Raging Skies: Development of a Digital Game-Based Science Assessment using Evidence-Centered Game Design

September 6, 2016

By: Man-Wai Chu, Werklund School of Education, University of Calgary

Angie Chiang, MindFuel (Science Alberta Foundation)

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 in 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.

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). However, 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 (2015) and NSF (2015) have indicated a need to develop STEM-proficient students through high-quality programs from pre-primary education through to secondary school. Their hope is that initial investments into 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. 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 require an investigation of the process students use to complete a task. In order to capture this evidence, new assessment formats that break the mould of traditional paper-and-pencil tests, are needed.

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 as formats and capabilities of assessment have been revolutionized as a result (Shute, Leighton, Jang, & Chu, 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 & Ventura, 2013). Assessments that take the format of digital tasks (e.g. 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, Persky, Weiss, & Jenkins, 2007). Many of these digital task assessments use

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

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. Digital game-based assessments (e.g. Physics Playground) have started to gain popularity in recent years (Shute & Ventura, 2013). Some of these assessments use existing commercial games (e.g. 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, & Ke, 2015). Critics of these game-based assessments have indicated 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). Therefore, making conclusions regarding students' skill levels based on the data collected from these games may possibly lead to weak and possibly inaccurate inferences.

Conversely, there are educational games, which focus more exclusively on teaching and assessing specific educational knowledge and skills than their commercial counterparts (Shute & Ventura, 2013). However, these games have been criticized for being a high-tech worksheet that does not utilize the evidence collected during the interactive portion of the task. Additionally, these games do not develop good game mechanics (e.g., in-game rewards, such as points or trophies, for high performance) leading to lower student engagement levels when interacting with these assessments. Hence, there is 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 on each. 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 uses a framework called evidence-centered game design (ECgD) that was designed to balance the entertainment and education elements during the development stages. This dynamic, digital game-based assessment is called Raging Skies, and aims to measure both science content knowledge and process skills. Raging Skies directly and purposefully embeds various outcome-informed tasks into gameplay. A more detailed description of Raging Skies is presented later. 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 offers a solution to the competing priorities of entertainment and education games.

This paper is split into three sections. First, the ECgD framework is described. Second, the developmental process of Raging Skies is explained, Finally, a discussion regarding the challenges faced during the development is presented.

Evidence-Centered Game Design (ECgD)

The analysis of the production framework is guided by the 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 – shown in Figure 1, into one unified process. On the right side of Figure 1 is the evidence centered design (ECD) framework that is often used to develop assessments based on evidentiary reasoning so that judgements on students' level of knowledge and skills may be made (for more details please see Mislevy, Almond, & 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, Mislevy, DiCerbo, & Levy, 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 (i.e., 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, pg. 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, ECgD designs the game's mechanics to suit the assessment and learning needs of interest during the initial planning stages. As such, it is 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 which 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 Stealth Assessment; Shute & Ventura, 2013).

This seamless integration allows the digital game environment to be highly immersive and engaging, thus helping to reduce test or evaluation anxiety (Shute et al., 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, & Baker, 2015; Rupp, Gushta, Mislevy & Shaffer, 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 (e.g., weather balloon and thermometer) located on their vehicle to collect information regarding the weather phenomenon (i.e., identify the type of storm). 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 emulate 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; Gierl & Leighton, 2007). Within the program, two types of learner outcomes were of particular interest – content knowledge and science inquiry skills. These two types of outcomes 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 outcomes 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:

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.

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, pg. 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, pg. 24)

Instead of selecting all 28 Weather Watch learner outcomes, only these eight were specifically targeted for Raging Skies (Alberta Education, 1996). 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 (AERA, APA, & NCME, 2014).

Just as the two types of learner outcomes are inter-connected, 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, 2013). The competency models that underlie ECgD typically include latent variables (e.g., science inquiry skills) as well as observable and measurable knowledge and skill variables (e.g., identify one or more possible answers to questions by stating a prediction or a hypothesis) within a content domain.

