2022-01

Measuring a Lack of Engagement in Raging Skies

Mattingly, Peter

UNIVERSITY OF CALGARY

Measuring a Lack of Engagement in Raging Skies

by

Peter Mattingly

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN EDUCATIONAL RESEARCH

CALGARY, ALBERTA

JANUARY, 2022

Abstract

A lack of engagement is bad for measurement via assessment (Kane, 2006). Traditional qualitative measures of a lack of engagement are error prone and expensive (V. Shute & Ventura, 2013). Thus, a quantitative approach is used in this study to attempt to address such issues. This study attempts to find a characteristic behaviour associated with a lack of engagement: Rapid-Guessing Behaviour (RGB; Wise, 2017). Data used during analysis comes from a project by Dr. Man-Wai Chu which studied learner data as they played a Game-Based Assessment (GBA; Chu & Chiang, 2018); A GBA is a game used as a platform for assessment (Ren, 2019). Analysis failed to reveal a strong link between RGB and a lack of engagement in this study. However, it is believed that learners with some behaviour patterns were exhibiting *Enjoyment Seeking Behaviour*; In short, they disengaged from the assessment to seek enjoyment elsewhere.

Keywords: Engagement, Assessment, Validity, Test Anxiety, Game-Based Assessment, Log Files, Rapid-Guessing Behaviour

Table of Contents

List of Figures and Illustrations

Introduction

This study aims to address a potential issue with measurement via assessment: a lack of engagement. A lack of engagement in this study has been defined in terms of not working with an assessment correctly, such that a learner cannot be measured by an assessment. Thus, detecting a lack of engagement is of prime importance, as the purpose of an assessment is ostensibly to measure what a learner "knows and can do" (Wise & Smith, 2016, p. 204). Traditionally, detecting a lack of engagement has been difficult and error prone (Fredricks, Blumenfeld, & Paris, 2004; V. Shute & Ventura, 2013). In addition, such efforts have been labor intensive and expensive (Fredricks et al., 2004; V. Shute & Ventura, 2013). Thus, the method in this study attempts a quantitative approach that is hoped to address these weaknesses, while also allowing the approach to scale with little cost.

To that end this study is broken up into traditional sections. First, background information is reviewed, followed by a review of the method used in the study, the method is then applied to the data and the results discussed. The construct of *engagement* is defined first, so that the construct of a *lack of engagement* can be considered. This leads into a review of how a lack of engagement threatens measurement by an assessment, the idea of *validity*. Validity is defined by how trustworthy the measurement from an assessment is believed to be (Kane & Mislevy, 2017; Kane, 2006). Following this, the construct of *test anxiety* is also argued to be a threat to validity. Test anxiety can be understood as anxiety that is experienced by a learner during an assessment (von der Embse, Jester, Roy, & Post, 2018; Von der Embse, 2012; Zeidner, 1998a). Then GBAs are considered, as they can be designed to asses while also collecting information about learner actions that may be used to infer a lack of engagement (Wise, 2017). GBAs are, roughly, games that can be used as a platform to assess (Ren, 2019). This study uses

data gathered from a larger project supervised by Dr. Man-Wai Chu, where learners were studied while playing the GBA *Raging Skies* (Chu & Chiang, 2018). The review then turns to examining traditional means of measuring a lack of engagement. This is done to illustrate how engagement might be studied with respect to GBAs, to point out issues with those measures, and to highlight why the approach used in this study might address those issues. For example, one common approach to measuring engagement uses a self-report instrument to ask the learner how engaged they thought they were during assessment (Shute & Ventura, 2013). This can have the issue of relying on the learner's imperfect recall and not capturing engagement as it happens (Shute & Ventura, 2013). Thus, the proposed measure for engagement in this study uses *log files*, records of learner actions (Kerr, 2015), to attempt to capture a lack of engagement as it happens and to remove learner recall from measurement. With the log files as the unit of analysis, the study then checks for a characteristic behaviour associated with a lack of engagement: *RGB* (Guo et al., 2016; Shute et al., 2015; Wise, 2017). RGB can be visualized as a learner rapidly guessing responses to tasks on a multiple-choice assessment (Wise, 2017). This analogy captures the essence of why RGB is associated with a lack of engagement: The learners are not paying attention to what they are doing and certainly not working with the assessment as it was intended (Kong, Wise, & Bhola, 2007; Wise, 2017).

With the goal of detecting a lack of engagement, and this review in mind, a method of inquiry can then take shape. Given the drawbacks of attempting to measure engagement in a traditional manner, learner actions as gleaned from log files can be used. With the definition of RGB, characteristics actions can be checked for in the log files. Then as RGB implies a lack of engagement, learners that display RGB can be labeled as lacking in engagement. In sum, this is the first step of the method: Scanning log files of learner actions and checking for actions that

can be identified as RGB. The second step of the method then seeks to corroborate this finding of a lack of engagement based on the construct of test anxiety. One of the measures used in this study are learner survey responses from Dr. Chu's project (Chu & Chiang, 2018). One of the items on this survey relate to test anxiety. Test anxiety is argued to be related to a lack of engagement, and so the task of corroborating findings of a lack of engagement then becomes seeking significant differences in test anxiety between learners who are believed to lack engagement and those that did not display the unusual behaviour of RGB. Given this background information, the research questions can then be formally stated. The first step of the method addresses this question:

- Can evidence of Rapid-Guessing Behaviour be found in the records of learner play?

Then the second step addresses this question:

- Given evidence of a lack of engagement in Group1, can survey responses related to test anxiety provide corroborating evidence for a lack of engagement?

Where *Group1* is the group of learners with evidence of a lack of engagement in the form of RGB.

Unfortunately, the method did not find the results that were hoped for. That is, while evidence of RGB was found, there was no significant relationship with test anxiety. Conversely, the method did turn up some interesting patterns of behaviour during the analysis. More specifically, RGB was present, but there was not sufficient evidence to conclude that it represented a lack of engagement in this case. The possibility remains that the learners were avoiding assessment in order to enjoy some other activity. That is, in the case of RGB, it is assumed that learners are using RGB to avoid working with the assessment (Wise, 2017). For example, perhaps they are bored with the assessment, or they feel anxious and want to avoid the

assessment as much as possible. But this overlooks the possibility that the assessment might be an activity the learners enjoy. Recall that the learners were being assessed by a GBA. A GBA is, roughly, a game used to assess and so it stands to reason that the learners might enjoy the game-like aspects and avoid the aspects that behave like an assessment (Ge & Ifenthaler, 2018; Ghergulescu & Muntean, 2012). This is an interesting possibility that warrants further study, but for now this narrative can turn to introducing the study in earnest.

Literature Review

Introduction

This review begins by defining engagement, and consequently a lack of engagement, for this study. Validity is then discussed, followed by another potential threat to the validity of an assessment: test anxiety (Ackerman & Heggestad, 1997; Zeidner, 2007). Test anxiety is also used to help find additional corroborating evidence of a lack of engagement during analysis. The term GBA is then defined, and then measures of a lack of engagement are explored. This culminates with outlining a measure for a lack of engagement: RGB (Guo et al., 2016; Shute et al., 2015; Wise, 2017).

Engagement and Assessment

This review begins with an examination and definition of engagement. This is done to make a usable definition of a *lack of engagement* for use throughout this study. More specifically, this section will begin with a general definition of engagement, and then narrow the focus to defining engagement with respect to assessment and GBAs. This step, of defining engagement, is necessary as the construct of *engagement* is difficult to define (Fredricks et al., 2004; Henrie, Halverson, & Graham, 2015; Sinatra, Heddy, & Lombardi, 2015). During research for this study, it was commonly observed that researchers synthesized their own definitions of

engagement; This practice is continued in the subsequent sections. This synthesis-of-engagement begins with one of the most common works referenced when developing a definition of engagement for research; The work of Fredricks and colleagues (2004). Then subsequent sub-sections elaborate on further aspects of engagement that are not present in the initial definition from Fredricks and their colleagues (2004). These aspects are, namely: the *scope* of engagement and how engagement might be observed.

To elaborate, the *scope* of engagement concerns the scale of engagement that is of interest to the study. For example, it is possible to research engagement at the scale of a single learner as they work with a classroom task (de Vreede et al., 2019; Filsecker & Kerres, 2014; Sinatra et al., 2015; Wang & Degol, 2014). It is also possible to research engagement at the scale of how all learners in a school might engage with their community (de Vreede et al., 2019; Filsecker & Kerres, 2014; Sinatra et al., 2015; Wang & Degol, 2014). The discussion on the scope of engagement narrows the focus of inquiry to a specific scale of engagement for the purposes of this study; That of learner engagement with the tasks of a GBA. The review then turns to how engagement might be observed. That is, given the variety of phenomena that a learner might take part in while engaged, which phenomena might act as a good indicator of engagement? This is answered by assuming that learner's actions are such an indicator (Fredricks & McColskey, 2012; Henrie, Bodily, Larsen, & Graham, 2018; Lu, Zhang, Li, Chen, & Zhuang, 2019; Wang & Degol, 2014). Finally, given this extended synthesis a definition of engagement, and subsequently the construct of a *lack of engagement*, for this study is made.

The challenges of defining engagement. As previously mentioned, engagement is difficult to define (Fredricks et al., 2004; Henrie et al., 2015; Sinatra et al., 2015). The research community surrounding engagement has established several reasons for this difficulty. First, it

can seem like there are as many definitions of engagement as there are articles discussing it (Fredricks et al., 2004; Henrie et al., 2015; Sinatra et al., 2015). Second, to complicate this plethora of definitions, many authors do not provide a definition of engagement in their work (Fredricks et al., 2004; Henrie et al., 2015; Sinatra et al., 2015). Third, to add to the confusion there are many terms that are often associated with or conflated with engagement; Such terms include: motivation, flow, and effort (Azevedo, 2015; Henrie et al., 2015; Hookham & Nesbitt, 2019; Lumsden et al., 2016; Shute & Ventura, 2013). Fourth, definitions for engagement found in research can draw on both scientific and unscientific definitions (D'Mello, Dieterle, & Duckworth, 2017). For example, a working definition of engagement might be synthesized based on the views of non-scientists that are involved with education (D'Mello et al., 2017). Finally, there is little agreement about how to measure engagement (Fredricks et al., 2004; Henrie et al., 2015; Sinatra et al., 2015).

These difficulties imply that there is likely no conventional definition of engagement to draw upon when establishing a definition of engagement for this study. Because of this lack of consensus, a definition of engagement must be constructed and argued for, in order to justify its use during this study. Thus, the following sections follow this novel course of developing and establishing evidence for a definition of engagement for this study. As such, what follows are strong arguments drawn from researchers that work with engagement, which are then synthesized into a singular definition of engagement. Finally, this definition is used to infer a definition for a *lack of engagement* used in this study.

Engagement in three dimensions. As promised, the review begins with the most common definition of engagement as a construct with three-dimensions (Filsecker & Kerres, 2014; Fredricks et al., 2004; M.-T. Wang & Degol, 2014). Fredricks and colleagues (2004)

define engagement as “how students behave, feel, and think” (p. 60) about an assessment or learning activity (Balasooriya, Mor, & Rodríguez, 2018; Gobert, Baker, & Wixon, 2015; Henrie et al., 2015; Sinatra et al., 2015). Behavioural engagement is defined as what actions a learner takes as they work with an assessment or learning activity. For example, behavioural engagement can be seen in a learner completing tasks, producing a work product, or selecting an response for a multiple-choice assessment item. Further, affective engagement is the emotional dimension of learner interactions with a learning activity or assessment. For instance, a learner might feel anxiety when working with an assessment. Finally, cognitive engagement is the mental labour a learner goes through during their work with a learning activity or assessment. For example, cognitive engagement is displayed in maintaining attention, problem solving, or mastering skills (Ding, Kim, & Orey, 2017; D’Mello et al., 2017; Fredricks et al., 2004; Fredricks & McColskey, 2012; Kim et al., 2017; Sinatra et al., 2015).

The scale of engagement. While the three-dimensional definition of engagement is well-supported, it does not speak to all potential aspects of engagement (Fredricks et al., 2004; Sinatra et al., 2015; Wang & Degol, 2014). That is, while it does define the types of learner actions which can be included with engagement, it does not define a *scale* for them. This idea of scale can be seen from the definition of the three-dimensional conception of engagement; That of how a student behaves, feels, and thinks about an assessment or learning activity (Fredricks et al., 2004). The term *learning activity* may mean many things: from a single in-class assessment to participation in a class more generally. There is a significant difference in what a learner does when working on an assessment, as compared to completing an entire class. A way to operationalize a sense of scale with respect to these different activities is the time spent by the learner while taking part in them. More specifically, a learner may spend a few minutes working

on an assessment, while they might spend weeks or months working on completing a class. Thus, there seems to be a need to differentiate different types of engagement, with respect to scale and consequently time spent by the learner. Sinatra and colleagues (2015) elaborate on this idea with their conception of *grain-size*. They conceptualize the scale of engagement on a continuum from *micro* (small) to *macro* (large) grain-size (Sinatra et al., 2015). Going back to the examples, an instance of engagement at the micro grain-size might be a learner working on an assessment. Further, engagement at the macro grain-size could be the work a learner puts into complete a class. Using the operationalization of time spent by the learner, engagement at the micro grain-size could be realized as a learner spending a small amount of time engaged with an assessment or activity. Similarly, a macro grain-size can be understood as engagement by a learner over a larger span of time.

Defining engagement, first synthesis. Given these first two different conceptions of engagement, a first effort at definition for engagement can be attempted. The reason for this initial synthesis is to introduce the reader to the developing definition of engagement used in this study. That is, the argument will be presented for the synthesis of the first two conceptions of engagement as an introduction to a more complete definition at the end of the section.

The first choice when synthesizing the two previous definitions is to choose which type of engagement to focus on: behavioural, cognitive, or affective. As will be argued in a subsequent section, this study will focus on what a learner does (their behaviour), as opposed to what they feel and think, as an indicator of engagement. As such, the developing definition of engagement for this study will adopt the *behavioural engagement* definition. Further, as the focus of this study is detecting a lack of engagement as learners work with the tasks of a GBA, the scope (grain-size) of engagement will be defined as *micro*. More specifically, this study is

concerned with how learners respond to assessment tasks, which may take only a few seconds per-response. As a succinct label, this initial definition of engagement can be termed *micro-behavioural engagement*. Then, moving forward, this initial definition will be revised with another conception of engagement, before a final definition is synthesized.

Engagement in action. Returning to the topic of how to define engagement, Fredricks and colleagues (2004) define categories of learner activities in which engagement might be found (behaviour, cognition, and affect), their work does not speak of how to measure engagement. A common way that some researchers measure engagement is through learner actions (Fredricks & McColskey, 2012; Henrie et al., 2018; Lu et al., 2019; Wang & Degol, 2014). That is, engagement is inferred and measured from what learners do. Consider an illustrative example of a learner working with a GBA. A GBA may contain game-like elements for a learner to interact with (Ren, 2019). For instance, such elements could be buttons for learners to press as they work with the GBA. Thus, learner actions could be inferred from learners pressing these buttons while they play in the GBA. Then, learner actions could be deduced by monitoring what buttons the learner pressed during their time with the GBA. Therefore, assuming engagement can be inferred from learner actions, learner interactions with a GBA could be a reasonable proxy to measure engagement. Thus, the review continues with determining which type of learner actions can be used to measure engagement.

Suitable and unsuitable actions. As the focus of this study is measuring engagement during assessment, specifically GBAs, the review continues with examining which actions can be taken during assessment that might be useful for measuring engagement. Learners take actions during assessment, but not all such actions are suitable for assessment (Henrie et al., 2018; Wang & Degol, 2014). More specifically, some learner actions are suitable for assessment

as they derive from what a learner knows and can do, and some actions do not. The existence of such suitable actions implies the existence of unsuitable actions; Or actions that learners might take during assessment that are not suitable for assessment. An example of unsuitable actions might be the actions of a learner responding randomly to assessment tasks. Such actions do not reflect what the learner knows and can do, but rather random chance. Consequently, as the aim of an assessment is to measure what a learner knows and can do, using responses that result from unsuitable actions as evidence of what a learner knows and can do is dubious at best (Wise & Smith, 2016). This argument also extends to GBAs, as GBAs are a type of assessment (Ren, 2019).

Suitable actions have been associated with the term *engagement* by researchers into assessment design (Kane & Mislevy, 2017; Kane, 2006); To avoid confusion, this conception of engagement can be labeled as *assessment-engagement.* More specifically, a learner taking suitable actions is said to be assessment-engaged. Similarly, unsuitable actions imply a lack of assessment-engagement. With the goal of determining which type of actions (suitable vs. unsuitable) may be used to measure engagement in this study, a useful argument is presented by Dr. Wise (2017) in their discussion of RGB. They define at least two types of response-behaviour for learners: solution behaviour and RGB (Wise, 2017); where they discuss *behaviour* in terms of collections of actions learners take while working with an assessment (Wise, 2017). These behaviours are defined in more detail in a later section, but an important distinction between them is that RGB is defined in terms of a learner avoiding interaction with an assessment (Wise, 2017). A learner avoiding interaction with an assessment could be reasonably assumed to produce responses that were unsuitable for assessment. That is, as a learner avoiding interaction with an assessment, is likely not using what they know and can do to respond, thus,

they are performing unsuitable actions. With this connection between assessment-interaction-avoidance and unsuitable actions, consider the definition of engagement set forth by Fredricks and colleagues (2004); Recall their definition as: “how students behave, feel, and think” about an assessment or learning activity (Fredricks et al., 2004, p. 60). A learner that is avoiding interaction with an assessment, can be assumed to not be behaving, feeling, or thinking much with respect to the assessment. Thus, it appears that the behaviour of avoiding interaction with an assessment, can be labeled as lacking engagement as per Fredricks and colleagues definition (2004). Then, recalling the connection between avoiding interaction with an assessment and unsuitable actions, suitable actions can be logically associated with engagement via this argument. In summary, for the purposes of this study, suitable actions can be assumed to be indicative of engagement.

A definition for engagement and a lack of engagement. Given the previous review, a full definition for engagement for this study can be synthesized. Then, given this definition, a definition for a *lack of engagement* for this study can be made. During the first synthesis in this section, engagement was defined as micro-behavioural engagement. Given the concepts reviewed in the subsequent sections, a means of observing and measuring engagement can be added; That of observing *suitable learner actions*. Thus, the complete definition of engagement for this study can be stated as: micro-behavioural engagement as measured by suitable learner actions as they work with a GBA. Then, logically, a lack of engagement can be defined as: micro-behavioural engagement as measured by unsuitable learner actions as they work with a GBA. To summarise, a lack of engagement in this study can be defined as learners working with a GBA in such a way as their actions cannot be used for measurement.

Validity and a Lack of Engagement

In the last section a lack of definition was defined for this study. This definition encapsulates the idea of being unable to measure what a learner knows and can do because of their actions. This idea is elaborated on with the idea of *validity*.

Validity and the interpretive argument. Validity is the amount of trust placed in the inferences drawn from an assessment (Kane, 2006). In other words, validity is the amount of confidence placed in the measurements of what a learner knows and can do based on the results of an assessment. This confidence is supported by the *interpretive argument* (Kane, 2006). In simple terms, the interpretive argument provides a set of arguments supporting a process detailing how learner work could be interpreted to make claims about what learners know and can do (Kane, 2006). That is, assessments are often designed to link learner work on their tasks to claims about what they know and can do (Kane, 2006). This link is defined by the *interpretive argument* (Kane, 2006). For example, a process for interpreting learner work on an assessment to make claims may follow this structure:

1. Learner work is analyzed for evidence.
2. This evidence is used to draw inferences about learner performance.
3. Then the inferences are used to argue for claims about what a learner knows and can do.

Each step in this process then could be supported by argument, which, in sum would be the interpretive argument. For instance, the argument support Step 1, might show support for how a particular means of analysis applies to the evidence gathered by the tasks of the assessment. The argument for Step 2 might detail inferences that the assessment claims to make, and the supporting evidence needed to substantiate those claims. Finally, the argument for Step 3 might detail how the inferences from Step 2 support claims about what a learner knows and can do. The

interpretive argument is a chain of reasoning, that maps evidence from learner work, to claims about what they know and can do. An interpretive argument whose arguments are supported by theory and evidence can be used to argue strongly for its claims (Kane, 2006). For example, an interpretive argument based on research and testing could present a strong argument about its associated assessment claims. Claims from a strong interpretive argument are said to have strong *validity*. Strong validity implies that the claims of an assessment likely closely match what a learner knows and can do (Kane & Mislevy, 2017; Kane, 2006; Lehman et al., 2019; Wise & Smith, 2016).

A lack of engagement and assessment. One way to weaken validity is a lack of engagement. With an interpretive argument there is commonly an assumption of engagement by learners with assessments (Kane & Mislevy, 2017; Kane, 2006). Engagement, with respect to validity, is defined in terms of how a learner uses an assessment. This definition of engagement is closely related to the idea unsuitable actions. That is, a learner is said to lack engagement when they working with an assessment as it was assumed to be used (Kane & Mislevy, 2017; Kane, 2006); They are using unsuitable actions. Conversely, a learner that is engaged with an assessment (with respect to validity) is working with an assessment as it was assumed they would be (Kane & Mislevy, 2017; Kane, 2006); They are using suitable actions. It is assumed, for the purposes of this study, that this definition of engagement associated with validity, is compatible with the definition of engagement used in this study.

If a learner is not engaged, then it is difficult to have confidence in the claims of an assessment about what a learner knows and can do. This means that the validity of the claims of the assessment are weakened. This is because the assumption of engagement by the interpretive argument does not hold. For example, a learner guessing the answers to tasks on an assessment

lacks engagement. Thus, the assessment claims about what this learner knows can do are suspect. Put another way, in this example, the assessment claims about what a learner knows and can do may reflect random chance. This implies that detecting a lack of engagement is important for validity. Additionally, if validity is weakened because of a lack of engagement, then claims about what a learner knows and can do can be distorted.

Test Anxiety

Another construct of interest to this study is *test anxiety* as it is used to support conclusions about detecting a lack of engagement. *Test anxiety* is defined as a collection of negative emotional, behavioural, and cognitive responses from a learner while being assessed (Ergene, 2003; Hembree, 1988; Seipp, 1991). These responses are presumed to be triggered by concern about the assessment or the consequences of the assessment (von der Embse et al., 2018; Von der Embse, 2012; Zeidner, 1998a); For instance, a learner could be concerned about failing an assessment and the consequences of that failure in the future. In other words, test anxiety is anxiety that occurs during assessment (Seipp, 1991).

Test anxiety can be modelled as potentially two types of anxiety: *state anxiety* and *trait anxiety* (Von der Embse, 2012; Zeidner, 1998b). State anxiety refers to experiencing anxiety when being assessed, as if the state of being assessed acts as a trigger for anxiety (Von der Embse, 2012; Zeidner, 1998b). Conversely trait anxiety refers to anxiety that is a more permanent state of affairs, as if a learner had a heightened potential to be anxious and this anxiety is exacerbated during assessment (Von der Embse, 2012; Zeidner, 1998b). Current research favours the *trait* interpretation of test anxiety (Zeidner, 1998b).

There are also two widely described components of test anxiety: *worry* and *emotionality* (Hembree, 1988; Seipp, 1991; von der Embse et al., 2018; Von der Embse, 2012; Zeidner,

1998b, 1998d). Worry refers to the cognitive aspects of test anxiety, these could include thoughts about failure or the consequences thereof (Hembree, 1988; Seipp, 1991; von der Embse et al., 2018; Von der Embse, 2012; Zeidner, 1998b, 1998d). Emotionality refers to the physical symptoms of test anxiety (Hembree, 1988; Seipp, 1991; von der Embse et al., 2018; Von der Embse, 2012; Zeidner, 1998b, 1998d). A large meta-analyses indicates that the *worry* component of test anxiety is responsible for many undesirable aspects related to test anxiety (Seipp, 1991).

Over the past 10 years there appears to be few works that discuss both the construct of test anxiety and engagement (Fredricks et al., 2004). In total there appear to be four works of research that discuss both constructs. More specifically, these four works at least mention the constructs of *test anxiety* and *engagement* and cite Fredricks and colleagues (2004). Three of the works examine how test anxiety and engagement vary with respect to another construct, while the fourth focuses solely on engagement and how test anxiety might affect engagement (Brallier, 2020; Liu, Yao, & Li, 2020; Raufelder, Hoferichter, Ringeisen, Regner, & Jacke, 2015; Shoemaker, 2017). Shoemaker (2017) discusses engagement and test anxiety in the context of how the constructs vary with respect to *mindful awareness*. Similarly, Raufelder and colleagues (2015) study *parental support and pressure* while observing how engagement and test anxiety vary. Then also, Liu and colleagues (2020) examine variations in test anxiety and engagement and how they relate to *achievement goal profiles*. Finally, Brallier (2020) studies how engagement relates to *academic achievement* among college students while briefly examining the relationship between test anxiety and emotional engagement. This set of research articles has few unifying themes, and the definitions of engagement vary from work to work; As previously discussed, this is to be expected, given the plethora of definitions engagement has in the field (Fredricks et al., 2004; Henrie et al., 2015; Sinatra et al., 2015). This lack of guidance as to a

relationship between test anxiety and engagement leaves little direction on how to model their relationship. Thus, a relationship and definition with respect to this study must be synthesized.

