UNIVERSITY OF Southampton

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Clare Walsh (2020) "Scoring games fairly: Biases and interference in games based assessment data", University of Southampton, Education School, Faculty of Social Sciences, PhD Thesis.

University of Southampton

Faculty of Social Sciences

Web Science

Scoring games fairly: Biases and interference in Games Based Assessment

by

Clare Elizabeth Walsh

ORCID ID 0000-0002-7757-2301

Thesis for the degree of Doctor of Philosophy

January 2020

University of Southampton

Abstract

Faculty of Social Sciences

Web Science

Thesis for the degree of Doctor of Philosophy

Scoring games fairly: Biases and interferences in game based assessment data

by

Clare Elizabeth Walsh

ORCID: 0000-0002-7757-2301

Gaming is an interactive medium that has much in common with education. Both games and good classroom practice are learning environments, with overall objectives, scaffolded progression, checks along the way, and regular, purposeful feedback. Games also provide a space to practice complex skills such as collaboration, or managing a system. These skills are rarely directly assessed in compulsory education because they are difficult to evidence efficiently. Games are fun learning environments for many children, and they could provide a means to resolve this problem. However, the structure of gaming data is not aligned to many assessment analysis methods. Gaming data is conditionally dependent, there are continuous variables as well as categorical and dichotomous responses, there is often more than one possible proxy for ability, and there are very large amounts of data missing. Aspects that assessors traditionally force to be constants, such as the number of attempts or the response time, become variables in games, and it is important to know their limitations and worth as variables.

This interdisciplinary study looks at these problems in scoring performance in games. It uses a quantitative methodology, with a case study secondary data set from MangaHigh. MangaHigh is a website with a range of dynamic maths games for primary and secondary aged learners, and over a million children were using the site at the time of data extraction. Using a sample data set, chosen by criterion sampling, the impact of missing data, response times and additional attempts was explored through insights and methods from Item Response Theory (IRT) and other quantitative analysis techniques. Demographic data also helped to contextualize the findings and inform decision-making.

In the analysis, choice of game mechanics were found to have an impact on the extent and nature of missing data, which was found to have a complex relationship with the target variable, ability. The choice of measure, such as mean, recency-weighted mean, high score or most recent score was found to be central to determining the grade. Several issues when the child competed against a human or bot competitor or collaborator were identified. Response time functioned as a context variable to define valid attempts, helping to identify non-targeted behaviours such as browsing, conceding or wandering off. As gamers have suggested, response time appeared to also function as a proxy for ability, but there does not seem to be a linear relationship between ability

and time. Instead, ‘speed’ seems to be the proxy, and this was found to be a function of the response time, the child, the game and also the band score and game mechanics. Outside of an optimal range, short response times could act as a confounding variable. There was evidence that some stability of performance may also act a proxy of ability. Finally, adding a familiarity weighting when a child comes back for a second attempt proved problematic, but a novelty weighting for early attempts can work. Having said that, although games became easier with each subsequent attempt, evidence from the first attempt playing appears unreliable, and the data has features that are characteristic of guessing behaviour.

Although a large number of problems were identified, this analysis also found some clear ways forward to adjust the assessment and games design, and the collection of data to make scores from games more meaningful and reduce bias in the scoring process. On the basis of this study, there are many design choices that could improve or deteriorate the quality of data gathered in gaming environments.

Table of Contents

Table of Tables

Table of Tables

Table of Figures

Table of Figures

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Clare Elizabeth Walsh

Title of thesis: Scoring games fairly: Biases and interference in games based assessment data

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signature: Date:

Acknowledgements

The author would like to thank Blue Duck Education for the provision of the data set and their support for the aims of this project. I would like to thank the EPSRC and the Digital Economy Network for funding the research. I would like to thank all of the staff at the Web Science Institute, and both the Electronics and Computer Science and Education Departments for their help and patience throughout this project, with special thanks to my two supervisors, Dr Christian Bokhove and Dr Su White.

Definitions and Abbreviations

AI	Artificial Intelligence
CAT	Computerized Adaptive Testing
CDA	Cognitive Diagnostic Assessment
CEMθ	Conditional Error of Measurement
DAG	Directed Acyclic Graph
DIF	Differential Item Functioning
ECD	Evidence Centred Design
ECgD	Evidence Centred games Design
EPPI	Evidence for Policy and Practice Information
ETS	Educational Testing Services
GCSE	General Certificate of Secondary Education
GBA	Game Based Assessment
GDPR	General Data Protection Regulation
ICC	Item Characteristic Curve
IEEE	Institute of Electrical and Electronics Engineers
IRT	Item Response Theory
LTSA	Learning Technology Systems Architecture
MAR	Missing At Random
MCAR	Missing Completely At Random
ML	Machine Learning
MNAR	Missing Not At Random
NaN	Not a Number
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PTEA	Pearson Test of English (Academic)
SCC	Skills Challenge Certificate
SEMθ	Standard Error of Measurement
TIMSS	Trends in International Mathematical and Science
UTME	Unified Tertiary Matriculation Examination
www	World Wide Web
W3C	World Wide Web Consortium

3PL	3, 4, 5 Parameter Logistic Model

Chapter 1 Introduction

Technology continues to offer a range of new options to study and learn. The appearance of serious educational games is one example of a shifting learning environment. Online games are fun and popular and may provide a less stressful environment for the child to practice and to evidence their learning. Games have many things in common with educational environments, they have participants, and sets of rules and challenges to overcome, with obstacles along the way. In addition, commercial online games often require skills which have been largely absent from educational assessment, but which are valued and commonplace in the modern work place.

Computers are eliminating the need for routine processing skills. Most processes that rely on rule based logic can now be programmed. This means that valuable work skills have been changing. The US Secretary of Labor launched the Secretary's Commission on Achieving Necessary Skills for the future as far back as 1991, and they found evidence then that the need for routine skills in the job market had already declined sharply. Skills such as collaboration, complex problem solving, digital literacy and managing more complex systems were still valued, and growing in demand. President Obama gave an additional push for these skills to become a priority in American education (Obama, 2009) and in 2010 Common Core State Standards were called on to integrate more expert thinking types of skills (Koenig, 2011). These calls for change were echoed internationally by the push the Organisation of Economic Cooperation and Development (OECD) (Ananiadou, 2009).

There seems to be broad agreement that success in social, public and business sectors is increasingly reliant on more complexity, and yet education systems, and particularly exam systems, are remarkably similar today to the exams of the 1990s. The OECD trialled a

collaborative element to the Programme for International Student Assessment (PISA) national comparative performance exams in 2015 (Gurria, 2016). The Welsh Government introduced the Skills Challenge Certificate (SCC) to assess these skills in all 16 and 18 year olds, through project work in 2016 (Qualifications Wales, 2019). These are relatively small changes given the political rhetoric around these new skills.

As serious educational games appear to provide an environment to evidence the skills, they seem to be a potential platform to allow more complex skills to be assessed. The technical infrastructure to record activity in these kinds of tasks efficiently already exists in the form of games. However, there has been very little research on making the automated processes of scoring in games fair from an assessment perspective. Fair scoring in education is a complex decision-making process. Major test providers often produce several hundred pages of documentation to justify the choices they make. There are a range of statistical analysis techniques to assess the reliability of the scoring, and there is a large amount of substantive knowledge to support how scores are qualitatively interpreted. The problem with game scoring is that the data gathered are not readily adaptable to those current analysis techniques, and there is very limited substantive knowledge available. Without a persuasive argument that the scores are fair, games are unlikely to be treated as a serious form of assessment.

So although the hope for Game Based Assessment (GBA) may be the eventual scoring of those more complex skills, there are considerable barriers at the moment. Scoring in any kind of game environment is not well documented. There is also limited consensus on how to score more complex skills, as a survey by the OECD found. Among the 16 national government representatives that stated that soft skills were valued in their education systems, very few had clear definitions of those skills (Ananiadou and Claro, 2009). With the exception of digital literacy, these are not new skills. At times the terminology around

them can suggest personal qualities, rather than baseline skills (Kirschner, 2018), hindering their adoption. There has been some progress in introducing soft skills into universities, which perhaps have the flexibility to more sensitively respond to what employers think of the readiness of their students (Short, 2019). But, in general, three decades on from the first US commission, soft skills assessment still seems to be an emerging field.

For this reason, for this study, I decided that isolating the game performance factors in assessments of mathematics skills was an appropriate next step in the research. The game play environment already introduces a number of new concerns and complexities that would be easier to understand and document without the additional challenge of simultaneously introducing challenging skills into the research design. This research was therefore intended to be a bridging study. If we are to use games to seriously assess new skills, we first need to know what gameplay data look like from an assessment perspective. The aim was to recommend actions at the end to improve the game design to get better data, or to suggest steps to refine the resulting scoring process with the data that come out.

This is a quantitative case study, looking at a series of dynamic mathematics games used with primary, or Key Stage 1 and 2, mathematics learners. Mathematics has a well-established history of teaching and testing, and it is therefore easier to isolate new factors and variables introduced by the game delivery platform, which is a field with very little published literature within assessment and education.

This is a Web Science PhD. Web Science aims to bring the full range of academic disciplines, from health to the humanities, together with computer science to produce truly interdisciplinary researchers (Berners-Lee, 2006). Both the fields of games design and assessment are highly specialized academic disciplines. On reflection, having reached the end of this research, it is now clear that the fundamental assumptions, approaches, glossary and methodologies between the two disciplines are not readily interpreted or understood,

but that is precisely the kind of problem Web Science hopes to tackle. This study is positioned at the intersection of computer science, education and assessment, to see which field has solutions that can help to resolve some of the issues. The findings show that there are opportunities for the use of games as a platform to deliver serious assessments. There is evidence of reliability in the data that were collected from these games, but there are also many aspects that need to be factored into the scoring process. This, however, is not necessarily unique to game delivered assessments.

From an education perspective, the data also gave evidence on a broad range of children in terms of ages, geographic location and ability. Despite the obvious parallels of objectives between games and educational systems, games design is fundamentally different from assessment design and pedagogy. From a computer science perspective, the resulting data analytics require original approaches. From an assessment perspective, there are opportunities to adapt the assessment and game design in the early stages to improve the quality of data and potentially refine the scoring process. The conceptualization of some of the new variables introduced in gameplay data, in particular, can have a major impact on scores of ability, and many of these are unique to game environments.

This study makes an original contribution to the field by:

1. Evidencing the need for a stochastic approach to response time that is band specific. In other words, it is not appropriate to use the known value ‘response time’ obtained from the time on the clock. Optimum ‘speeds’ need to be conditioned on the child, the game, and the child’s position in the grading scale.

2. Showing that very short response times can evidence non-targeted behaviour and should not be rewarded.

3. Demonstrating a way to use assessment techniques to estimate the reliability of

first, second and additional attempts at the game to decide which attempt to use.

4. Using assessment techniques to quantify precisely how much easier subsequent attempts become.

5. Recommending and justifying the introduction of a 'novelty weighting' for early attempts, rather than the familiarity weighting used in assessment for subsequent attempts.

6. Scoping the range of actions available to deal with missing data in game data sets with appropriate recommendations.

7. Identifying possibilities for bias when children are in different geographical locations and have varying chances of playing another human or bot.

8. Showing the role bot competitors and collaborators can play to reduce the number of variables during multiplayer games, and in turn, make estimation more stable.

As this is an early study in this field, there were features to factor in which had no direct pre-existing guidance on how to deal with them. There are, therefore, a number of recommendations at the end of this thesis for possible future developments.

1.1 The background

Formal tests matter. Success in school leaving exams is a major goal of education and formal test scores are used as a performance measure for educational professionals. Teachers, students, parents, universities and future employers, in general, focus on the students' ability to write a good answer, unless there is a major error in test construction. Conversely, an assessment expert focuses on the test-writer's ability to write a good question in the first place, and whether the test was fair.

Assessment design is a highly specialised field. Despite the important role of professional

assessment, the practice itself is not often discussed in public, and it is given limited coverage in Initial Teacher Training courses. The Carter report found that:

> 'In England there are world leading organisations in educational assessment and yet this expertise is not always utilised in [Initial Teacher Training] programmes. (Carter, 2015, page 34)'

The new core content framework for Initial Teacher Training in the UK, for example, requires teachers to study formative and peer or self-assessment practices (Twiselton, 2019), but not necessarily be told how the school leaving tests might be constructed and validated. The issue of fair testing has troubled assessment experts for many decades.

In the UK, the structure of the exam system may have been an additional imperative to develop the discipline of assessment. There is no single, annual university entrance exam. In June 2018, there were 759,670 entries for an A level qualification in a specific subject (Ofqual, 2018), which is the principle qualification taken in England. Students generally choose three or more subjects, and there are over 45 subjects to choose from. A child taking a mathematics A-level, for example, has a choice between mathematics, statistics, pure mathematics and use of mathematics. There are four main exam boards, and so there are multiple versions of the same qualification in the same subject (Baird, 2018).

It is important that every exam has the same level of challenge, so that universities and employers can make sound judgments on whom to accept. In England, the Office of Qualifications and Examinations Regulation has been responsible for regulating the different exam boards. The exams themselves, though, are developed by credited awarding bodies, who have relatively little contact during the test development phase beyond a shared framework for subject criteria and shared regulatory requirements (Isaacs, 2010). It is a system that has maintained fairly stable standards since the 1990s, at least in mathematics (Jones, 2016). Whenever there is more than one version of a paper, it is inevitable that each year there will be an 'easiest test', a 'hardest test' and all the levels in

between. There will be some questions with design faults or poor wording or structuring that only reveal themselves once the answers have been collected in (Mislevy, 2018). These margins of error in an exam as economically and socially important as an A-level are problematic to ignore.

When the results are published 6 weeks later, using a range of statistical evidence and judgment, (Baird, 2018) sufficient work will have been carried out to make the claim that the grades awarded, and the position of the grade boundaries for each version of the test, are fair. This means that they should be equivalent at every grade boundary and represent the ability of the students, not the meanness or leniency or competence of the test writer or a human judge (Messick, 1987). This is not a simple process, but there has been considerable work done to refine and improve judgments and decisions.

One of the fundamental principles of assessment is the acknowledgement that any test score is a combination of the child's true ability and an error in measurement (Lord, 1980). Some children will guess, or run out of time, and some students simply behave erratically on the day as a result of stress and pressure. Any number of things could also go wrong in writing a question. It may be poorly worded, or too complex, or too reliant on literacy levels. The test writer may make socio-cultural assumptions that are not as universal as they believed, or not accessible to the learner's age group (Mislevy, 2018). Measurement systems are characterised by a degree of imprecision and in education assessment in particular, there could be overlap between questions (Andrich, 1988) or questions that test a very narrow section of the syllabus and therefore unfairly skew the results. The child's true score, therefore, is impossible to know precisely because it always has to pass through the filter of imperfect tests. Wiliam, in fact, argued that the scope for error was so great that standardised testing should be abandoned (2001). An appetite for standard assessment, however, remains in the public sphere, and while that is the case, there are techniques to

uncover and reduce the scope for error and get closer to the true score.

The difficulty of each test is also of major concern. Although the aspect of the syllabus under examination can give an indication of the level, not every question that covers the same aspect of the syllabus has the same level of difficulty. Assigning a value of '1 unit of cognition' to each tick can be just as arbitrary as using a random value. Some questions need a huge leap in cognition to get the answer correct, but still get a value of one. Once the social role of assessment had become established in the post war period (McArthur, 1987), much of the early work in assessment revolved around the joint challenges of error identification and estimating difficulty through statistical analysis (Rasch, 1960, Lord, 1980).

Although there are many potential approaches to assessment standardisation, Item Response Theory (IRT), a suite of statistical techniques, has emerged to help analyse performance data when it is in numerical format. The analytics address some of the concerns mentioned above. These techniques have been used with a wide variety of test delivery formats.

In more recent years, games have become a popular form of entertainment, rivalling more traditional media like television and film. It did not take long for researchers to see the potential in games to deliver educational content (Squire, 2005). Educational games aim to create immersive and engaging environments to deliver educational content through hands-on tasks (Squire, 2008). Sims games provide experience of managing a city or farm or other complex environment (Fullerton, 2014), many games offer online collaboration, hero's quest games offer complex problem solving (Adams, 2010). In general, the soft skills are very much built into these environments

The data from educational games are gathered in numerical format, and seems suited to statistical IRT analysis. However, there are some inconsistencies in the data structure

produced in games which makes it challenging to use assessment analytics on it (Mislevy *et al.*, 2012). If assessors' tools will not work, and games analysis tools do not address the concerns of assessors, there is something of an impasse on using games to deliver educational assessments.

1.2 Conflicting interests of game design and assessment

While in principle games have many things in common with tests, the data produced are very different. Looking at the literary output of assessors and games designers, there are some very fundamental differences in focus. Assessors assume that the same version of the test has been delivered to all of the students, an assumption that is so fundamental that it is rarely stated. Games, on the other hand, offer choice and so children see different selections of tasks (Mislevy, 2012). Some assessors might believe that the data set is complete and that missing responses are indicative of an inability to answer (Mislevy, 1996, Ludlow, 1999). This cannot be the case in games, where individual pathways through an overarching narrative arc are fundamental (Fullerton, 2014). Players are forced to miss out certain pathways. The element of choice also has the effect of producing chains of actions (DiCerbo, 2016). This is in contradiction to the isolated responses favoured by IRT practitioners (Linacre, 2006). And then finally, a large number of new factors are introduced, such as variable response times, direct competitors, and second and third attempts (Adams, 2010). There were times in the early stages of this study when gameplay data seemed too resistant to analysis.

From a personal perspective, the environment that the results were collected in was fascinating, though. There was a variety of tasks and the gaming environments and mechanics were varied. Sometimes they raced to a finish line, at other times they shot down robot attackers, or rotated crafts in space. They had collaborators and competitors of the human and robot kind. Sometimes true collaborators were disguised as competitors,

inviting a friend to a challenge game, and then conceding and allowing them to win. There were records of primary school children on opposite sides of the world working together on the same homework tasks. All of their wrong turns, preferences and persistence were evidenced in the data set.