Figure 4 shows the competency model used to guide the development of Raging Skies. The competency model shows how the learner outcomes are connected to each other through the latent variables – content knowledge and science inquiry skills. Additionally, 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 to inform the observable and measurable variables is then used to support claims of proficiency of the learner outcomes and their related latent variables. For example, during the assessment, students are asked to collect information on six observable and measurable variables which represent different aspects of the storm; they are highlighted using a box in Figure 4. Students' performance on these six variables can be 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. After the eight learner outcomes were identified, the production team started to write a story that would be realistic to students while ensuring 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 that so 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, Ventura, Bauer, & Zapata-Rivera, 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 & Ventura, 2013).

One way that digital games keep players engaged is the use of a reward system such as points or in-game money that can be exchanged for upgrades or customizations. These rewards are typically contained within the game, that is no physical real-world money or upgrades to the players' materials are provided by the game. Usually players are provided with an account to track their points or in-game money. Players are able to access their account frequently to track the amount of points or in-game money they have as well as how much more they need to reach their next level of upgrade or customization. To get players motivated to start the game, many designs have guided tutorials during their first few administered tasks so that enough points and in-game money can be earned to keep the player engaged.

During the introduction, students find themselves stranded on the side of a road after their car has broken down. Storm chasers pass by and offer to drive the student to their home; however, they receive news that a storm is starting to form nearby. The team decide to make a quick detour so that they may be the first to reach the storm and identify it; this leads the student to their first storm which happens to be a tutorial designed to help them understand the tasks they are trying to complete (e.g., measure different aspects of the storm and identify the storm type using this information) and locate all the resources (e.g., location of storm log button). As the team approaches the storm, the student helps the team by taking their measurements and identifying the storm. During this tutorial, students are prompted to click on all the tools so that they may use the resources (e.g., weather balloon) and record the measurements (e.g., wing speed). Using the collected information, students are asked to identify the storm type by comparing their measurements with those shown in the storm log. The storm log, shown in Figure 5, provides students with hints in terms of the types of measurements associated with each storm type. At the beginning of the game, this storm log will be fully populated with all the

necessary information. However, as students show evidence of proficiency in terms of measuring storm elements and identifying storm type, the information in the log will disappear. Students will need to rely on their previous experiences instead of the storm log to identify future storms. On the other hand, if students proficiency begins to decrease, the amount of information in the storm log will start to increase.

After the identification of the tutorial storm, the team continue their journey of driving the student home. The storm chasers suggest that they chase a few more storms along the way. The student agrees, and is provided with a map with low and high pressure readings, as shown in Figure 6, so that they may select an area that they believe will have storm activity (i.e., area with low air-pressure) as they drive towards their home. Students who select a high air-pressure area will lose money from their account for wasting gas to travel to a location without a storm. Alternatively, if students select the low air-pressure locations, which are often indicative of a storm, they will be shown footage of a storm through the vehicle dashboard from the perspective of a passenger. The selection of a possible storm location using the low- and high-pressure map is used to collect evidence for learner outcome 5.8.10. This first storm encountered by students will be moderately difficult.

Upon arriving at the first storm, students are prompted to make a hypothesis in terms of the storm type. This action of making a hypothesis is used to provide evidence for the learner outcome 5.1.2 which asks students to identify a prediction. As students progress through the storm, they are asked to use the measurements they collect to make an informed decision regarding the type of storm they are chasing. The measurements students are asked to collect are aligned with four of the five content knowledge outcomes (i.e., learner outcomes 5.8.2, 5.8.3, 5.8.5, and 5.8.8) in the competency model.

A competitive element, racing against computer generated storm chasers, is added to increase student engagement. If students are the first to identify the storm (when playing against the computer), they earn extra game money which allows them to purchase gas and upgrade/customize their equipment. The amount of game money rewarded to students is proportional to their performance during the storm task. As such, the amount of money received by the student is a form of formative feedback regarding their performance during the storm task. This feedback is elaborated in the next section of this paper.

Formative feedback during game. After students identify the storm type, they are provided with 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 7. The feedback report indicates students' level of performance using money as their reward system. The better the student performs during the storm task, the more money they will receive. There are either two or three levels of performance for each of the variables. For example, the first variable of identifying cloud types has only two performance levels because students either correctly or incorrectly identify the type of cloud in the storm. Conversely, the second variable of identifying the wind speed is split into three levels of performance to inform students of their accuracy in terms of identifying the correct wind speed.