To begin this synthesis, a primary concern of test anxiety with respect to this study is its association with impaired assessment performance (Hembree, 1988; von der Embse et al., 2018; Von der Embse, 2012; Zeidner, 1998c). There is a robust negative linear relationship between test anxiety and assessment performance (Hembree, 1988; von der Embse et al., 2018; Von der Embse, 2012; Zeidner, 1998c). That is, if test anxiety is high, then test performance tends to be low. This relationship between test anxiety and performance has led several researchers to conclude that this relationship can colour or confound measurement via assessment when test anxiety is present (Ackerman & Heggestad, 1997; Zeidner, 2007). In other words, as test anxiety relates to poor performance, test anxiety can be viewed as inducing construct irrelevant variance (Ackerman & Heggestad, 1997; Zeidner, 2007); Or, more simply, test anxiety is associated with performance that may not represent what a learner knows and can do. This is very similar to a lack of engagement. To go one step further, it may be the case that test anxiety is a similar threat to validity as a lack of engagement. Recall also the working definition of a lack of engagement for this study: learners working with a GBA in such a way as their actions cannot be used for measurement. If a GBA can be recognized by learners as an assessment, then test anxiety may also be assumed to be affecting such learners; As test anxiety is seemingly triggered by the act of a learner working with an assessment (von der Embse et al., 2018; Von der Embse, 2012; Zeidner, 1998a). If it is the case that some learners experience test anxiety when working with a GBA, then it would seem to follow that their performance might be negatively affected. If low learner performance is related to test anxiety and not what a learner knows and can do, then it seems dubious to accept such performance as evidence for measurement. Thus, for the purposes

of this study, the constructs of test anxiety and a lack of engagement are assumed to be closely related. Given this presumed close association between test anxiety and a lack of engagement, it will be assumed that test anxiety can function as a covariate for a lack of engagement. Then also, measurement of test anxiety will be assumed to also indicate the presence of a lack of engagement. This review now turns to considering how a lack of engagement with assessment might be addressed via the application of GBAs.

Game-Based Assessment

A definition for a GBA is an assessment that uses a game as a "platform" (Ren, 2019, p. 6) for assessment to help improve engagement (Kiili & Ketamo, 2018; Wang et al., 2016). For example, consider an example based on some of the latest work related to GBAs (Ge & Ifenthaler, 2017; Ren, 2019; V. J. Shute & Sun, 2019). For instance, a method involving an assessment, followed by gameplay, followed by another assessment can be considered a GBA. Such an assessment might begin with a pre-test to measure a learner construct. This could be followed by the learner playing a game. Finally, they could be given a post-test. The outcome of the post-test could then be used to argue that changes to the construct were the result of gameplay (Ge & Ifenthaler, 2017; Ren, 2019; Shute & Sun, 2019). Another example of a GBA might be a game that is designed to assess. Such a game could be designed to use learner actions from play as evidence for assessment (Ren, 2019). This definition of GBA stands in contrast to the related term *gamification*. Gamification is the practice of appending game-like features to an assessment or task (Lumsden et al., 2016). While a GBA is a game, gamification does not necessarily make a task into a game.

These examples show how GBAs are assessments that use a game as a platform for assessment. However, the way they assess differs. The first example shows an assessment with

instruments that are separate from gameplay. This can be defined as a GBA with external assessment (Ren, 2019). The second example shows an assessment whose gameplay is designed to be an assessment instrument. This example is of a GBA with internal assessment (Ren, 2019). Together, the examples show that the assessment associated with a GBA can take place at different times (Ren, 2019). This definition implies that there can be many types of GBAs. Thus, a GBA is specified by the goals and methods of assessment (Ren, 2019).

This broad definition has arisen because of how GBAs have developed over the years (Kim & Ifenthaler, 2019). At first, to make a GBA meant designing a game to be an assessment (Kim & Ifenthaler, 2019). This approach uses the Evidence-Centered Design (ECD) or Evidence-Centered Game Design (ECgD) frameworks to design GBAs (Kim & Ifenthaler, 2019). A similar influential approach is Stealth Assessment (Kim & Ifenthaler, 2019; Shute & Ventura, 2013). This technique leverages ECD to design GBAs such that learners would likely not be aware they were being assessed (Shute & Ventura, 2013). These type of GBAs are designed to collect information from learner actions as they play for the purpose of analysis to find evidence that can be used for assessment. More detail about ECD and ECgD is available in the works of Dr. Robert Mislevy (see, e.g., Mislevy, Behrens, Dicerbo, Frezzo, & West, 2012; Mislevy, Oranje, Bauer, Davier, & Hao, 2014). More recently researchers working with GBAs have leveraged the tools of "learning analytics and data mining techniques" (Kim & Ifenthaler, 2019, p. 7).

GBAs designed with ECD or ECgD are of specific interest to this study. This is because these GBAs use learner actions as evidence by design. Therefore, it can be assumed the information about those actions would be available for analysis to find evidence of a lack of

engagement. Thus, the definition of GBA for this study adopts the definition of GBAs that use internal assessment, specifically those designed with ECD or ECgD.

Measures of a lack of engagement

This section examines traditional ways of measuring a lack of engagement. This is done by examining common methods used to measure a lack of engagement. Then, weaknesses of these approaches are addressed. This section is intended to act in contrast to the computational method prosed for measuring a lack of engagement used in this study.

Self-report and observational protocols. A common way to measure a lack of engagement are self-report instruments (Shute & Ventura, 2013). These instruments are surveys or other similar devices that allow the learner to rate their own lack of engagement (Shute & Ventura, 2013). Typically, these instruments are used in two ways. The first way is to embed the instrument into an activity (Shute & Ventura, 2013). This involves having the learner begin an activity and interact with it for a period of time, then interrupt the learner to use the instrument before returning to the activity (Shute & Ventura, 2013). This approach attempts to measure a lack of engagement while an activity is happening. This approach has the flaw of interrupting the activity and having the learner switch tasks to the instrument. This interruption may have a negative effect on measurement (Shute & Ventura, 2013). The second way these instruments are used is retrospectively. In this approach the learner is asked to complete an activity before completing the instrument (Shute & Ventura, 2013). That is, in this scenario, the learner is asked to recall their lack of engagement as they remember it. This has the flaw of asking the learner to interpret their lack of engagement with an activity after they have completed it. This measurement based on what a learner remembers about their lack of engagement can also have a

negative effect on measurement (Shute & Ventura, 2013). Finally, a shortcoming of both approaches is that they depend on a learner responding faithfully.

Another common method of measuring a lack of engagement is an observational protocol (Fredricks et al., 2004). This method uses raters to observe learner actions and then infer a lack of engagement from those actions (Fredricks et al., 2004). Advantages of this measure include: that it may find a lack of engagement as it happens, it avoids disrupting learner work, and it does not depend on honest responses from learners (Fredricks et al., 2004). However, a drawback is the expense of hiring and training raters. In addition, raters must use their judgement to identify evidence of a lack of engagement from learner actions, which biases their decisions.

Variance in measurement. As reviewed in the previous section there are drawbacks to traditional methods of measuring a lack of engagement. That is, both self-report and observational protocols introduce construct irrelevant variance into the measurement of a lack of engagement (Shute & Ventura, 2013). Both methods involve making personal judgments while measuring a lack of engagement. The self-report measure asks the learner to interpret their own lack of engagement. Also, observational protocols depend on a rater's judgement of learner actions. These interpretations call for making assumptions about evidence of a lack of engagement. Then these assumptions make for biased decisions and add unwanted variance in measurement. Also, self-report measures can introduce variance by interrupting learner work. This variance is unwelcome while measuring a lack of engagement (Henrie et al., 2018; Shute & Ventura, 2013). Thus, other measures that may avoid this variance are warranted.

Recording Actions in Log Files

An approach that may address these sources of variance is to use learner actions as evidence of a lack of engagement. GBAs can record learner actions in *log files*. For the purposes

of this study, a log file is defined as a record of learner actions as they work with a GBA (Kerr, 2015). Using log files as a source of evidence of a lack of engagement has several benefits as compared with more traditional means. More specifically, a GBA can record learner actions without interrupting the learner. Also, this recording can be made without personal interpretation. Finally, making the recordings is a cost-effective alternative to raters (D'Mello et al., 2017; Gobert et al., 2015; Shute & Ventura, 2013). However, using log files also comes with drawbacks.

There are several common issues with gathering evidence from learner actions as recorded in log files. These issues are their tendency to be sparse, noisy, and difficult to interpret by hand (Kerr, 2015, 2016; Oranje, Gorin, Jia, & Kerr, 2017). Sparseness is defined as the penchant for two log files of learner actions to be quite different (Kerr, 2015, 2016; Oranje et al., 2017). For example, if two learners were to work with a GBA, one might understand how to play quickly while another might have to learn them. Because of this difference in initial behaviour, the log files of the learners' actions would be quite different. Noisiness is the propensity for evidence, as found in log files, to be outnumbered by other information that would not be relevant to investigation (Kerr, 2015, 2016; Oranje et al., 2017). Or, succinctly, evidence about a lack of engagement may be buried in a lot of other information. Put another way, noisiness can be thought of as the problem of finding the signal in the noise. For example, a learner may guess responses to a few tasks in a GBA while responding faithfully to the others. Lack of engagement is present in the log files for this learner, but there are many more engaged responses. The issue of analyzing log files by hand can be seen in the scale of information involved. A single learner may produce hundreds or thousands of records. The problem quickly escalates when many

learners are considered. Analyzing many learner log files by hand quickly becomes impractical (Kerr, 2015, 2016; Oranje et al., 2017).

Rapid-Guessing Behaviour

One method of inferring a lack of engagement from learner actions, as found in records in log files, is RGB. RGB is defined in terms of haphazard actions that a learner takes while working with multiple-choice assessment tasks. These actions are defined as learner responses to multiple-choice assessment tasks that are considered to be too fast for the learner to have been able to read and understand the task (Guo et al., 2016; Shute et al., 2015; Wise, 2017).

Researchers that study RGB define it as one of two common behaviours exhibited by learner actions as they work with multiple-choice assessment tasks (Kong et al., 2007; Wise, 2017). The first behaviour is RGB, while the second is termed *solution behaviour.* In terms of the theoretical framework surrounding RGB, these two behaviours are associated with engagement; RGB with a lack of engagement, and solution behaviour with engagement (Kong et al., 2007; Wise, 2017).

These researchers define engagement in terms of how learners interact with assessment tasks. That is, a learner is said to be engaged when they are work with multiple-choice assessment tasks to the best of their ability (Kong et al., 2007; Wise, 2017). In other words, the learners are using what they know and can do when using solution behaviour on these assessment tasks. Thus, a learner using solution behaviour is said to be *engaged* with working on assessment tasks, by using what they know and can do to address these tasks. In contrast, a learner using RGB is conceptualized as lacking in engagement, and they are assumed to not be using what they know and can do to work on assessment tasks (Kong et al., 2007; Wise, 2017). Thus, in sum, in the conception of theoretical framework associated with RGB, engagement and

these behaviours are tightly coupled. The actions learners display, indicate how engaged they are with an assessment. To avoid confusion this conception of engagement is termed RGB-engagement for the remainder of this section.

It follows, in this conception, that learners doing RGB (and lacking engagement) pose a problem to assessment, as their responses likely do not contain information usable for measurement. That is, a lack of RGB-engagement is a threat to the validity of assessment (Kong et al., 2007; Wise, 2017). Recall the definition of validity: validity is defined as the confidence in the claims of an assessment (Kane & Mislevy, 2017; Kane, 2006; Lehman et al., 2019; Wise & Smith, 2016). Thus, if a learner were doing RGB, their responses would not draw on what they know and can do, and thus would likely not contain information usable for measurement. Assessments are used to measure what a learner knows and can do (Wise & Smith, 2016), thus it follows that responses that lack RGB-engagement, are inappropriate for measurement and thus threaten the validity of an assessment.

Given this examination of RGB, RGB-engagement, and validity, it can be argued that the RGB-engagement and the definition of engagement for this study are compatible. Recall that the definition of engagement used in this study is: micro-behavioural engagement as measured by unsuitable learner actions during work on a GBA. Examining this definition closely shows parallels with the definition of RGB-engagement. First, both definitions use actions to indicate or measure a lack of engagement. Second, the idea of unsuitable learner actions maps closely to those actions associated with RGB. That is, just as actions that are RGB are unusable for measurement in an assessment, unsuitable actions are also inappropriate for measurement in a similar way. Finally, both definitions of engagement make a link between a lack of engagement and threats to validity. Thus, as the two definitions for engagement share so many similarities,

RGB will be assumed to be appropriate for detecting a lack of engagement as defined in this study.

Thus, as RGB will be adopted as a tool of analysis for this study, a short discussion is warranted about how RGB can be detected. RGB is conceptualized as responses to multiple-choice assessment tasks that are considered to be too fast for the learner to have been able to read and understand the task (Guo et al., 2016; Shute et al., 2015; Wise, 2017).A common way that researchers determine a timeframe that could be considered *too fast* is visual inspection of the distribution of response times (Wise, 2017). Another way is to find the average response time to an assessment item and then labeling responses with a fraction of the average response time as RGB (Wise & Ma, 2012). For example, consider a task that has an average response time of one minute, a learner that takes one second to respond can be assumed to be using RGB. Another common feature of RGB is the correct response rate. More specifically, those responses believed to be RGB tend to have an aggregate correct response rate near random (Wise, 2017). For example, consider a set of four assessment tasks, each of which having possible four possible responses. A learner responding correctly to only one of these tasks (a ~25% aggregate correct response rate) would be suspected of using RGB. The method in this study uses a hybrid approach of examining the average response time and the aggregate correct response rate; Following from Wise (2017).

Finally, using RGB as a measure of a lack of engagement has several advantages as compared with other common methods of measuring engagement. Detecting RGB is based on learner actions, rather than interpretation of those actions. RGB can be captured in log files, and so measuring it does not interrupt the learner. Also, RGB does not have the bias seen in traditional approaches. That is, there are no raters involved to interpret which actions may

indicate a lack of engagement. Finally, this approach is cost-effective as it avoids using expensive raters.

Research Questions

The purpose of this study is to measure a lack of engagement in learners playing a GBA. The reason for measuring a lack of engagement is to ensure the validity of claims drawn from assessments. The literature reviewed up to this point supports attempting to infer a lack of engagement as evidenced by RGB. This inference could then be strengthen by finding additional evidence of a lack of engagement as implied by test anxiety. Two research questions follow from this initial approach. There are several units of analysis for the RQs. Recall, that while playing a GBA, learner actions can be recorded in log files, which contain individual records of learner play. This leads to the first research question:

RQ1. Can evidence of RGB be found in the records of learner play?

If there are records of learner play that indicate RGB, then learners with these records of play can be labelled as Group1. Further, learners that did *not* have indications of RGB in their records can be labeled as Group2.

RQ2. Given evidence of a lack of engagement in Group1, can survey responses related to test anxiety provide corroborating evidence for a lack of engagement?

If RQ2 can be answered in the affirmative, then this lends support to the conclusion that the detection of RGB is associated with a lack of engagement in a GBA.

Method

Introduction

This chapter details how the research questions are to be approached in this study. This begins with a dialog on the rationale of picking a quantitative approach for this study. Then how the information used in the study was gathered. This covers how learners were recruited to the Raging Skies project, and how the information about their actions were collected. This project sought to gather information from learners who played the GBA Raging Skies. The GBA Raging Skies is a game where the learner is cast in the role of a storm chaser. The gameplay involves taking measurements of storms and attempting to identify their type. More detail on the gameplay and the project is reviewed later in this section. Following the examination of the sample from the Raging Skies project, the sources of information used with the method are examined. Specifically, these are the log files that contain information about learner actions, and the surveys the learners took during the project. The ethics of the Raging Skies project are also considered. Finally, the steps of the method are detailed and how they relate to the RQ's.

The information for the current study came from a project led by Dr. Man-Wai Chu on the validity of the claims made from the GBA Raging Skies (Chu & Chiang, 2018). While the author of the current study did not participate in data collection, this study serves as data analysis for the larger Raging Skies project. The author of this study has been granted permission to work with anonymized data from the Raging Skies project.

Rationale

This section discusses why a quantitative approach for measuring a lack of engagement is used in this study. To review, a lack of engagement is important to measure because of its effects on validity. If a learner lacks engagement, then it is difficult to trust assessment claims about

what they know and can do (Kane, 2006). Thus, to measure a lack of engagement, either a qualitative, mixed method, or quantitative approach is called for.

A difficulty with a qualitative approach to measuring a lack of engagement, is that it can be expensive (Fredricks et al., 2004). With such an approach, raters must be acquired, trained, and compensated in order to judge a lack of engagement. This can be troublesome at large scales. A qualitative approach might be appropriate for a single classroom or study. However, qualitatively judging hundreds or thousands of classrooms may be prohibitive.

The drawback of scaling that effects qualitative approaches may also affect mixed method approaches. Mixed method approaches are a combination of qualitative and quantitative methods (Creswell, 2014). Thus, adopting a mixed method approach may invite the difficulties highlighted for qualitative approaches. A qualitative method that is part of a mixed method approach may suffer from problems related to the scale of information. Thus, the interest in a quantitative approach.

An advantage of a quantitative approach is that such a technique can conducted automatically on computers. This addresses one of the drawbacks of using qualitative methods. That is, if the method can be carried out automatically by a computer once, then it can be carried out on many computers. This implies that quantitative methods that can run automatically on computers can analyze the actions of many hundreds or thousands of learners for little cost. Such approaches also avoid the bias seen in qualitative approaches. Given the choice of a using a quantitative approach, one must then choose among quantitative methods. There are many quantitative methods for studying engagement in learners (Fredricks & McColskey, 2012). However, the choice of a method is constrained by the definition of engagement in this study. Engagement has been defined such that inquiry is constrained to learner actions. This definition

excludes those methods that gather information from other sources beyond learner actions. For example, such excluded methods might be: measuring the learner's pulse, or tracking their gaze. Thus, the method in this project is chosen to gather information from learner actions.

Ethical Considerations

The project of Raging Skies went through ethical review. Permission for the project was granted by the Ethical Review Board of the University of Calgary (ID: REB17-0242). In addition, the school board(s) involved in the project had their own ethical review processes. The school board(s) also granted permission to access the schools involved in the project.

Sample

This section examines the sample of learners that were involved in the Raging Skies project. First the context of the project is considered. This is followed by a description of how the sampling was done. Then the learners are examined: their demographics and the number of their log files that can be used for analysis.

Project setting. The Raging Skies project was conducted across several schools within a large city in Alberta Canada. All schools were a part of the same school board that granted ethics approval for the project. The classes involved in the project were grade five classes. The classes that participated had covered the topic of weather. The topic of the GBA Raging Skies was weather. The topic of weather was chosen for Raging Skies to coincide with the curriculum of the classes in the project. Thus, Raging Skies could be used to measure the curricular outcomes for weather. In sum, great care was taken to integrate Raging Skies into the curriculum and instruction of the classes that participated.

Sampling procedure. The sampling procedure recruited 464 grade 5 (approximately age 10) learners. There were several groups of people who were a part of the sampling procedure.

These groups included: the school boards that oversaw the schools, elementary or junior high school principals of those schools, teachers of classes within those schools, parents / guardians of learners, and the learners themselves. This procedure required several types of forms for people in these groups. These forms included: informational letters, consent forms, and an assent form. These forms are included in the appendices. Note that each appendix may contain more than one form. The informational letters were for: principals, teachers, parents / guardians, and learners. These informational letters can be found in Appendix A. The consent forms were for principals, teachers, and parents / guardians. The consent forms are in Appendix B. The assent form was for the learners and is included in Appendix C. There was also a script used for recruiting learners, see Appendix D. Finally, there are two additional documents that were used in the project that are included in the appendices. These documents are: a pre-test on the topic of weather, and a post-survey given after play (which is reproduced in Appendix e).

The sampling procedure began by approaching school boards. The school boards were asked for permission to contact the principals of the schools in their purview. After obtaining permission from the school board, elementary principals of those schools were invited to participate in the project; an invitation letter is shown in Appendix A. The elementary and junior high school principals that consented to participate were asked for their permission to contact teachers in their schools. See Appendix B for the principals' informed consent forms. Given a principal's permission, teachers in their schools were invited to participate in the project (see Appendix A for their informational letters). Those teachers that consented to join the project, were asked for their permission to introduce the project to their grade 5 students who had completed their weather unit (the teacher informed consent form is shown in Appendix B). Given teacher permission, this introduction took the form of an informational 10-minute class

meeting (the script for this meeting appears in Appendix D) as well as an informational letter. Appendix A shows the informational letter for learners and their parents / guardians. During this meeting students were invited to take home an informational letter about the project, for both their parents / guardians and themselves, as well as a consent form for their parents / guardians (see Appendix B for the parent / guardian consent form). These forms, the informational letter and the consent forms, contained language to inform both parents / guardians and students that the project's researchers would collect the log files of learner actions as they played Raging Skies. Two weeks after the first meeting a 90-minute class meeting was held. This two-week gap was planned for to allow enough time for students to take home the consent forms, have them signed, and then returned. In this meeting the signed parent / guardian consent forms were collected; consent forms are found in Appendix B. Those students that had parent / guardian permission were invited to participate in the project. This participation involved permitting researchers to study the log files of learner actions that the learners produced during the project. All learners could play Raging Skies, whether they participated in the project or not. The learners that signed their assent forms became participants; the assent form is in Appendix C. During the second meeting learners and participants could play Raging Skies. The second meeting began with a pre-test that gauged participant knowledge on weather. After the pre-test the participants and other learners played Raging Skies. After playing, the participants were asked to complete a short survey about their: feelings while playing, technology use, demographic information, and to elicit feedback about the game itself; the post-survey is in Appendix e.

To add more detail as to why both learners and participants could play: their teachers insisted on it. The reasoning of the teachers was understood to be: If Raging Skies is a benefit, then all learners, not only participants, should be allowed to take advantage of it. Thus, the

sampling procedure resulted in a convenience sample of participants who had parent / guardian consent and who assented to have their log files of their play of Raging Skies studied.

Participants. The participants were 464 grade 5 students (age ~10). Of the participants 226 identified as male (~49%), 214 identified as female (~46%), eight other (~2%), and 16 preferred not to state (~3%).

Sample size. The sampling procedure resulted in 462 log files. Each of the participants, associated with these log files also completed the post-survey. In total five logs were discarded, leaving a total of 457 for analysis. Four of the log files had to be discarded because of a technical issue that left them blank. One log file had to be discarded because it was corrupted. The log file was considered corrupted as it was a perfect duplicate of another participant's log file. Thus, the total number of usable log files was 457. These log files were used in the analysis.

Measures

There were four sources of data for the Raging Skies project, two of which will be used in this study. The sources of data for the project were the: pre-test, Raging Skies log files, post-test, and post-survey. The pre-test and post-test were not used as a source of data for this study. This is because the focus of this study is to measure engagement, rather than being concerned with test marks.

id	action	result	correctAnswer	Stormid	timestamp	DateOfAssessment
10254253	HudResultsInterface	Results_Interface_Continue	null	1	07/06/18 11:24	07/06/18
10254253	Map_Chase_Storm	btn_chase	null	null	07/06/18 11:24	07/06/18
10254253	HudPrecipitationType	Rain	Rain:Hail	8	07/06/18 11:25	07/06/18
10254253	HudPrecipitationAmount	Medium	Light	11	07/06/18 11:12	07/06/18
10254253	StormLog	triple_cloud	null	9	07/06/18 11:03	07/06/18
10254253	StormLog	double_cloud	null	9	07/06/18 11:03	07/06/18
10254253	HudWindDirection_Open	null	null	3	07/06/18 11:16	07/06/18
10254253	Map_Weather_Icon	icon_pressure_low	null	null	07/06/18 11:22	07/06/18

Figure 1. A portion of an anonymized log file.

The first source of data for this study is the Raging Skies log files, or simply *log files*, that contain records of participant actions as they played Raging Skies. The information in the log files was analyzed to detect RGB. Figure 1 shows a sample anonymized log file and participant actions. The second data source was the post-survey (termed the *survey*) which contained responses from the participants. These survey-responses were used to find differences among participant groups about how they felt about playing Raging Skies. Both sources of data, log files and the survey, were analyzed in the method.