Much of the efforts of computer scientists have been directed at producing the large body of research in a distinct field of assessment, Intelligent Tutoring Systems. Intelligent Tutoring Systems have been in use since the 1980s and aim to replicate the role of a human tutor, observing, evaluating and recommending next steps through a range of Artificial Intelligence (AI) techniques (Nkambou, 2010). In fact, Intelligent Tutoring Systems are so dominant in the field of applying modern computing techniques to educational problems that Nkambou et al use them as synonymous with the term AIEd (Artificial Intelligence Education). There has also been considerable interest from computer scientists in Games for Learning, which uses online game environments as a means of either instruction or application of skills taught (Gee, 2005). Both of these fields do assessment-like work, but do not have the aim of precisely measuring the child's ability. Both are more focused on identifying and recommending the next best steps in learning. Some researchers from computer science have looked into the possibility of using Artificial Intelligence (AI) analysis approaches, such as fuzzy cognitive logic and k-means algorithms (Baron, Salinas and Crespo, 2014; Baron *et al.*, 2015), or the possible use of semantic ontologies to model skills (Vendlinski *et al.*, 2010).

There is a sense that new data analytics capabilities from computer science might be able to address some of the assessment concerns, but a cautionary approach might be advisable. It seems that there has rarely, if ever, been a point when standardised assessment has been widely embraced as a positive addition to educational processes. Rather it became essential to satisfy concerns about standards (McArthur, 1987). In assessment, when computational

processing power made IRT approaches more practical and more widespread, Messick (1987) pointed out that a strictly mathematical approach to scoring was not adequate. A score has a social meaning and interpretation and a large number of qualitative factors have to be considered alongside any quantitative analysis.

This is true today, when AI approaches aim to do fairly similar things as early IRT applications. They both aim to produce strictly probability based judgments on human behaviours. The problems of refining and improving the data collection process, and clarifying the conceptualisation of key variables and behaviours have not gone away with sophisticated AI approaches. This study explores the possibilities to improve the quality of data coming into the model, through a range of quantitative analysis and IRT techniques.

If scoring performance through games delivery is a challenge that lies somewhere at the intersection of assessment and a computer science, a basic mastery of both disciplines would be helpful to understand the affordances and limitations of gaming and assessment environments. This is a Web Science PhD, and is therefore inter-disciplinary by design. Web Science is the study of online behaviour, informed by both a knowledge of the technology of engineered online environments and understandings of human behaviour from the social sciences (Berners-Lee *et al.*, 2006). It is the recognition that the Web is a socio-technical environment, and a new type of researcher, trained in both fields, is needed to make sense of much of what is happening. Using games as educational assessments seems to be a strong example of the kinds of challenge that Web Science aims to address, as the results are shaped by both engineered structures (choice and narrative pathways, allowing repetitions and recording response times) and social behaviour (letting your friend win, experimentation etc.). Web Science could be seen as reflective of Messick's considerations around validity (1987), but within much wider applications of research.

It is desirable that people working in assessment and computer science collaborate on the

issue of Games Based Assessment (GBA). Games designers will likely see their efforts frustrated if they cannot address the serious concerns of those who work in high stakes proficiency testing. At the same time, the assessment landscape today is fairly familiar to that of the 1990s. Many other aspects of the life we are preparing children for, how they will work, how they socialise and communicate, are radically different. Games could unlock some of the barriers to moving forwards from knowledge based assessments that served in the last century, and introduce more complexity into the assessment landscape. Assessors have by and large ignored GBA as a potential means to aid the transition to more complexity in assessment, which, as mentioned at the beginning of this century, many influential groups and individuals have called for. Understanding GBA requires a degree of engagement in some of the problems computer scientists have faced making the games work. Another part of the original contribution of this project is to look at the areas where there has been an apparent breakdown in communicating different interpretations of success between these two groups, which Web Science aims to do.

1.3 Research questions

This thesis is a quantitative case study of Key stage 1 and 2 mathematics skills assessed through a series of games delivered in a dynamic environments. The maths skills are based on the primary curriculum, and the games environments are varied, rich and dynamic. Children playing the games had control over what they played and how often.

A quantitative approach was taken because it allowed more broader questions around what is happening to be answered (Walliman, 2017). In addition, games are data rich environments, and quantitative approaches might be able to exploit this (Ifenthaler, 2014). A case study was chosen because just one data set contained a huge amount of detail. There were no agreed procedures from the published literature, as the Literature Review in Chapter 3 will show, but where insights emerged, it came from case studies of individual

games. Case studies also allow for depth of research to understand complex systems (Yin, 2003), and games can become very complex.

The research questions are:

1. **What counts as a valid attempt on a game task?**

When children are allowed to repeat the games many times, it is not clear if their first attempt should be treated as an example, or a trial, or if it should be included in a permanent record of the estimation of the child's ability. The unit of measurement that qualified as a 'task' was not clear, either. Success requires a series of actions, rather than one isolated 'item'. Other factors also influence the validity of attempts. There was clear evidence of children conceding a win to another, for example, when they gave up after a few seconds.

2. **Does missing data impact the final score?**

As mentioned above, there was a vast amount of missing data forced by the structure of the games. Where children had an element of choice, there was missing data, and there was a lot of choice in these games. How should we interpret this missing data? Was it in any way related to the target variable of ability?

3. **How can the game-specific variables of response time and iterations be conceptualized and scored?**

In the wider field of assessment, response times and iterations are treated as constants, even if they are variable in practice. Every child has one hour to complete the test, only the top-most answer is accepted, even if there are five or six alternatives crossed out. In commercial games, varying response times are a central indicator of ability, but is that fair? Is it fair to allow children additional attempts, and if so, what do we do with the old data on

earlier attempts?

4. **Within the game, how reliable do the results appear to be?**

Ultimately, the purpose of this study is to investigate whether there might be an argument for the more widespread deployment of games as assessments. To do this, some statistical indication of reliability is helpful going forward.

There were other issues around validity that would have been interesting to explore, but required a different research design. It was also something of a chicken and egg conundrum. The quantitative approach highlighted some interesting patterns that needed insight from qualitative approaches, but a purely qualitative approach was unlikely to focus on problematic areas without insight and direction from the quantitative approaches. In the end, a satisfying number of patterns and possibilities were found while answering these questions, and there is still opportunity for both more focused and broader investigation of the findings.

1.4 Organisation

Chapter Two begins with a general outline of the current state of relevant research in the fields of computer science, education and assessment. The audience for this PhD, the supervisors and examiners at each stage, came from either the Electronics and Computer Science department or the Education department. It was essential to summarise the state of research and position the problem from different perspectives. The approaches, methodologies, aims and standards that are common in either of the disciplines are broadly outlined in this chapter, and there are many relevant considerations to highlight from both fields.

Chapter Three will present the results of the systematic literature review of the current findings on Games Based Assessment (GBA). Before carrying out a pilot scoping of the literature, the intention was to carry out a quantitative meta-analysis of the effectiveness of each approach using the reported reliability statistics. On further investigation, the field of GBA is still far from that stage. None of the studies had got to the point where they felt reliability statistics could be presented. The criteria for inclusion in the literature search was broad, but still only produced 41 texts, and there were very few commonalities among those. It is perhaps a benefit at this early point in the research that so many different problems have been identified and so many approaches have been presented. No one method to score games has emerged as dominant.

While hugely insightful, the initial literature review offered very little guidance or structure to answer the specific research questions posed here. In particular, much of the published literature offered insights into which methods did not work well, and why they did not work, but very little in terms of what might yield more promising results. Chapter Four presents the results of a second scoping of the literature by snowballing approaches. The aim of this was to understand how researchers have tackled problems that have parallels with some of the concerns that I had from a pilot investigation of the data set. The insights came from both assessment and computer science. While the systematic literature helped to frame and contextualise the problem more clearly, this second literature review informed more precisely the methods used in this study.

This project will take data from a mathematics gaming environment, used in primary and secondary schools in the USA, Brazil and the UK, among other countries. It prepares students in aspects of the Key Stage 2 and 3 Maths curriculum, or US Common Core for the same age group, and the main purpose of the games is to encourage homework practice. It has dynamic gaming interfaces, with cartoon avatars, bright colour washes and

varied sound tracks and special effects noises. The game mechanics are integral to success in the games, as well as the skill. Children play a range of games, such as steering submarines and hot air balloons towards values, or using knowledge of angles to create spiders' webs. The games include single and multiplayer modes, competition and collaboration. The data set had over a billion data points and over a million users worldwide at the time of extraction.

Chapter Five describes the data set used in this case study in more detail. There were many new features included in the data, and so the exploratory phase was fairly detailed. There are also ethical considerations as this is a very young population and they were required to play these games in order to do their homework. This chapter also presents the methods chosen. The scores from the games in this case study were always intended to be used by the teacher and child as an alternative to a classroom progress test. They were designed for a low stakes assessment environment and for formative purposes. Because the games were intended for formative purposes, they provide longitudinal insights into children's performance.

It was quickly evident from a pilot review of the data that even in relatively straightforward maths games, the data evidenced soft gaming skills as well as the classroom delivered lessons, such as transformations or geometry. The methodology used here is intended to reduce test difficulty variation and remove error of measurement, and it is important to note that the psychometric methods chosen are more commonly used with high stakes proficiency testing, not progress tests. The reasons for taking this approach will also be discussed in Chapter Five.

In Chapter Six, the results are presented. The study found that assessment analysis literature can shed light on how to conceptualize and introduce new variables in a computing environment. There are opportunities in the assessment and games design to

improve the quality of the data collected during calibration phases. The games design helps to explain some of the surprising results. Reliability statistics, though not complete on their own, do suggest that performance in these games could, in fact, be scored fairly despite the noise caused by the fact of delivering the assessment through a game.

Chapter Seven provides the discussion and reflection on the results, and in particular how specific results might be interpreted. Some of the methods chosen could work with gameplay data under some circumstances, but there were also many unexpected but revealing data patterns in the results. On inspection and reflection, much more was involved in the measurement process than initially anticipated when designing the methods and there was statistical evidence of features of gameplay that were not anticipated because they have not been described in the literature before. Robots and other humans, and new decisions over what to hold constant appeared and adapting some of the successful methods from studying these games with other types of games of other skills was hard.

Chapter Eight summarises the implications for practitioners and returns to the wider issues going forward. The methods and findings described in this study were aimed specifically at improving scoring by targeting and refining the evidence gathering and collation stages, and there were some clear recommendations around the data collection to be made. Design choices could be more intentional and findings could inform Games Based Assessment (GBA) design and research in future, even if other statistical methods are used.

There are considerably wider questions around scoring game play that go beyond this study, and in particular, whether the best future design of GBA research might be an AI-based design. The overall conclusions of this study are very much in line with Messick's observations several decades ago (1987), that mathematical judgments delivered by computer are helpful, but they are not enough on their own to produce understanding of ability.

Chapter 2 Bringing together computer science, education and assessment in Games Based Assessment

There are many considerations that impact on how games might be used for assessment. The purpose of this chapter is to position Games Based Assessment within the field of education. It will introduce the implications for stake holders in the educational sphere, but also outline the main concerns of assessors and games designers, and the assumptions that might be behind their different methodologies. These disciplines are brought together under a Web Science lens, and this is a Web Science PhD, and so the chapter begins with a brief outline of what Web Science is.

2.1 Positioning Games Based Assessment as a Web Science challenge

The Web Science Institute was established in 2008 by a team at the University of Southampton, UK, but it now has a global presence. Researchers in the Electronics and Computer Science department had observed over the years that the most technically appropriate and optimised models they had built were often unsuccessful. Humans had a habit of choosing bad models, or misusing the good ones, and there was limited information within the field of computer science research that could account for much of the impact of their work (Berners-Lee, 2006). The http protocol that made the World Wide Web (www) possible is an example of this. Few coders at the World Wide Web Consortium (W3C) foresaw the impact of their tweaks to the html protocol code to make

uploading easier around 2002. Nonetheless, those small changes to the global standard code and a huge amount of social invention resulted in the explosion of smart technologies. Social media has had far reaching impacts in almost every sphere of public and private life. From a computer science perspective, the adoption of new technologies seemed quite random. From time to time, people have turned to Professor Sir Tim Berners Lee, the creator of the http protocol and one of the founders of Web Science, to see if he could account for its misuse and offer solutions. Web Science was created to address those two tasks (Halford, 2010).

Web Science brings together people with backgrounds in any discipline, and provides them with training in the social sciences, computer science and interdisciplinary work to tackle the rapidly changing problem spheres around the Web and its uses. At a fundamental level, Web Science is based on the premise that the web is constantly being co-created by different groups of people. There are inventors who make the technological innovations possible, business and economic leaders who find money-making uses for those technologies, and humans who make choices as users (Halford, Pope and Carr, 2010). There is a long history of interdependence between new technologies and the social structures they emerge in (MacKenzie and Wajcman, 1999). Even those who actively choose to be a non-user can sometimes become a determining factor in how the web evolves (Oudshoorn and Pinch, 2003), and they, in turn, are affected.

The founding Web Scientists felt that a new type of researcher was needed, with a new set of tools to understand this interplay between technology and society. It was a means to get better, fairer technologies in future (Berners-Lee *et al.*, 2006). The potential to use games to deliver assessments is an example of this kind of problem. It is argued in this paper that the technological environment of games, which emphasises online real time collaboration or competition, state machines and complexity, already encourages complex skills, such as

collaboration and problem solving or management to be practised. The evidence we can collect is determined by the engineered environment of games and it is helpful to understand why certain technical restrictions on what children can do have been imposed. The fact that the game is non-linear, it is a hypertext, means that games are fundamentally unlike any other kind of test. It also dictates the kinds of data that we can get out. It is therefore helpful to understand engineered design decisions and recognise what can and cannot be altered or adapted. At the same time, the actions and the scores have a social meaning and intention. Assessment processes bring a range of socio-cultural assumptions into play. There is a motivation to find a way to assess soft skills is in order to make a shift in education systems from knowledge-retention to skills-based a realistic choice. It needs to be considered from every side of the problem.

My background coming to the Web Science PhD is from an education background, and particularly, assessment, but with training in computer sciences through the Web Science MSc programme.

2.2 The educational and social context of assessment

2.2.1 The increased role of testing in education systems

In the UK, the role that tests play in education and society has changed. Converging trends in society and, more specifically education, have given exam regimes an influential role in the last 20-30 years.

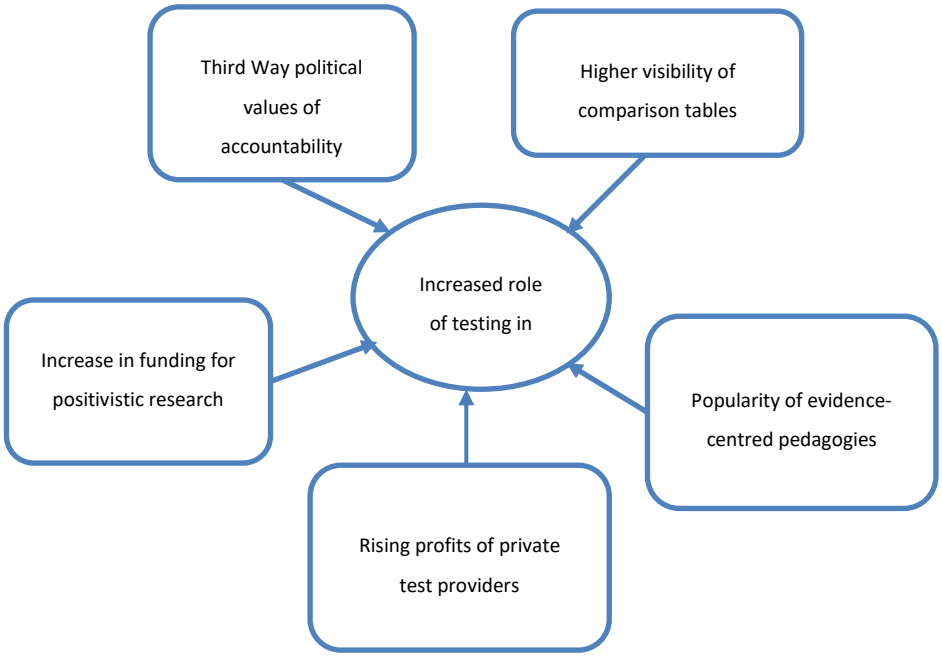


Figure 1 Influences on the rise of exam regimes

Statutory assessments of the National Curriculum were introduced in the UK at 7 and 11, as well as 16 and 18, in 1989, and political changes soon helped to consolidate the status of examinations. Since then the UK, for example, has seen a range of trends converge to give examinations a privileged status in education systems, summarised in Figure 1.

Third Way politics, which began in the late 1990s, emphasised accountability in public services (Giddens, 2009). As a result, summative test results became used as a vehicle to evidence responsible use of government funding. This was not just a trend seen in the UK. In the US, the 2001 No Child Left Behind project was an attempt to raise educational outcomes through the Common Core programme (Baird *et al.*, 2018). It failed to raise outcomes for poor children, but established a lasting legacy of standardised testing in US schools.

Internationally, pressure to perform in standardized tests came with the Organisation for Economic Co-operation and Development (OECD) assessment of 15 year olds in member and non-member nations through the Programme for International Student Assessment

(PISA) in 2000 (Ananiadou and Claro, 2009). The purpose was to produce international comparison tables, supplementing the pre-existing Trends in International Mathematical and Science Study (TIMSS) and Reading Literacy Study (PIRLS). Success in these competitive tables has been see as a proxy for the broader success of education systems globally.

Changes in the funding of educational research may also have raised the profile of standardised tests. Since the 1990s, as part of the Third Way approaches, and in response to the rise of data science possibilities, there has been an increase in government funding for research that evidences learning in a quantitative way (Nind, 2006). Quantitative approaches need robust assessment tools.