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 game money (e.g., variables in which the student received only $50) to review their answers and re-select their choice. Students are provided with one opportunity for each variable for which they did not receive the full amount of game money to re-select their choice. However, a correct selection during the second attempt does not result in additional 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 game money earned during each storm will allow students to purchase upgrades to and customizations of their equipment (e.g., change the color of their dashboard) as well as gas for driving to future storms. The game money issued during each storm task is proportional to the difficulty levels. This means the more difficult a storm task is rated, the more game money a student may earn.

The design of Raging Skies is based on both game design (e.g., presenting easy tasks first and then increasing the difficulty) and assessment principles (e.g., 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 game money 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 game money for gas. Less 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 incentivised to reach further storms (i.e., storm tasks with a higher difficulty rating) because they will be able to receive more money during those tasks. For example, students may receive a maximum of $1250 during a moderate difficulty level storm task, but they may receive up to $1875 during a high difficulty level 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

storms tasks (i.e., 7-11 storm tasks) so that enough evidence may be collected to ensure a relatively high reliability (i.e., internal consistency) based on their performance on each learner outcome.

The previous sections discussed the first three phases of the ECgD framework in terms of game development (i.e., concept and pre-production) and assessment (i.e., 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 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 in the production stage of the game development (i.e., left side of Figure 1) portion of the ECgD framework. The digital-game developers are creating the assessment so that alpha and beta testing may be completed during the fall of 2016. These testing phases are important for both the game development and assessment because the game developers may use the feedback to ensure the game maximises the flow, while assessment researchers may use the opportunity to validate the assessment. The feedback from these testing phases will ensure no technical issues exist with the assessment, which can easily detract from an engaging environment in which students immerse themselves.

Validation is critical to good assessment development, and is important for ensuring the intended purposes and goals are measured (AERA, APA, & NCME, 2014). The process of validation requires the collection of relevant evidence to provide a sound scientific basis for the proposed interpretation of scores (Kane, 2013). The validity results of Raging Skies will provide good psychometric evidence for the game as an assessment of the selected learner outcomes. The feedback and results received during the initial testing phases will allow for changes to be made to the assessment by both the digital game developers and educational assessment researchers before it is fully implemented and released to the public.

Although the development of this game-based assessment is not complete, the process of using the ECgD framework to guide this project has presented the team of researchers and digital-game developers with some challenges. 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

The development of Raging Skies was guided by the ECgD framework. Although the 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 were both during Phase 1 of the ECgD framework – defining the constructs. Specifically, the development team had difficulties identifying proper methods to collect the evidence needed to support the observable and measurable variables and 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 (e.g., creativity) were problematic when trying to define the construct and when identifying observable and measurable evidence (Kim & 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 (i.e., 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 students to know at Grade 5. By developing a game-based assessment that is guided by specific learner outcomes, it allows classrooms 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 only provided with three choices: (a) straight, (b) clockwise, and (c) 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 9, 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; as such, this is an area that will hopefully be enhanced with future iterations of this assessment.

Another challenge faced by the development team was ensuring 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 only represented a small sub-set of this construct. Additionally, the selected sub-set of outcomes were typically quite closed-ended. For example, learner outcome 5.1.2 indicated student should "identify one or more possible answers to questions by stating a prediction or a hypothesis" (Alberta Education, 1996, pg. 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 prevents the open-ended nature of science inquiry to be assessed.

During the beta testing phase and 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. There is also a necessity for future research 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.

Digital game-based assessments, although starting to become more popular, are still in their infancy (Shute & Venture, 2013). In order to address some of the criticisms of entertainment and education game-based assessments, it is important that new games are created with proper game designs 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 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 of 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.