Raging Skies and Log Files. This section discusses the GBA Raging Skies and the log files that contain records of participant click actions as they play. Here the gameplay of Raging Skies is reviewed. In addition, the structure and content of the log files are examined. The log files record information as participants clicked on elements of Raging Skies. The elements include virtual instruments, the Storm Log, and the graphic used for completing a hypothesis or conclusion task concerning a virtual storm; as seen in Figure 2. The virtual instruments are

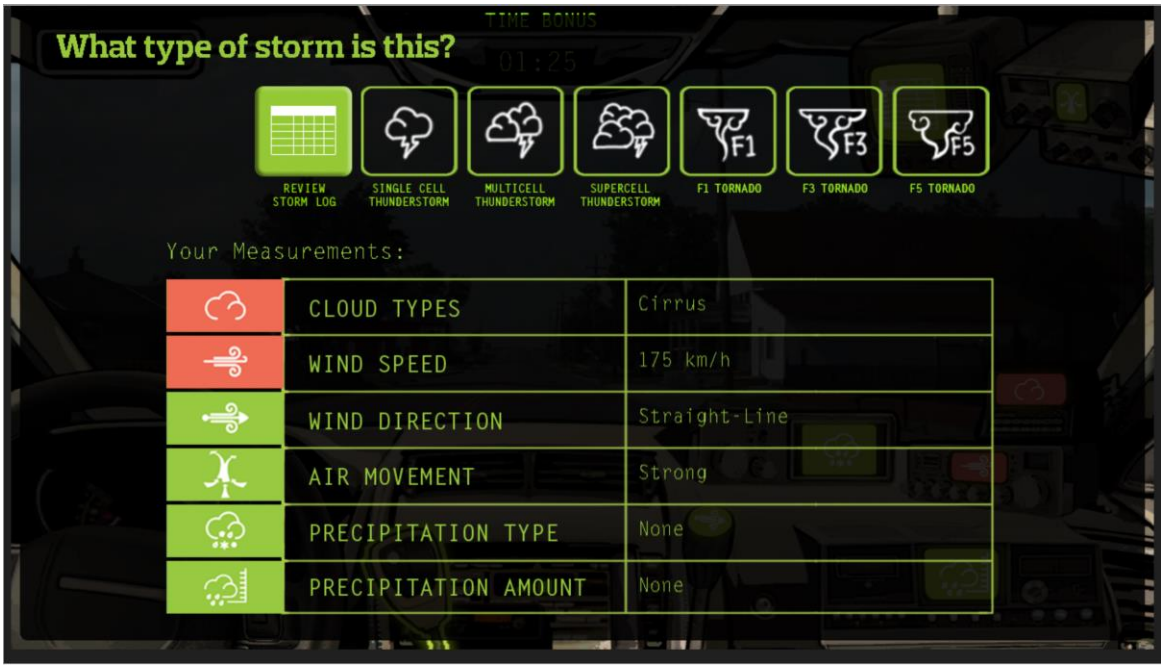


Figure 2: Making a hypothesis or conclusion about a storm.

shown in Figure 3. Figure 3 shows the graphic of the dashboard of the car that participants are shown during play. The Storm Log is shown in the Storm Log section in Appendix f. These elements are defined and discussed in the following paragraphs. Raging Skies is presented as a game that puts the participant in the role of a Storm Chaser.

Raging Skies. Raging Skies begins with a picture of a mighty tornado destroying a barren landscape with the title *Storm Chasers Raging Skies* overlayed. A large green start-button invites a participant to press it to begin the game (Figure 3).

The next scene is a dirty garage, tools and computer equipment take center-stage, while a beat-up car can be seen off to the side festooned with tools (Figure 4). After a few moments an alarm sounds, the screen flashes, and a voice (the narrator) says “We've got some storm activity; Let's hit the road." This invites the participant to begin the gameplay proper.

Figure 3. Raging Skies splash screen.

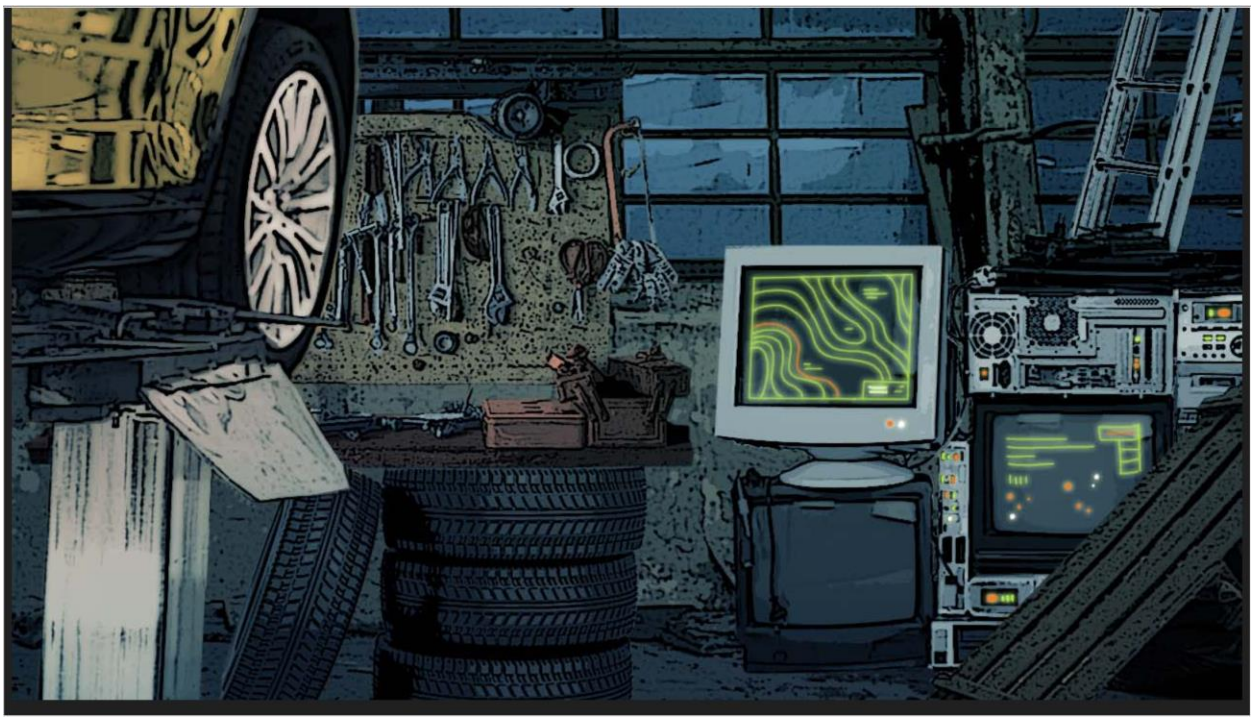


Figure 4. The garage.

After clicking on the computer, the scene changes to the inside of a car, while a fierce storm rages across the landscape (Figure 5). The narrator continues and explains the gameplay, the instruments used to take measurements of the storm, the *storm log*, and how many attempts the participant can be expected to make when making measurements; The storm log is an artifact within the game that contains the characteristics of storms as they might be measured by the in-game instruments. In sum, the participant is tasked with *chasing* a storm, taking measurements, and using the information in the storm log to determine the type of storm being chased. There are six types of storms in the game: Single Cell Thunderstorm, Multicell Thunderstorm, Supercell Thunderstorm, F1 Tornado, F3 Tornado, and F5 Tornado.

There are six different measurement tools: Wind Direction, Precipitation Type, Precipitation Amount, Wind Speed, Cloud Type, Updraft Speed / Air Movement.

Figure 5. The first storm.

The tools are labelled in Figure 6 with the measurement tasks they can be used for. A click on a measurement tool will show the participant a unique graphic for that tool; The graphics for the measurement tools can be found in Appendix f. Each tool prompts the participant with a task related to the storm. Specifically, each task is to infer a characteristic of the storm they are chasing, and to select a response that fits one of these characteristics. For example, clicking on the Precipitation Amount tool would task the participant to choose between the options of *Light, Medium, Heavy,* or *None* relating to the amount of precipitation that can be seen in the graphic of the storm; See Appendix g for more information on the options available for each task. The first time a participant clicks on a measurement tool a *simulated* version of its associated task is presented, along with a short written and verbal description of what the tool is and an explanation of what each of the responses relates to with respect to the storm.

*Figure 6***.** Measurement tools and the Storm Log.

This special instance of each task is labeled with the term *simulator*; For example, clicking on the Precipitation Amount tool for the first time would show the task of deducing the precipitation amount of the storm, but labeled with: *Precipitation Amount Simulator*. The simulated version of each task allows the participant to interact with the task freely, so they experience how the task works without consequence. After the participant interacts with the simulated version of the task, the simulation is replaced by the normal task; For example, if a interacted with the *Precipitation Amount Simulator* task, after the interaction the label would change to *Precipitation Amount* indicating that the participant would then be interacting with the normal task.

Each task can be answered correctly or incorrectly. A correct answer is indicated with the graphic of the measurement tool highlighting in green, while an incorrect answer is shown by the graphic being highlighted in red (Figure 7).

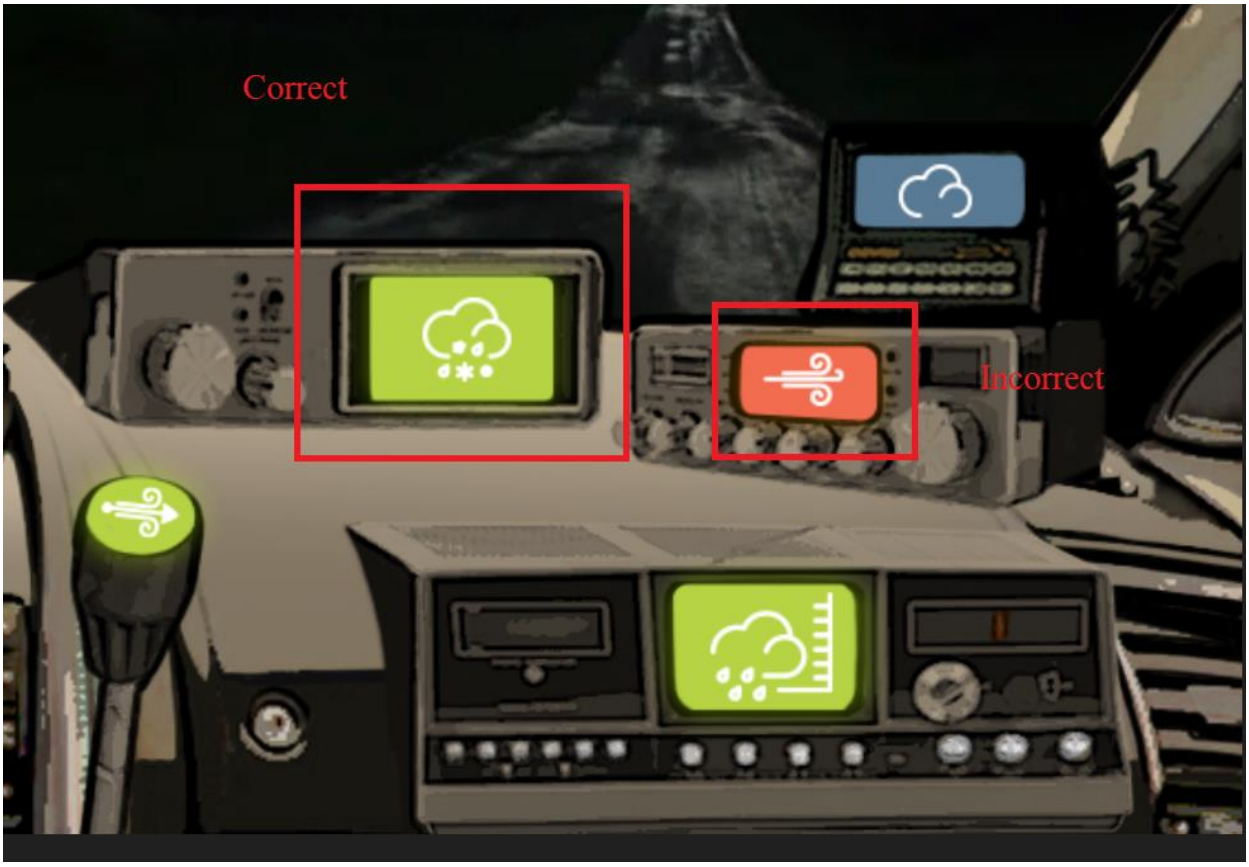


Figure 7. Correctly and incorrectly answered measurement tasks.

During the first storm the participant can attempt each measurement tool task three times before they are not allowed any further guesses. On subsequent storms there is a limit of two attempts. Correct and incorrect responses to these tasks vary with the type of storming the participant works with.

A participant is prompted to make a hypothesis about the type of storm they are working with after attempting three tasks. This hypothesis takes the form of a prompt that appears on the screen and tasks the participant to choose a storm type (Figure 8). They are prompted to guess at the type of storm they are chasing, given the information they have gathered thus far. They are not penalized or rewarded for completing this task.

Figure 8. Making a hypothesis.

Similarly, after the participant has exhausted their attempts at all measurement tool tasks, they are prompted to come to a conclusion about what type of storm they are working with (Figure 9). This conclusion-task is scored and the participant is shown a graphic indicating if they were correct or incorrect. After this, a *results* graphic is displayed showing overall performance on measurement tasks and drawing conclusions about the storm (Figure 10). Note that money is used here as a reward and punishment depending on performance (Figure 11).

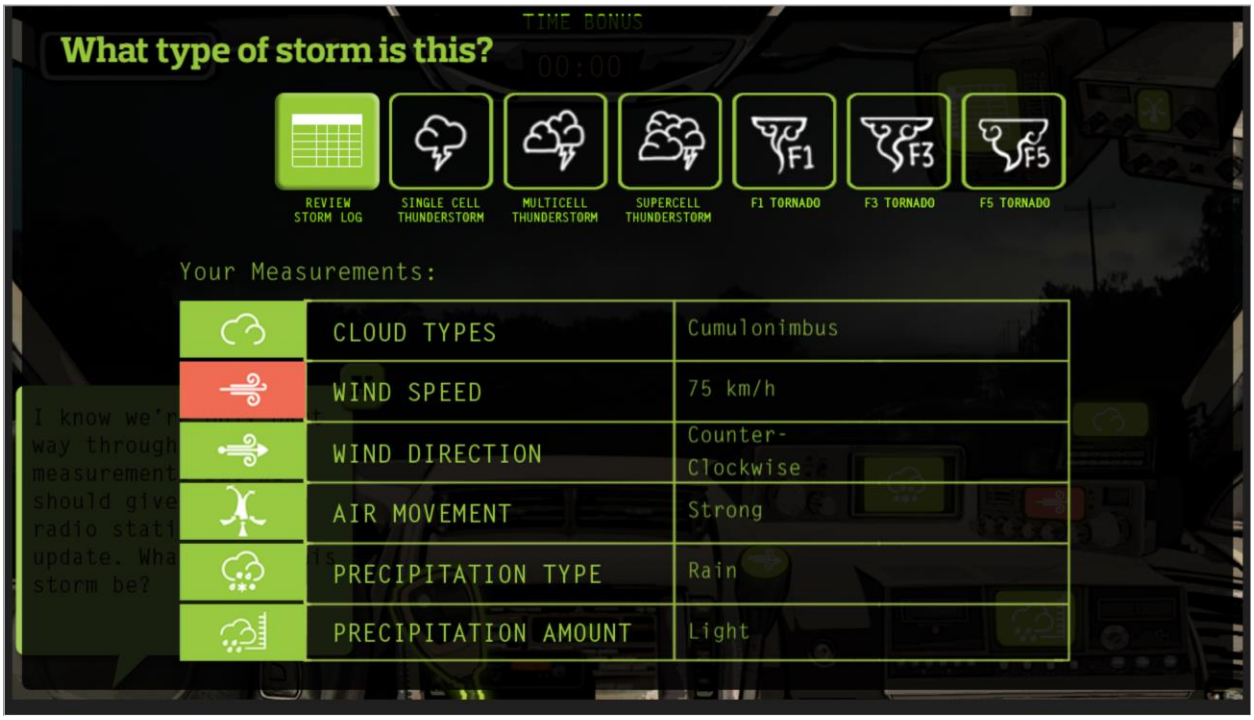


Figure 9. Coming to a conclusion.

Figure 10. Results!

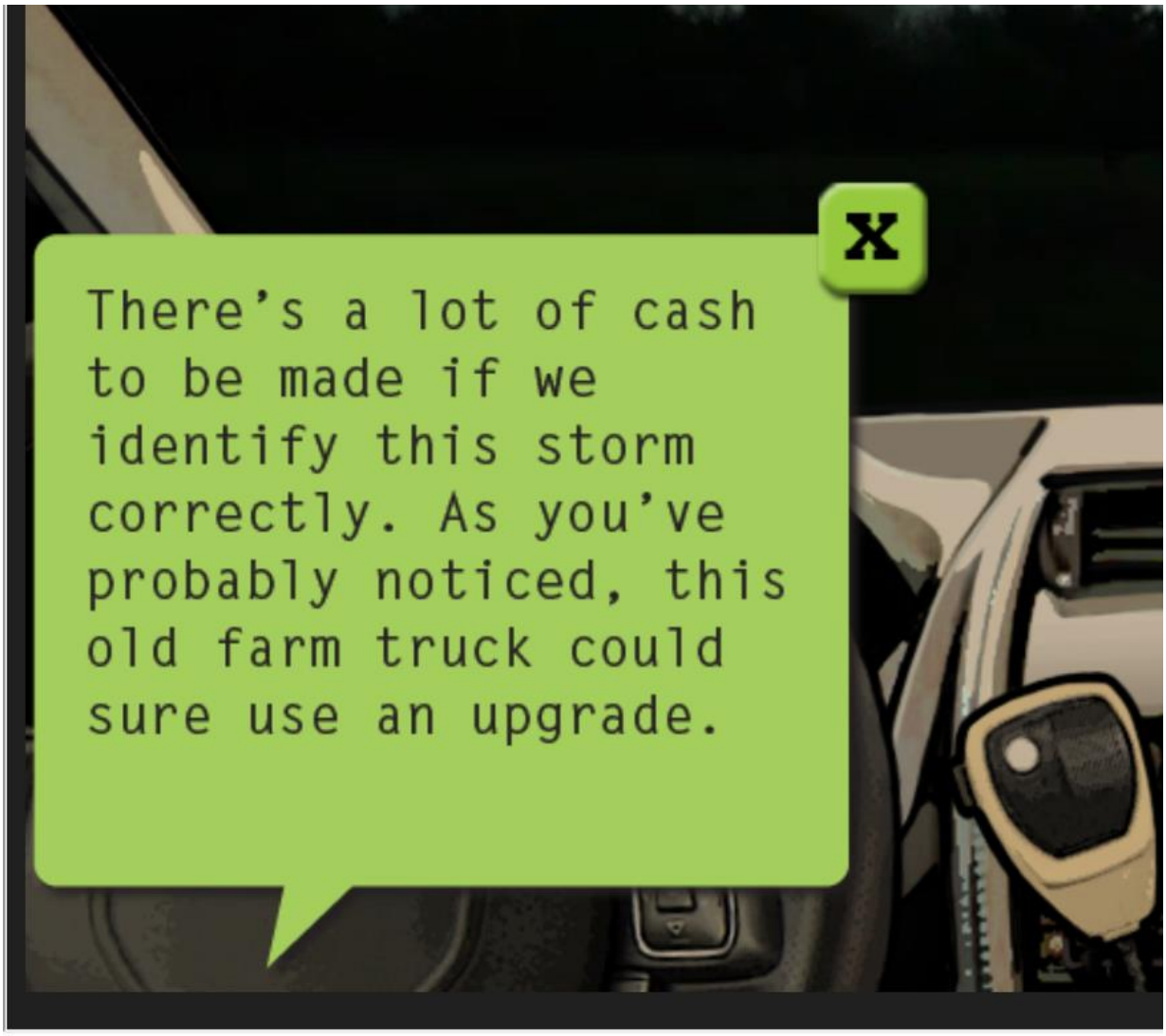


Figure 11. Money matters.

The participant can earn money by successfully completing tasks while working with a storm. The participant is not awarded any money for unsuccessful completion of such tasks. Money is spent on in-game *gas* which is used to chase storms. In the mechanics of the game this translates into some money being spent each time a participant chooses a storm to work with.

Once the first storm is finished the participant is shown a map from which to pick further storms to choose (Figure 12).

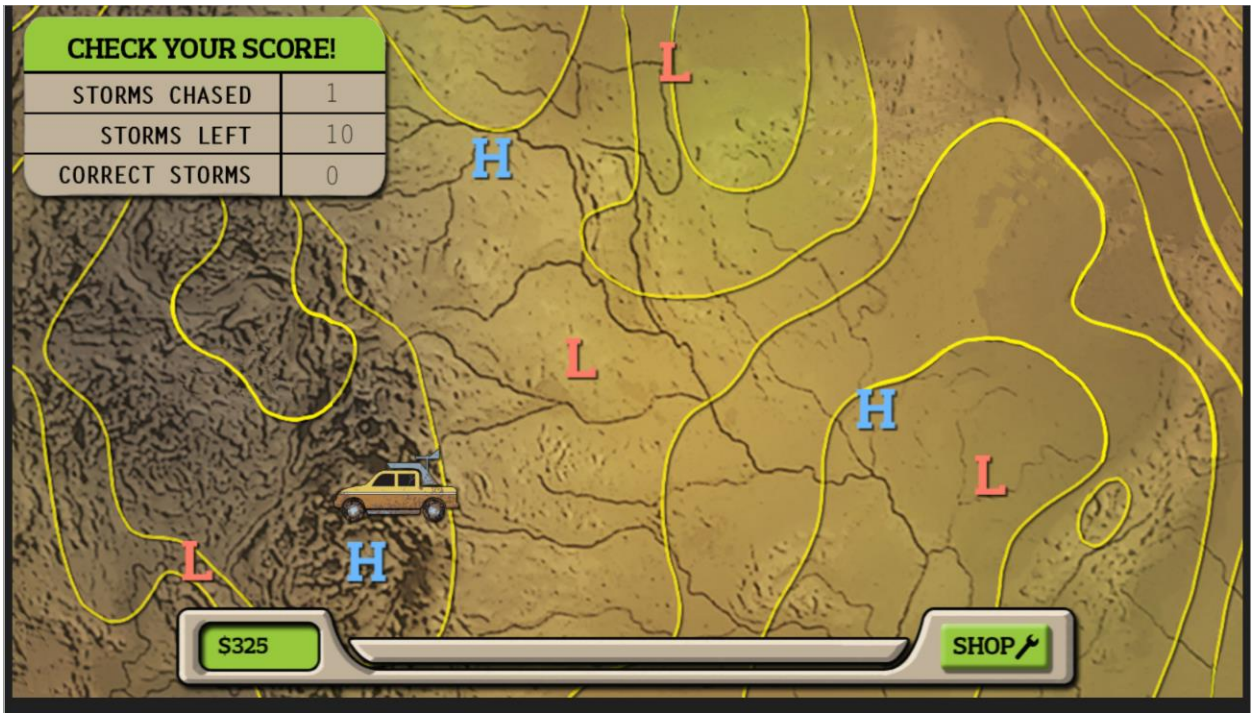


Figure 12. Chasing storms.

When choosing a storm to work with the participant can choose between storms labeled with an L (indicating a low-pressure system), and an H (a high-pressure system). The participant is cautioned against choosing storms labeled with an H as these only lead to penalties for the participant. The game explains that such high-pressure weather events are not *storms* of interest to the player, rather the weather appears quite nice (Figure 13). In contrast selecting an L (a low-pressure system) to chase shows the familiar graphic of a storm and the participant is prompted to take measurements, make a hypothesis, and come to a conclusion about the type of storm. In addition to the L and H symbols on the map, the participant can also click on a button labelled *shop* to purchase upgrades for their cars (Figure 14). The upgrades for the cars offer an aesthetic change, as well as additional time to correctly deduce storms and a better chance at earning more money.

Figure 13. Chasing a high-pressure weather event.

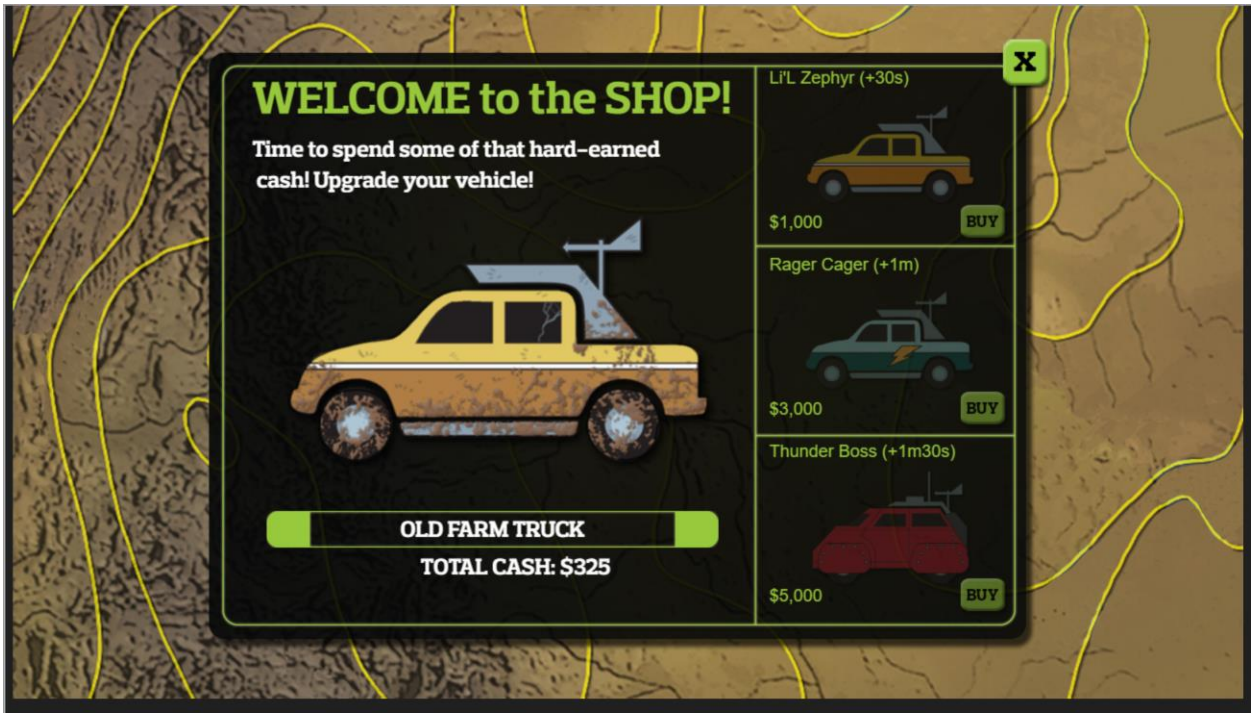


Figure 14. The shop.