As a commodity, testing has also become subject to market forces. Sophisticated analysis techniques have appeared along with greater computational powers, and demand for higher standards has increased the cost of test production. Testing is now a business that generates approximately $2.5 billion in the US alone (Reingold, 2015). The appeal process against exam boards has increasingly moved to the law courts[1]. In fact, litigation against assessors is now so common that there are solicitors who specialise in prosecuting these cases, and higher education appears to be suffering in particular[2]. The tolerance of error in assessment judgments is gradually narrowing.

---

[1] see Hill vs OCR, 2002, or the law suit brought in 2013 by an alliance of unions, students and councils against the reform of GCSE grading, among many others

[2] In the UK, the Office of the Independent Adjudicator for Higher Education saw an increase in complaints of 20% over the 2017 – 2018 period, and found in favour of the student in 20% of cases in 2018. They awarded £639,515 in compensation to students, and degree classification and marking accounted for over half of their complaints.

Exams have also been found to serve to maintain the position of the elite in society. A very specific example is the Advanced Placement test, which was introduced by Educational Testing Services (ETS) in the USA in 1956 at a flat fee of $100. It now charges per exam, up to ten exams. At the time of writing this would be a cost of $940 per child, which usually has to be met by the family. It has been argued that the test now serves to filter out those who cannot afford it from the benefits that a good education concurs (Camara *et al.*, 2000).

2.2.2 Testing and formative assessment

The increasing role of standardized tests in education has not been without controversy and problems. Exam results have a history of attracting attention. In 1988, before Third Way politics, Pearson observed that,

> 'It is generally accepted that public examinations influence the attitude, behaviour and motivation of teachers, learners and parents,' page 693, cited in (Pan, 2009).

Some hope that this influence could be positive, with higher motivation to achieve and an overall raising of standards to the benefit of everyone. With that comes many negative consequences. Risk factors for mental health issues among children and young people are complex, but a fairly comprehensive study in the UK found that academic pressure, and in particular exam stress, was the most common antecedent to suicide and was a factor in 27% of the cases they studied (Rodway *et al.*, 2016). Stobart (2018) has argued that backlash against assessment driven culture was a key driver of theories of multiple and emotional intelligences.

The assessments discussed so far are summative assessment. The distinction between formative and summative assessment is often made in terms of their intended use. The intention of formative assessment is to help learning, while the purpose of conducting summative assessment is to report on learning (Black, 2009). There is, of course, in reality,

a multitude of intended and actual uses of any assessment. A poor result in a summative assessment might spur a child on to work harder. In general, formative assessment is a cycle of events that allows the results of the assessment to be fed back to the learner and used as a basis to plan the next steps in their learning (Harlen, 2006). The games used in this study were intended for formative purposes.

2.2.3 Online assessment and new skills

The first online assessments were launched by Cambridge Assessments in 2000. Since 2016, in Nigeria, the Unified Tertiary Matriculation Examination (UTME), the main university entrance exam, is now only delivered online. Nigeria had turned to computer delivery to address the problem of exam leaks and malpractice (Sanni and Mohammad, 2015). Egypt moved their end of year exams for 600,000 15 and 16 year olds to tablet delivery in May 2019 (Al-Masry, 2019). Both countries are reported to have experienced street protests in response to technical problems delivering tests this way.

Digitalisation of exams has had some impact on the qualitative content of tests and performance. It has allowed new skills to come through. In 2015, the PISA exams were successfully delivered entirely online for the first time. In 2018, they introduced a collaborative problem solving task for the first time, where students were required to interact with two intelligent agents in a chat room to solve a task (Graesser *et al.*, 2018). In the UK, Qualifications Wales have introducing the Skills Challenge Certificate (SCC) as an element of the revised Welsh Baccalaureate Certificate, has made digital literacy compulsory for all 14-18 year olds, and a performance measure, along with a range of other skills previously excluded from UK examination systems (see the Donaldson Report on https://gov.wales for more information).

In a report on the change from paper based delivery of the PISA exams to computer based delivery, Jerrim (2016) found that a number of qualitative differences were experienced

between the two types of delivery. Some different types of questions became possible for the first time, particularly through interactive features such as drag and drop, and so it might be expected that performance in such tasks might also differ. But even when the nature of the question was essentially the same, there were still factors that impacted the results. Different cognitive processes were needed just to read on computer compared to paper. Factors such as the screen size and resolution could also affect results. In addition, the novelty of delivering tests on computer, and the need for basic computer skills changed the nature of otherwise unchanged tasks.

The argument for change in education is growing, but change in assessment needs to come first. Fullan, Langworthy and Barber pointed out that:

> 'One of the biggest systematic challenges to the spread of the new pedagogies is that they are not yet being measured in any coherent way. Unfortunately, most systems that we have seen simply do not yet have ways to measure the new pedagogies and deep learning outcomes. Assessment is of paramount importance to policymakers, leaders, teachers and parents, not just because public accountability demands it, but because all parties need to know what works in order to achieve the new aims' page 9, (Fullan, Langworthy and Barber, 2014).

The pedagogies of deep learning that they describe can be varied. The OECD describe the need to move away from knowledge based exams, and towards the assessment of skills and competencies (Ananiadou and Claro, 2009). Various frameworks for deep learning these have appeared, under different names. In Wales, they are the Essential and Employability skills (Qualifications Wales, 2019), in the USA they are the 21st Century Skills (P21, 2002). Employment job sites often use the term soft skills. What they have in common is a degree of complexity (Sellar and Lingard, 2014). Games allow a range of more complex actions to be tracked, but there are several steps between tracking behaviours and establishing a calibrated means to measure those behaviours. If measurement is taken as the process of linking abstract concepts to empirical indicators, there needs to be an

explicit plan to classify and possibly also quantify the data (Carmines, 1979). Its mere existence is not enough. Although assessment is a looser term than measurement (Newton, 2007), it nonetheless requires theoretical considerations as well as empirical ones.

2.3 Assessment concerns

Standardisation is the process either to equate scores and demonstrate a degree of comparability across different test forms, for example online and paper delivered versions of the same test (Davier and Halpin, 2013). The kinds of questions that assessors ask revolve around standardisation issues. There is no one method to answer these questions, and in the UK, for example, a trusted status system for exam boards is used instead. The USA has produced guidelines for answering these questions, and many of the approaches they recommend involve a range of statistical analysis to calibrate the test before the more mechanical process of scoring the child takes place (AERA, 1999).

Concerns around standardisation often influence the kinds of research questions assessment experts asked:

1. Are the test results reliable?

   Reliability is the first of the two hallmarks of any empirical measurement process, and of educational assessment in particular: reliability and validity. Reliability addresses the concern that the test would produce the same result if it was repeated. Reliability aims to reduce the extent of error in any measurement process (Carmines, 1979).

2. Is the test valid?

   Validity is the second of the two hallmarks, and it has been argued that without validity, there can be no such thing as an assessment process (Newton, 2007). In

practice, validity can be difficult to define, and many, many variations of the term 'validity' exist in assessment. At a fundamental level, it is the extent to which you are measuring the thing you think you are measuring, and only that thing. The subtexts behind the concept of validity may encompass a large number of different assumptions: That it is possible to assess a concept from test scores; that it is useful to make decisions on the basis of the test score; that it is socially and ethically justifiable to implement a programme of tests from the scoring process. Nonetheless, the principles of identifying the target of assessment remain a core part of any formal testing process

3. Is the tool biased?

Discrimination in assessment is often used synonymously with reliability, and means that the assessment was able to separate the strong from the weak learners. If there is a randomness to the results, the test does not discriminate. It is distinct from Differential Item Functioning (DIF), which has the meaning that the test performed differently with groups of learners depending on demographical factors, such as gender or age (Wilson, 2004, Bond, 2015). DIF is also sometimes called bias and is an undesirable quality.

The way that these questions are explored in professional assessment could be summarised as a two stage data pipeline, shown in Figure 2. Before any mechanical scoring of learners can take place, there is a more analytical calibration phase when these questions are answered using a range of methods and techniques.

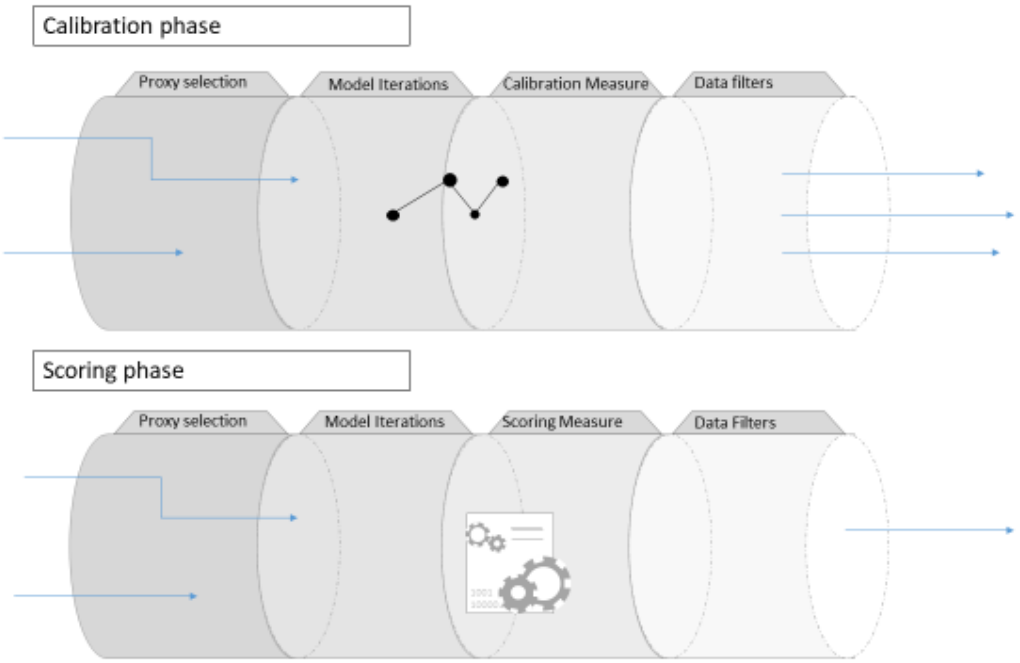


Figure 2 Two-stage data pipeline. The first is an analytical calibration phase, and values from this process are sent to the more mechanical scoring phase.

The data pipeline shown here might be applied to Games Based Assessment. It is necessary before beginning any scoring process to decide what proxy for ability will be used, accumulated actions in the game or response time (Figure 2). There needs to be a decision about how second or third attempts might be modelled. What values will be assigned to each action? When there are several attempts, there needs to be a choice of the measure to be used, the maximum, the mean, a weighted mean etc. Does every attempt count, or, for example, is there a minimum length of time the child needs to be play for it to be classified as a real attempt? These are some of the examples of theoretical considerations for games, and they will be elaborated on further in this thesis.

The data analytics used in the calibration stage have typically come from Item Response Theory (IRT). IRT emerged as a suite of applied statistical techniques to facilitate equating processes (Raykov and Marcoulides, 2011) and identify sources of non-targeted behaviour in tests. The origins of IRT can be traced back to Galton's work in 1869, but approaches emerging in the last part of the 20th century were more heavily influenced by the Rasch

model (Wilson, 2004, Bond, 2015). Psychometric measurement was emerging as a discipline, and in a series of lectures in 1960, Rasch introduced a probabilistic measurement model of phenomena in psychology and education. He wanted to address the problem that most measurements in the social sciences were much too sample dependent (Rasch, 1960). If the questions changed, the estimation of the child's ability was also likely to change. He wanted an invariant measure of ability, just as the physical sciences used, one that was not sample dependent (Bond and Fox, 2015). He proposed estimating two parameters from the performance data, a value for the difficulty of each question, and a value for the ability of each child.

Another important influence was the work in 1968, of Birnbaum, Lord and Novick (1968). They argued that the score of any person in a test was the sum of both the test-taker's true ability, and an error term (Lord, 1980). Many non-targeted behaviours such as confusing questions, guessing, cheating, or mean and lenient teachers are very common in testing scenarios. A score is therefore an estimation of the child's true ability + an error term. The Rasch method, producing both item difficulty estimates and student ability estimates created an expected behaviour: We expect that the strong students will get the hard questions and the easy questions correct, and the weak students will only manage the easy questions. We could compare this expected behaviour to the observed behaviour and use that to generate an error term (Bond, 2015). In this way, problem questions could be isolated and removed. The new statistical techniques were an important development towards fairer testing, but as Messick, (1987) another key influence on modern day testing pointed out, ultimately, a score is a social construct and performs a social function, and a wide range of qualitative considerations, as well as quantitative measurement tools need to be used.

Today, where IRT is the chosen standardisation methodology, the professional practice of

test development is heavily influenced by the principle of Evidence Centred Design (ECD) (Mislevy, 2003). In the late 90s, a team from ETS led by Mislevy drew attention to the major disadvantages to carrying out IRT statistical analysis after a test had been administered (Mislevy, Almond and Lukas, 2003). They acknowledge Birnbaum, Lord and Novick's point (1968) that some questions create more error than others, and should therefore be removed. But as long as IRT was only carried out after the test had been delivered, those questions would still be in the data set. The exam board could only adjust the evidence that they had to remove error and improve precision. If they pre-tested with a small sample of learners, and carried out IRT analysis on that sample data set, in a pre-test calibration phase, they could then adjust the questions to get complete and better evidence once the test was used in a high stakes environment. Many national and international exam boards now maintain a 'bank' or large store of pre-tested questions.

2.3.1 Psychometrics and gameplay data

To use psychometric analysis on scoring data from traditional tests, there are generally three key requirements:

1. Questions are categorically scored.

   When a question is dichotomously scored, it means that each question is scored either as correct or incorrect. If it is partially scored, responses are put into more than one grade, for example '3 out of 5'. (Wilson, 2004). A Partial Credit Scoring model allows scores that are holistic judgments, such as bands for an essay, or marks out of 10 in the Olympic High Diving competitions, to be validated quantitatively. It is possible, for example, to carry out an inter-rater reliability analysis to measure the extent to which different judges agree (Gwet, 2014). The assumption, though, is that it is a bounded categorical variable.

2. Data must be conditionally independent

For psychometrics to work well, questions should be discrete (Mislevy, 2014). In other words, getting the answer to Question 4 wrong, for example, should have no impact on the test-taker's chances of getting Question 5 correct (Bond and Fox, 2015).

3. The same set of questions are presented to all test takers.

For most scoring models, the assumption is that the students in each data set used in calibration have all seen the same questions (Bond and Fox, 2015; Mislevy *et al.*, 2012). This does not necessarily mean that they see the same questions in the scoring phase as some tests select questions more randomly from a pre-calibrated bank of items. Any form of Computerized Adaptive Testing (CAT), for example, selects questions in the live test depending on the previous response (Van der Linden and Glas, 2000). But at some point before that scoring phase, a plausible value for those questions will have been generated in a pre-testing calibration phase, so that the computer has information on the level of difficulty to be able to select an easier or more challenging next question (Graf, 2014). This pre-testing calibration still tends to be on complete data sets.

Gameplay data is different. The size of the data sets produced during gameplay tends to be much larger. Researchers have drawn an analogy between what they call an ocean of data about the learner which is available in games, in other words there is a large data trail of every single step they take in the process of reaching an answer. DiCerbo et al (2016) described conventional tests as producing a sparse desert of data, with very limited information. Early hopes for Games Based Assessment were that more data meant better evidence. This also causes problems though, as, for example, the full MangaHigh data set which was sampled for this project had over a billion data points in it at the time of extraction. Only specialist computers can process data sets of

that size.

There are other problems, too. Instead of using categorical responses to questions as the measure of ability, often a continuous variable is used as a proxy for ability instead. Response time is regularly used in commercial video games to judge success. There are also many actions where there is no obvious right or wrong choice, because students are completing tasks, not answering discrete questions. Often the data set is also full of conditional dependencies (Mislevy *et al.*, 2012). A related problem is that, because of the element of choice, not all of the children follow the same path through the game, and that creates very large amounts of missing data.

Because of these differences between assessment data and gameplay data, it is not possible to just run a psychometric analysis over the data set to understand what is happening in the game. Psychometric analysis has a set of requirements outlined just above, which game play data do not meet. A different approach is needed.

2.3.2 Assessment methodologies

There are a variety of methodologies available to assessors, and no one accepted international norm. Many tests around the world are still standardised by a form of cross moderation, in other words two or more judgements are made by qualified humans and then they are compared (Baird *et al.*, 2018). Some institutions have started to use Natural Language Processing techniques in place of human judges to score evidence in the form of free discourse. The international university entrance exam, the Pearson Test of English Academic (PTEA), or the Harvard Business School entrance essays both use this technology, and results are validated by cross moderation with human judgments. Judging the effectiveness of these technologies is hampered a little by the lack of algorithmic transparency, but these sample tests have passed validation processes.

Unlike essay scoring, which has a long history of human scoring, it is unclear how a human might judge performance in game play either, and so this approach to validation is not an option.

Many of the approaches to scoring gameplay data have approached the problem using some other form of Artificial Intelligence, as will be discussed in the Literature Review in Chapter 3, but this still leaves the question of how to validate the results.

At its most fundamental form, scoring starts with a process of identifying correct answers. This original statistical data that has not been transformed in any way is usually called a 'raw score'. Raw scores are problematic, because the data is ordinal, but not interval (Bond, 2015). It is traditional to assign a value of 'one mark' for each tick, and then sum them up. Raw score percentages look like a fair measurement, but are not. They look like the Ruler A in Figure 3 below, when in fact, they actually resemble Ruler B more closely.