Reference

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Alberta Education. (1996). Science (elementary). Edmonton, Alberta: Alberta Education. Retrieved from http://www.education.alberta.ca/media/654825/elemsci.pdf

Barab, S. A., Gresalfi, M. S., & Ingram-Goble, A. (2010). Transformational play: Using games to position person, content, and context. *Educational Researcher, 39*(7), 525-536. Retrieved from http://ase.tufts.edu/DevTech/courses/readings/Barab_Transformational_Play_2010.pdf

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2010). *An evidence centered design for learning and assessment in the digital world* (CRESST Report 778). Retrieved from The National Center for Research on Evaluation, Standards, and Student Testing website: http://files.eric.ed.gov/fulltext/ED520431.pdf

Bennett, R.E., Persky, H., Weiss, A.R., and Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP technology-based assessment project* (NCES 2007–466). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from the Institute of Education Sciences website: http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp

Council of Canadian Academies [CCA]. (2015). *Some assembly required: STEM skills and Canada's economic productivity: The expert panel on STEM skills for the future*. Ottawa (ON): The Expert Panel on STEM Skills for the Future, Council of Canadian Academies. Retrieved from http://scienceadvice.ca/uploads/ENG/AssessmentsPublicationsNewsReleases/STEM/STEMFullReportEn.pdf

Gee, J. P. (2007). *Good Videogames + Good Learning: Collected Essayson Videogames, Learning and Literacy.* New York: Peter Lang Publishing.

Leighton, J.P. & Gierl, M.J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. Educational Measurement: Issues and Practice, 26, 3-16. Doi: 10.1111/j.1745-3992.2007.00090.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. Doi: 10.1111/jedm.12000

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in gamebased assessment. *Computers & Education, 87*, 340-356. Retrieved from http://myweb.fsu.edu/vshute/pdf/creativity.pdf

Mislevy, R.J., Almond, R.G., & Lukas, J. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). Retrieved from Educational Testing Service website: https://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Mislevy, R.J., Oranje, A., Bauer, M.I., von Davier, A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., & John, M. (2014). Psychometric considerations in game-based

assessment. GlassLab: Institute of play. Retrieved from http://www.instituteofplay.org/work/projects/glasslab-research/

National Science Foundation [NSF]. (2015). *Revisiting the STEM workforce: A companion to science and engineering indicators 2014 Retrieved from* http://www.nsf.gov/nsb/publications/2015/nsb201510.pdf

Rieber, L. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Education and Technology Research & Development, 44*, 42-58. Doi: 10.1007/BF02300540

Rowe, E., Asbell-Clarke, J. & Baker, R. (2015). Serious game analytics to measure implicit science learning. In C.S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), Serious Game Analytics: Methodologies for Performance Measurement, Assessment, and Improvement. New York, NY: Springer Science+Business.

Rupp, A.A., Gushta, M., Mislevy, R.J., & Shaffer, D.W. (2010). Evidence-centered Design of Epistemic Games: Measurement Principles for Complex Learning Environments. Journal of Technology, Learning, and Assessment, 8(4). Retrieved from http://edgaps.org/gaps/wp-content/uploads/rupp2010.pdf

Shute, V. J., & Ventura, M. (2013). Measuring and supporting learning in games: Stealth assessment. Cambridge, MA: Massachusetts Institute of Technology Press. Retrieved from http://myweb.fsu.edu/vshute/pdf/white.pdf Shute, 2011, 2013

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education, 80*, 58-67. Doi: 10.1016/j.compedu.2014.08.013

Shute, V., Leighton, J. P., Jang, E. E., & Chu, M-W. (2016). Advances in the science of assessment. *Educational Assessment, 21*(1), 34-59. Doi: 10.1080/10627197.2015.1127752. Retrieved from: http://myweb.fsu.edu/vshute/pdf/asstPPF.pdf

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295-321). Mahwah, NJ: Routledge, Taylor and Francis.

U.S. Department of Education. (2015). *Science, technology, engineering and math: Education for global leadership* Retrieved from http://www.ed.gov/stem