After playing through 10 storms, the participant finishes the game. At the end of the game the participant is shown a graphic that summarizes their performance. On the left side of the graphic information about how well the participant was able to identify different types of storms is shown. Then on the right an overall score is given with a title and stars indicating performance (Figure 15). After the participant clicking the button labelled *What's Next?* they are prompted to play again.

Log Files. When a participant clicks on an element of Raging Skies a record of that action was made in their log file. These log files resemble a spreadsheet, they are divided into rows and columns. There are seven columns in a log file, and one row for each click action that a participant took during play. The columns are labeled: *id, action, result, correctAnswer, Stormid, timestamp,* and *DataOfAssessment*. The *id* column records a unique identifier for each participant.

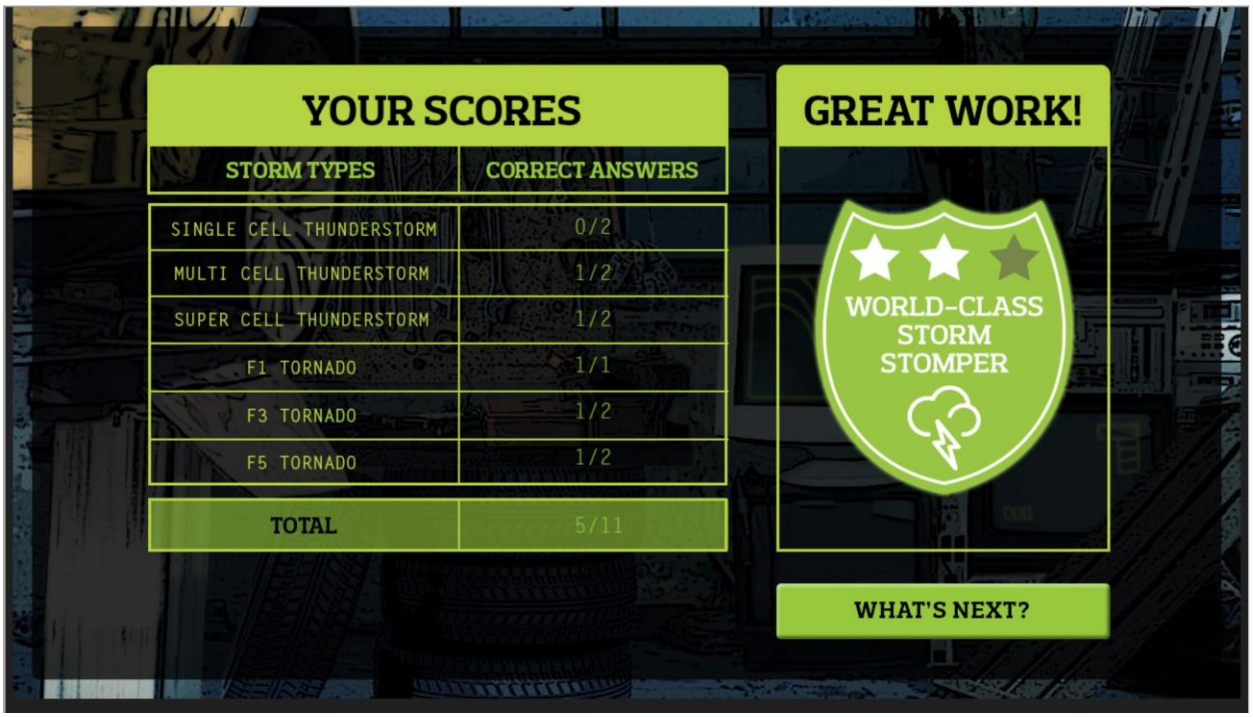


Figure 15. End of the game.

The *action* column contains information that corresponds to participants clicking on elements in Raging Skies. The *result* column has information about when a participant made a choice with an element in Raging Skies. For example, when doing a Precipitation Type measurement task of a storm, they might choose the Rain option. This choice of Rain is then recorded in the log file under the *result* column. The *correctAnswer* column records the correct answers for elements that participants could make choices with. For instance, a Precipitation Type measurement task has several options to choose from. When doing a measurement task only one or two of these options are correct. If the correct option for a Precipitation Type measurement task was either Rain or Hail, then this information would be recorded in the *correctAnswer* column. The *Stormid* column holds a value to show which storm chasing experience the participant is playing. Values in this column correspond to which storm a participant chooses to chase. The *timestamp* column records the time that each record is made. These times correspond to when participants click on elements in Raging Skies. Finally, the *DataOfAssessment* column records the date when the participant played Raging Skies and when the log file was made. See Figure 1 for a sample log file. More details about the information that is recorded in the log file is shown in Appendix g.

Given the records contained in the log file, it may be possible to use them to measure a lack of engagement. More specifically, when a participant plays Raging Skies, they are taking actions to click on the elements of the game. Clicks on these elements result in records being made in the log file. Thus, the records in the log files are sufficiently detailed to show the actions of clicking on elements as a participant plays Raging Skies. More detail about the relationship between actions and records that appear in the log files can be found in Appendix g. Additionally, as the times of these click actions are recorded (in the *timestamp* column of the log

files) then these actions can be analysed in the order they were taken. Thus, the records in the log files may be used to infer when actions were taken, and so may be used to infer RGB.

Now that the structure of the log files has been examined, the organization of the content can be considered. Each log file contains records of participant actions as they played Raging Skies. These records can be broken down into two classes. The first-class of records are those that indicate a participant is playing. These records-of-play are shown by records of participants completing tasks: measurements, retrying incorrect measurements, forming hypotheses, and coming to conclusions about the storm. The second-class of records are related to a participant choosing which storm to chase and play with. The first-class of records can be used for measurement of what a participant knows and can do. The second-class is gameplay unrelated to what a participant knows and can do with respect to storms. Thus, a first step when analyzing participant log files is to gather all the records that are related to play (the first-class records) and that can be used for measurement. These sets of records are termed *Play(s)*. Each Play can be identified by a numeric value in its Stormid column. Conversely, those records that are not Plays are identified by the value *null* in their Stormid column.

Each participant log file contains one or more Plays. Each Play includes one or more records related to playing with a single storm. Such records are of the various tasks that make-up play: measurements of the storm, opening a measurement tool, retrying an incorrect measurement, at least one record of making a hypothesis, and at least one record of forming a conclusion. These records of actions of a participant doing measurement tasks are labelled *Measurement*(*s*). Also, there are records of participants opening measurement tools before completing measurement tasks or retrying those tasks. These records can be labelled as *Open*(*s*). Thus, when a participant opens a measurement tool to do a measurement task, these actions are

reflected in records in their log files as an Open and a corresponding Measurement. Furthermore, when a participant does a measurement task incorrectly, they can try again at the same task. These actions can be labelled as *Retry* / *Retries*. Records of Retries are also associated with their corresponding Opens. Finally, records indicating one or more hypothesis or conclusion task can be labelled as *Hypothesis* / *Hypotheses* and *Conclusion* / *Conclusions* respectively. These labels will be used when reporting the results.

Survey. One survey question focusing on test anxiety was selected from a larger general survey that had been administered to all students:

- When I take a test I worry about failing.

This question was taken from the first set of survey items related to participant emotions. This survey question was adapted from question 14 in the Motivated Strategies for Learning Questionnaire (Pintrich, 1991, p. 15):

- When I take a test I think of the consequences of failing.

This question was chosen because no other questions on the survey appeared applicable to this study's focus on a lack of engagement. Other survey questions with some tangential relationship to the focus of the study appeared to be related to *cognitive engagement*. While this may appear to be a good match for the topic of engagement, the survey did not include a definition for cognitive engagement. As previously discussed definitions of engagement vary widely (Fredricks et al., 2004; Henrie et al., 2015; Sinatra et al., 2015), and so it was deemed inappropriate to assume that the construct of cognitive engagement as defined in the survey would match the definition provided for this study. In sum only the single question *When I take a test I worry about failing* was used during analysis.

Analysis

This section reviews the steps of the method, as well as practical implementation details of the method. The method attempts to measure a lack of engagement based on participant actions as they played Raging Skies. The method starts by gathering evidence of a lack of engagement. It then concludes by attempting to support the claim of a lack of engagement via participant response patterns on the survey. Participants that whose actions imply RGB are labeled as *Group1*. Those participants that lack evidence of RGB are labeled as *Group2*.

The practical implementation details provide specifics as to how each step can be accomplished. There are six practical details to review for Step 1. These details include: extracting participant Plays from the log files, error checking, and finding participant responses that were both suspiciously fast and that had low aggregate correct response rates. The last two details may be used to infer RGB and are discussed in more detail in the following section. The practical detail for Step 2 includes how to find the survey responses of participants associated with RGB.

Method overview. The method has two steps. The method begins by finding evidence of a lack of engagement in participant actions. Then it finishes by analyzing survey responses related to test anxiety with the goal of finding corroborating evidence for a lack of engagement.

Step 1. The first step is to find evidence of a lack of engagement in the form of RGB. Completing this step will answer RQ1. This is done by locating RGB in the records of participant actions. RGB may be found by looking for participant responses to assessment tasks that were done in an unusually short amount of time (Guo et al., 2016; Shute et al., 2015; Wise, 2017). Also, additional evidence for RGB may be found by searching for an aggregate participant correct response rate near random (Wise, 2017). Thus, finding RGB may be done by

inspecting participant response times and correct response rates. This dual approach is used because this method was used to validate measuring RGB with this approach on multiple-choice assessment tasks (Wise, 2017). Thus, it was deemed prudent to use this approach, given that this study is exploratory with respect to measuring RGB in a GBA.

A drawback of using RGB for this study is that the assessment tasks from the GBA Raging Skies may not appear to be traditional multiple-choice tasks. Evidence for RGB is found in the responses to multiple-choice assessment tasks (Guo et al., 2016; V. J. Shute et al., 2015; Wise, 2017). However, only some items in Raging Skies can be considered multiple-choice tasks. Measurement tasks, and retrying such tasks, prompt a participant to choose between several options. Doing a hypothesis or a conclusion task prompts the participant to choose a type of storm. An additional complication is that retrying tasks within Raging Skies does not conform a multiple-choice assessment task or the definition of RGB. This is because a second response to a task (i.e., retrying a task) implies two things that differ from traditional multiple-choice assessment tasks. The first difference is that the participant has already read or understood the challenge the task was posing. This first difference also varies from the definition of RGB. Recall the definition of RGB: multiple-choice assessment tasks that are considered to be too fast for the participant to have been able to *read and understand them* (emphasis added; Guo et al., 2016; V. J. Shute et al., 2015; Wise, 2017). More specifically, a participant retrying a task they had already encountered may answer quickly because they have already read and understood the task from their previous encounter. The second difference is that a participant has additional information they would not have in a traditional multiple-choice assessment task: their previous incorrect answer. Put another way, participant behaviour likely changes when encountering a

task for the second time, as they already know at least one response is incorrect. Thus, evidence for RGB may be found only in Measurements, Hypotheses, and Conclusions.

Initially, participant response times were examined for evidence of RGB. Afterwards, the aggregate correct answer rate of participant responses was examined. Those responses that were both unusually short and that had an aggregate participant correct response rate near random were labelled as RGB. To begin, the Normative Threshold (NT) approach was adopted to detect very short response times (Wise & Ma, 2012).

Normative Threshold. The NT approach begins by finding a mean response time. This is followed by calculating a fraction of that mean response time. This fraction of the mean response time is then labeled as a *time threshold*. The time threshold is used to label responses as either potentially being RGB or not. That is, those responses with response times less than the time threshold are labelled as possibly being RGB. Wise and Ma (2012) tried several percentages of the mean response time: 10%, 15%, and 20%, with a 10% threshold having the best performance in their analysis. The 10%, 15%, and 20% percentages will all be used in this analysis, to cast the widest net for RGB. The thresholds that correspond to these percentages will be labeled as NT10, NT15, and NT20, respectively.

Correct response rate near random. Once a time threshold has been established, then the aggregate participant correct response rate could be examined. Examining the aggregate participant correct response rate meant finding the aggregate correct response rate for the suspicious-responses of each participant. As will be discussed in more detail later, there are several types of responses. Also, a correct response rate can be calculated as if a participant were answering at random. Thus, the two correct response rates can be compared. If the observed

aggregate correct response rates for a participant falls below the random response rate, then this can be used as further evidence of RGB.

Using both approaches, NT threshold and correct response rate, casts a wide net for responses that might be RGB. Once evidence of RGB has been accumulated for a response, then that response can be labelled as RGB. Further, the records associated with that response can be labeled as containing RGB. Records of participant actions that contain RGB, may be marked as lacking engagement.

Step 2. Given these two sets of records from Step 1 of the method, Group1 and Group2 can be formed. The survey responses related to test anxiety of these groups can then be compared to attempt to find corroborating evidence that a lack of engagement is present in participants in Group1. First, participants associated with records that contain evidence of a lack of engagement (RGB) can be assembled into Group1. Second, all participants that do *not* have evidence of RGB, and a lack of engagement, can be placed in Group2. Third, the survey responses related to test anxiety are measured.

Previously, it was argued that test anxiety and a lack of engagement are closely related in the context of this study; So much so that it is assumed that the constructs are covariates. Thus, it follows, that if a lack of engagement varies between Group1 and Group2, then there should be a difference in survey responses related to test anxiety between the groups. This hypothesized difference can be tested for with a T-test. Formally, the hypothesis being tested can be stated as:

- H0, null hypothesis, survey responses related to test anxiety from participants in Group1 and Group2 do not have significant differences.
- H1, survey responses related to test anxiety from participants in Group1 and Group2 differ significantly.

Rejecting H0 and failing to reject H1 then implies support for the assumption of covariance between test anxiety and a lack of engagement. Further, if the assumption of covariance can be supported, then this lends support to the conclusion that a lack of engagement is present in participants in Group1.

Practical considerations. This section enumerates practical implementation details for the method. These practical details are provided for each step in the method.

Step 1.In order to implement Step 1 of the method, there were six practical considerations to address. The first was to find all Play records. Second, from those records, ensure that the data was error-free. Third, locate the response times of participants to tasks in Raging Skies. Fourth, checking for errors related to the response times. Fifth, calculating NT thresholds for all such tasks, in order to find tasks that were completed suspiciously quickly. Finally, sixth, finding the aggregate correct response rate of these tasks. Addressing these considerations resulted in a set of records that were suspected of harbouring RGB.

Finding plays. The first stage of Step 1 was to find the total number of usable Plays from the participant logs. A participant's log file would reflect a period of play, followed by a participant choosing another storm to chase / play with, then more play with the new storm, etc. This formed a pattern in the log files of records with numeric-Stormids followed by records with null-Stormids, then numeric, and so on. Thus, isolating the Plays was a matter of collecting all records which had numeric Stormids, and which were interleaved with records that had null-Stormids.

Errors in plays. Given these Plays, the next step was to check them for errors. These errors could be broken down into two categories: Plays with more than one Stormid and Fragmented Plays. A Play with more than one Stormid are a set of records containing several

Plays that appear one after the other without any intervening records with null-Stormids. These Plays were complete, but they are figuratively *stuck-together*. The Plays are complete in the sense that they contain Measurements, Retries, Hypotheses, and Conclusions. Such Plays have all the records that one would expect a non-erroneous Play to have. In contrast, Fragmented Plays are records of parts of Plays that are attached to a complete Play. Such fragmented Plays are metaphorically broken. It proved impossible to re-assemble a complete Play from several fragmented Plays.

Finding response times. Given a set of Plays, it was then possible to find the first indicator of RGB. Recall the definition of RGB: participant responses to multiple-choice assessment tasks that are considered to be too fast for the participant to have been able to read and understand them. To review, the multiple-choice assessment tasks in Raging Skies are: measurements, making hypotheses, and forming conclusions. Records of participant responses to such tasks could be found in the Plays. The records were: Measurements, Hypotheses, and Conclusions, respectively. To find RGB in such responses, one must measure the time a participant takes to respond. These response times could be found by measuring the time from when a participant was presented with a task to when they responded to it. For measurement tasks, the participant was presented with it when they opened the measurement tool. Thus, the response time for a measurement task could be inferred from the time between an Open and its related Measurement. A complication is that a participant may open one or more measurement tools many times before responding. That is, there may be several Opens for each Measurement. Thus, a reasonable approximation for the response time for a measurement task could be found by finding the time between the Open that directly precedes the Measurement. Response times for hypothesis tasks and conclusion tasks were found differently. In contrast to measurement

tasks, a participant does not choose when to view a hypothesis or conclusion task, these tasks are presented to the participant once they have completed several measurement or retry tasks. As a consequence, hypothesis and conclusion tasks do not have Opens. Therefore, measuring the response time for a hypothesis task could be found by finding the time between a Hypothesis and a record that directly preceded it. For example, a participant may have just finished making a measurement when the graphic for the hypothesis task appeared. When this graphic appears, the participant can then take some time to respond. In this example, it is the time between a participant's last measurement and their response to the hypothesis task that could be used to infer a response time. Responding to a conclusion task is similar. The response time for a conclusion task could be found by measuring the time between a Conclusion and the record directly preceding it.

Errors in response times. Some response times were found to be negative, and therefore erroneous. This error was related to Open records, and thus only response times for measurement and retry tasks were affected. This type of error came in two forms. The first form was a Measurement without an associated Open record. The second was an Open record appearing after a Measurement was recorded. This issue was likely caused by faults in the system responsible for writing the logs. Negative response times were dropped from analysis. Given the remaining response times NT thresholds could be calculated.

Finding normative time thresholds. Recall that a part of the definition for an NT threshold is a percentage of the mean response time for assessment tasks. Raging Skies has several types of assessment tasks included in this analysis; Measurement, hypothesis, and conclusion tasks. Thus, NT thresholds must be calculated for each type of task. An additional complication is that there are several types of measurement tasks. There is one type of

measurement task for each type of virtual measurement tool a participant can use. Thus, each type of measurement task must have their own NT threshold. Given the process of gathering response times for the different types of tasks, the mean of each type of response time can be calculated. Then, with the mean response times, NT thresholds can be calculated. Thus, any response time that fell below the threshold for their class of responses could be labelled as a *suspected-response*. That is, such responses were suspected of being RGB, given how quickly they were responded to.

Finding aggregate correct response rates. The next step in finding evidence of RGB was to determine the aggregate correct response rate of suspected-responses for each participant. If the aggregate correct response rate was less than or equal to random chance, then this is further evidence that the participant was using RGB. Calculating the aggregate correct response rates was simple. If the participant response matched the correct response listed in the *correctAnswer* column of the record of the response, then this was counted as correct. If the participant's response did not match, then the response was incorrect. It was also possible for some measurement tasks to have two correct answers. In addition, the Wind Speed measurement task prompted the participant to find a correct response in a range of values. Thus, a correct response, for a Wind Speed measurement, counted if the participant was in the correct range of values. The hypothesis and conclusion tasks had only single correct answers.

In contrast, calculating the probability of answering correctly by random chance was more difficult. There were two steps in this process. The first step was to calculate the probabilities of correctly guessing the answer to each type of task at random. The second step was finding the expected correct answer rate for guessing at random over several tasks. The set

of suspected-responses could vary by participant, thus the expected correct answer rate for guessing at random might differ between participants.

The first step of calculating individual probabilities was relatively straightforward. Starting with the measurement responses, the probability of a correct answer by chance was: the number of correct responses divided by the total number of responses. However, there were a variety of possible responses for the measurement tasks. For example, the Precipitation Amount measurement task had a probability of a correct response at random of 1/4. On the other hand, the Wind Speed measurement task accepted a range of values, and so the probability would be a more complex fraction. Calculating the probabilities for addressing hypothesis and conclusion tasks was simpler. Both hypothesis and conclusion tasks had 6 possible responses, and participants were only (ideally) allowed a single response. For more details on the probabilities of a correct response to various tasks at random, see Tables 5 & 6.

A difficulty arose with some participants attempting hypothesis and conclusion tasks multiple times. In some cases, participants had five or more attempts at a hypothesis or conclusion task. In consultation with Dr. Chu, the design of Raging Skies was explicit that only a single attempt at these types of tasks were allowed. However, information from the log files showed a pattern of multiple responses. These multiple responses to hypothesis and conclusion tasks were thus treated as aberrant behaviour and dropped from analysis.

The second step was more labour intensive. As the goal was to find the expected correct response rate for guessing at random, each participant with suspected-responses had to be scrutinized. That is, the expected value (the mean) of the probabilities of answering a response at random had to be calculated for each participant and their responses. Then, this expected-probability was compared to the observed correct response rate. If the observed correct response

rate for the suspected-responses fell below or equal to the expected correct response rate for guessing at random, then the suspected-responses could be labelled as RGB.

Step 2. The second step of the method required two steps in order to complete. First, the participants had to be grouped into two sets, one set that was associated with RGB and the other that was not. Second, the survey responses for the groups had to be quantified.

Grouping. This began with the records that contained evidence of RGB. Further, the compliment of the set of records containing RGB shows records that do not contain RGB. Each of these sets of records is associated with a set of participants. Participants associated with records that contain RGB were labeled as Group1 and the other participants were labeled as Group2. The identification numbers of these participants were then used to find the survey responses for these participants. Specifically, those participants in Group1 had survey responses labeled G1 and those in Group2 were labeled as G2. These two sets of survey responses could then be quantified for analysis.

Quantification. Quantification of the survey responses involved recording the value of each response and comparing these sets of data with a T-test. As can be seen from Appendix e survey responses were defined on a Likert scale, in-turn these responses were associated with numerical values. Specifically, the Likert scale responses for the survey question of interest was associated with the natural numbers 1 through 5. Once the survey responses were coded, by matching the survey response with the value associated with the Likert Scale, the numerical data was analyzed.

Summary

In summary, RGB may act as evidence for a lack of engagement. Support for using RGB as a measure for a lack of engagement with the GBA Raging Skies, comes from finding backing

evidence in the patterns of participant survey responses. In addition, the method and practical considerations for its implementation were presented.

Results

This section presents an overview of the data analysis for this study and a presentation of the findings from the method. This is done by going through each step from the method. The first step (Step 1) of the method involves finding evidence of RGB. Step 2 then analyzes the differences in survey responses between Group1 (participants with some RGB) and Group2 (non-RGB participants) to attempt to find corroborating evidence of a lack of engagement. Data analysis was completed with the Scipy Python package (SciPy 1.0 Contributors et al., 2020), while the data was manipulated and stored with a combination of the Pandas Python package and a Python programming environment (Reback et al., 2020).

Step 1 Results

Recall that there were 457 usable log files gathered from participant play of Raging Skies. Those log files contained 6094 Plays. There were errors that affected 30 Plays. Some of these errors were correctable, but some were not. Some of these erroneous Plays could be corrected to form valid Plays. Ideally this process of correction would have fixed every erroneous Play, but some were so corrupted they could not be recovered. Thus, correcting these erroneous Plays resulted in 52 additional (corrected) Plays. A total of six Plays were discarded due to unrecoverable errors. After addressing these errors there were 6114 usable Plays for analysis. The response times from each of these Plays were then retrieved from the Plays. Recall that there were three types of assessment tasks in Raging Skies that could be used for analysis: measurements, hypotheses, and conclusions. The total number of response times of each type of task were: 37014 (measurements), 6171 (hypotheses), and 6305 (conclusions). Of the measurement task response times, two were found to be negative and were discarded. Thus, the number of usable response times was: 37012 (measurements), 6171 (hypotheses), and 6305

(conclusions). The average response times for measurements, broken down by type of in-game instrument, are shown in Table 1. The average response time was computed with the arithmetic mean. As this type of mean can be sensitive to outliers in the data, Appendix h contains an analysis of potential outliers. No outliers in the data were found that significantly impacted the means. Also, in Table 1, the NT10, NT15, and NT20 thresholds for different types of measurement responses are shown. Tables 2 contains the mean response times as well as the NT10, NT15, and NT20 time thresholds for hypothesis and conclusion responses. These tables show the first test for detecting RGB. That is, any response times that fall below the NT thresholds can be labeled as suspected-responses. Given these NT thresholds, Tables 3 and 4 show how many responses had times that fell under an NT threshold for measurement, hypothesis, and conclusion task responses respectively. Also, these tables show the ratio of responses that fell beneath which type of NT threshold.

