullReportEn.pdf (accessed January 18, 2018).

Gee, J P. 2007. *Good Videogames + Good Learning: Collected Essays on Videogames, Learning and Literacy.* New York: Lang.

Leighton, J P, and M J Gierl. 2007. "Defining and Evaluating Models of Cognition Used in Educational Measurement to Make Inferences About Examinees' Thinking Processes." *Educational Measurement: Issues and Practice* 26, no 2: 3–16.

Kane, M T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50, no 1: 1–73.

Kim, Y J, and V J Shute. 2015. "The Interplay of Game Elements with Psychometric Qualities, Learning, and Enjoyment in Gamebased Assessment." *Computers and Education* 87: 340–56. Available at http://myweb.fsu.edu/vshute/pdf/creativity.pdf (accessed January 18, 2018).

Mislevy, R J, R G Almond and J Lukas. 2003. *A Brief Introduction to Evidence-Centered Design* (Research Report No RR-03-16). Princeton, NJ: Educational Testing Service. Available at www.ets.org/Media/Research/pdf/RR-03-16.pdf (accessed January 18, 2018).

Mislevy, R J, A Oranje, M I Bauer, A von Davier, J Hao, S Corrigan, E Hoffman, K DiCerbo and M John. 2014. *Psychometric Considerations in Game-Based Assessment*. Np: CreateSpace.

National Research Council. 2006. *America's Lab Report: Investigations in High School Science*. Washington, DC: National Academies Press.

———. 2014. *Developing Assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.

National Science Foundation (NSF). 2015. *Revisiting the STEM Workforce: A Companion to Science and Engineering Indicators 2014*. Arlington, Va: NSF. Available at www.nsf.gov/nsb/publications/2015/nsb201510.pdf (accessed January 18, 2018).

Rieber, L. 1996. "Seriously Considering Play: Designing Interactive Learning Environments Based on the Blending of Microworlds, Simulations, and Games." *Education and Technology Research and Development* 44, no 2: 43–58.

Rowe, E, J Asbell-Clarke and R Baker. 2015. "Serious Game Analytics to Measure Implicit Science Learning." In *Serious Game Analytics: Methodologies for Performance Measurement, Assessment, and Improvement,* ed C S Loh, Y Sheng and D Ifenthaler, 343–48. New York: Springer.

Rupp, A A, M Gushta, R J Mislevy and D W Shaffer. 2010. "Evidence-Centered Design of Epistemic Games: Measurement Principles for Complex Learning Environments." *Journal of Technology, Learning, and Assessment* 8, no 4: 4–47. Available at http://edgaps.org/gaps/wp-content/uploads/rupp2010.pdf (accessed January 18, 2018).

Shute, V J. 2011. "Stealth Assessment in Computer-Based Games to Support Learning." In *Computer Games and Instruction,* ed S Tobias and J D Fletcher, 503–24. Charlotte, NC: Information Age. Available at http://myweb.fsu.edu/vshute/pdf/shute%20pres_h.pdf (accessed January 18, 2018).

Shute, V, J P Leighton, E E Jang and M-W Chu. 2016. "Advances in the Science of Assessment." *Educational Assessment* 21, no 1: 34–59. Available at http://myweb.fsu.edu/vshute/pdf/asstPPF.pdf (accessed January 18, 2018).

Shute, V J, and M Ventura. 2013. *Stealth Assessment: Measuring and Supporting Learning in Games*. Cambridge, Mass: MIT Press.

Shute, V J, M Ventura, M I Bauer and D Zapata-Rivera. 2009. "Melding the Power of Serious Games and Embedded Assessment to Monitor and Foster Learning: Flow and Grow." In *Serious Games: Mechanisms and Effects*, ed U Ritterfeld, M Cody and P Vorderer, 295–321. Mahwah, NJ: Routledge, Taylor and Francis.

Shute, V J, M Ventura and F Ke. 2015. "The Power of Play: The Effects of Portal 2 and Lumosity on Cognitive and Noncognitive Skills." *Computers and Education* 80: 58–67.

US Department of Education. 2015. *Science, Technology, Engineering and Math: Education for Global Leadership*. Available at www.ed.gov/stem (accessed January 18, 2018).

Weiss, D J. 1982. "Improving Measurement Quality and Efficacy with Adaptive Testing." *Applied Psychological Measurement* 6, no 4: 473–92.