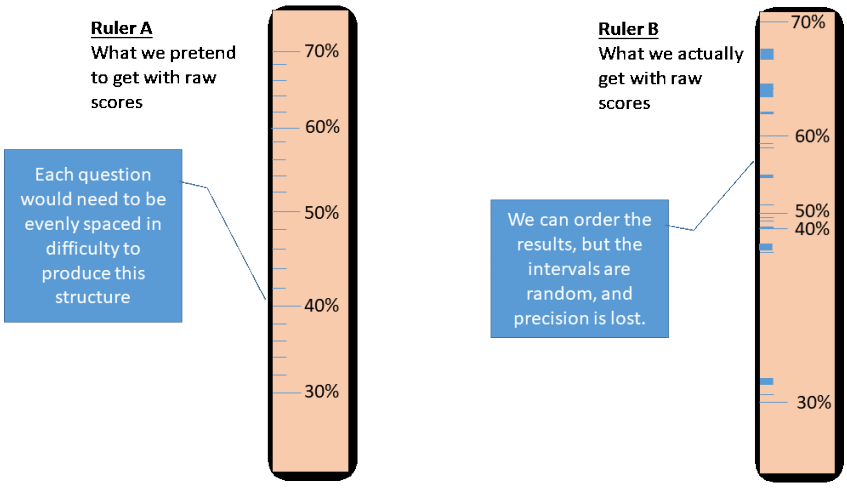


Figure 3 Measuring ordinal and interval data (ruler A) and just ordinal (ruler B)

In reality, a measure using just raw scores contains data that is ordinal, but not interval (Ruler B in Figure 3).

If there is no reason to assume that each question required exactly the same unit of conceptual understanding to get the answer correct, then an interval data approach is

needed. One method of standardizing scores so that the data is presented as interval data is to distribute the scores for each child over a normal curve (Harvill, 1991).

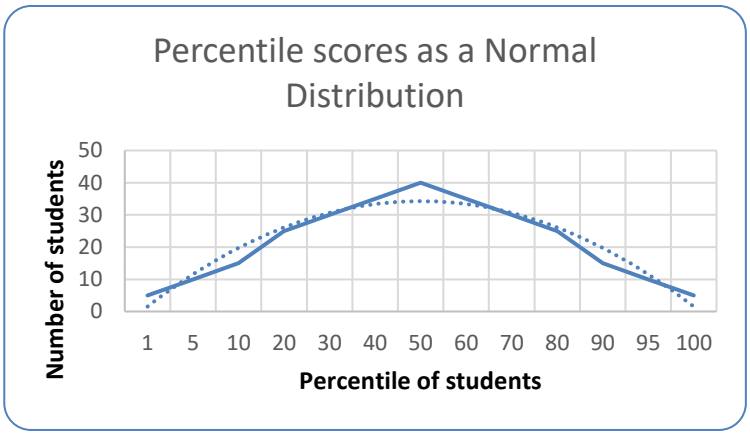


Figure 4 Gaussian or normal distribution of the scores

Figure 4 shows how children's scores can be distributed over a normal or Gaussian curve. The child with a score at the mean is located in the central point. All of the raw scores are converted into z-scores to give a more accurate indication of the size of gaps in attainment between the children (Harvill, 1991).

Rasch (1960) proposed an alternative approach that involved two parameters: The ability of the child and the difficulty of the task. To estimate the difficulty of each task, a fail to success rate is generated. For example, Task 1 was given to 100 children, and 64 of them got the task right and 36 got the task wrong. This produces a fail to success ratio for each task of 36:64. Say Task 2 was harder and had a fail to success ratio of 55:45. Rasch recommended distributing these probability values using a natural logarithmic transformation.

Task 1: Ln (36/64) = -.58

Task 2: Ln (55/45) = +.20

This has a similar effect of distributing the difficulty estimates over a normal curve, but the results are centred around a mean of zero, with negative numbers associated with easier

tasks and vice versa. A similar process is carried out with the child, but as a success to fail, not fail to success ratio. So a child who managed to get 27 out of 37 questions correct would have their ability level calculated as 27 successes to 10 fails:

Child A: Ln (27/10) = +.99

Child B: Ln (15/22) = -.38

In this way, negative ability estimates are also associated with lower ability.

Rasch (1960) developed this idea further by using the item difficulty estimate, not the raw score, to estimate the person's overall ability. The Rasch equation (Equation 1) shows how the overall ability parameter (θ) of person (j) can be calculated by looking at their performance on each item (i), and using the difficulty parameter (b) as the value to be escalated for that item.

$$P_{ij}\left(\theta_j, b_i\right) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$

Equation 1 The Rasch model to estimate the child's ability (Rasch, 1960).

In this way, the child is rewarded less for getting easy questions correct, and more for getting harder questions correct.

Like the standardised scores, Rasch's model also produces interval data, but it has the added benefit that the results are no longer item-sample dependent. If you just presented the child with 2 out the 10 questions, the difficulty estimate, and therefore the ability estimate of the child, would still be the same. You would perhaps have reservations about the representativeness of such a small estimate pool, but the difficulty estimate would be stable. By removing the problem that the scores are dependent on the sample of questions asked, the results are invariant.

There are other advantages, too. With an estimate of the ability of the child, and an

estimate of the difficulty of the tasks, there is enough data to produce an expected and observed behaviour, which in turn allows a statistical test that is very similar to a Chi Square analysis to be performed. Using the calculation above, if a child had an overall ability level of, say, +2 logits, then they should logically be able to answer a task if it had a difficulty level of, say, -.1 logits. Her ability level is higher than that of the task, and if she got it wrong, the observed behaviour would differ from our expectations. Equally, a less able child should struggle to answer a task that was above their level of ability. If there are a lot of unexpected observed behaviours, it suggests that there might be an alternative explanation, not ability. The most likely cause is poor question design. By treating the difference between the child and difficulty values as a residual, the sum of squared errors for each residual on the item and child parameters allows the error to be quantified. Rasch called this the discrimination parameter, or the power of the question to consistently separate the strong from the weak test takers.

$$P_{ni}\ (x = 1) = f(B_n - \ D_i) = \frac{e}{1 + e^{(B_n - D_i)}}$$

Equation 2 The Rasch model to estimate the discrimination parameter (Rasch, 1960)

Equation 2 shows the Rasch calculation for this, where P is the probability of n person answering i item correctly when the ability of the learner B is known, and the difficulty of the item D is also known. This probability is equal to the constant e, or natural log function, raised to the difference between a person's ability and an item's difficulty (Rasch, 1960).

A good graphical representation of these values and error terms is a Wright Map, shown in Figure 5.

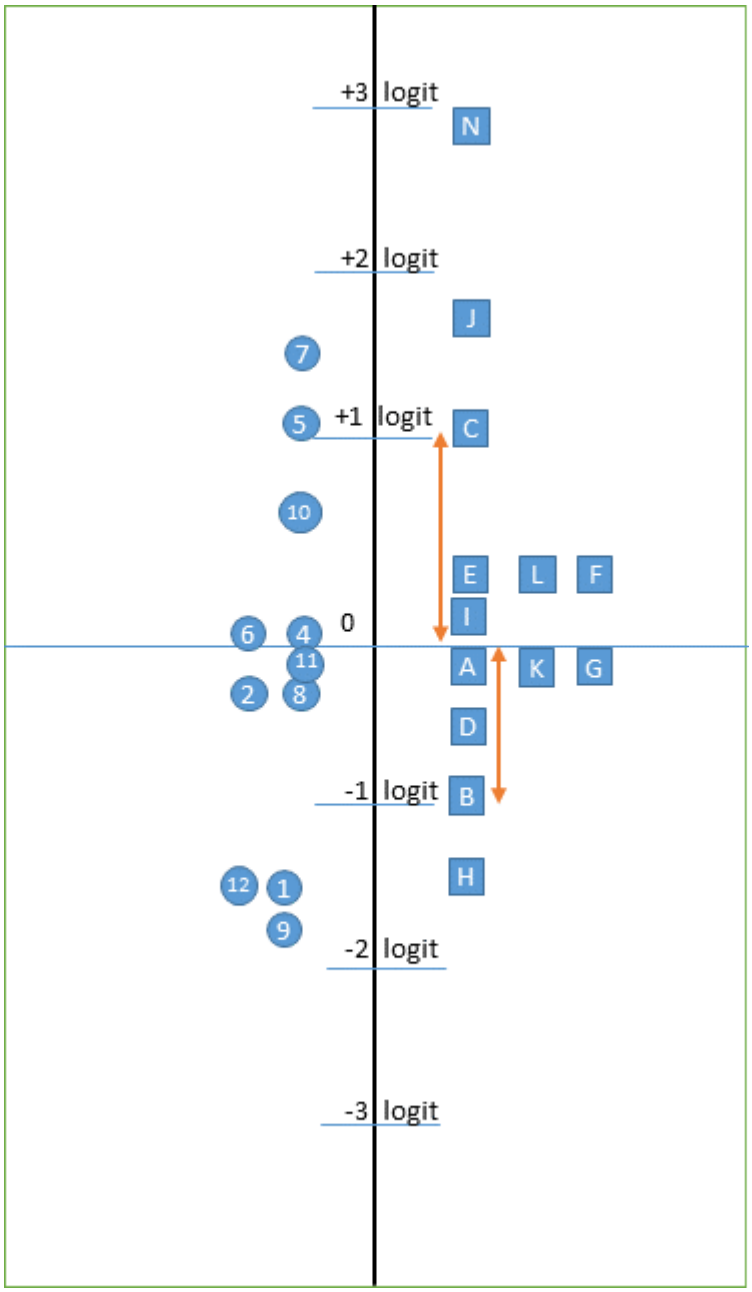


Figure 5 Wright Map, showing the values of student ability (squares) and item difficulty (circles) on the same scale (Wright and Masters, 1982)

The difficulty parameter (D) of each question is shown by the location of the circles vertically along the logit line, around a mean of zero. The ability parameter (B) of each person is shown by the location of the squares (Figure 5). If student C got question 4 wrong, and student B got question 4 correct, it would produce residuals between the expected and unexpected behaviours of -1 and +1 respectively.

Another common visual representation in assessment is the Item Characteristic Curve

(Lord, 1977). Rather than distribute the results over a bell-shaped curve, each question is represented by using a cumulative distribution function.

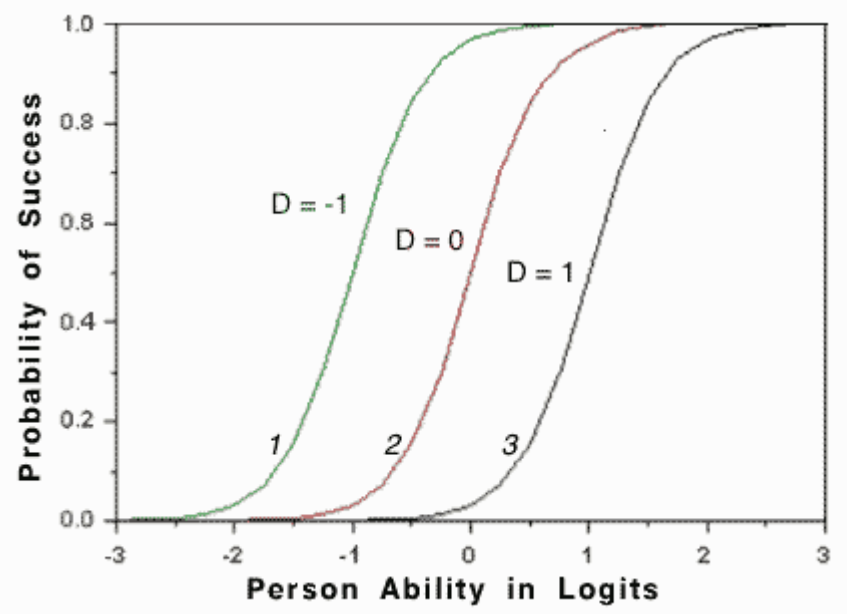


Figure 6 The Item Characteristic Curve showing different levels of difficulty

In Figure 6, the vertical axis represents the probability of getting a correct answer (bounded by 1). The b parameter, or location of the curve on the horizontal axis shows the difficulty of the task, which is located at the point where the curve crosses the .5 threshold on the y axis. In this case, the green curve 1 has a difficulty level of -1 logits, the red curve 2 crosses the 50-50 line at exactly 0, and the black curve 3 has a difficulty level of +1 logits. Only the strongest students, those with the highest estimated ability, had a high probability, close to 1, of getting the black question right. A second feature is parameter *a*, or the slope, which shows the discrimination power of the question as described by Rasch

above.

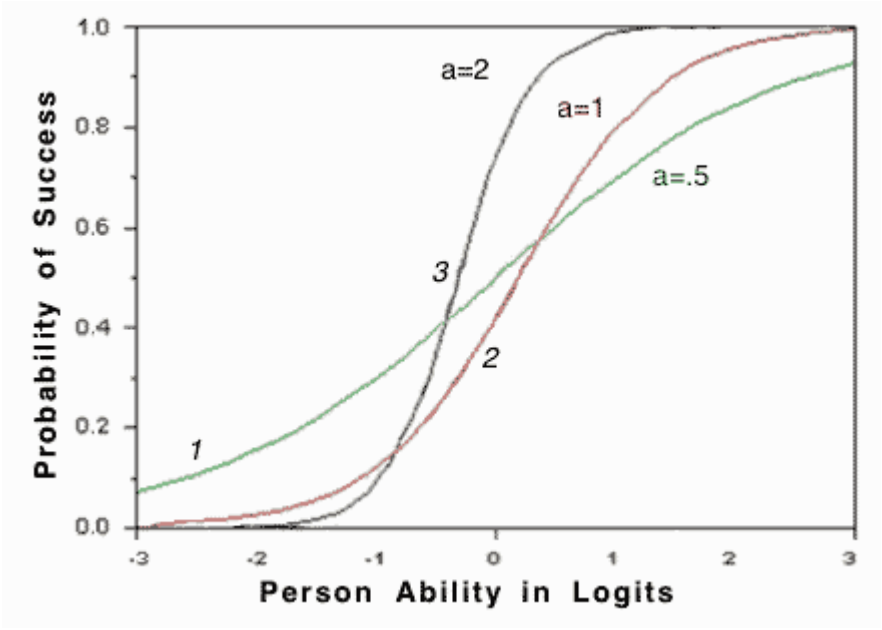


Figure 7 Different discrimination parameters in the Item Characteristic Curve

In this second image (Figure 7), the black line 3 has strong and steep discrimination power. The green line 1 is much flatter and has much weaker discrimination, in other words, there are more errors in measurement (Andrich, 1988). This is known as the 2-parameter IRT model.

The discrimination measurement is often performed twice: Once giving the same weighting to the whole data set, which produces a set of values called Outfit; and then a second time, where errors that occur around the threshold of either the child or the task are given more weight, creating a set of values called Infit (Bond and Fox, 2015). Infit and Outfit statistics can help the assessment design team to recognize kinds of error such as poor question design, a narrow coverage of the curriculum, or the fact that children seem to have run out of time. This is especially helpful when viewed in combination with score patterns.

Table 1 Lineaker's suggested interpretation of Infit and Outfit patterns (Lineacre and Wright, 1998)

Infit	Outfit	Score pattern					Interpretation
1.1	1.0						Rasch Model Ideal
1.3	0.9						Specialised knowledge / poor syllabus coverage
1.0	3.8	0	1	1	0	1	Guessing / poor question design
		0	0	1	1	1	Carelessness / poor wording
		1	1	1	0	0	Running out of time
2.3	4.0						Miscode

For example, it is very rare to have no error, and a Rasch score of around 1 in both values is expected (Table 1). When the infit is higher than the outfit, it usually suggests that either the child has specialist pockets of knowledge, or that the test did not cover the whole syllabus. When outfit is higher, there is a random guessing parameter, and looking at which questions students got wrong can help the diagnosis. If they are towards the beginning, it is probably cruising through the early questions or not paying attention. If they are at the end, the child probably ran out of time. If they are randomly spaced, it would appear to be a problem with guessing. This is a guide only, and many substantive considerations need to be taken into account (Messick, 1987). There is no optimal value for for infit and outfit, as it can vary depending on the type of task and social role of the test, among other things.

Considerations around sample size do affect confidence in the accuracy of the estimates. Small pools of questions, and small numbers of children will also affect the precision of the measure and so a reliability index is also calculated (Bond and Fox, 2015). It estimates the fraction of the observed response variance that can be reproduced if the analysis were to be repeated (Andrich, 1988). As Lord (1980) posited, a score is a combination of the child's true ability and an error term. The reliability index can be used to suggest a range of

marks within which we can feel confident that a learner's true score lies. This is often expressed as a Standard Error of Measurement (SEM θ) value, which is a form of confidence interval. Often this is also broken down at the different threshold levels, or grade boundaries for a test and called the Conditional Error of Measurement (CEM θ) (Harvill, 1991). This is because the distribution of the data can often be unbalanced, with a large amount of information and children around the mean, and sparser, and therefore less accurate data, at the upper and lower boundaries of a data set (Wilson, 2004).

These are the basic building blocks and principles of IRT, and Chapter Five will look more specifically at additional work within the field that has informed the methods in this study.

2.4 The emergence of educational online video games

Online gaming has grown from nothing in the early 1980s to an industry estimated to be worth over US$1billion worldwide annually, and it is going up all the time. Video games have their principles in ancient games design (Fullerton, 2014), but more recently they have been heavily influenced by aesthetic and technical considerations. Newer genres of video games have emerged to take advantage of innovations in the games mechanics - the way the player moves around and interacts with objects in the gaming environment (Adams and Rollings, 2010). Games designers mostly use analytics to refine the quality of the game play experience to make them more engaging and perhaps to optimise the opportunities to market paid-for services during game play (El-Nasr, Drachen and Canossa, 2016).

Simulation games have a long history of use in training in the military, going back to the 1920s (Mead, 2013). Computer-enhanced versions have gradually been added.