Table 1

Measurement task responses by type

	Mean response time (seconds)	NT10 (seconds)	NT15 (seconds)	NT20 (seconds)
Precipitation Type	2.91	0.29	0.44	0.58
Cloud Type	3.57	0.36	0.54	0.71
Wind Direction	3.04	0.30	0.47	0.61
Updraft Speed / Air Movement	2.42	0.24	0.36	0.48
Precipitation Amount	3.06	0.31	0.46	0.61

Wind Speed	5.17	0.52	0.78	1.03

Table 2

Hypothesis and Conclusion task responses

	Mean response time (seconds)	NT10 (seconds)	NT15 (seconds)	NT20 (seconds)
Hypothesis	6.46	0.65	0.97	1.29
Conclusion	6.96	0.70	1.04	1.39

Table 3

Measurement task responses under NT thresholds by type

	NT10 (number of responses)	NT15 (number of responses)	NT20 (number of responses)
Precipitation Type	55	55	55
Cloud Type	36	36	36
Wind Direction	73	73	73
Updraft Speed / Air Movement	49	49	49
Precipitation Amount	42	42	42
Wind Speed	2	2	81

Table 4

Hypothesis and Conclusion task responses under NT thresholds

	NT10 (number of responses)	NT15 (number of responses)	NT20 (number of responses)
Hypothesis	37	37	199
Conclusion	7	602	602

Table 5

Random chance of success for measurement tasks

	Correct answer rate for random chance, one answer	Correct answer rate for random chance, two answer
Precipitation Type	1/6	2/6
Cloud Type	1/4	2/4
Wind Direction	1/3	2/3
Updraft Speed / Air Movement	1/3	2/3
Precipitation Amount	1/4	2/4

For the NT10 threshold, there were 126 participants with at least one response that fell under that threshold. For NT15, there were 252 participants. Finally, for NT20, there were 280 participants under that threshold. Then, given the records whose response times fell under each NT-threshold (NT10, NT15, and NT20), the next step was to compare their observed correct

response rate to the expected correct response rate as if they were answering at random. Tables 5 and 6 list the correct answer rate for random chance for measurement tasks. The correct answer rate for hypothesis and conclusion tasks are simply 1/6; There were only six types of storms could chase, and thus only six responses for these tasks. For Table 5 recall that the measurement tasks listed there could either have a one or two correct answers. Each column of Table 5 lists the correct answer rate for the one-correct-answer case and the two-correct-answer case of a measurement task.

Table 6 shows the expected correct answer rates for random chance for the Wind Speed measurement task. Recall that the Wind Speed measurement task prompted the participant to choose a range of values which might contain the correct wind speed value. Participant answers could be set in increments of 5. For example, a participant could try to answer within a range of 25 to 30 but not 23 to 30. The table is organized to relate ranges of possible responses to their probability of being chosen at random; The format of these ranges are: low, high. The ranges chosen to be shown in Table 6 were ranges that appeared during the analysis. In total there were 81 possible values to choose from in the Wind Speed measurement tool.

Table 6

Wind Speed

Values	20, 70	30, 100	100, 130	130, 200	200, 320	325, 400	320, 400
Correct answer rate	10/81	14/81	6/81	14/81	24/81	15/81	16/81

Given these correct response rates based on random chance, it was possible to detect participant correct response rates that were at or below random chance. For the NT10 threshold

there were 37 participants whose suspected-response correct response rate was comparable to random chance. For NT15, there were 82 participants with suspected-responses that had a correct response rates near random. For NT20, there were 93 participants whose suspected-responses had a correct response rate comparable to random chance. In total there were 106 unique participants that showed evidence of RGB in their responses. These results indicate that that RQ1 can be answered affirmatively: RQ1 asked whether evidence of RGB can be found in the records of play, and there is evidence that a proportion of participants (106 of 457) employed RGB.

Step 2 Results

Given the 106 participants isolated in Step 1, Group1 can be formed. Recall that Group1 are the participants whose recorded actions contained evidence of RGB. As such, these participants form Group1. Thus, given that there were 457 participants with usable log files, there were 351 participants that did not have evidence of RGB in their recorded actions, and so these 351 participants form Group2. Analysis showed no statistically significant differences in mean survey responses between Group1 and Group2 for the survey overall ($p > .89$). Table 7 shows descriptive statistics of the two groups of survey responses, while Table 8 summarizes the results of the Step 2 analysis. This result indicates a failure to reject the null hypothesis (H0) for Step 2 that Group1 and Group2 do not have significant differences. This lack of significant differences in the mean scores of Group1 and Group2 implies a false response to RQ2. That is, there were not significant differences in the survey responses between Group1 and Group2, on the question that relate to test anxiety (i.e., as related to engagement).

Table 7

Descriptive statistics

	Mean	Variance

Group1	3.18	1.79
Group2	3.16	1.87

Table 8

T-test results

	T	p
Assuming equal variance	0.13	0.90
Assuming unequal variance	0.13	0.89

Note. Two T-tests were run, one assuming equal variance between Group1 and Group2, while the other assume an unequal variance.

Initial Summary

In summary, while Step 1 did locate evidence of RGB in some participants recorded actions, Step 2 did not locate significant differences in survey responses between those participants with evidence of RGB and those that did not. Thus, RQ1 is answered in the affirmative that evidence of RGB be found in the records of participant play, while RQ2 is answered in the negative in differences between the survey responses between Group1 and Group2 on questions that relate to test anxiety (i.e., as related to engagement) overall.

Additional Exploratory Analysis

This pattern of results was puzzling, and so additional analysis was undertaken. More specifically, while RGB appears to be present in the data, this is not reflected in the survey responses analyzed by the method overall; It seems logical that the survey responses would show some indication of the RGB found in the data, and so perhaps the effects are shown within a subsection of the survey responses. This leads to an informal research question:

Given the pattern of Rapid-Guessing Behaviour present in the data, are there any effects on the survey responses?

This additional informal research question, which can be labeled as IRQ1, can be addressed similarly to the method of addressing the primary research questions.

The method for conducting the exploratory analysis and addressing IRQ1 follows from the primary method. That is, the method consists of: comparing the survey responses for each group (Group1 and Group2) across each survey question (except question 7) with T-tests. This is done identically to the previous method. In other words, survey responses were coded with the value associated with their response on the Likert scale of the survey. These values were then analyzed with a series of T-tests. Then, significant differences in response patterns between the groups on each survey question can be marked by an abnormally low *p* values (i.e., $p <= .05$) for each test while accounting for the increased number of post-hoc tests. Unfortunately, no meaningful significant patterns were detected among subsets of the survey questions.

Results Summary

In overall summary, there did not appear to be any detectable relationship between RGB and the survey response from the Raging Skies study. The initial method did not produce any statistically significant results. In addition, the supplementary exploratory analysis found statistically relationships in survey responses that appeared to have no relationship with RGB or engagement. In sum, there is little support for using RGB as a method of detecting a lack of engagement in the GBA Raging Skies.

Discussion

As investigated in the previous section, this project could not establish a relationship between RGB and engagement in the GBA Raging Skies. That is, RGB-like actions were

isolated from the log files of participant play, but that behaviour was not significantly associated with survey responses that were theorized to be related to engagement. As there were no significant findings to discuss, this section focuses on discussing possible reasons for this lack of findings, and potential future directions for inquiry. This section is composed to two main threads. The first thread discusses a possible alternative to RGB that may have been detected during this project; Enjoyment-seeking behaviour. The second thread discusses how this behaviour might be detected in future studies.

Enjoyment-Seeking Behaviour

To refresh, RGB is defined as participant responses to multiple-choice assessment tasks that are assumed to be too fast for the participant to have been able to read and understand them (Guo et al., 2016; Shute et al., 2015; Wise, 2017). The reason for this behaviour is argued to be that participants are assumed to be responding without fully considering assessment tasks (Guo et al., 2016; Shute et al., 2015; Wise, 2017). Common assumptions about why a participant might not fully consider assessment tasks include: speeding through an assessment to avoid dealing with it and attempting to earn points by guessing when there is limited time (Wise & Smith, 2016).

These assumptions do not highlight the possibility that a participant might be avoiding considering an assessment item in order to quickly return to an enjoyable experience. This may be the case with GBAs as they include elements that are intended to elicit a sense of fun or enjoyment. This stands in contrast to traditional multiple-choice assessments, as traditional multiple-choice assessments do not commonly include such elements tailored to elicit enjoyment. Several authors have noted their concern about enjoyable elements in GBAs interfering with assessment (Ge & Ifenthaler, 2018; Ghergulescu & Muntean, 2012). Thus, it

may be the case that participants within the Raging Skies project were using actions similar to RGB to avoid parts of the GBA that they did not enjoy in order to return to more enjoyable experiences. It follows that the RGB-like actions isolated from the log files during analysis may have been a result of *enjoyment-seeking behaviour* (ESB). More specifically, ESB may be defined as RGB-like actions, whose underlying reason is to avoid unpleasant aspects of a GBA, while seeking enjoyable experiences from gameplay.

Assuming that ESB exists, this may explain the difficulties in associating RGB-like actions found in this study with a lack of engagement. The method of RGB detection appears to be sound: the RGB-like actions found during analysis were exceptional fast, and usually incorrect. Thus, it follows that the participants were likely not being assessed correctly while using these RGB-like actions. However, if the reason for this behaviour diverged from the assumptions of RGB, then it becomes dubious to assume that the behaviour detected was really RGB. In addition, this may also explain the puzzling results as derived from the survey information. To the knowledge of the author, there was no survey question in the Raging Skies project that could have been expected to detect ESB. That is, the survey-instrument did not appear designed to elicit responses of enjoying parts of Raging Skies, while disliking others. Thus, it becomes difficult to find evidence of ESB, but the possibility remains.

As an illustrative example, consider the following scenario as a possible explanation for a participant employing ESB. When beginning to play Raging Skies, the participant is presented with a novel story, exciting graphics, and interesting gameplay centered around storm-chasing. To the participant, this enjoyable gameplay is then interrupted by what appears to be an assessment task. This interruption may dampen their enjoyment somewhat. Then given extended gameplay these interruptions may be bothersome to some participants, such that they begin to

discard such interruptions as quickly as possible to resume the enjoyable gameplay. This process of discarding may then take the form of behaviour similar to ESB, as the participants attempt to skip the assessment portions to attempt to return to the gameplay they enjoy. It may be this dynamic of perceived gameplay interrupted by undesirable assessment that resulted in ESB.

As an alternative example, Dr. Chu mentioned in passing that some participants were fascinated with obtaining additional in-game vehicles. To be more specific, each storm-chase within Raging Skies could reward in-game money that a participant could use to purchase other types of vehicles to use during game-play. The in-game vehicles were purely cosmetic but were enough of a draw for some participants that they were fascinated by them. It seems reasonable to assume that such participants might feel bothered by tasks in Raging Skies that did not directly relate to their goal of obtaining these cosmetic vehicles. Thus, assessment tasks might be perceived as bothersome to some participants for this additional reason.

Thus, while the aim of this study was to isolate indicators of a lack of engagement, the unusual behaviour captured during analysis may highlight another behaviour that is unique to GBAs.

Modulating Assessment and Gameplay

Given this hypothesis of ESB, this leads to the question of how to detect it. In keeping with the theme of this project, a means of detecting ESB would be to monitor participant actions. As the expression of ESB and RGB is assumed to be tied to the amount of assessment present in a GBA, it seems reasonable to vary the amount of assessment across participants. More specifically, if the amount of assessment and gameplay were measured, then the sum of a participant's time spent in both modes is ~100% of the time spent with the GBA. Thus, the - amount of time in the assessment-condition versus the gameplay-condition can be quantified as

percentage of the total time spent with the GBA; For example, gameplay might be ~50% of this time, leaving the remaining time (~50%) for assessment. It follows, that by varying these percentages over a population of participants, would lead to different experimental groups of participants that experience different versions of the GBA that vary across these conditions; For example, a group might have ~60% gameplay and ~40% assessment, while another might have ~30% gameplay and ~70% assessment. Then, by varying these amounts to their extremes, one could arrive at different control groups for gameplay and assessment. That is, if an experimental group were 0% gameplay and ~100% assessment, this should function as traditional multiple-choice assessment. On the other hand, a group of 100% gameplay and 0% assessment, would function as if the participants were playing a traditional game. These two extremes, as well as other experimental groups with different values for gameplay and assessment should draw out differing behavior as a function of the amount of assessment and gameplay a participant experiences during a GBA. Then by measuring the amount of RGB-like behaviour, it may then be possible to find significant variance in these behaviours depending on the amount of assessment experienced by each group of participants. Finally, a survey instrument administered after gameplay may offer the opportunity for validation, just as in this project. This rough outline may serve as a template for future work in attempting to disentangle RGB and ESB in GBAs.

Technocentrism

One critique of this work is its apparent focus on technical analysis of participants behaviour during the Raging Skies study; Log files of their actions. Another modality of information of potential interest might have been participants physical behaviour during the study; Such information may have made a stronger case for engagement, if it was available. This critique touches on the broader idea of technocentrism as explicated by Papert (Papert, 1988). In

Papert's words technocentrism is "…the fallacy of referring all questions to the technology" (Papert, 1988, p. 4). In other words, Papert critiques methods that eschew the complexity of learning, to focus on questions related to technology. Papert's critique centers around the objection that learning is much more complex than scientific and technological methods are designed to study (Papert, 1987). Put another way, for example, a scientific experiment may involve working with an inclined plane, and the method of inquiry is assumed to apply to such simple constructs. However, with learning, a researcher must work with students, and the assumptions associated with methods intended for the former experimentation do not capture the complexity involved with the later; An inclined plane does not think.

The goal of this work was to attempt to make inferences based on admittedly limited and imperfect information. However, the results of this inquiry may not be limited to the application of the method; Rather, the results can include inferences based upon how the method failed or was imperfect. This approach adopts a philosophical approach similar to Kuhn. Kuhn noted that the role of methods in scientific inquiry is more complex than any single experimental work (Kuhn & Hacking, 2012). During what Kuhn termed *normal science* methods can be used to carry out experimentation. However, such methods can also lead to inferences based on how they fail (Kuhn & Hacking, 2012). In Kuhn's model such failures indicate what they termed a *crisis*, but in this work there are other inferences to be made (Kuhn & Hacking, 2012). That is, given that the method here failed, this implies that other more complex inquiry is likely called for in order to draw stronger and more successful conclusions. In other words, the failure of the method here inspires a more holistic approach; Just as Papert suggested in his critique of technocentrism ('Beyond Technocentrism', 2021; Papert, 1988). With that in mind, other improvements to the method and inquiry can be discussed.

Future Work

The first suggestion for improvement involves expanding the logging system associated with Raging Skies. The current logging system for Raging Skies focuses its information gathering on participant responses to the tasks within Raging Skies. As can be seen from the overview of Raging Skies, there are several more aspects to the GBA than these tasks; For example, participants could choose to interact with the shop to purchase upgrades for their cars. While not immediately obvious how this would relate to engagement, given the results of this project, the additional information may be helpful in making inferences.

The second suggestion relates to the methods used. The methods used in this project focused solely on log files to infer participant action. Given the results from this project it would seem beneficial to expand the modalities of observing the participants to gather more evidence of engagement. A specific recommendation of this type might be using this method in conjunction with other methods for studying engagement such as surveys and observations; Their combined evidence may make a stronger case for engagement, where information from log files was insufficient. Another possible modality might be to monitor participant facial expressions and behaviour with computer systems; Such computer systems might then be used to make inferences with the help of machine learning or computer vision.

Summary

In summary, while there is some doubt that RGB was detected in this project, ESB appears to be a reasonable alternative. ESB, in contrast to RGB, is the oscillation of a participant between two states: enjoyment of gameplay, and a negative experience with an assessment portion of GBA. Given this difference between RGB and ESB, an alternative method of detection was proposed for future work. Specifically, that of changing the amount of gameplay

and assessment in a GBA across several groups of participants and monitoring their actions for telltale signs of ESB and / or RGB.

Conclusion

The goal of this study was to attempt to link participant actions that were believed to imply a lack of engagement (random guessing behaviour or RGB) with other evidence that a lack of engagement was present (survey questions about test anxiety). The primary reason for this inquiry was that a lack of engagement can be a threat to validity, and thus casts doubt on assessment measurement. RGB was detected in some participant interactions with Raging Skies, but no well supported link could be established between RGB and a lack of engagement. However, it is hypothesized that a different construct was detected instead of a lack of engagement: enjoyment seeking behavior (ESB). More specifically, participants were found to take actions that were compatible with RGB, but they did not display the test anxiety that was expected. Thus, it seems plausible that such participants were not trying to avoid returning to the assessment, but rather were avoiding assessment so they could enjoy the Raging Skies gameplay.

Future work should focus on elaborating on building on the structure of the Raging Skies project. That is, while the data from the Raging Skies study was suggestive, the survey measure was not targeted specifically at detecting a lack of engagement. Thus, future work may profit by attempting to duplicate the Raging Skies project, while expanding the number of constructs represented in the survey measure. That is, in order to avoid introducing counterproductive variance into the findings, the Raging Skies project can be used as a template for future inquiry, while also expanding and modifying the survey instrument to include more constructs that can be related to a lack of engagement. These additional measures may then shed light on a connection between participant RGB and a lack of engagement. In summary, while the results of this study

are suggestive, further work is needed to attempt to connect participant actions with a lack of engagement.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*(2), 219–245. doi: 10.1037/0033-2909.121.2.219

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist*, *50*(1), 84–94. doi: 10.1080/00461520.2015.1004069

Balasooriya, I., Mor, E., & Rodríguez, M. E. (2018). Design of a microlevel student engagement data capture system. In F. Xhafa, S. Caballé, & L. Barolli (Eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing* (pp. 632–641). Springer International Publishing. doi: 10.1007/978-3-319-69835-9_59

Beyond Technocentrism: Coding as Experience. (2021). In P. Sengupta, A. Dickes, & A. V. Farris, *Voicing Code in STEM* (pp. 1–22). The MIT Press. doi: 10.7551/mitpress/11668.003.0004

Brallier, C. (2020). The Effects of Student Engagement on Academic Achievement Among College Students (PhD Thesis). Retrieved from https://ezproxy.lib.ucalgary.ca/login?url=https://www.proquest.com/dissertations-theses/effects-student-engagement-on-academic/docview/2397820506/se-2?accountid=9838

Chu, M. W., & Chiang, A. (2018). *Raging Skies: Development of a Digital Game-Based Science Assessment Using Evidence-Centred Game Design*. doi: 10.11575/PRISM/32941

de Vreede, T., Andel, S., de Vreede, G.-J., Spector, P., Singh, V., & Padmanabhan, B. (2019, January 8). *What is engagement and how do we measure it? Toward a domain independent definition and scale*. 749–758. doi: 10.24251/HICSS.2019.092

Ding, L., Kim, C., & Orey, M. (2017). Studies of student engagement in gamified online discussions. *Computers & Education*, *115*, 126–142. doi: 10.1016/j.compedu.2017.06.016

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist*, *52*(2), 104–123. doi: 10.1080/00461520.2017.1281747

Ergene, T. (2003). Effective Interventions on Test Anxiety Reduction: A Meta-Analysis. *School Psychology International*, *24*(3), 313–328. doi: 10.1177/01430343030243004

Filsecker, M., & Kerres, M. (2014). Engagement as a volitional construct: A framework for evidence-based research on educational games. *Simulation & Gaming*, *45*(4–5), 450–470. doi: 10.1177/1046878114553569

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. doi: 10.3102/00346543074001059

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 763–782). Boston, MA: Springer US. doi: 10.1007/978-1-4614-2018-7_37

Ge, X., & Ifenthaler, D. (2018). Designing engaging educational games and assessing engagement in game-based learning. In *Gamification in Education* (pp. 1–19). IGI Global. doi: 10.4018/978-1-5225-5198-0.ch001

Ghergulescu, I., & Muntean, C. H. (2012). Measurement and Analysis of Learner's Motivation in Game-Based E-Learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in Game-Based Learning* (pp. 355–378). New York, NY: Springer New York. doi: 10.1007/978-1-4614-3546-4_18

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist*, *50*(1), 43–57. doi: 10.1080/00461520.2014.999919

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183. doi: 10.1080/08957347.2016.1171766

Hembree, R. (1988). Correlates, Causes, Effects, and Treatment of Test Anxiety. *Review of Educational Research*, *58*(1), 47–77. doi: 10.3102/00346543058001047

Henrie, C. R., Bodily, R., Larsen, R., & Graham, C. R. (2018). Exploring the potential of LMS log data as a proxy measure of student engagement. *Journal of Computing in Higher Education*, *30*(2), 344–362. doi: 10.1007/s12528-017-9161-1

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*, 36–53. doi: 10.1016/j.compedu.2015.09.005

Hookham, G., & Nesbitt, K. (2019). A systematic review of the definition and measurement of engagement in serious games. *Proceedings of the Australasian Computer Science Week Multiconference*, 1–10. New York, NY, USA: ACM. doi: 10.1145/3290688.3290747

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of Score Meaning for the Next Generation of Assessments* (1st ed., pp. 11–24). Routledge. doi: 10.4324/9781315708591-2

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Portsmouth, NH: Greenwood Publishing Group, Inc. Retrieved from https://ucalgary-primo.hosted.exlibrisgroup.com/permalink/f/mtt0p8/01UCALG_ALMA21569166010004336

Kerr, D. (2015). Using data mining results to improve educational video game design. *Journal of Educational Data Mining*, *7*(3), 1–17.

Kerr, D. (2016). Visualizing changes in strategy use across attempts via state diagrams: A case study. *International Journal of Computer Games Technology*, *2016*, 1–9. doi: 10.1155/2016/8492312

Kiili, K., & Ketamo, H. (2018). Evaluating cognitive and affective outcomes of a digital game-based math test. *IEEE Transactions on Learning Technologies*, *11*(2), 255–263. doi: 10.1109/TLT.2017.2687458

Kim, S., Chang, M., Deater-Deckard, K., Evans, M. A., Norton, A., & Samur, Y. (2017). Educational games and students' game engagement in elementary school classrooms. *Journal of Computers in Education*, *4*(4), 395–418. doi: 10.1007/s40692-017-0095-4

Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-Based Assessment Revisited* (pp. 3–11). Cham: Springer International Publishing. doi: 10.1007/978-3-030-15569-8_1

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. doi: 10.1177/0013164406294779

Kuhn, T. S., & Hacking, I. (2012). *The structure of scientific revolutions* (Fourth edition). Chicago ; London: The University of Chicago Press.