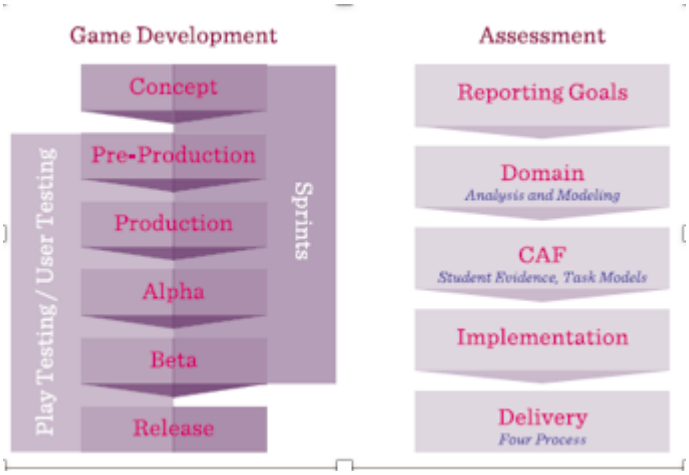

Figure 1. Design frameworks for games and assessments that are integrated using ECgD. Adapted from "Psychometric Considerations in Game-Based Assessment," by R. J. Mislevy, A. Oranje, M. I. Bauer, A. von Davier, J. Hao, S. Corrigan, … M. John, 2014, *White Paper*, p. 135. Copyright 2014 by GlassLab. Reprinted with permission.

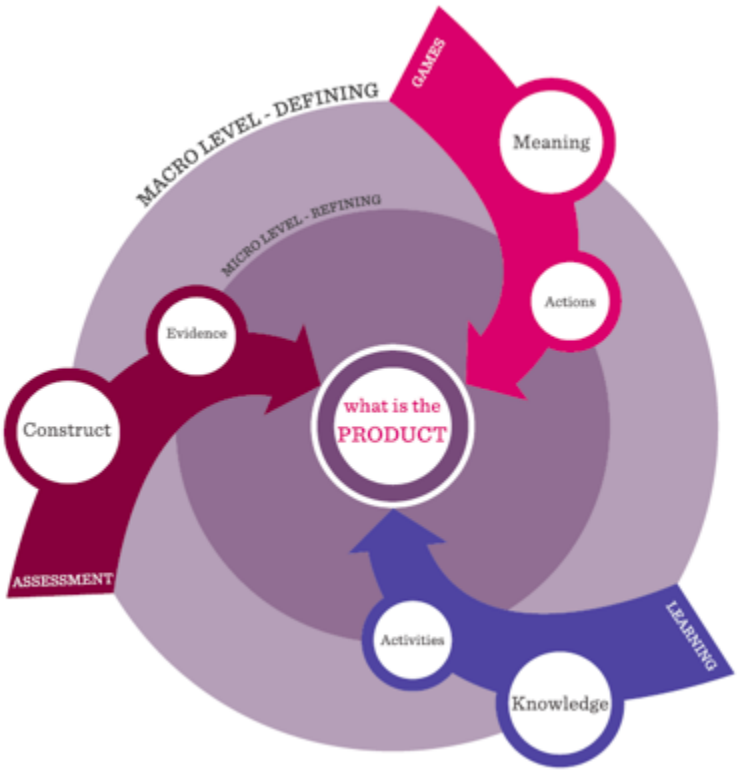

Figure 2. Model of unifying frameworks from the disciplines of games, assessment, and learning. Adapted from "Psychometric Considerations in Game-Based Assessment," by R. J. Mislevy, A. Oranje, M. I. Bauer, A. von Davier, J. Hao, S. Corrigan, … M. John, 2014, *White Paper*, p. 136. Copyright 2014 by GlassLab. Reprinted with permission.

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.