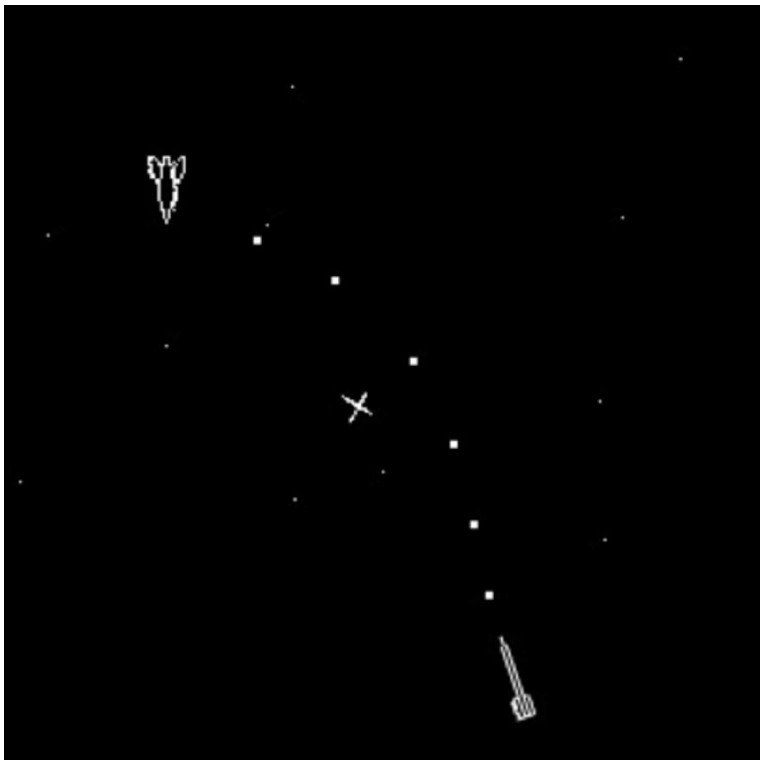


Figure 8 Screenshot of *SpaceWar!* ©MIT

Figure 8 shows a screenshot of arguably the first ever video game. *Spacewar!* was developed in 1962 with military finance from the Pentagon. The US military has used video games at almost every organizational level to raise standards. More recent games such as *Generation Kill* or *DARWARS,* closely reflect the complexity of surveillance monitoring, challenges in cross-cultural communication and the convoy operations that are used in modern warfare (Smith, 2010).

Military use of games may be significant. Corey Mead, a professor of English, has argued that the military has played an important role in innovation in education, and has been ahead of the mainstream educational curve in now-established practices such as large scale standardized testing, distance learning and vocational training (2013). Their use of games for teaching may also follow this trend.

Gaming is an interactive space that requires active participation, and that is where many observers see the potential for educational opportunities in gameplay environments (Gee, 2005). Games feature structures such has an overall objective to work towards; the

scaffolding of complex tasks; and the introduction of hurdles or tests along the way (Fullerton, 2014). The potential to make use of these features in education was convincing enough for the Federation of American Scientists (FAS, 2006) to issue an influential report encouraging private and state sponsors to develop applications of educational games. One attractive feature of games as learning tools is the set of skills usually required to succeed in games. Complex systems management, complex problem solving and collaborative work, however, are already a basic part of games design (Fullerton, 2014). Game play data may provide the evidence we need to score those skills efficiently (Eseryel, Ifenthaler and Ge, 2011). Early work on using games was to assess soft skills included Cisco's network fault detector for procedural problem solving (Frezzo, 2009), systems management (DiCerbo, 2016) and a range of other skills that will be discussed in full detail in the Literature Review Chapter in Unit 3.

Some simple app games, such as a simple Flappy Bird style game, have very basic coding and can be created in an afternoon. In general, though, the process requires many hours of attention from skilled programmers and there are many limitations to building games[3]. Any online game will have different programmes to establish the rules of play, the graphics, a renderer to display it on screen, music, artwork and instructions around what the input commands can and cannot do (Fullerton, 2014). Even a doorway or a rock that is displayed on screen may need a large vertex of data to show it from different angles, or to establish how another character interacts with it, such as walking through or hitting (Adams and Rollings, 2010). The data collection and storage process is just one of many layers in this design (Augustin *et al.*, 2011).

For any research looking at Games Based Assessment, the game itself is the first challenge.

[3] For a broader discussion of this see Kelleher and Pousch 2007, Buckingham Burn 2007

Four approaches to using games for serious educational purposes have been taken:

1. Use an existing game

The 2008 simulation game, *Spore*, was aimed at the entertainment market, but was developed in close consultation with evolutionary scientists. At one time it was suggested that the game could assess understanding of evolution, as the detail in the game, such as the terminology and cause and effect reactions were accurate. However, the game was within the 'shoot 'em up' genre, where a blast gun triggered reactions, perpetuating the erroneous belief that humans can exercise volition in evolutionary processes (Reese and Tabachnick, 2010). Another example might be Microsoft's Minecraft: Education Edition, which uses very similar games interface and mechanics to the original game (Kuhn, 2017).

2. Adapt existing games

Some big gaming companies have allowed their games to be adapted to suit a learning and assessment environment. PopCap Games allowed an adaptation of *Plants Vs. Zombies* to assess problem-solving (Shute *et al.*, 2016). Assassin's Creed: Origins is an educational version of the game aimed at encouraging historical exploration, although like Spore, it has also been accused of perpetuating erroneous models of history (Bondioli, 2019).

3. Design a new game

The most time-consuming approach is to develop a game specifically for assessment purposes. The *Virtual Performance Assessment Project* at the Harvard Graduate School of Education developed a virtual environment, with a *Choose your own Adventure* style narrative to explore students' understanding of science enquiry (Ferzli, Pigford and Black, 2015).

4. Use the game as a starting point

When it became obvious that the problems with *Spore* meant that the in-game scores were not necessarily good evidence that evolution students had understood their topic, it was suggested that rather than use the in-game scores, students could demonstrate their knowledge through a critique of the game. This did, however, just produce a traditional essay response, not the complex skills that gaming is meant to facilitate (Reese *et al.*, 2012). Overmars Games Maker is a variation on this, in that it allowed people to create their own games using drag and drop functionality, which is so easy it is used to teach computational thinking (Overmars, 2004).

2.4.1 Games design concerns.

The game environment will be designed to respond to the concerns of games designers and these are not necessarily aligned to the concerns of assessors as laid out in the first half of this chapter.

Games design tends to follow the rules of object-oriented programming, starting with clearly defining the nouns, such as ‘an input device’ (Adams and Rollings, 2010). The programmer then looks at the sub-classes of that noun, such as ‘the key board’ or ‘the back arrow’, and states the set of rules about what each subclass can do and how it interacts with similar classes of objects, such as the shift key.

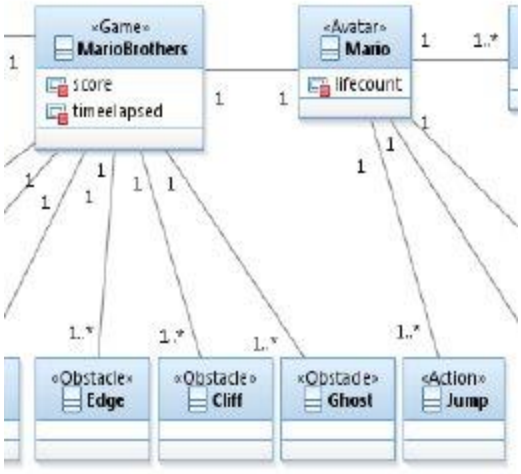


Figure 9 Fragment of the domain model from Mario Brothers ©Muhammad Uzair Khan

This model influences how performance data is then sent back to a scoring model. Figure 9 shows a small fragment from platform game, Mario Brothers. There are cliffs to fall off of, ghosts to run into, and all of these actions will have implications for the next stage of the graphics, the overall score between all of the players, and an interim 'lifecount' value specific to the Mario character, among other consequences, such as the next steps of the game.

After the domain model, the designer will start to define the user model – which are the goals, preferences, and knowledge of the person playing the game. Programmers often prefer to make adaptations to games to personalise the gaming experience (Adams and Rollings, 2010). In constructing the user model, games designers will often be challenged to overcome issues of limited data storage capabilities, and will need to make decisions over which objects, such as a coin collection, outfits or special abilities within the game could be stored and carried over to the next session (Koster, 2005).

When designing scoring mechanisms, games designers may be more interested in whether the game was fun and people kept playing, how scores can be tracked and stored from one play session to the next, how much weight early results should have, compared to more recent results, or how to make things work technically (El-Nasr, 2016).

Despite the very technical nature of games design, keeping the game enjoyable is a central concern. Games design has been heavily influenced by Cscikzentmihalyi's (1975) theory of flow and engagement in tasks. Flow is an optimal moment where the level of challenge in any complex ability is just above the level of ability of the person attempting the task (Csikszentmihalyi, 1997). This is very similar to Vigotsky's (1987) Zone of Proximal Development in the field of education, which suggests that a child working at a level just above their current ability will be fully engaged. In order to keep players engaged, games analysts tend to aim to optimise that state of flow (El-Nasr, Drachen and Canossa, 2016).

It has taken an unprecedented level of international co-operation and compliance across linguistic, commercial, legal and political borders to make it possible to make the web work in very different countries around the world. For the web as a whole, it is largely the outcome of the ongoing work of the World Wide Web Consortium (W3C), and this collaboration was by no means an inevitable step (Tinati *et al.*, 2014). To allow continuity of service, it is essential everyone in the world agrees to do things the same way, and so games designers will always be constrained by rather rigid standards, often referred to as protocol. In education, the Institute of Electrical and Electronics Engineers (IEEE) are a body that oversees international standards for online learning technology. The flow of performance data between different systems architecture is laid out in their agreed structures and standards, called 'Learning Design', along with the code that makes learning technologies possible. Figure 10 shows a small overview image of one such learning design protocol.

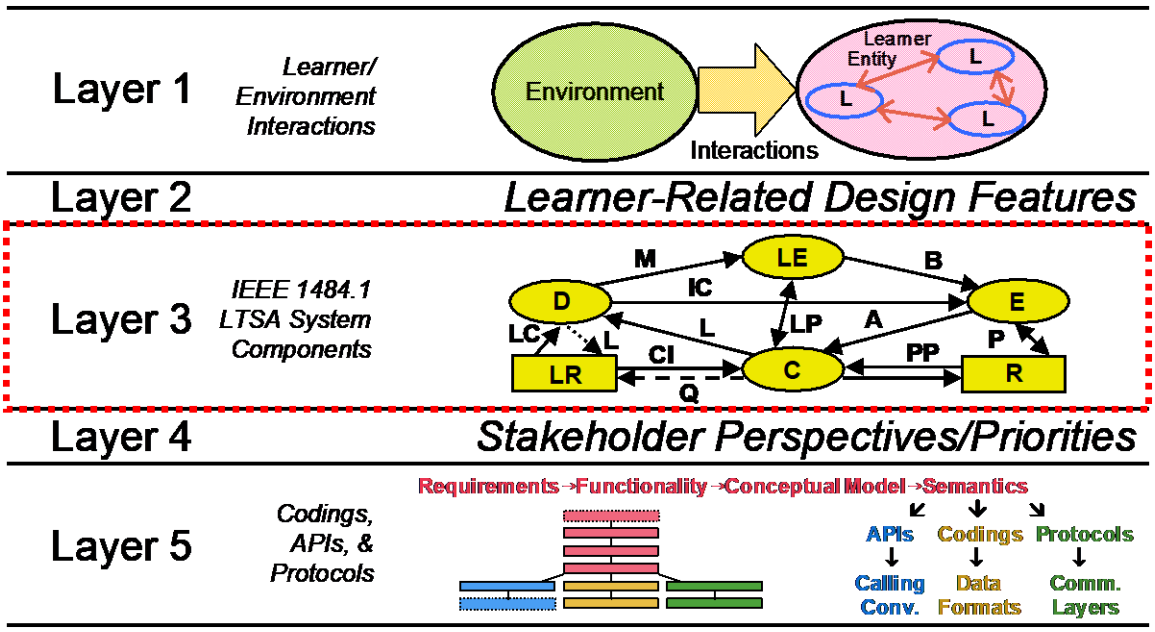


Figure 10 The layers of interoperability in scoring systems in learning technology, IEEE Standard #1484.1-2003 Learning Technology Systems Architecture (LTSA) ©IEEE

As Figure 10 shows, a large amount of consideration has already gone into the flow of data from the stored information to the user in the Learning Technology Systems Architecture (LTSA). Layer 1 would be the 'domain', or the rules and definitions of what can and cannot be done. Layers 2 and 4 are the human features, or the things that the assessor may want to happen during the gameplay. A character will move, or a child will input a response. Layer 5 covers the types of multimedia delivery. Layer 3, highlighted above, is of particular interest because this is where the mechanical scoring of games will take place. The first thing to note is that this is the most normative layer in this process, and there is little space for flexibility.

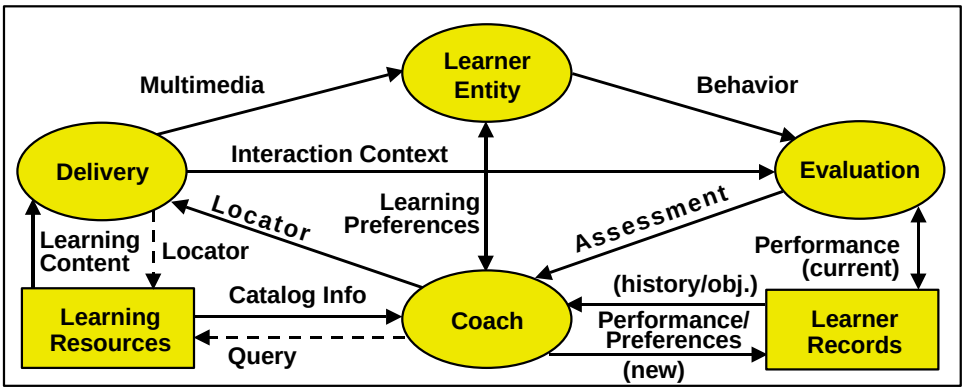


Figure 11 The data gathering layer, or LTSA system components layer, IEEE Standard #1484.1-2003 Learning Technology Systems Architecture (LTSA) ©IEEE

How the learner or *Learning Entity* behaves, The *Evaluation* of this, such as right or wrong, and any *Coach* interventions such as clues or hints to keep the player motivated are constantly updating and being added to the *Learner Records* for each individual leaner (Figure 11). It also determines which *Learning Resource* or teaching materials will be selected for the *Delivery* of the next scheduled steps in the game. Too much complexity in the volume of data flowing between these components can cause computational issues, and some researchers working with GBA have cited problems with complex models (Frezzo *et al.*, 2009).

Another model quite often cited by those specifically working in GBA is Evidence Centred games Design (ECgD) (Mislevy *et al.*, 2012). As discussed earlier in this chapter, Evidence Centred Design encouraged assessment designers to use IRT a priori of any high stakes assessment, to refine question design and improve the quality of the evidence gathering process (Mislevy, Almond and Lukas, 2003). Evidence Centred Games Design is an extension of the model for Evidence Centred Design.

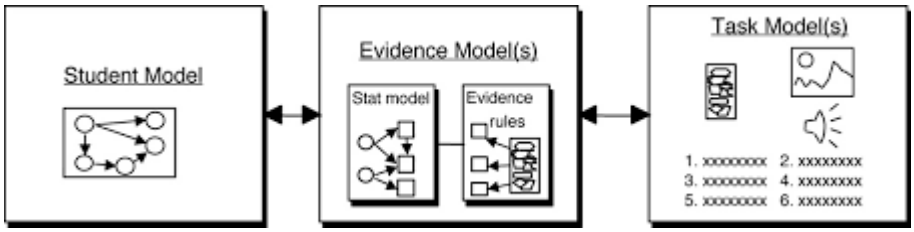


Figure 12 Evidence Centered Design page 2 (Mislevy, Almond and Lukas, 2003)

The image above, used to illustrate ECD, emphasizes that there must be an analytics layer in any assessment process to produce evidence that the link between student ability and their performance in the task was justified (Figure 12). Mislevy et al (2015) laid out the same three layers of evidence gathering within gameplay environments with a set of four criteria as to what should happen in that middle evidentiary layer, summarised below:

1. Competencies from a non-gaming environment must be identified a priori to game assessment design.

2. The designer must have a strategy for mapping those externally defined competencies to game play competencies.

3. Systems should integrate formative feedback.

4. There should be a means to link the fun and engaging nature of game environments, which encourage iteration, with an iterative assessment model which allows meaning and accuracy of measurement to emerge.

(Mislevy *et al.*, 2015)

ECgD is more a list of principles of design than a recommendation to use psychometric calibration in a pre-testing phase. The diagram shown in Figure 12 is very general, and to a games designer it may look like a summarised version of the LTSA layer laid out in the IEEE standards, which deals with the technical considerations, not the quality of the tasks in evidencing ability.

2.4.2 Gaming constraints on the assessment environment

The next steps in games may be determined by the most recent action, and so scoring often needs to take place in real time, and be constantly updated. This may not be the case with simulations, but if the hope of GBA is to allow more complexity to be assessed, it is an important consideration. To carry scores over, Bayesian probability approaches are more suitable than Frequentist statistics as Bayesian approaches allow evidence for scores to accumulate (Almond *et al.*, 2015).

Like Frequentists approaches, Bayesian statistical inference assumes that there is a parameter, θ, that we wish to estimate, such as the ability of the learner in a particular cognitive skill or the difficulty of a task. At a theoretical level, for Bayesian learning approaches, θ is seen as a random quantity, not a constant, with the effect that they assume that our opinion on the prior θ can change. Bayesian approaches also recognise that researchers rarely come to a topic knowing nothing (Leonard and Hsu, 2001; Gelman *et al.*, 2013). Biologists know, for example, that observed data from pregnancy tests carried out on men is unlikely to increase the probability estimate of any individual man being pregnant from 0%.

The personalised experience that is valued in games (Fullerton, 2014), means that games designers intentionally retain data on players so they can deliver more targeted content (El-Nasr, Drachen and Canossa, 2016). After running game data analytics during the last session, an opinion on the child's value θ of ability exists before any new data from this particular gameplay begins, unless the player is playing for the first time. Bayesian maths therefore differs from frequentist maths in that it allows a prior estimation or value for θ to be entered into the equation as P (θ) and then new data are added P (data), to estimate the ability from a combination of the old and new data available P (θ|data).