Lehman, B., Jackson, G. T., & Forsyth, C. (2019). A (mis)match analysis: Examining the alignment between test taker performance in conventional and game-based assessments. *Journal of Applied Testing Technology*, *20*(S1). Retrieved from http://www.jattjournal.net/index.php/atp/article/view/142699

Liu, H., Yao, M., & Li, J. (2020). Chinese adolescents' achievement goal profiles and their relation to academic burnout, learning engagement, and test anxiety. *Learning and Individual Differences*, *83–84*, 101945. doi: 10.1016/j.lindif.2020.101945

Lu, Y., Zhang, J., Li, B., Chen, P., & Zhuang, Z. (2019). Harnessing commodity wearable devices to capture learner engagement. *IEEE Access*, *7*, 15749–15757. doi: 10.1109/ACCESS.2019.2895874

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games*, *4*(2), e11. doi: 10.2196/games.5888

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in Game-Based Learning* (pp. 59–81). New York, NY: Springer New York. doi: 10.1007/978-1-4614-3546-4_5

Mislevy, R. J., Oranje, A., Bauer, M. I., Davier, A. A. von, & Hao, J. (2014). *Psychometric considerations in game-based assessment*. Redwood City, California: GlassLabGames. Retrieved from https://books.google.ca/books?id=EBi5oAEACAAJ

Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analyzing, and interpreting response time, eye-tracking, and log data. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of Score Meaning for the Next Generation of Assessments* (1st ed., pp. 39–51). Routledge. doi: 10.4324/9781315708591-4

Papert, S. (1987). Information Technology and Education: Computer Criticism vs. Technocentric Thinking. *Educational Researcher*, *16*(1), 22–30. doi: 10.3102/0013189X016001022

Papert, S. (1988). A Critique of Technocentrism in Thinking About the School of the Future. In *Children in the Information Age* (pp. 3–18). Elsevier. doi: 10.1016/B978-0-08-036464-3.50006-5

Pintrich, P. R. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* Retrieved from https://eric.ed.gov/?id=ED338122

Raufelder, D., Hoferichter, F., Ringeisen, T., Regner, N., & Jacke, C. (2015). The perceived role of parental support and pressure in the Interplay of test anxiety and school engagement among adolescents: Evidence for gender-specific relations. *Journal of Child and Family Studies*, *24*(12), 3742–3756. doi: 10.1007/s10826-015-0182-y

Reback, J., McKinney, W., Jbrockmendel, Bossche, J. V. D., Augspurger, T., Cloud, P., … Gorelli, M. (2020). pandas-dev/pandas: Pandas 1.2.0rc0 (Version v1.1.5). Zenodo. doi: 10.5281/ZENODO.3509134

Ren, X. (2019). Stealth assessment embedded in game-based learning to measure soft skills: A critical review. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-Based Assessment Revisited* (pp. 67–83). Cham: Springer International Publishing. doi: 10.1007/978-3-030-15569-8_4

SciPy 1.0 Contributors, Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. doi: 10.1038/s41592-019-0686-2

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research*, *4*(1), 27–41. doi: 10.1080/08917779108248762

Shoemaker, K. (2017). Teaching Mindful Awareness Skills to Middle School Students and its Relationship to Student Engagement with School and Student Test Anxiety (PhD Thesis). Retrieved from https://ezproxy.lib.ucalgary.ca/login?url=https://www.proquest.com/dissertations-theses/teaching-mindful-awareness-skills-middle-school/docview/1901892218/se-2?accountid=9838

Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., … Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education*, *86*, 224–235. doi: https://doi.org/10.1016/j.compedu.2015.08.001

Shute, V. J., & Sun, C. (2019). *Games for assessment*. Retrieved from http://myweb.fsu.edu/vshute/pdf/GBA.pdf

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. The MIT Press. doi: 10.7551/mitpress/9589.001.0001

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*(1), 1–13. doi: 10.1080/00461520.2014.1002924

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493. doi: 10.1016/j.jad.2017.11.048

Von der Embse, N. P. (2012). *High-stakes accountability: Examining student and teacher anxiety within large scale testing*. Michigan State University. School Psychology. Retrieved from https://d.lib.msu.edu/etd/1830/datastream/OBJ/download/High-stakes_accountability___examining_student_and_teacher_anxiety_within_large_scale_testing.pdf

Wang, A. I., Zhu, M., & Saetre, R. (2016). The effect of digitizing and gamifying quizzing in classrooms. In T. Connolly & L. Boyle (Eds.), *Proceedings of the 10th European Conference on Games Based Learning* (pp. 729–737). Nr Reading: Acad Conferences Ltd. Retrieved from http://hdl.handle.net/11250/2426374

Wang, M.-T., & Degol, J. (2014). Staying engaged: Knowledge and research needs in student engagement. *Child Development Perspectives*, *8*(3), 137–143. doi: 10.1111/cdep.12073

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. doi: 10.1111/emip.12165

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. *Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada*. Retrieved from https://pdfs.semanticscholar.org/85b7/c58ebacb707c1d30f97d37370deb315ebd39.pdf

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204–220). Routledge. Retrieved from https://www.taylorfrancis.com/books/e/9781315749136

Zeidner, M. (1998a). An Introduction to the Domain of Test Anxiety. In *Perspectives on Individual Differences*. *Test Anxiety* (pp. 3–27). Boston: Kluwer Academic Publishers. doi: 10.1007/0-306-47145-0_1

Zeidner, M. (1998b). Models and Theoretical Perspectives. In *Perspectives on Individual Differences*. *Test Anxiety* (pp. 61–91). Boston: Kluwer Academic Publishers. doi: 10.1007/0-306-47145-0_3

Zeidner, M. (1998c). Test Anxiety and Cognitive Performance. In *Perspectives on Individual Differences*. *Test Anxiety* (pp. 207–236). Boston: Kluwer Academic Publishers. doi: 10.1007/0-306-47145-0_9

Zeidner, M. (1998d). The Nature and Phenomenology of Test Anxiety. In *Perspectives on Individual Differences*. *Test Anxiety* (pp. 29–60). Boston: Kluwer Academic Publishers. doi: 10.1007/0-306-47145-0_2

Zeidner, M. (2007). Test Anxiety in Educational Contexts. In *Emotion in Education* (pp. 165–184). Elsevier. doi: 10.1016/B978-012372545-5/50011-3

Appendix A

This appendix contains informational letters that were a part of the Raging Skies project. These letters provide information about the project to various parties involved in the project. These parties include: principals, teachers, parents / guardians of participants, and the participants themselves.

Principal and Teacher, Informational Letter

To Principals and Teachers,

Digital game-based assessments have been gaining in popularity over the past few years and Mindfuel in partnership with the University of Calgary have designed a unique learning experience based on the weather outcomes in the grade 5 Alberta Program of Studies. The interactive game-based assessment, Raging Skies, is an innovative video game where 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. The study will help teachers and students better understand their mastery of the learning outcomes in an engaging and educational way. The game is a simulation of storm trackers on their mission to classify different weather phenomena. The game has the potential of enhancing the grade 5 science curriculum through its real-time footage of storms across North America. The purpose of this study is to measure the validity of the video game, Raging Skies, as a means to assess specific and general outcomes related to the grade 5 Alberta Program of Studies.

You will be asked to allow researchers in your school to work with your teaching staff and their students. At the end of the game interviews with your staff will be conducted and an online survey will be administered to their students. Students' participation, as well as your

teachers, in this study is voluntary and is not a requirement; however, their participation will provide direct feedback that can be used to inform future delivery and uses of the effectiveness of this simulation game.

Your teachers' and their students' involvement in the study will not be an increase to their workload. Non-participants will complete the unit using the enhanced feedback, but the researchers will not collect any of their materials.

Consent is voluntary and individual. We kindly ask that consent forms be signed and returned to the researchers by [DATE], 2017. If your teachers and their students agree to participate in the study, they have the right to remove themselves at any time without any consequence until one month after the last day of data collection/end of the game.

This study has been approved by the University of Calgary Conjoint Faculties Research Ethics Board. A University of Calgary researcher, Man-Wai Chu, is the principal investigator of the study and may be contacted at any time to answer questions. If you have any questions regarding this study or the informed consent, please contact the researcher at manwai.chu@ucalgary.ca.

Parent / Guardian and Participant, Informational Letter

To Student and Parents,

Digital game-based assessments have been gaining in popularity over the past few years and Mindfuel in partnership with the University of Calgary have designed a unique learning experience based on the weather outcomes in the grade 5 Alberta Program of Studies. The interactive game-based assessment, Raging Skies, is an innovative video game where 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. The study will help teachers and students better understand their mastery of the learning outcomes in an engaging and educational way. The game is a simulation of storm trackers on their mission to classify different weather phenomena and has the potential of enhancing the grade 5 science curriculum through its real-time footage of storms across North America.

The purpose of this study is to measure the validity of the video game, Raging Skies, as a means to assess specific and general outcomes related to the grade 5 Alberta Program of Studies. The researchers are asking you and your child's permission for them to participate in testing out the game to see if it is engaging as well as educational. Researchers will collect the computer log files and compare them with a standard multiple choice test and then ask the students to complete a short survey on whether or not they enjoyed the game and found it useful.

Your child's involvement in the study will not be an increase to their workload. Non-participants will still play the game, but the researchers will not collect any of their data.

Consent is voluntary and individual. We kindly ask that consent forms be signed and returned to the researchers by [DATE], 2017. If your teachers and their students agree to participate in the study, they have the right to remove themselves at any time without any consequence until one month after the last day of data collection/end of the game. This study has been approved by the University of Calgary Conjoint Faculties Research Ethics Board. A University of Calgary researcher, Man-Wai Chu, is the principal investigator of the study and may be contacted at any time to answer questions. If you have any questions regarding this study or the informed consent, please contact the researcher at manwai.chu@ucalgary.ca.

Appendix B

This appendix contains reproductions of informed consent forms that were a part of the Raging Skies project. These forms provided a way to secure permission for participation in the study from various parties. These parties include: principals, teachers, and parents / guardians of participants.

Principal, Informed Consent Form

Title of Project: Validation of Raging Skies as a means of science assessment

Background

Digital game-based assessments have been gaining in popularity over the past few years and Mindfuel in partnership with the University of Calgary have designed a unique learning experience based on the weather outcomes in the grade 5 Alberta Program of Studies. The interactive game-based assessment, Raging Skies, is an innovative game where 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. The study will help teachers and students better understand their mastery of the learning outcomes in an engaging and educational way. The game is a simulation of storm trackers on their mission to classify different weather phenomena and has the potential of enhancing the grade 5 science curriculum through its real-time footage of storms across North America.

Purpose of the Study

The purpose of this study is to measure the validity of the video game, Raging Skies, as a means to assess specific and general outcomes related to the grade 5 Alberta Program of studies.

What Will You Be Asked to Do?

You will be asked to allow researchers in to your school to work with your teaching staff and their students. At the end of the study, interviews with your staff will be conducted and questionnaires will be distributed to their students. An interview will also be conducted with

you so that the researchers may better understand your views of these assessments in your school.

What Type of Personal Information Will Be Collected?

Students' names and Grade 5 science mark will be collected. The educational researchers will be collecting the computer logs from each student that demonstrates their knowledge of the outcomes. All participants shall remain confidential. The educational researchers will be

collecting students' pieces of evidence that document their achievement (e.g., computer logs and test scores). Elementary science experts will also view some of the log files collected. Your willingness to allow the educational researchers to collect information regarding your

student's achievement during this study will be documented with this consent form. Your feedback on these assessments will also be collected through an interview.

Are there Risks or Benefits if I Participate?

There are no risks to participate in this research. Your data and your students' data will be treated confidentially and reported using pseudonyms and/or in aggregated form. The decision to

participate or not participate will have no bearing on your employment. This research offers no paid compensation for participation, and you will incur no cost to participate. The benefit of participating in this study is you have the opportunity to inform developments of future formative feedback activities.

What Happens to the Information I Provide?

The information provided will be kept confidential. Additionally, any hard copies will be kept in a locked location and any electronic data files will be kept on password protected computers in a locked environment. Permission slips and the master student list will be destroyed after 5 years. Participation in the study is voluntary and you may withdraw from the study at anytime until one month after the end of the study. Only you, other teaching staff associated with this study, and educational researchers will have access to this information. The results of this study will be shared with other educational researchers through papers and professional conferences. If you are interested in the publications of this study, you may contact the researchers for a copy of the study once it has been published.

This consent form is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Take the time to read this carefully and to understand any accompanying information. A copy of this form has been provided for you to keep.

The University of Calgary Conjoint Faculties Research Ethics Board has approved this research study.

Teacher, Informed Consent Form

Title of Project: Validation of Raging Skies as a means of science assessment

Background

Digital game-based assessments have been gaining in popularity over the past few years and Mindfuel in partnership with the University of Calgary have designed a unique learning experience based on the weather outcomes in the grade 5 Alberta Program of Studies. The interactive game-based assessment, Raging Skies, is an innovative game where 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. The study will help teachers and students better understand their mastery of the learning outcomes in an engaging and educational way. The game is a simulation of storm trackers on their mission to classify different weather phenomena and has the potential of enhancing the grade 5 science curriculum through its real-time footage of storms across North America.

Purpose of the Study

The purpose of this study is to measure the validity of the video game, Raging Skies, as a means to assess specific and general outcomes related to the grade 5 Alberta Program of Studies.

What Will You Be Asked to Do?

You will be asked to allow researchers in to your classroom to work with students on the digital video game and then administer a short pre-and post- computer administered multiple choice test. This project will be a review of Grade 5 material. The outcomes are linked to the

weather unit so you will be able to use the data as part of your assessment of the unit. At the end, you will be asked to complete a short survey to indicate your satisfaction with the game.

What Type of Personal Information Will Be Collected?

Students' names and Grade 5 science mark will be collected. The educational researchers will be collecting the computer logs from each student that demonstrates their knowledge of the outcomes. All participants shall remain confidential. The educational researchers will be collecting students' pieces of evidence that document their achievement (e.g., computer logs and test scores). Elementary science experts will also review some of the computer logs generated. Your willingness to allow the educational researchers to collect information regarding your student's achievement during this study will be documented with this consent form. Your feedback on these assessments will also be collected through an interview.

Are there Risks or Benefits if I Participate?

There are no risks to participate in this research. Your data and your students' data will be treated confidentially and reported using pseudonyms and/or in aggregated form. The decision to participate or not participate will have no bearing on your employment. This research offers no paid compensation for participation, and you will incur no cost to participate. The benefit of participating in this study is you have the opportunity to inform developments of future formative feedback activities.

What Happens to the Information I Provide?

The information provided will be kept confidential. Additionally, any hard copies will be kept in a locked location and any electronic data files will be kept on password protected computers in a

locked environment. Permission slips and the master student list will be destroyed after 5 years. Participation in the study is voluntary and you may withdraw from the study at anytime until one month after the end of the study. Only you, other teaching staff associated with this study, and educational researchers will have access to this information. The results of this study will be shared with other educational researchers through papers and professional conferences. If you are interested in the publications of this study, you may contact the researchers for a copy of the study once it has been published.

This consent form is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Take the time to read this carefully and to understand any accompanying information. A copy of this form has been provided for you to keep.

The University of Calgary Conjoint Faculties Research Ethics Board has approved this research study.

Parent / Guardian, Informed Consent Form

To Parents,

Digital game-based assessments have been gaining in popularity over the past few years and Mindfuel in partnership with the University of Calgary have designed a unique learning experience based on the weather outcomes in the grade 5 Alberta Program of Studies. The interactive game-based assessment, Raging Skies, is an innovative game where 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. The study will help teachers and students

better understand their mastery of the learning outcomes in an engaging and educational way. The game is a simulation of storm trackers on their mission to classify different weather phenomena and has the potential of enhancing the grade 5 science curriculum through its real-time footage of storms across North America.

Purpose of the Study

The purpose of this study is to measure the validity of the video game, Raging Skies, as a means to assess specific and general outcomes related to the grade 5 Alberta Program of studies.

What will your child be asked to do?

All grade 6 students will participate in playing the game and take a brief pre- and post- multiple choice quiz on the computer. This project will help your child review Grade 5 material. This information will aid teachers in measuring the science outcomes as well as validating the game as an assessment tool. Your child's involvement in the study will not be an increase to their workload. At the end of this activity, your child will be asked to complete a short survey to rate how they enjoyed the game. Non-participants will still play the game, but the researchers will not collect any of their materials.

What type of personal Information will be collected?

Students' names and Grade 5 science mark will be collected. The educational researchers will be collecting the computer logs from each student that demonstrates their knowledge of the outcomes. Your willingness to allow the educational researchers to collect information regarding your child's achievement during this activity will be documented with this consent form. Please submit this completed consent form to your child's teacher. Survey data will be gathered via Google Docs/Forms. The online survey is being administered by Google©, an American software company. As such, your responses are subject to U.S. laws, including the USA

Patriot Act. The risks associated with participation are minimal, however, and similar to those associated with many email programs, such as Hotmail© and social utilities spaces, such as Facebook© and MySpace©.

Are there Risks or Benefits if I Participate?

There are no risks to participate in this research. Your child's data will be treated confidentially and reported using pseudonyms and/or in aggregated form. The decision to participate or not participate will have no bearing on your child's academic standing. This research offers no paid compensation for participation, and you will incur no cost to participate. The benefit of participating in this study is enhanced feedback on the mastery of the science outcomes.

What Happens to the Information I Provide?

The information provided will be kept confidential. Additionally, any hard copies will be kept in a locked location and any electronic data files will be kept on password protected computers in a locked environment. Permission slips and the master student list will be destroyed after 5 years. Only teachers and educational researchers will have access to this information. The results of this study will be shared with other educational researchers through papers and professional conferences. Parents interested in the publications of this study may contact the researchers for a copy of the study once it has been published. Participation in the study is voluntary and you, and your child, may withdraw from the study at anytime until one month after the end of the study.

This consent form is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your child's participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Take the time to read this carefully and to understand any accompanying information.

The University of Calgary Conjoint Faculties Research Ethics Board has approved this research study.

Appendix C

This appendix contains a reproduction of the participant assent form to join the Raging Skies project.

Title of Project:

Raging Skies: A Digital Game-Based Science Assessment

What is a research study?

A research study is a project to find out new information about something. You do not need to help with this research study if you do not want to.

Why are you being asked to be part of this research study?

You are being asked to take part in this research study because we are trying to learn more about how students learn from video games. Your class has been chosen along with a few hundred other students.

If you join the study what will happen to you?

You will get to play a fun and adventurous science video game to see how well it supports the weather unit.

Will the study help you?

The results of this study will help you better understand your level of knowledge and skills in the weather unit.

Will the study help others?

Understanding if the game is fun as well as educational, may help improve how to best support the learning that occurs in a classroom.

Do your parents know about this study?

This study was explained to your parents and they said that we could ask you if you like to be a part of this study.

Who will see the information collected about you?

The information collected about you during this study will be kept safely locked up. Researchers will be collecting your science marks from your teachers. Nobody will know it except the people doing the research. The study information about you will be given to your teachers. The researchers will not tell your friends or anyone else. Survey data will be gathered via Google Forms; the online data will be deleted from Google Forms after the researchers have downloaded it. The permission slips and master list will be destroyed after 5 years.

What do you get for being in the study?

You will gain insightful information about learning knowledge and skills so that you may improve areas of weaknesses.

Do you have to be in the study?

Participation in this study is strictly voluntary. Your decision to participate or not participate will have no effect on your academic standing.

What if you have any questions?

Should you have any questions, please ask your teacher or contact Man-Wai Chu.

Other information about the study.

If you decide to be in the study, please write and sign your name below. You can change your mind and stop being part of it at any time (up until one month after the data has been collected.)

Appendix D

This appendix contains a script, used in the Raging Skies project, for recruiting participants to the project.

Script for Recruiting Students

Two weeks before start of study (send permission slips home) Hello! My name is Man-Wai and this is Rebecca and we are researchers from the University of Calgary. We are here to test a new video game that our team developed in order to make the weather unit more fun as well as educational. The game is a new way to learn about storms and a way for your teachers to see how well you are learning about them. The purpose of the study is to make sure that the game is fun as well as matched to the grade 5 learning outcomes. The results of the study will help you and your teachers to see how well you have learned about the weather outcomes and see if you liked playing the game as a way to learn about weather. We need your parents to consent to allow us to collect some of your test and task scores as data for this research project. Please bring home this consent form [hand out parent consent form] so that your parents may consider whether they would allow us to collect your information.

At the start of the study. Hello again! We are back to investigate how fun and educational this science game is in helping you learn about storms. As we explained last time, the game is a new way to learn about storms and a way for your teachers to see how well you learning about them. The purpose of the study is to make sure that the game is fun as well as matched to the grade 5 learning outcomes. The results of the study will help you and your teachers to see how well you have learned about the weather outcomes and see if you liked playing the game as a way to learn about weather.

Your participation in this study is voluntary and is not a requirement; however, your participation will provide direct feedback that can be used to inform future delivery and uses of this video game. Although your parents may have agreed to let you participate in this study, you also have a choice to participate or not participate in this study. So we have a consent form for you to sign as well. [hand out student assent form] Let's go through the permission slip so that you will understand this study better. At any time please stop me if you have any questions.

What is a research study? A research study is a project to find out new information about something. You do not need to help with this research study if you do not want to.

Why are you being asked to be part of this research study? You are being asked to take part in this research study because we are trying to learn more about how students play video games to improve learning. Your class has been chosen and 400 students will be involved.

If you join the study what will happen to you? You will be in the study for session at your school and will then be asked to complete a short survey on whether you enjoyed the game and whether it was helpful.

Will the study help you? The results of this study will help you better understand your level of knowledge and skills in the weather unit.

Will the study help others? Understanding students' use of this video game will help to decide if it will be used in other grade 5 classrooms and may help improve how to best support the learning that occurs in a classroom.

Do your parents know about this study? This study was explained to your parents and they said that we could ask you if you like to be a part of this study.

Who will see the information collected about you? The information collected about you during this study will be kept safely locked up. Nobody will know it except the people doing the research. The study information about you will be given to your teachers. The researchers will not tell your friends or anyone else. Survey data will be gathered via Google Forms; the online data will be deleted from Google Forms after the researchers have downloaded it.

What do you get for being in the study? You will gain insightful information about learning knowledge and skills so that you may improve areas of weaknesses.

Do you have to be in the study? Participation in this study is strictly voluntary. Your decision to participate or not participate will have no effect on your academic standing.

What if you have any questions? Should you have any questions, please see your teacher.

Other information about the study. If you decide to be in the study, please write and sign your name below. You can change your mind and stop being part of it at any time.

Do you have any other question? [researcher will answer questions]

Please print your name and then sign the consent form.

Appendix E

This appendix includes the content from the post-survey given during the data collection portion of the Raging Skies project. This survey was presented to participants through Google Docs/Forms. The formatting has been changed to be more concise. Survey items are noted with their header and the Likert scale used.

The first items on the survey collected demographic information:

1. Name
2. Gender
3. Language spoken at home
4. How many hours in a day do you spend with technology? (1 or less, 2 to 3, and 4 or more).

First group of questions. They were introduced with:

> "Using the scale below and thinking about your experience of play Raging Skies and tests in general, please rate the following items. Please answer all items, even if you are not sure. Please select only a single rating for each item."

Their Likert scale was:

1. *Never* (1),
2. *On Occasion* (2)
3. *Some of the time* (3)
4. *Most of the time* (4)
5. *All of the time* (5).

The items were:

1. *I feel bored when playing the Raging Skies video game.*

2. *I feel excited when playing the Raging Skies video game.*
3. *I like the Raging Skies video game.*
4. *I understood how to play the Raging Skies video game.*
5. *Did the Raging Skies video game feel like a test?*
6. *I feel that I learned about different storms while playing the game.*
7. *When I take a test, I worry about failing.*
8. *I feel frustrated sometime when playing the Raging Skies video game.*

The second group of questions had no introduction. The Likert scale was:

1. *Strongly Disagree (1)*
2. *Disagree (2)*
3. *Agree (3)*
4. *Strongly Agree (4)*

The items were:

9. *I believe that I have potential to learn something new by myself if using such games.*
10. *I set up* (sic) *due date or time to finish tasks.*
11. *I make a plan by myself before studying something.*
12. *The content of the Raging Skies Video Game represents what I have learned in my class well.*
13. *I can learn technology easily.*
14. *I often play around with technology.*
15. *Technology enables studying more interesting* (sic).
16. *I would like to do games like this in school and outside of school.*

17. *I believe the contents of the game is comparable to the content learned during the Grade 5 Weather Watch science unit.*

The third group of questions was introduced with:

To what extent do you do the following on a computer? Include things you do in school and things you do outside of school.

The scale used was:

1. *Not at all* (1)
2. *Small extent* (2)
3. *Moderate extent* (3)
4. *Large extent* (4)

The items were:

18. *Play computer games*
19. *Write a word document using a computer program*
20. *Make drawings or art projects on the computer*
21. *Make tables, charts, and graphs on the computer*
22. *Find information on the Internet for a project or report for school*
23. *Use e-mail and social networking site/apps to communicate with others*
24. *Talk in chat groups or with other people who are logged on at the same time*.

The fourth group had no introduction. Its scale was:

1. *Never or hardly ever* (1)
2. *Once every few weeks* (2)
3. *About once a week* (3)
4. *Two or three times a week* (4)

5. *Everyday* (5)

The items were:

25. *How often do you use a computer at school?*

26. *How often do you use your own mobile device/tablet at school?*

27. *How often do you use a computer outside of school?*

The fifth, and final, group of questions was introduced with:

Please indicate the extent to which you AGREE or DISAGREE (sic) *with the following statements.*

The scale included:

1. *Strongly disagree* (1)
2. *Disagree* (2)
3. *Agree* (3)
4. *Strongly agree* (4)
5. *I never use a computer.* (5)

The items were:

28. *I am more motivated to get started doing my schoolwork when I use a computer*

29. I have more fun learning when I use a computer

30. I get more done when I use a computer for schoolwork.

The survey concluded with a single question:

31. How would you evaluate your performance during the game?

With a scale that had options:

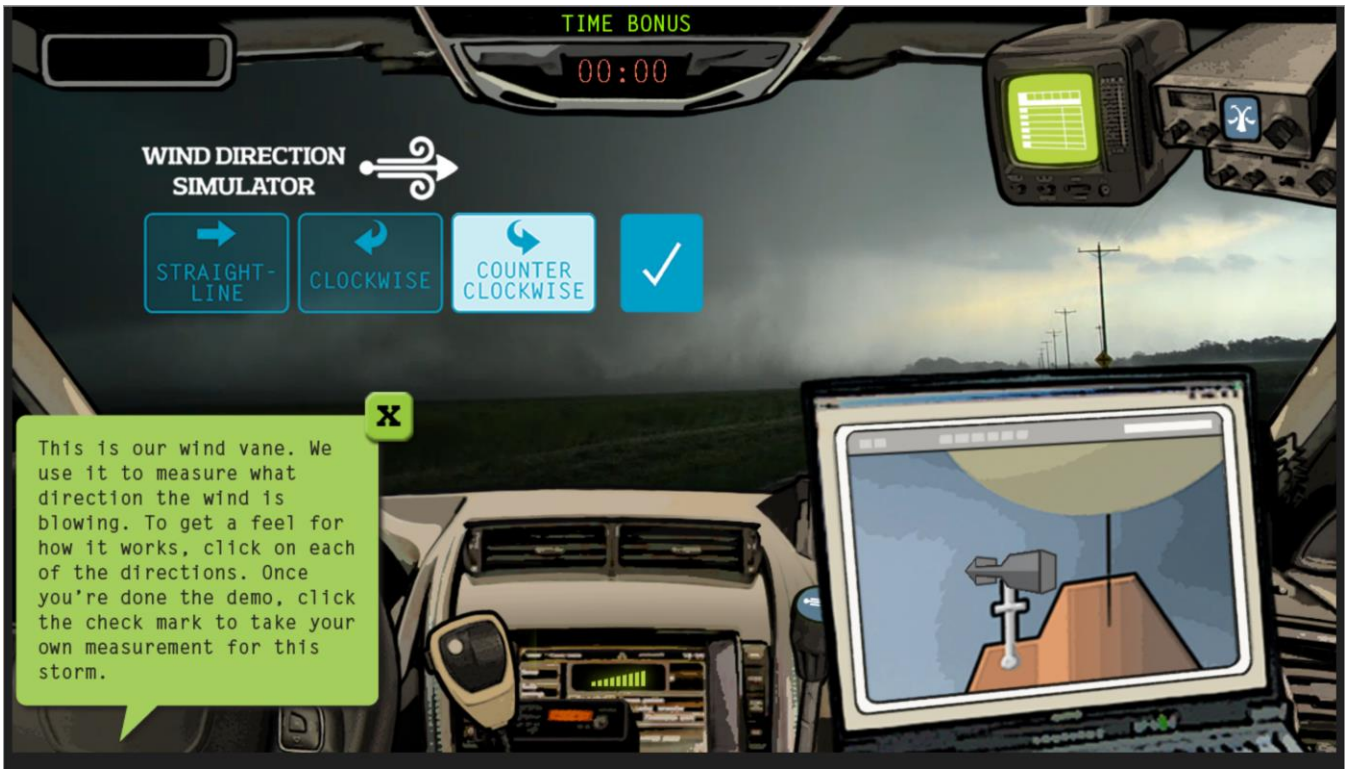
1. *Unsatisfactory*
2. *Improvement needed*

3. *Meeting my teacher's expectations*
4. *Exceeds my teacher's expectations*
5. *I don't know*

Appendix F

This appendix includes pictures and explanations of each of the instruments that participants can make measurements with when playing Raging Skies. Each instrument and measurement is detailed individually.