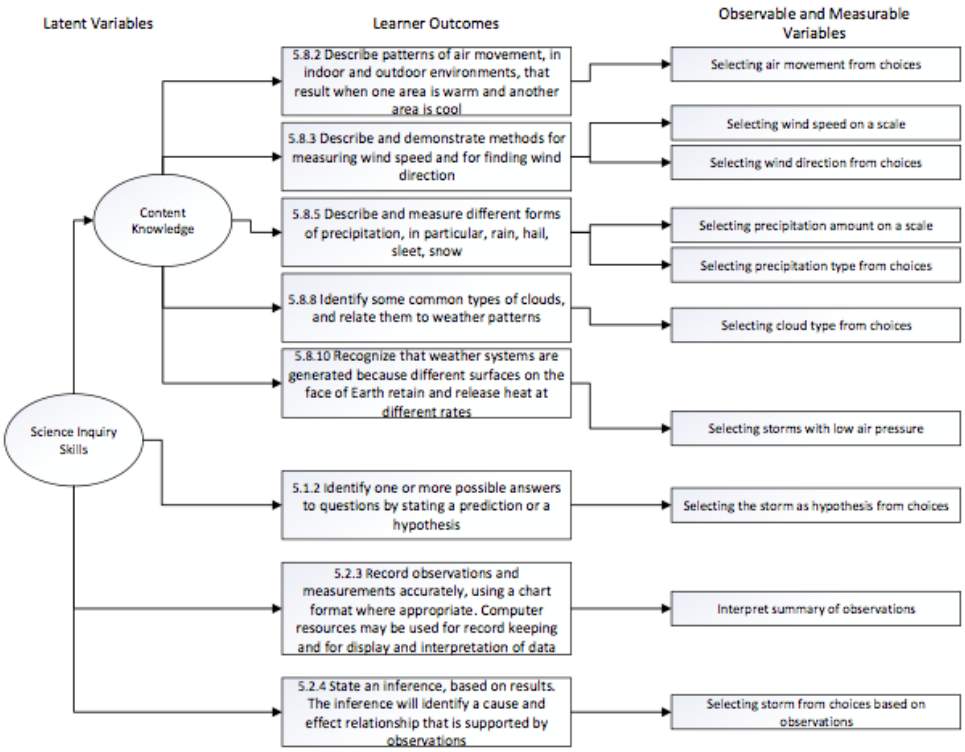

Figure 4. Competency model of learner outcomes from Alberta's Grade 5 program of study's Weather Watch unit designed specifically for the game-based assessment Raging Skies.

Figure 5. Screen shot of the storm log provided during the tutorial storm. During the initial stages of the game, this storm log is fully populated with information, as shown. The amount of information in the storm log will be inversely proportional to students' proficiency throughout

Raging Skies. This means as students' proficiency increases, the amount of information decreases. Copyright 2016 by MindFuel™. Reprinted with permission.

Figure 6. Map of North America with low and high pressure readings so that students may select the location (i.e., low pressure areas) of the next storm to chase. Copyright 2016 by MindFuel™. Reprinted with permission.

Figure 7. Screenshot of the formative feedback report students receive after a storm task. Students are rewarded with game money so that they may purchase upgrades/customize their equipment (e.g., change the color of their dashboard) and purchase gas to reach the next location.

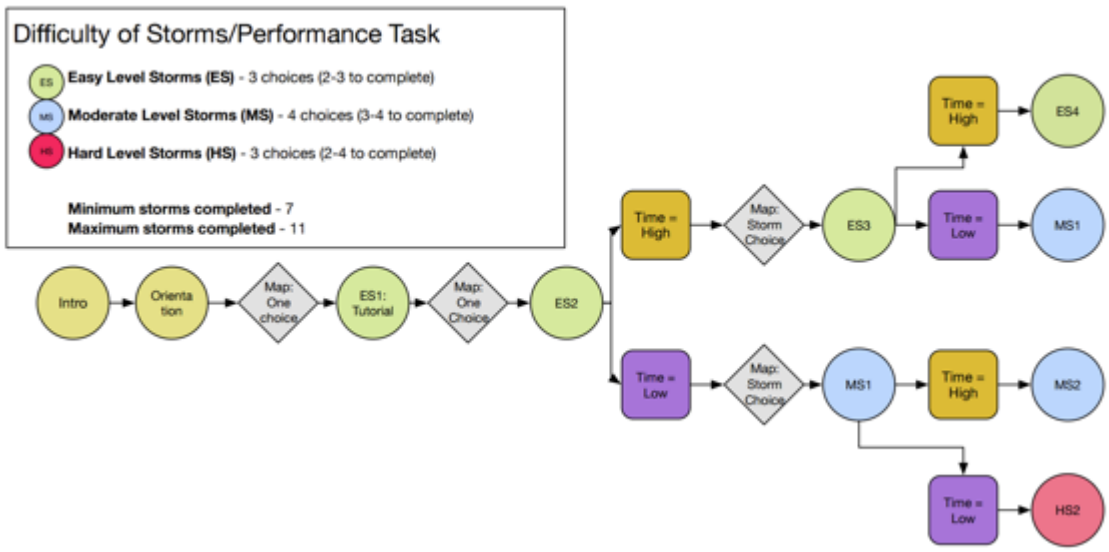

Figure 8.

Figure 9.