In some game types, it is essential to carry data from one gameplay session through to the next, for example, a city management game. A version of Bayesian analysis is the mathematical model used in many of the studies on GBA that look at in-game scoring of learner ability (Mislevy *et al.*, 2012; Mislevy *et al.*, 2014; DiCerbo, 2014; Almond, 2015). Bayes is compatible with IRT principles and approaches (Almond *et al.*, 2015), and there may be additional benefits. Although at the time of writing there have been no published papers, Stata have offered code to introduce a 3 parameter logistic (3PL) Bayesian model of the ICC which allows different guessing parameter for each question, a 4PL model that that frees up the asymptote at the top end of the curve and a 5PL model that allows asymmetry in the ICC[4].

While there has been some progress on refining the mathematical scoring, this study focuses on an aspect that has received very little attention in the published literature, which is the quality of the data going into the model in the first place. There may be opportunities to be more selective of the evidence used, and that can also refine the quality of the scoring process. For example, one major challenge with Bayesian approaches at the moment is that many of the software packages that make Bayesian estimation easier to use, such as OpenBUGS or WinBUGS, dictate how missing data will be handled. In the case of the BUGS software packages, missing values have to be imputed (Lunn *et al.*, 2012). This is common in Machine Learning analysis approaches, where Not a Number (NaN) values in an array are not tolerated and either a value of 0 or another value needs to be imputed. As will be developed in Chapter 4, there are reasons in assessment to be wary of randomly assigning missing values a value of 0.

Another possible problem with Bayesian approaches may be that the line between learning

[4] See https://www.stata.com/stata-news/news31-1/bayesian-irt/ for the code and discussion

and performance data becomes very unclear. In education, Soderstrom and Bjork (2015) made a distinction between learning, a resilient and flexible ability to apply skills and knowledge after instruction has taken place; and performance, the behaviours that we observe and measure while instruction or training is taking place. Soderstrom and Bjork concluded that the two do not always follow a predictable or sequential path, and that performance could be a fleeting and highly imperfect index of learning. Tracking everything may not be desirable, and games are environments that have always attracted a degree of subversive behaviours. A curious student might, for example, bulldoze down a city they are supposed to be carefully managing, just to see what happens.

These reservations around Bayesian approaches will, at some point, need to be balanced with the many affordances it offers. The discrete, unidimensional data that are a requirement of frequentist-based IRT approaches are rare in games that require complex skills (Almond *et al.*, 2015; Almond, 2015), and so as well as allowing scores to be carried from one game to another, the prior also allows conditionally dependent data to be incorporated more readily. Bayes has become popular in the wider field of web analytics precisely because it is more suited to the hypertext environment of the Internet. The term ‘hypertext’ refers to non-linear ways of navigating through information by clicking on links (Nelson, 1982). The route the player takes through the game may be constrained or influenced by the earlier paths that they have taken to get to that point, and this is a challenge for assessment designers.

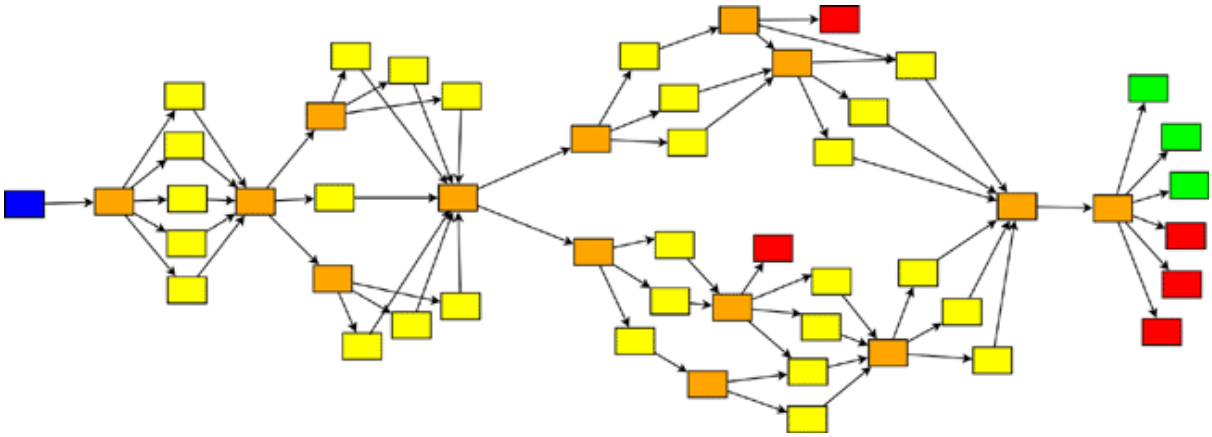


Figure 13 Narrative structure of choice in games © https://heterogenoustasks.wordpress.com

Figure 13 shows an example of a pathway through a complex game narrative, where earlier choices limit the possible next steps, but the players are brought back to central points in the process. Not all games have this kind of structure, but where it is used, it can be problematic for assessors. Games contain complexity, which is not easy to incorporate into assessment.

As well as a 'score' for favourable choices or actions, there may be response times, or items collected. There may health records, or additional powers acquired (Fullerton, 2014). Games also generate a large amount of incidental data, or paradata, recorded during gameplay (DiCerbo, Mislevy and Behrens, 2016; Mislevy *et al.*, 2012).This may include things like time stamps, or information on competitors, which may also be added as a conditional dependency.

Many traditional tests aim to line students up in order from the weakest to strongest. Games, on the other hand, tend to take a multi-dimensional approach. One player might be more accurate at completing mathematical sums, but another is faster. The term 'Cognitive Diagnostic Assessment' (CDA) is used to describe how multiple dimensions to any one complex skill, are reported back. In games, it requires a table, or array, called a Q-Matrix, to tally up dichotomously scored points that are believed to represent specific skills. These

skills are expressed as variables, and the degree of proficiency in each variable is estimated and reported to stake holders, rather than one overall score or grade band (Almond *et al.*, 2015). It has been used in wider assessment to give students are more comprehensive understanding of strengths and weaknesses (Leighton and Chu, 2016). Many of the texts cited in the literature review in Chapter Three created domain models which take a multi-dimensional approach to scoring.

2.5 Summary of the interdisciplinary challenges

This chapter has shown that both education, assessment, and games designers have already engaged with questions around how to score performance, and what those performances might tell us. Assessment and games analysts have already developed their own set of guiding principles and methods to improve on scoring processes. Despite the fact that scoring ability is central to the work of professionals from both fields, there are perhaps not as many parallels between them as might be hoped. They both bring very unique specializations and expertise to the table, though, and so it is an environment ripe for innovation when they come together. Assessments are a controversial issue and an important aspect of modern public life, and it is important that the views of all three professions are considered. The next chapter will look at the work specifically in the field of GBA to date, in the light of the questions and concerns highlighted in this chapter.

Chapter 3 Systematic Literature Review on Games Based Assessment

After considering what an educational researcher, an assessment researcher and a games designer might consider to be the hurdles to effective scoring in video games, a more refined idea of the challenges in this field emerged, and there were some questions to be answered. These were used to direct a systematic literature review on the state of research in the more specific field of Games Based Assessment (GBA). The review was carried out from February to March 2017, with the aim of answering four questions around GBA. It is a relatively new field of research and so there were few published texts that directly addressed the challenges of scoring in this new and complex environment. A number of guidelines were consulted to identify the best approach. The first choice of approach, a quantitative meta-analysis of performance statistics proved impossible, from a pilot study of the literature. This chapter will cover the methods chosen to carry out the review, following that initial scoping and pilot of the available literature, and discuss the results and findings of the systematic review of the literature on GBA.

3.1 Pilot study

The original intention for the literature review was to carry out a quantitative meta-analysis of the reliability statistics reported in the published literature, to see which approaches or which game mechanics have produced more stable results in the scoring process. In reality, there were far fewer published texts directly relevant to any of the questions that I initially expected to find and statistical output on the effectiveness of scoring systems was varied and limited. The majority of the articles were methods articles, and these offered some insights into the analysis process. They were clearly not, however, suited to a quantitative comparative approach, and published results from scoring systems were still very rare at

the time of carrying out this review.

In the pilot study, after an initial scoping of 15 potential sources chosen by convenience sampling, there were few commonalities between the papers. There were also very few statistical outputs reported. In particular, SEMθ values or any reliability measures were absent from all 15 sources from the pilot. Of those initial sources, 2 reported statistically significant correlations that had been found in the data sets they were working with. On inspection, the reason for the lack of assessment style outputs became clear after looking at the journals and conference papers where those articles had been published. They had all come from publications in the field of computer science, not education.

3.2 Literature review methodology

A different methodological approach to the literature review was needed. Revisiting the guidance, a qualitative snowballing approach to the literature search was rejected early, as research suggests that this does not suit research areas where there is a large degree of specialisation (Baumeister, 2013), and the texts in the pilot mostly used correlation analysis or AI techniques. Although this study positions scoring games as an interdisciplinary study, where some knowledge of education, assessment and games design is essential (Repko, 2008), the texts in the pilot seemed trans-disciplinary, with authors very heavily influenced by common texts in their own field. The research on literature review processes also suggests that such topics are unsuited to snowballing techniques (Baumeister, 2013; Finfgeld-Connett and Johnson, 2013; Baumeister and Leary, 1997)

Instead, a semi-exhaustive narrative literature review was selected. A number of targeted data bases were searched until no more relevant sources were found for 5 consecutive search pages. Research suggested that this approach was more suited to trans-disciplinary study (Levy and Ellis, 2006).

A narrative approach to reporting the findings was selected. As the pilot found few commonalities in the articles on GBA, a narrative approach suited the topic of research (Baumeister and Leary, 1997). In addition, Nind (2006) reported on the many affordances of carrying out a literature review on a subject that had attracted a broad range of researchers, and that a narrative approach better did justice to the many views represented. Although a relatively small number of texts appeared in the search for key words around GBA, it appears that no one method has yet emerged, and the research covers a fairly broad scope.

The reporting protocol that appears in Appendix A was drawn up following guidelines from Okoli and Schabram (Okoli and Schabram, 2010) and Riesenberg and Justice (Riesenberg and Justice, 2014). The guidelines from the University of London's Evidence for Policy and Practice Information (EPPI) Co-ordinating Centre website was also consulted in the reporting of data collection, keyword identification, mapping and results (EPPI, accessed January 2017). The literature was tagged in NVivo with the terms laid out in Appendix A. Given the small number of texts that have been published in this field (the final literature set was n=41) all available texts were manually inspected.

3.2.1 Literature review questions

From the review of the three fields and their approach to gaming, I had the following four key questions when looking at the literature. From the previous literature:

1. What is technically possible for games mechanics and games analysis?
2. How is domain knowledge or skill being modelled?
3. What scoring models have been used?
4. What work has been done to validate the current findings?

3.2.2 Pre-existing literature reviews

There was no desire to repeat the work of others, and in searching the literature, five of the texts were literature reviews, and so these were looked at first. On reading, they appeared to be tangentially relevant to the current study, and in particular, a sense of what has been tried and found to fail or succeed was absent from all of the analyses, with all five reviews concluding that this is an area that deserves attention but is still in its infancy.

1. The National Research Council (National Research Council . Committee on Science Learning: Computer Games, Honey and Hilton, 2011): Learning science through computer games and simulations

The National Research Council at looked games as a means of increasing learner engagement. They found that there was sufficient evidence to justify further research in games-based learning. They did not make reference to research carried out on assessment in games but argued that gaming data appeared to have beneficial uses. They concluded that more research was needed.

2. Eseryel et al, (2011): Alternative assessment strategies for complex problem solving in game-based learning environments

This literature looked more specifically at the issue of validating assessment of tasks that required several steps to complete and concluded that there was not sufficient evidence in the literature at that time of how to assess this. The review was limited to the assessment of complex problem-solving in games.

3. Hainey et al. (2012): Assessment integration in games-based learning: a preliminary review of the literature

Hainey et al (2012) carried out a review of the literature for the purposes of informing policy, but they too largely concluded that the field was in its infancy and more work was needed.

4. Ifenthaler et al (2014): Challenges for education in a connected world: Digital Learning, Data Rich Environments, and Computer-Based Assessment

   In a special issue of Technology Knowledge and Learning, a review of major trends and challenges in digital learning was carried out. They concluded that more work was needed to develop games-based assessments, as part of a range of other research developments proposed.

5. deKlerk et al (2015) Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example.

The report on a project to score games began with a brief systematic review of the literature. The team concluded the only psychometric approaches that had proven successful with gaming data were Bayesian-based approaches. Only one alternative to Bayes was found in the literature, which was pre-test, post-test analysis, using data gathered externally to the game.

The five reviews concurred that more research in this field is needed, and that no clear understanding of how to score games had yet emerged. As the studies were a few years old and inconclusive, it justified looking again to see if progress had more recently been made, and the findings are reported below.

Since carrying out the initial review in 2017, there has been development in terms of measuring the effectiveness of games in promoting learning. A meta-analysis more recently published found mixed results statistically, with a number of factors contributing to success (Sailer, 2019), but greater clarity on what kinds of games are more effective are

beginning to appear. No published literature reviews specifically on scoring have been found.

3.2.3 Text selection

The search of the existing literature was carried out between 22/2/2017 and 6/3/2017 on two main data bases, Web of Science (https://clarivate.com) and ERIC (https://www.eric.org.uk). In addition, further texts were sought in the grey literature on Google Scholar, opendoar.org and opengrey.eu.

A full list of the inclusion and exclusion criteria can be found in Appendix A1. The EPPI guidelines for funded literature reviews were consulted because considerable research on quality control has gone into refining the standardised approaches they recommend. However, full adherence to EPPI procedures was not adopted as the studies did not met their requirements for 'quality research', which needed large sample sizes among other things (EPPI, accessed January 2017). Although valuable as an insight into procedural rigour, it would have been impossible to extract any value from the body of evidence on GBA following their guidelines for inclusion rigidly. Petticrew and Roberts (2006) created inclusion criteria more specific to a narrative review approach, and with more lenient inclusion criteria, and their guidelines were also consulted. The inclusion and exclusion criteria can be found in Appendix A2 and A3.

The pilot study indicated that requests using the term 'validation' produced very few relevant sources even in the first few pages, and so, although central to this study, the term was not used. Instead, the key words in Table 2 were used in the search. Search engines in the two main databases returned over 50,000 entries, but many of these were unrelated to the field of interest.

Table 2 Word combinations used in the search

Words used in the text search	
GAMES + BASED + ASSESSMENT	Also GAMES + ASSESSMENT
GAMES + BASED + TESTING	Also GAMES + TEST

During the pilot stage of the literature review, refinement by adding the criteria '+ Education' did not appear to introduce any new sources that met the selection criteria. In fact, some key studies were no longer appearing, and so the words in Table 2 were used.

In both databases, the search continued until no new relevant cases were found over four consecutive pages. The main inclusion questions at the title level were:

1. Does is it describe a game?
2. Does it measure a cognitive skill?
3. Does it have a scoring structure that is reported?

The first filtering on the basis of the title of the article looked at whether the study reported both games and outcome of assessment of a cognitive ability. Around a third of the texts scoped in the first few pages came from the field of clinical assessment within medicine, and these were rejected. Around another third referred to optimal coding practices to build a game, and these were also rejected. Broad interpretations of terms like 'game', 'cognitive skill' and 'scoring structure' were used at this stage to keep as many texts as possible in the review. This resulted in an initial selection of 202 texts for analysis at the level of the abstract.

Table 3 Number of texts found in different data bases after the initial filtering at the title level

INITIAL TEXT SELECTION

DATABASE	Number of texts	Date searched	Terms used
WEB OF SCIENCE	64	22/2/2016	Games + Based + Assessment
	35	06/03/2016	Games + Testing
ERIC	25	03/03/2016	Games + Based + Assessment
	7	03/03/2016	Games + Testing
GOOGLE SCHOLAR	43	23/02/2016	Games + Based + Assessment
	12	23/02/2016	Games + Testing
OPENDOAR	14	03/03/2016	Games + Based + Assessment
	1	03/03/2016	Games + Testing
OPENGREY	1	03/03/2016	Games + Based Assessment

Table 3 shows where these texts were found. Duplicate texts were removed in referencing software Endnote. ERIC produced considerably fewer texts than Web of Science. Some of the texts sourced from Google Scholar did appear in peer reviewed publications or books which had been peer-edited, but did not appear in the key word search from the two main databases. The grey literature was included because it has been argued that research without definite results might be unappealing to journals in terms of publication, but they could still produce valuable insight into what does not work (Begg, Cooper and Hedges, 1994). Cumming (2013a) has even argued that this publication bias has been found to distort the impression of the effectiveness of approaches. The pilot and the pre-existing literature reviews on GBA broadly concurred that GBA is challenging and successful

results papers are still scarce, and therefore might be more susceptible to publication bias than other areas where there are established methods and results, and so unpublished literature was also consulted where available.

After an initial scoping of the abstracts, the following 2nd stage criteria were applied.

1. The text must have some kind of skill or knowledge or educational outcome that is cognitively-based.

   For example, measures of motivation to learn and nothing else were excluded. At times, a broad interpretation of this was taken. For example, the use of magic spells as evidence of creative problem-solving in the game *Elder Scrolls* (Shute and Ke, 2012) was accepted on the grounds that it attempted to say something about problem-solving.

2. The text must have some kind of assessment element as the focus of the research.

   Texts which validated findings within the game by using external measurement tools and scales, for example, results from standardised tests completed in a classroom setting were still included.

3. There were no time limits on the date of publication.

   The earliest text found was dated 2005, and there have been no major developments since that date that would invalidate previous findings.

4. The definition of 'gaming features' were considered to be delivery of the assessment that involved aspects of competition or collaboration, goals, rules of play, and resources to help them solve conflict, adapted from Fullerton's (2014) recognisable features of games.