Wind Direction

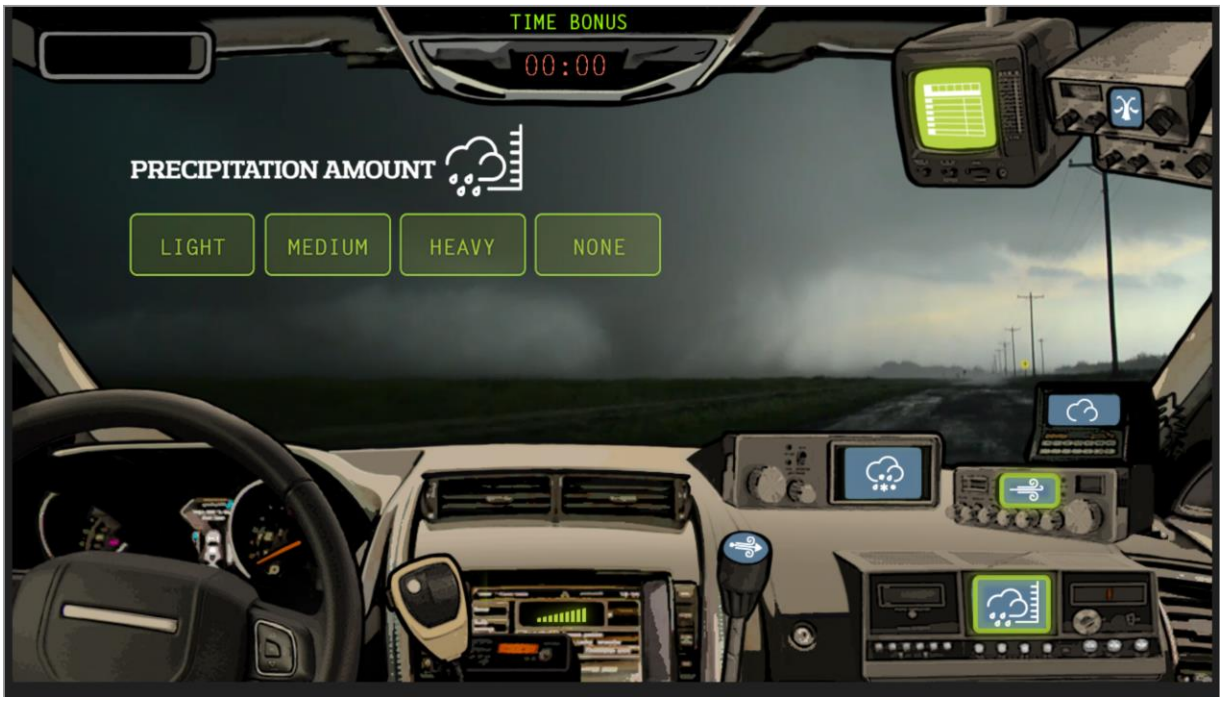


This instrument (the wind vane) measures the wind direction. The measurement is made by the participant via the graphic of a wind vane (in the lower right). The participant is prompted to infer the direction the wind from the storm is blowing. Possible measurements the participant can make for the wind direction are: Straight-Line, Clockwise, and Counter Clockwise.

Precipitation Type

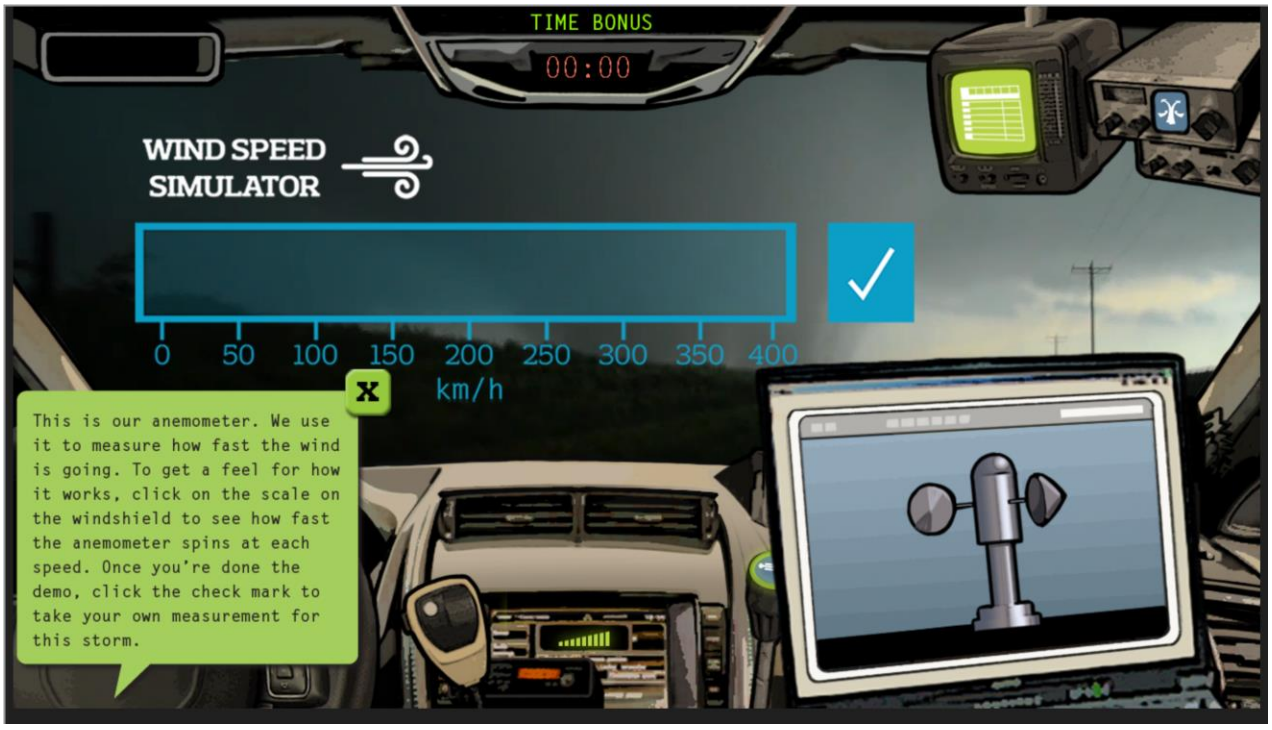
This instrument measures the type precipitation. The participant is invited to infer the type of precipitation from the graphic of the storm through the front wind shield of the vehicle. Possible measurements include: Rain, Hail, Rain/Snow, Ice Pellets, Snow, and None.

Precipitation Amount



This instrument (a rain gauge) measures the amount of precipitation from the storm. The participant can attempt to infer this amount from the graphic of the storm. Possible values for this measurement are: Light, Medium, Heavy, and None.

Wind Speed



This instrument (an anemometer) measures the wind speed of the storm. The participant can infer the wind speed from the graphic in the lower right-hand corner. The graphic appears to spin with the wind. For example, when the instrument is spinning quickly this is meant to imply that the wind speed is high. Possible values of the wind speed vary from zero to 400 (kilometres per hour).

Could Type

This instrument measures the type of cloud seen from the storm. The could type is inferred from the graphic of the storm. The participant can also click on the *cloud guide* marker (seen in the middle left of the picture) to be taken to then display information in the Storm Log about cloud types. Possible measurements to make include: Cumulus, Cirrus, Cumulonimbus, and Wall Cloud.

Updraft Speed/Wind Direction

This instrument measures speed of the updrafts of wind related to the storm. The graphic in the lower right serves as a guide for the participant. The orange arrows in the guiding graphic will animate slowly or quickly to attempt to imply the speed of the updraft. Measurements with this instrument are referred to as Wind Direction in the log files. This is why this instrument is also referred to as an instrument for measuring the wind direction. Possible values for measurement include: Light, Medium, or Strong.

Storm Log

Finally, the Storm Log is an element of Raging Skies that is related to measurement within the game, but does not measure by itself. The Storm Log contains information about different types of storms. Information about the different types of storms is displayed when the participant clicks on the graphics of the different types of storms across the top of this figure. The storms included in the Storm Log are (From left to right): Single Cell Thunderstorm, Multicell Thunderstorm, Supercell Thunderstorm, F1 Tornado, F3 Tornado, F5 Tornado.

The information contained in the Storm Log, about the types of storms, corresponds with the measurements that a participant can make with the measurement instruments. This information is displayed in the table shown in the figure. The information appears in the column to the right of the column labeled *Your Measurements*. Each of the rows of the table correspond to a property of a storm. These properties are related to the measurements that participants can take with the instrument present in the game. For example, the row in the graphic labeled with *Cloud Type* displays what types of clouds are associated with a particular storm. The types of

clouds displayed correspond to the values that are present in the Cloud Type instrument (i.e., Cumulus, Cirrus, Cumulonimbus, and Wall Cloud). Thus, clicking on various graphics of storms shows the different properties of each storm. These properties appear next to information about the participant's own measurements. This format allows for easy comparison between what the participant has measured and the properties of the storms. This hopefully allows for the participant to infer which storm they are making measurements from.

Appendix G

This appendix shows the relationship to participant clicks on different elements of Raging Skies and the information that appears in a log file. The elements of Raging Skies include: virtual instruments to make measurements, the Storm Log, and the graphic that prompts participants to make hypothesis or conclusions about storms.

This appendix begins with the information added to the log file when a virtual instrument is clicked and the participant is asked to make a measurement. Different information is recorded when the participant is prompted to measure, and when they enter information to make a measurement. The information shown here is the former, rather than the latter. Also, recall that participants can have incorrect responses for measurement. After an incorrect response, participants can attempt to measure a second or a third time. The information recorded in the log file differs if the participant is measuring for the first time or they are trying again. The information in this table corresponds to the record made in the log file for the first measurement.

Each measurement type is labeled under the “Element” column. The corresponding information recorded in the log file is then included in the “Log file record” column. The information in the log file for clicking on an instrument and being prompted to make a measurement appears in the *action* column of the log file. When recording this information, the other columns in the log file show other information. The *id* column records the participants unique identifier, *result* is set to *null*, *correctAnswer* is set to *null*, *Stormid* records the identifier of the storm that is being chased, *timestamp* records the time of the click, and *DataOfAssessment* records the date.

First measurements, prompting participants to measure

Element	Log file record
Wind Direction	HudWindDirection_Open
Precipitation Type	HudPrecipitationType_Open
Precipitation Amount	HudPrecipitationAmount_Open
Wind Speed	HudWindSpeed_Open
Cloud Type	HudCloudType_Open
Updraft Speed / Wind Direction	HudAirMovement_Open

After being prompted to measure, new information is added to the log file when the participant makes a measurement. This information appears in the *action*, *result*, and *correctAnswer* columns of the log file. The *action* records what type of measurement is being made. The *result* column shows which option the participant selected when making the measurement. The *correctAnswer* column then records the correct response or responses for that measurement. Also, the *id* column records the participants unique identifier, *Stormid* records the identifier of the storm that is being chased, *timestamp* records the time of the click, and *DataOfAssessment* records the date.

This information is shown in two tables. This first table shows the information that appears in the log that denotes that the participant is making a measurement. The second table shows the information recorded when a selection is made by the participant when taking a measurement. The information that is recorded in log files about correct responses does not need an additional table. For each measurement there are either one or two correct responses. For a

single correct response, the information recorded is identical to the information recorded when a participant chooses an option. For example, if the correct response to a measurement of Wind Direction was Straight-Line, then the correct response would be recorded as Straight-Line. However, the information recorded differs if there are two correct responses. Such correct responses are multiple pieces of information separated by a colon character. For example, if a measurement of Wind Direction had a correct response of Straight-Line or Clockwise, then the correct response would be recorded as: *Straight-Line:Clockwise*. Finally, the information that appears in these tables are for first measurements made by the participant.

First measurements, participant is measuring

Element	Log file record
Wind Direction	HudWindDirection
Precipitation Type	HudPrecipitationType
Precipitation Amount	HudPrecipitationAmount
Wind Speed	HudWindSpeed
Cloud Type	HudCloudType
Updraft Speed / Wind Direction	HudAirMovement

Recall that there are several options to choose from for each measurement. These options are recorded in the Option column. Thus, this table records the various options that a participant can choose, and the information that appears in the log file.

First measurements, participant is selection an option

Element	Option	Log file record
Wind Direction		
	Straight-Line	Straight-Line
	Clockwise	Clockwise
	Counter Clockwise	CCW
Precipitation Type		
	Rain	Rain
	Hail	Hail
	Rain/Snow	Rain Snow
	Ice Pellets	Ice Pellets
	Snow	Snow
	None	None
Precipitation Amount		
	Light	Light
	Medium	Medium
	Heavy	Heavy
	None	None
Wind Speed		
	0 through 400	0 through 400
Cloud Type		
	Cumulus	Cumulus

	Cirrus	Cirrus
	Cumulonimbus	Cumulonimbus
	Wall Cloud	Wall Cloud
Updraft Speed / Wind Direction		
	Light	Light
	Medium	Medium
	Strong	Strong

The next table shows what information is recorded when a participant clicks on an instrument to make an additional measurement beyond the first. That is, the participant is retrying a measurement after a first incorrect response. This information is identical to the first measurement, except for the information recorded to indicate what measurement the participant has selected.

Additional measurements beyond the first, participant is measuring

Element	Log file record
Wind Direction	HudWindDirection_Retry
Precipitation Type	HudPrecipitationType_Retry
Precipitation Amount	HudPrecipitationAmount_Retry
Wind Speed	HudWindSpeed_Retry
Cloud Type	HudCloudType_Retry
Updraft Speed / Wind Direction	HudAirMovement_Retry

The next table shows what is recorded in a log file when a participant clicks on the Storm Log element and selects a type of storm to get information about. Information is also recorded in the Storm Log about measurements the participant has made on the current storm. When a participant clicks on the part of the storm log that reports that information, *current* is recorded in the log file.

Storm Log

Element	Option	Log file record
Storm Log		
	Single Cell Thunderstorm	single_cloud
	Multicell Thunderstorm	double_cloud
	Supercell Thunderstorm	triple_cloud
	F1 Tornado	f1_tornado
	F3 Tornado	f3_tornado
	F5 Tornado	f5_tornado

Next, when a participant is prompted to make a hypothesis about the type of storm they are chasing, the information *HudHypothesis* is recorded into the *action* column of the log file. When a participant is prompted to conclude what storm they are chasing, *HudPickStorm* is recorded into the *action* column. Additional information, listed in the table below, is also recorded into the *result* and *correctAnswer* columns. The information contained in the *result* and *correctAnswer* columns are shown in the *Log file record* column of the following table. If the participant makes an incorrect selection, then the values in the log file will differ. Otherwise, the values in the *result* and *correctAnswer* columns will be the same. For example, if the participant selection is an F1 Tornado but the correct answer is an F3 Tornado, then the *result* column will record F1 while the *correctAnswer* column will record F3. Also, the *id* column records the participants unique identifier, *Stormid* records the identifier of the storm that is being chased, *timestamp* records the time of the click, and *DataOfAssessment* records the date.

Storm hypotheses and conclusions

Hypothesis / Conclusion	Log file record
Single Cell Thunderstorm	SingleCell
Multicell Thunderstorm	MultiCell
Supercell Thunderstorm	SuperCell
F1 Tornado	F1
F3 Tornado	F3
F5 Tornado	F5

Once the participant has come to a conclusion about the storm they are chasing, and made a selection, a graphic is shown to summarize their performance on that storm. This summary is indicated in the log files by *HudResultsInterface* in the *action* column and a *Results_Interface_Continue* in the *result* column. The *correctAnswer* column is set to null. Also, the *id* column records the participants unique identifier, *Stormid* records the identifier of the storm that is being chased, *timestamp* records the time of the click, and *DataOfAssessment* records the date.

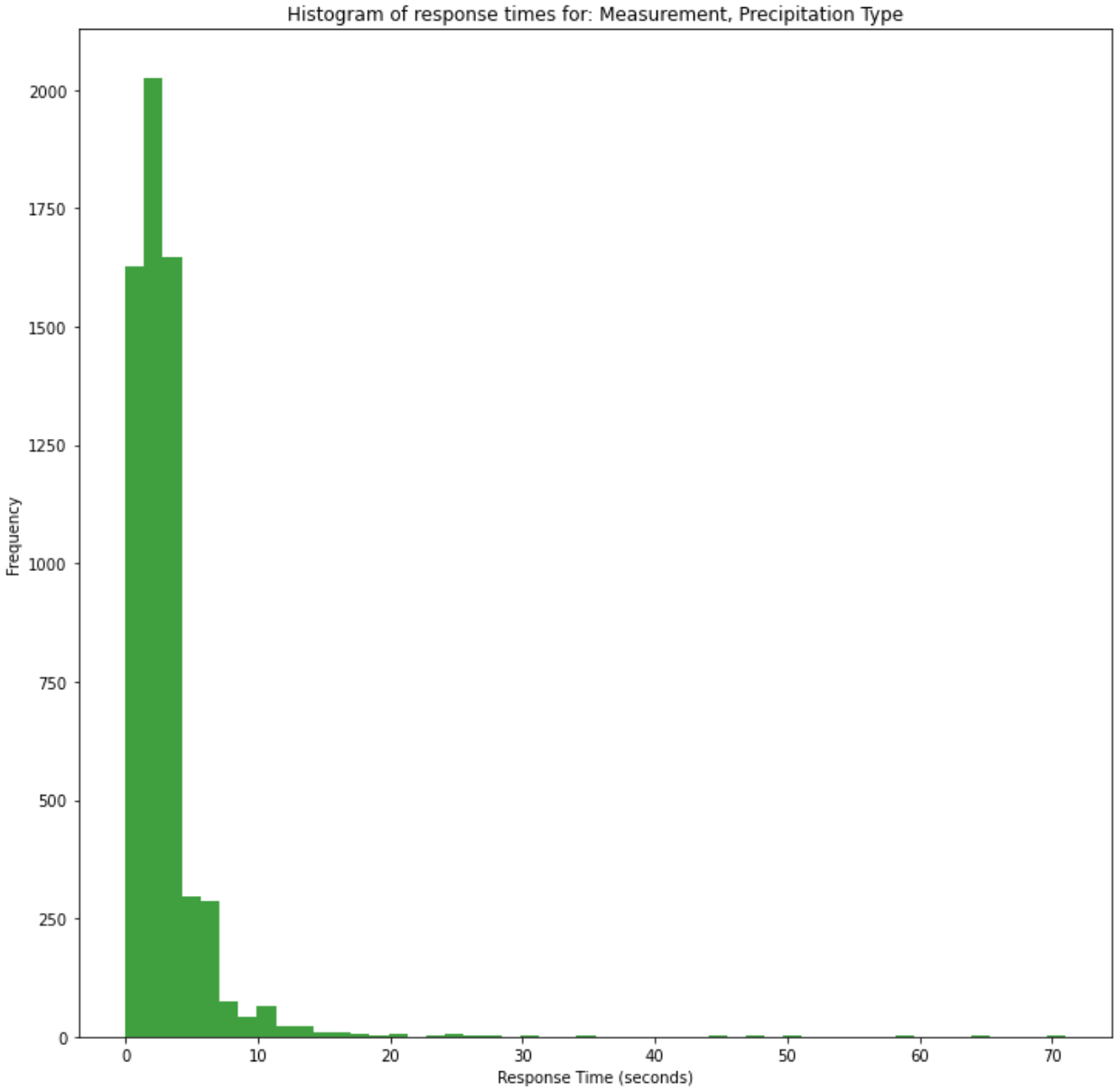
Finally, when a participant is choosing a storm to chase, information is recorded in the log files. When a participant is shown a selection of storms to chase, this is reflected in the log file with *Map_Weather_Icon* in the *action* column. Also, either *icon_pressure_low* or *icon_pressure_high* is shown in the *result* column. The *correctAnswer* column is set to null. When a participant makes a selection of which storm to chase *Map_Chase_Storm* is recorded in the *action* column. Also *btn_chase* is recorded in the *result* column. Also, the *id* column records the participants unique identifier, *Stormid* is set to *null, timestamp* records the time of the click, and *DataOfAssessment* records the date.