Given the limited information on this topic, all sources from the published literature and

the grey literature were reviewed, but with more attention given to those texts that have been peer reviewed. Table 4 gives the codes used, and the full list of texts and their coding of texts at the abstract stage of the literature review can be found in Appendix A.2

Table 4 Coding for rejected texts

Coding to indicate why a text was rejected	
Code	**Reason**
Med	Text refers to assessment used in medicine or psychological assessment
Tech	Text refers to assessment and evaluation of the effectiveness of the technology itself, not the skill of the player
Non-ass	Text offered little or nothing about assessment
Non-game	Text was not about game-delivery of assessment
Non-source	The full text could not be sourced

Of the initial 202 texts, 41 met this second set of criteria. Initially, NVivo was used to code those remaining sources automatically with key terms for type of game mechanics, type of cognitive skill and the type of analysis and reported statistics. However, this approach yielded little meaningful information and few patterns. There was considerable heterogeneity in the literature, and many of the relevant findings of other published texts were described in terms that could not be anticipated a priori to manual inspection. All texts were manually inspected instead. There was some homogeneity in terms of the justification and background of the issue, and some recurring key names such as ‘Bloom’, ‘21st century skills’, or ‘Evidence centred Games Design’, but in the sections where the actual scoring and data selection were described, very different approaches were taken. Each published study was therefore treated as a separate case study, and, in general, each

identified a different aspect of the range of challenges using games as assessments. Together, their contributions created a more comprehensive picture of the problem.

3.3 Findings

Despite the fact that researchers often took very different approaches to the challenge of measuring performance in games, some themes emerged. Bloom's taxonomies of learning were a common justification for the use of games as assessments, with the use of noun forms from his original taxonomy[5] dominating. Overall, 9 of the studies referred in their introduction to the need for new kinds of tests to meet the changing needs of the economy. Of the final 41 texts, 9 studies specifically aimed to measure a complex skill, such as managing a city, measuring creativity or timing and comparing approaches in procedural problem solving. Over half of the research was carried out in the USA.

There was less concurrence on matters that were more directly related to the questions posed before carrying out this literature review. The following sections will outline the findings with regard to the questions I had before conducting the literature review.

3.3.1 Question 1: What is technically possible for games mechanics and game analysis?

Computer connection in school classrooms places limits on the kinds of game that can be created and processed in a school environment. Game development is also costly and time-consuming and several researchers reported on the steps they have taken to overcome these issues. *PopCap Studios*, a major games developer, allowed their commercially successful game *Plants vs Zombies* to be adapted to measure latent problem-solving skills (Shute, 2011; Shute *et al.*, 2016). *Cisco Systems*, a multinational technical conglomerate, have

[5] See Krathwohl (2002) for a comparison of the original and revised Bloom taxonomies.

been able to invest in more complex simulations as part of their training programs (Frezzo *et al.*, 2009). Elsewhere, simple mechanics and dynamics dominated, but there seemed to be scope to create characters, encourage role-playing and simulation and experiment with narrative forms.

The preferred model for analysing games were Bayesian Nets and Bayesian Knowledge Nets. In their literature review de Klerk et al (2015) found that Bayesian nets were the dominant analytical approach to game scoring reported in their literature. Almond (2015) also strongly recommended Bayesian approaches as a computer-compatible form of IRT. Several researchers argued that Bayes is more flexible with conditional dependencies, and that makes a strong case for the use of Bayes to model scores (Almond, 2015; de Klerk, Eggen and Veldkamp, 2014).

Several teams of researchers, however, commented on the exponential growth in complexity of the coding needed to calculate and store Bayesian scores. Cisco's complex trouble-shooting simulation was not functioning reliably at the time of publication (Mislevy *et al.*, 2014; Frezzo *et al.*, 2009) suggested that monitoring any more than 4 parameters at a time was challenging. Shute (Shute *et al.*, 2016) recommended the use of several discrete Bayes nets, rather than one large model. Artificial Neural Networks were used to construct probability based mappings of ability to skills (Lamb *et al.*, 2014), but the researchers also noted that the number of parameters that can be included in such models is also restricted by the exponential growth in processing required.

3.3.2 Question 2: How is domain knowledge or skill being modelled?

There were some findings on the process of data gathering. Frezzo et al (2009) were using Bayesian probability nets to model data, but made a firm case for employing theoretically and empirically based models of assessment. As mentioned above, 9 of the texts aimed to assess complex psychological skills: procedural problem solving (Frezzo, 2009, Mislevy,

2014); creativity (Graf, 2014, Shute, 2016, Leighton, 2016); systems management (DiCerbo, 2016); and complex problem solving (Shute, 2011, Shute, 2016).

Graf (2014) pointed out that research in the field of GBA has been hampered by the fact that games are producing emergent behaviours that are little understood. Others observed that where there is substantive theory, it may not be adequate to account for behaviours. In assessing creativity, Leighton and Chu (2016) observed that although the substantive theory suggested that quantity of ideas indicated creativity, sometimes a simple one-step solution could be more creative than a whole list. They observed that there are so few off-the-shelf models of many of the cognitive competencies games are seeking to measure creates additional complexity in the process of scoring games.

The principles of Evidence Centred games Design (ECgD) (Mislevy *et al.*, 2012) appear to have had some influence on the field of GBA, as 20% of the studies made reference to this set of principles. ECgD is more a set of principles of assessing in games, rather than an approach to analysis.

A first criteria of ECgD is that competencies from a non-gaming environment must be identified a priori to game assessment design. Those working in the field of Bayesian mapping in particular worked to incorporate competencies into their domain model (de Klerk, Eggen and Veldkamp, 2014; Almond, 2015). In some cases, adaptations were made to a pre-existing game, *Plants vs Zombies*, in order to evidence a predetermined skill (Shute *et al.*, 2016). Sometimes the mapping was not done within the game. Working on the game *Spore*, one author suggested playing the game, and then using a critique of the science behind it as a form of assessing whether principles of evolution had been internalised (Mitgutsch and Alvarado, 2012), although this last study was found among the Grey Literature.

A second criteria of ECgD was that the designer must have a strategy for mapping those

externally defined competencies to game play competencies (Mislevy *et al.*, 2012). There were suggestions that modelling complex constructs may be facilitated by technological solutions used in the wider field of information management.

Authors recommended a range of AI approaches to model success criteria. While there is not space in this thesis to explain each of these technologies in detail, it may be worth noting the approaches trialled by AI researchers up to now. Semantic Ontologies were recommended as a possible means to identify and model the interrelations between ideas using a priori assigned formal logic (Vendlinski *et al.*, 2010). Analytic Hierarchy Process and Analytic Network Process have been used in other decision-making contexts to identify independent and co-dependent factors. This approach was used to identify dependency ratios, which in turn established the importance of each criteria to the final score in an assessment task (Bolivar Baron *et al.*, 2015). Bolivar *et al.* (Bolivar Baron, Castillo Salinas and Gonzalez Crespo, 2014) created a learning assessment system using fuzzy cognitive maps to model causality, and Structural Equation Modelling. Various clustering algorithms, such as fuzzy clustering algorithms or K-means algorithms were also proposed to categorise independent factors associated within learning objectives and construct a relational graph (Bolivar Baron *et al.*, 2014). Almond argued the case for using factor analysis with Bayesian nets and a pilot set of data to work from (2015). From the Grey literature, Lamb also used factor analysis to identify facets of skills, but quite often the facets were related to expertise with the game mechanics (2014). These technologies were all recommended to substitute or aid the process of human-mapping the cognitive or psychological function to the evidence. They all appear to be in an experimental phase with inclusive success rates with games data.

Few studies reported on steps to identify areas of error in mapping. Cross moderation and expert judgment dominated. Only two studies (Graf, 2014; Lamb *et al.*, 2014) reported

carrying out tests to identify and remove individual items or players that were not performing within adequate boundaries of confidence and reliability. Vendlinski (2010) reported using empirically based methods to edit the number of items but did not state what these empirical methods were. Lamb carried out a 2-parameter Item Response Analysis of his initial data sets and reported removing items.

In the majority of studies, though, variables were included without empirical support. In many cases, a spirit of trying different analyses to see what happened dominated, which, reflects the lack of consensus in the field.

Games for Learning, the coding of games for learning purposes (Gee, 2005) rather than assessment, seemed to influence much of the research. Games for Learning collects and processes assessment data for formative purposes. It is distinct from the summative purposes characteristic of Game Based Assessment (Mislevy, 2015). Explicit reference to formative assessment was absent, but the term ‘feedback’ was used in the majority of the texts. Integrating formative feedback in real time into the assessment process was the fourth criteria for ECgD (Mislevy *et al.*, 2012). Shute and Ke (2011) recommended stealth assessment, which has the intention of evidencing ability covertly. They suggested that gameplay could be monitored without the consensual awareness of game players, and their candid performance data could be used for assessment purposes. The intention was to remove exam stress and provide a more positive and fairer assessment environment for those people who respond poorly under formal testing conditions but do well in less high pressured environments. Since their article on stealth assessment was written, discussions around privacy have changed, particularly leading up to the introduction of the GDPR in Europe in 2018. Their approach may no longer be considered ethical, but it had gained traction among the GBA community for a while.

A final criteria in ECgD is that, because the fun and engaging nature of games encourages

iteration, this in turn needs to be linked to an iterative assessment model which allows meaning and accuracy of measurement to emerge (Mislevy *et al.*, 2012). Aside from those texts already cited above that recommended Bayesian approaches as a technical means to store and update data, discussion on the iterative nature was absent. None of the researchers explicitly described the conceptualisation of repetition that they were using. Do all attempts count equally? Is there a point when old data are discarded? Was ability stable throughout gameplay? If not, did this indicate fluctuating ability or something else? Was the recency weighting common in many Bayesian applications suitable in assessment? An iterative model of assessment is a significant departure from typical delivery of assessments in snapshot form. It seemed that a meaningful iterative assessment model would need to start with a conceptualisation of the relationship between additional attempts and ability, especially in the light of Soderstrom's (2015) work on developing expertise through repetition, which will be given more consideration in Chapter 4.

3.3.3 Question 3: What scoring models have been used?

Four articles provided some insight into the steps taken to choose what to score, and how this was judged to be representative of latent cognitive ability. These are:

3.3.3.1 *A Packet Tracer,* Cisco's Networking Academy Programme (Frezzo *et al.*, 2009)

Cisco wanted to train and assess engineers with the responsibility of identifying faults in network systems, and so developed the procedural problem-solving simulation *Packet Tracer* game. Frezzo *et al.* (2009) compared patterns of game play between experts and novices for similarities. A novice that followed similar decision-making patterns to the expert was deemed to have identified the most efficient approach to solving the problem. They included a table of observables used in the scoring system (Table 7).

Table 5 Observables used as evidence of learning in assessing procedural problem solving in Packet Tracer (Frezzo *et al.*, 2009)

EXAMPLES OF OBSERVABLES IN THE EVIDENCE MODEL		
DESIGN	**Implement**	**Troubleshoot**
CORRECTNESS OF OUTCOME	*Correctness of outcome*	*Correctness of outcome*
FUNCTIONALITY OF DESIGN	*Correctness of Procedure*	Error identification
CORE REQUIREMENTS	Efficiency of procedure	Error over-identification
PERIPHERAL REQUIREMENTS	Help usage	*Correctness of procedure*
	IOS syntax	Efficiency of procedure
	Volume of actions	- Help usage
	Procedural sequence	- IOS syntax
		Volume of actions
		- Procedural sequence
		- Sequence of actions
		- Sequence of targets

Table 5 gives some indication of observable tasks scored. It seems that next steps analysis was used, and in the full description, the authors talk of pattern matching with expert users. 'Help usage' would delimit the problem space and presumably lead to quicker solutions, but it is unclear how this was measured. A large volume of actions presumably means a less efficient solution. The table appears to be more a framework for statements about performance which were used to prompt tips, such as 'Check your diagram, you are missing a connection between two networking devices,' rather than producing a proficiency judgment.

3.3.3.2 *Sim City Edu* pollution challenge (DiCerbo, Mislevy and Behrens, 2016)

This project looked specifically at a city management simulation game. Young players had to manage a city. Bayes Nets were used to escalate scores in real time. Scores were expressed in terms of 3 key variables: low pollution rates, high population rates and low unemployment rates. The research team found that these three statistics were limited in

terms of the information that they conveyed to stake holders, and that other evidence from contingent work products was also needed. It was noted that sometimes simple measures, such as scatter plots, gave validity feedback on whether the assessment seemed to be evaluating targeted or non-targeted actions. For example, the main intervention that players could make was to re-zone an industrial area into, for example, housing. It was discovered in early iterations of the assessment model that they trialled that re-zoning had minimal impact on the final score.

Following up on grey leads directly with DiCerbo about this work, she mentioned that one of the challenges to scoring this game was that the enjoyable nature of gameplay was sometimes in conflict with the assessment aims. Completely bulldozing the city you were meant to be managing turned out to be great fun. This is not mentioned in the published article, but illustrates a significant problem with GBA.

3.3.3.3 *Newton's Playground* (Kim, Almond and Shute, 2016)

Newton's Playground was a qualitative physics game, testing the laws of physics by manipulation of a ball using planes, ramps, pendulums, levers and springboards. The scoring was decided a priori to play testing, with responses categorized as 'Gold Trophies' if they solved the problem with no more than 3 objects, and Silver Trophy for a solution requiring more than 3 objects. Kim *et al.* (2016) included a sample of the log file variables that were recorded, but this was limited to time stamps, a ball trajectory, and records of Boolean true/false values for silver, gold or unsolved performances. They used an agent identification system to determine how the ramps or levers were being applied by looking at the location of various objects, identifying what objects were touching each other and velocity of movement. Successful and unsuccessful solutions were identified and tagged by human judges, and then new data were compared for similarity.

The findings focused on the impact that the experience of playing the game had on

learning in a pre- and post- test assessment. It appears that the work on scoring was a necessary stage to create an engaging and helpful game, rather than a goal in itself.

3.3.3.4 *Elder Scrolls IV: Oblivion* (Shute and Ke, 2012)

Oblivion is a first person role-playing game. It assesses creative problem-solving ability, through the manipulation of tools, spells and weapons. The player needs to overcome various challenges and stay alive, and this study proposed a system of assessment that rewards a range of abilities such as knowledge, for example, the most likely place to find necessary items, or creativity, experienced in this game as choices (Shute, 2012). It could be argued that this is just a commercial game, not educational, but it was included in the literature review because of their approach to assessing the skill of creative problem solving, which may have wider applications. The researchers proposed a measure of how much the problem space had been delimited by the player's actions, using performance data from a set of players. The novelty parameter, for example, is calculated by the number of people who chose this option in relation to the total number of people in the data set. It is a little reflective of Guilford's ideas of divergent thinking (1959). An example of this is shown in the table below.

Table 6 Sample of scores in *Elder Scrolls IV: Oblivion* for creative problem solving (Shute and Ke, 2012). Players were scored on their techniques to overcome a river obstacle.

Example of action model with indicators for novelty and efficiency

Action	**Novelty**	**Efficiency**
Swim across river filled with dangerous fish	n= 0.12	e = 0.22
Levitate over the river	n= 0.33	e=0.70
Freeze the water with a spell and slide across	n = 0.76	e = 0.80
Find a bridge over the river	n = 0.66	e = 0.24
Dig a tunnel under the river	n = 0.78	e = 0.20

The efficiency parameter came from how many people succeeded in crossing the river successfully out of the total number of people who attempted this method (Table 6). These values were then loaded as priors into a Bayesian model, which classified the performance into one of two states, high or low ability. Although the game itself is unconventional in education, basing scores on probability estimates from observed performance data has some similarities to IRT approaches. Treating it in this way, to understand the delimitation of a problem space, seems to have potential in games with choice as a feature. One possible concern might be the extent to which the results were sample dependent.

One conclusion from looking at these four reported scoring models side by side is that a particular challenge of scoring games seems to be that different games mechanics need a different scoring approach.

3.3.4 Question 4: What work has been done to validate current findings?

Ten studies made reference to IRT in the first scoping of the literature, but eight of those were filtered out because these did not refer to game delivery, and in general claims of validity were rare. One paper cited using IRT in validating a pre- and post- external test of students, not on the telemetry data itself (Bressler and Bodzin, 2013). One study used IRT in the item selection process (Lamb *et al.*, 2014) but it was particularly interesting. That article came out of PhD research, which meant that, although it was found in the grey literature, a thorough description of the research had been documented. Using data from science games, factor analysis was used to separate the array of data into three categories: game control, flow and science processing knowledge. The author then ran a 2PL IRT model on student and item data to remove items that were not functioning within an adequate range of stability. The main purpose of the study was to implement an Artificial

Neural Net to score performance, and this allowed exactly three attributes, such as 'visual attention' to be assigned to each task. Presumably, precisely three were necessary to be able to parse the data with no missing values. One possible concern about this study was that his technical requirement seemed to force a rather arbitrary selection of skills at times. It was unclear, for example, how 'visual attention' or 'attention switching' might be attributed to one task but not another. Another concern with this study was that although the model seemed technically thorough, from a qualitative perspective, it suggested that sometimes technical solutions only work provided that the data are forced to fit the model, rather than making the model fit the data.

Around a fifth of the texts which met the selection criteria referred in their introduction to the inadequacy of current methods of assessing the skills of modern economies. Texts that addressed these skills more explicitly looked at systems management (DiCerbo, Mislevy and Behrens, 2016), complex procedural problem solving (Eseryel, Ifenthaler and Ge, 2013) and multi-player collaborations (Gosper and McNeill, 2012). Over half of the studies introduced some form of post-game enjoyment survey, particularly in the form of Likert Scales. They found that players reported positively on their experience, and a strictly empirical analysis would overlook this important aspect of game-based assessment.

Evidence Centred games Design (Mislevy *et al.*, 2012) and Cognitive Diagnostic Assessment (Leighton and Chu, 2016) were proposed as task analysis models. What seemed missing from the literature was a process of trialling and refinement, or calibration, from many of the models. It may simply be that those phases were not fully reported because many of the articles were for publication in computer science journals. Other insights into refining cognitive task analysis models were found and these are reported below in no specific order.