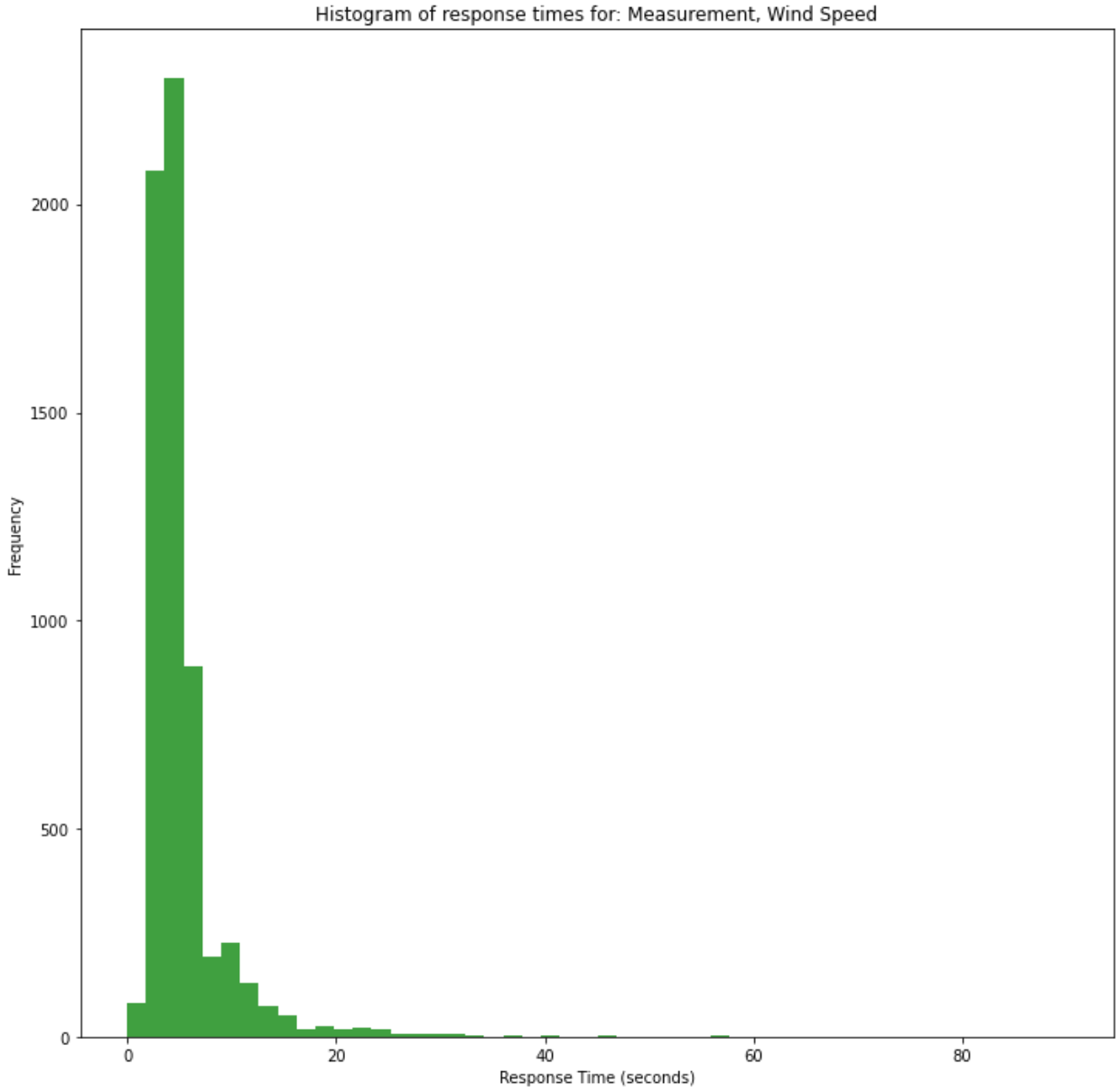
Appendix H

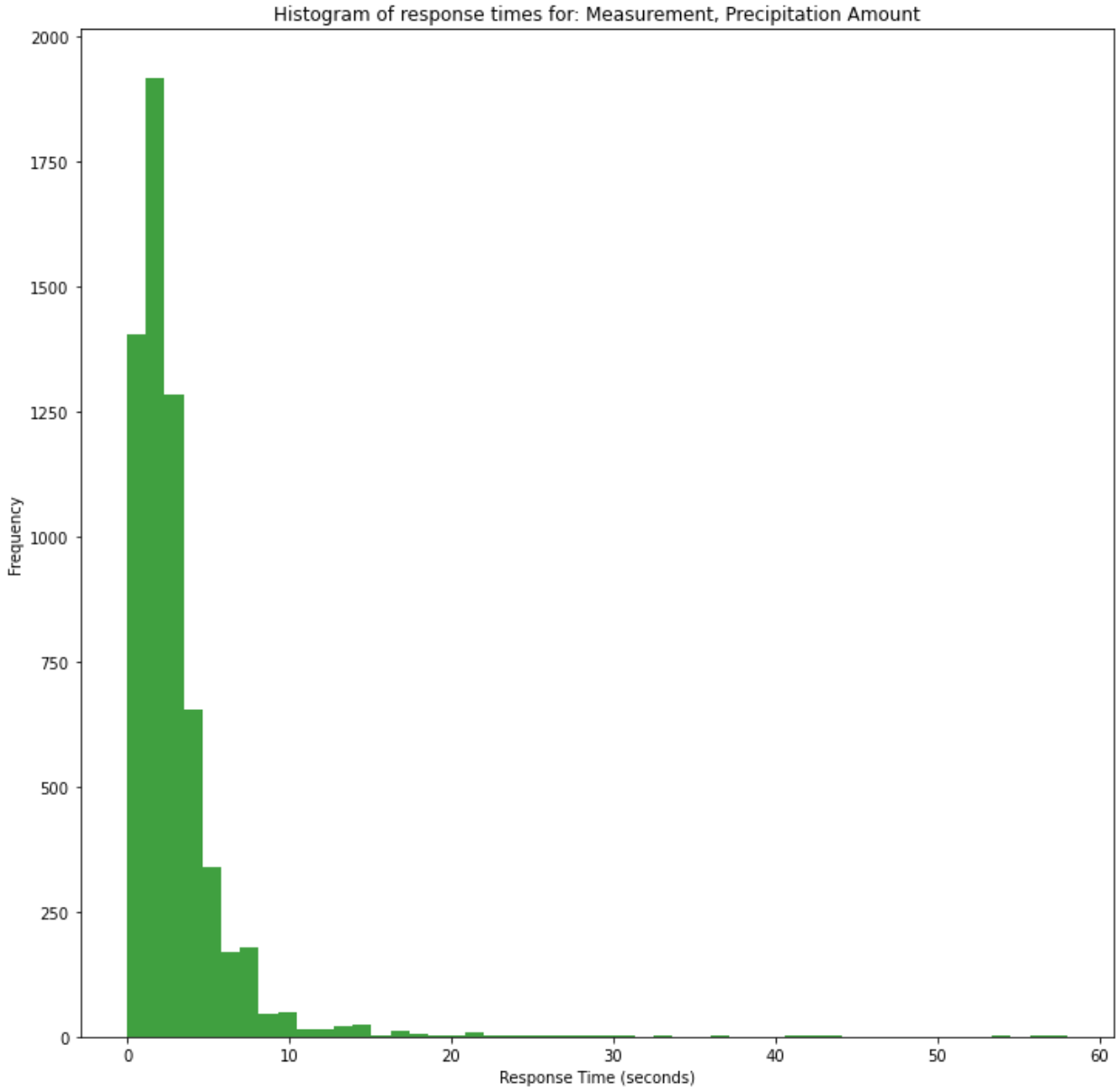
This Appendix illustrates an outlier analysis of response times for measurement, hypothesis, and conclusions tasks from Raging Skies. There were few if any outliers in the response time data. The distributions of the response times followed a, so-called, *long-tailed* pattern. That is, there were many response times grouped around the mean with many unique outliers with little if any grouping. It was concluded that removing the outliers was unnecessary as they did not appear to effect the value of the mean very much. To support this conclusion, alternative means are shown with outliers removed that were beyond 3 and 4 standard deviations above the mean. Also, the distributions of the response times for the various tasks (measurement, retry, hypothesis, and conclusion tasks) are shown, in several additional graphs..

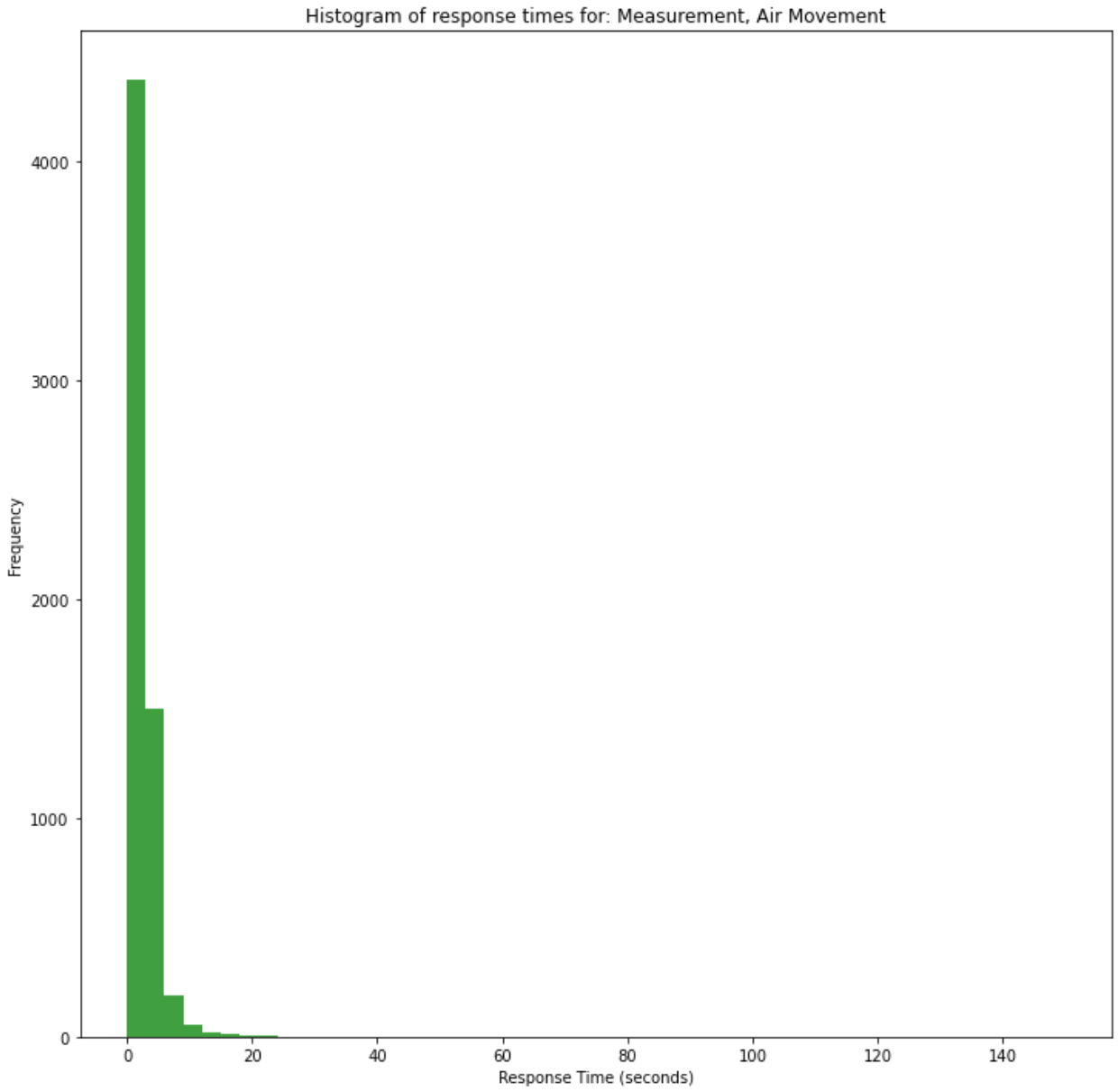
Task	Arithmetic Mean	Mean with outliers above 3x standard deviations removed	Mean with outliers above 4x standard deviations removed
Measurement, Precipitation Type	2.9276913099870296	2.6808475689881734	2.7356979405034325
Measurement, Cloud Type	3.5687418936446176	3.183767733421313	3.308585445625511
Measurement, Wind Direction	3.0390473104342193	2.8224406224406224	2.865230869636156
Measurement, Air Movement	2.4185971164749716	2.2867850741404596	2.2570261437908496
Measurement, Precipitation	3.067552243641665	2.7904793279525615	2.871648136036625

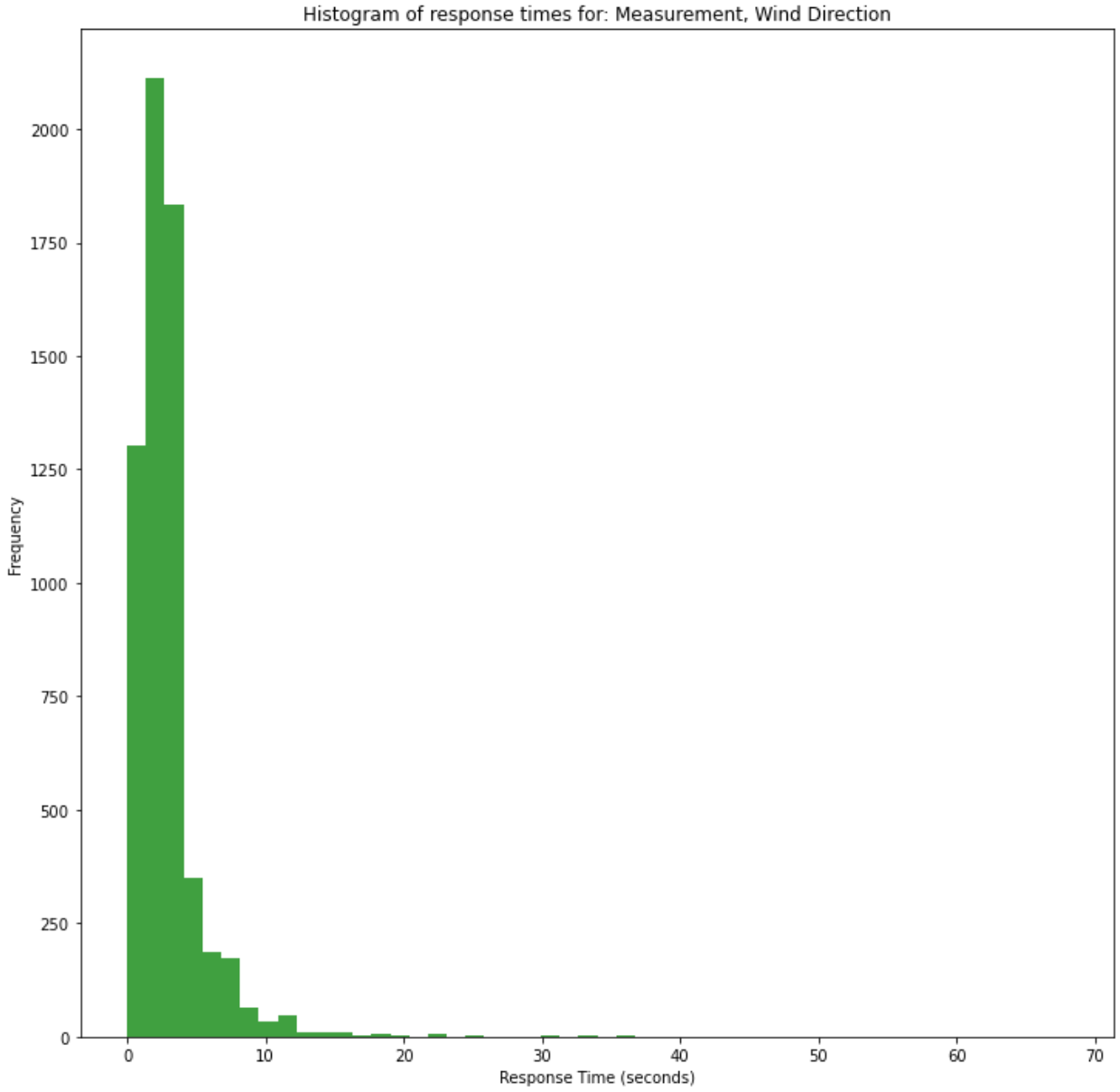
Amount			
Measurement, Wind Speed	5.178282009724473	4.751943112287084	4.885611274991806
Hypotheses	6.471569739186781	6.0109800065552275	6.1713355048859935
Conclusion	6.9438255138867095	6.39529745936563	6.51227959697733

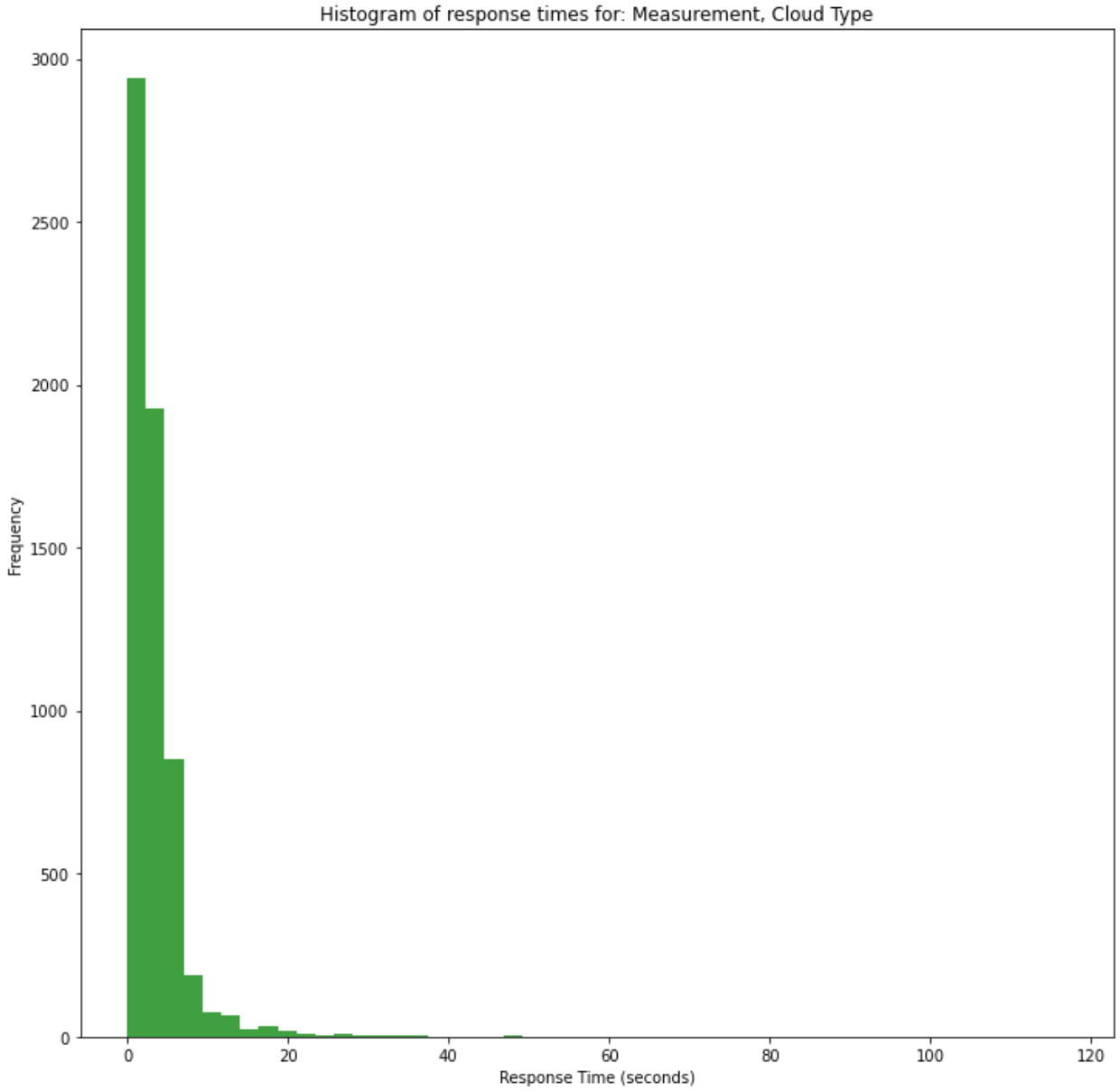

Histogram of response times for: Measurement, Precipitation Type
Frequency
2000
1750
1500
1250
1000
750
500
250
0
0
10
20
30
40
50
60
70
Response Time (seconds)

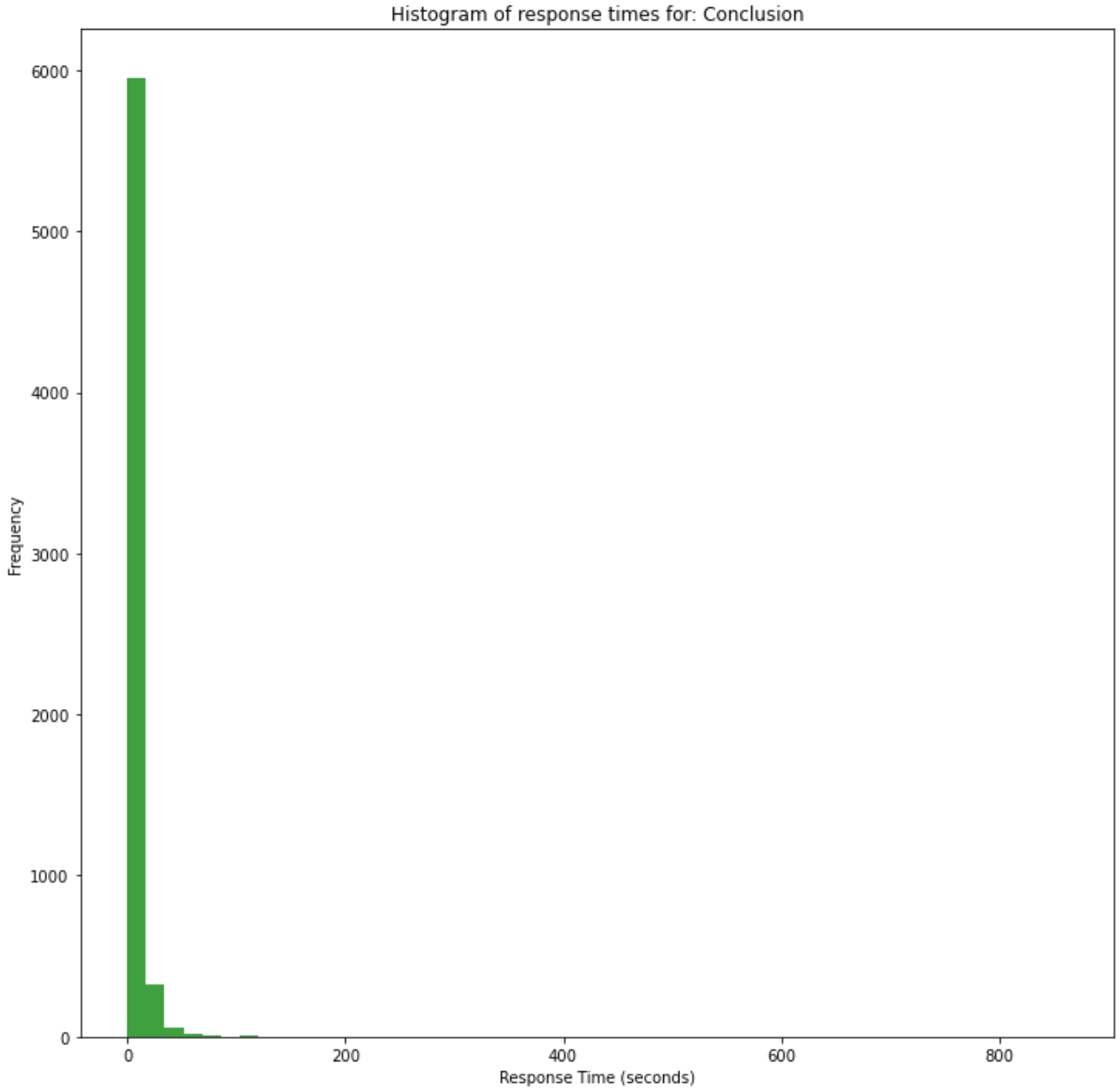

Histogram of response times for: Measurement, Wind Speed
Frequency
2000
1500
1000
500
0
0
20
40
60
80
Response Time (seconds)

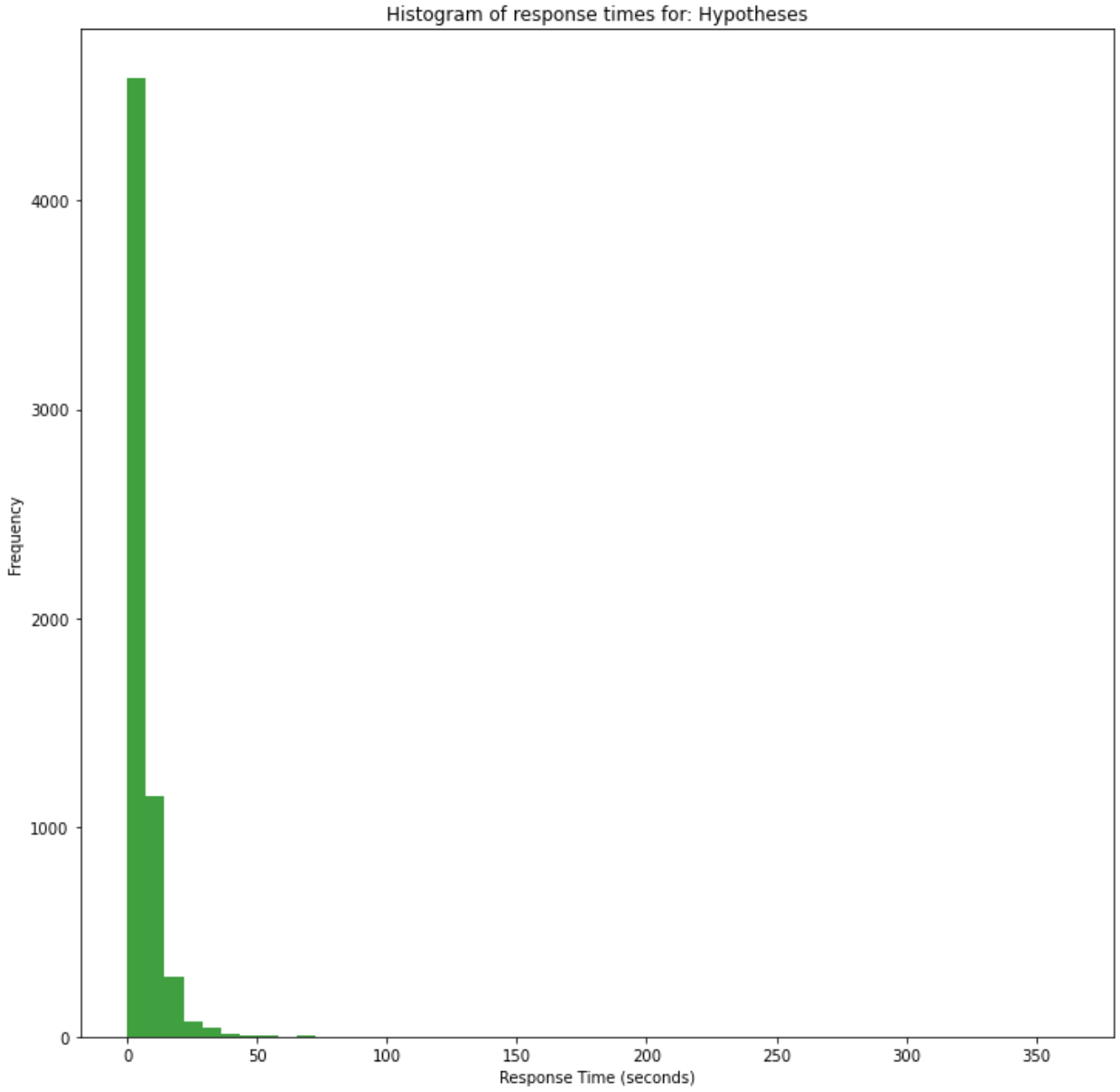

Histogram of response times for: Measurement, Precipitation Amount
Frequency
2000
1750
1500
1250
1000
750
500
250
0
0
10
20
30
40
50
60
Response Time (seconds)


Histogram of response times for: Measurement, Air Movement
Frequency
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
Response Time (seconds)


Histogram of response times for: Measurement, Wind Direction
Frequency
2000
1750
1500
1250
1000
750
500
250
0
0
10
20
30
40
50
60
70
Response Time (seconds)


Histogram of response times for: Measurement, Cloud Type
Frequency
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
120
Response Time (seconds)


Histogram of response times for: Conclusion
Frequency
6000
5000
4000
3000
2000
1000
0
0
200
400
600
800
Response Time (seconds)


Histogram of response times for: Hypotheses
Frequency
0
1000
2000
3000
4000
0
50
100
150
200
250
300
350
Response Time (seconds)