Two studies (Lamb *et al.*, 2014; Graf, 2014; Vendlinski *et al.*, 2010) reported carrying out

empirical analysis to identify and remove individual items or players whose behaviour was atypical from further analysis and development of the scales. In other cases, when error of measurement was discussed as an issue, no approaches to dealing with it were offered.

Where score data were described, the use of raw scores dominated. Standardising scores through z tests, t-test values, or logarithmic transformation would be more usual before attempting to make claims of difficulty levels, ability or equivalency of different tests (Bond and Fox, 2015; Harvill, 1991). The reporting of significant correlations at times appeared exploratory and serendipitous rather than confirmatory and intended. For example, Chin *et al.* (2016).reported that regression analysis showed that time spent on task correlated with some 'likes' awarded by peers to an end product in selected circumstances.

Only one text referred to variance in how their model performed among children who were in different grade bands in the assessment scale (Vendlinski *et al.*, 2010). They found that once players passed their own ability threshold, their patterns of behaviour differed depending on whether they were playing a linear and non-linear narrative form of a game. The suggestion was that narrative form could sometimes be seen as a confounding variable. The majority of studies only attempted to classify learners into mastery or non-mastery, but often they did not declare what limen, or threshold, they used to determine mastery, such as a 50-50 or 80-20 probability of getting an answer correct.

3.4 Conclusions from the review of the literature on GBA

The wide scope of topics in this rather limited number of texts shows the very wide range and number of considerations involved in scoring games fairly. Little appears to have been done on estimating how challenging each task might be, or the possibility that untargeted data are being introduced into scoring models. More work is needed to understand the

domain model in terms of what can be evidenced. Bayesian modellers tend to produce a graphic similar to the one in Figure 14.

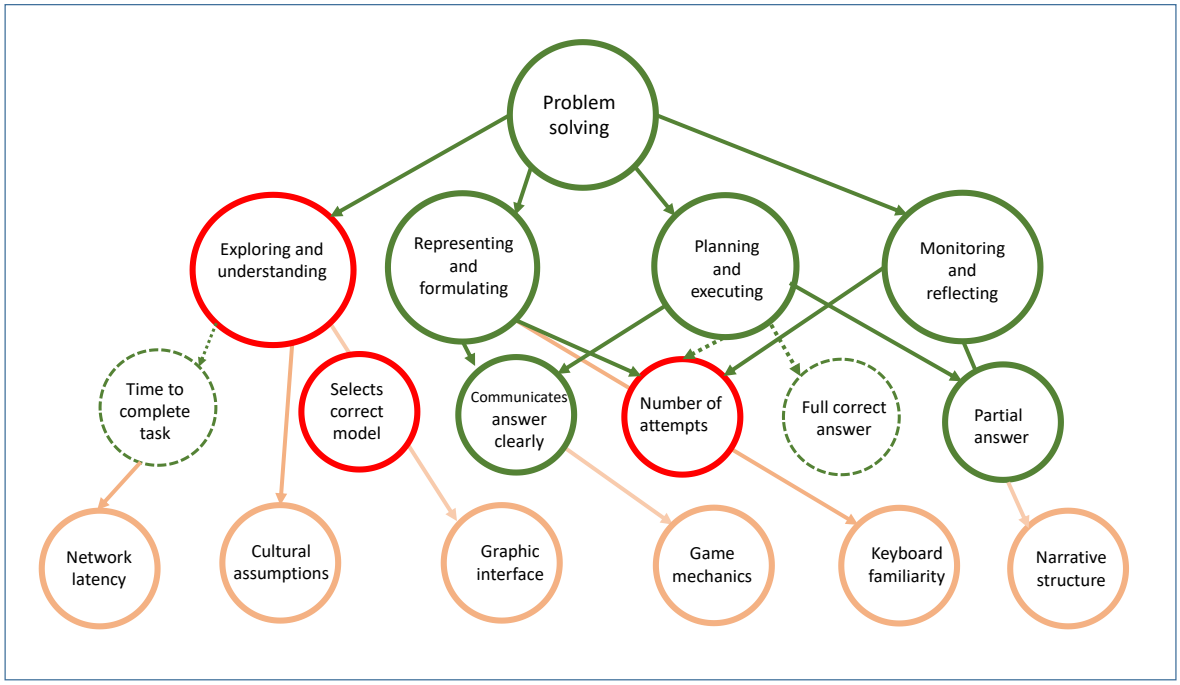


Figure 14 Cognitive task analysis model showing possible over- and under-inclusion of nodes

Figure 14 shows a summary derived from the findings of this literature search. It is a form of cognitive task model, similar to the ones that an AI researcher might produce. There is an overarching latent skill at the top, the θ value that the researcher would like to estimate, such a problem solving. The overarching cognitive ability is broken down into subskills, and in constructing this diagram, the PISA problem-solving framework (Ananiadou, 2009) was used for illustrative purposes only. A framework such as this might inform the construction of a cognitive task analysis model, but more analysis is needed. In this second layer, however, there may be assumed connections illustrated by the red circles, which actually have no evidence. The third layer contains the actions in the game that generate evidence for these subskills. Any one action might evidence more than one subskill, but there is still a need to test whether the links are justifiable, not just assumed. The dotted lines around ‘time to complete tasks’ and the ‘number of attempts’ are variables that seem under-explored or under reported in the literature, but seem significant. The literature

review also hints at other interfering variables that are not on the 3-layer model. In this diagram, these have been suggested by the fourth orange layer. In short, there seems to be potential for considerable over- and under-conclusion of nodes in the domain model that games designers are drawing up.

Some form of concurrent validity test on the mapping from game task to syllabus needs to be introduced. The test syllabus is the list of tasks that children are expected to perform, and the knowledge and skills that they are expected to demonstrate in completing these tasks. It may force the test designer to remove supposed edges from their original domain model if there is no evidence of a relationship as there may be redundant concepts or tasks, or some may not discriminate accurately or fairly. While technical advances have been made by the computer science community, these are assessment concerns and there is much work to be done in this area.

As few methodological approaches emerged from the literature review, further texts were consulted to identify possible approaches to be taken in this study, and considerations to bear in mind. This second scoping aimed at identifying a methodological approach to match the research questions, and that will be discussed in the following chapter.

Chapter 4 What other research sheds light on modelling gaming data?

The systematic review of the literature on Game Based Assessment (GBA) in chapter 3 did not reveal any established methodologies for looking at gaming data. However, there has been a large body of research that has been subject to peer review in terms of dealing with many of the aspects of assessment that seem pertinent to GBA, namely handling missing data in assessments, modelling variable response time in testing, and other literature that sheds light on how to delimit the problem space when testing skills in assessment. These had not been used in game environments but tackled similar challenges to those raised by GBA. This chapter was therefore necessary to show the broader body of published literature that informed the choices for the methodology used in this particular thesis.

This chapter will start with a review of Vygotsky's Zone of Proximal development (1987), closely linked to Cziskzhentmihalyi's (1975) work on flow, which are themes that unite both games designers and educators. There is general consensus among those already working in the field of GBA that Bayesian approaches (Almond, 2015), and in particular Bayesian Knowledge Nets (de Klerk, Eggen and Veldkamp, 2014) are better aligned to scoring these new skills, and so there will be a brief overview of Bayes, and considerations around when to calibrate the difficulty levels of the tasks and score ability. Gameplay data produces a large amount of missing data, and ways of conceptualizing missing data in assessment modelling will be discussed. The last part of this chapter draws on research from the broader field of assessment on how to use time as an alternative proxy for ability in game scoring, and how to deal with familiarity with the task through iterations.

4.1 Changing educational landscapes

Both games designers and educational researchers are familiar with the concept that work can be optimised when the level of challenge of any task is just above the person's current level of ability. In education, this was expressed in the foundational work by Vygotsky on the Zone of Proximal Development (ZPD) (1987). The ZPD was an optimal learning space for children, where the resources they were presented with were just above their existing level of ability. Anything easier risks inducing boredom, anything too challenging risks introducing too much anxiety. This same idea was developed with wider social applications by Cziskzentmihalyi (1997) in his work on 'flow', which will be more familiar to games designers. Flow is also an optimal state where the level of difficulty of a task is slightly above the level of ability of the person and the person becomes fully absorbed. It is an emotional state that good games designers hope their players will enter.

In the wider sphere of Vygotsky's (1987) work, he believed that education was socially constructed. Within the ZPD, there was a Zone of Collaboration, which is a point where other humans, such as teachers, parents or classmates, have an opportunity to socially impact on the learner. Within the Zone of Collaboration, there were two more sub-categories, the Zone of Available Assistance (ZAA), or all the possible resources available, and the Zone of Proximal Adjustment (ZPA), which are those resources, such as tools, people, the environment and pre-existing knowledge that are actually perfectly appropriate to the learner. Reading Vygotsky in the UK in the late 80s and early 90s, around the time of writing, teachers were often the main source of input and there was perhaps one course book for an entire year and a few shared supplementary materials in libraries. Optimising the ZPA was challenging to implement because there were such limited available resources.

Luckin (2018) revisited Vygotsky's work, drawing attention to the fact that Vygotsky

always acknowledged that there were these filters on the ZAA. Great teachers are in limited supply, and there are financial or geographic barriers.

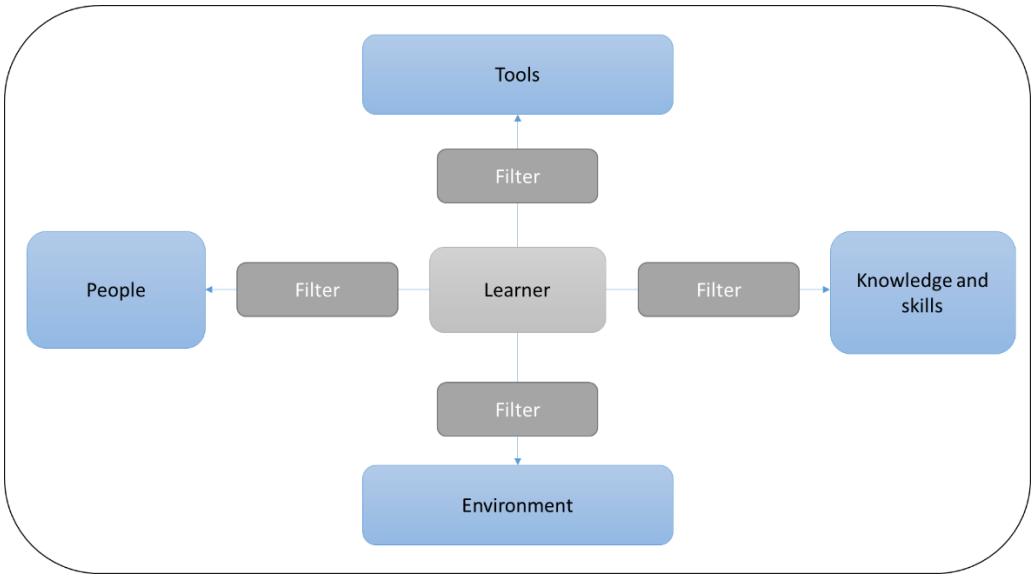


Figure 15 The ecology of resources model of context, (Luckin, 2018)

In response to technological changes, Luckin (2018) proposed a model called the Ecology of Resources Model, shown in Figure 15. Technology has dramatically removed some of the traditional filters that Vygotsky identified in the ZAA. Top universities offer some of their best courses online through MOOCs, and even learning tools such as Khan Academy are available to children in the most remote and disadvantaged locations now that resilient low energy solar powered servers are available (Walls *et al.*, 2015). As these zones of learning change and filters drop away, new challenges emerge.

Games are an example of a new resource in the teacher's tool kit. Engagement is a common metric used among games analysts to provide evidence that the game is popular (El-Nasr, 2016), but it does not provide evidence of ability. The methods to identify Cziskzentmihalyi's (1997) concept of flow are not the same as using the gameplay data as evidence of ability.

4.2 Modelling assessment practices and some early decisions

The existence of games which collect data on performance does not automatically make them suitable or useful for assessment purposes. In any gameplay data set there is likely to be a mix of useful and impeding data. Toulmin's taxonomy of reasoning diagrammed the process of identifying which data or evidence aids the decision process, in this case the measurement of a latent skill, or which impedes it.

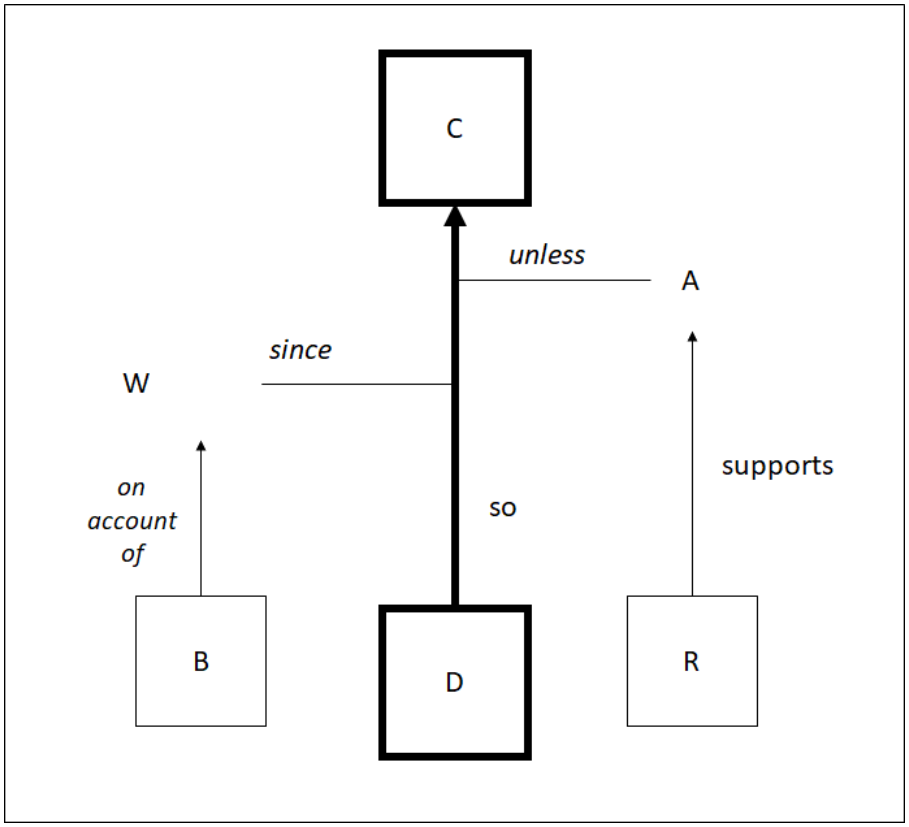


Figure 16 Toulmin's general model of argument. Reasoning flows from data (D) to claim (C) supported by a warrant (W), which in turn is supported by backing evidence (B). This may be qualified by an alternative explanation (A), which is in turn supported by rebuttal evidence cited in (Mislevy, 2018).

Figure 16 lays out the fundamental purpose and process of calibration in assessment. The aim is to get from an untreated, raw telemetry data set (D) to any claims (C) that the games had evidenced ability fairly and we have measured only that. The analyst therefore needs a warrant (W), or a justified reason to make these claims. That will come from backing evidence (B), which will need to be understood and interpreted in the context of the data

collection process. With assessment data, there may also be rebuttal evidence (R), which supports an alternative explanation (A). In assessment, this alternative explanation is that there were compounding variables captured in the data collection process.

In games, reasoning from the perspective of just one of the three disciplines of education, assessment and games design may result in limited success. Educators ultimately need to make inferences about the results. Assessment designers are more likely to understand the warrant for the use of gameplay data, but they tend to approach the decision-making process with a range of socio-cultural concerns (Messick,1987, Mislevy, 2018) as well as statistical support (Wilson, 2004, Bond, 2015). The environment where the backing evidence is collected, however, is the specialist domain of the computer scientists. Their chain of reasoning is more likely to be expressed in machine logic terms, in order to make the most basic implicit assumptions about relationships clear. For example, computer science literature makes use of symbolic structures such as Directed Acyclic Graphs (DAGs) (Almond, 2015), which will be described in more detail below. These are not part of initial teacher training preparation. Given the complexity of game design and assessment data analysis, it is tempting to collect the data and interpret it in completely separate silos, but this seems restrictive.

DAGs are best illustrated for assessors in terms of Bayesian models. Assessment starts with the assumption that there are many known variables that contain possible knowledge about a target value we are trying to estimate (Leonard and Hsu, 2001). Some of these will be constant, and response time is an example of this in the wider field of assessment. What this means is that, although time management undoubtedly impacts on scores, we hope that is captured in the overall performance, and all children are recorded as completing in the same constant time set for the test, such as an hour and so it is rarely included in the assessment model.

Other variables, such as the total number of marks, will be variable. Some values of variables are known, meaning that they come directly from the data, such as how many children answered a particular question correctly. Others are unknown and will need to be estimated as a mathematical function of the data. How difficult the question was is a likelihood function, drawn from the data, and this was the process introduced in chapter two.

DAGs consist of a finite number objects known as vertices. These are always nouns that represent either the constants or the variables. They also have a number of edges, represented by lines and these are usually verbs to describe a relationship between the nouns (Levy and Mislevy, 2017).

Figure 17 Simple evidence model DAG, where performance on task x contributes to the score y

Figure 17 shows a DAG for many teacher-scored tests. The evidence from questions (x) is used to estimate the ability of persons (y), shown by the direction of the arrow. This is done by summing up the number of correct answers. It is also a model that is typical in many games design domain models with evidence of ability assigned to a skill by expert judgment, for example a task might require the skills of collaboration and addition of fractions.

A Bayesian approach in game scoring tends to carry records of scores for the child over from previous games, and in some games, such as city management, this is a starting point, in the form of a prior. This is because in games, play tends to take place over a matter of days, weeks or months (Frezzo *et al.*, 2009). The impact of using a prior is shown in the first DAG, (a) in Figure 18. The prior θ, or the previous estimate of ability, acts as a

control against the new data, $x_1 \ldots x_4$, that is collected in the current gameplay session.

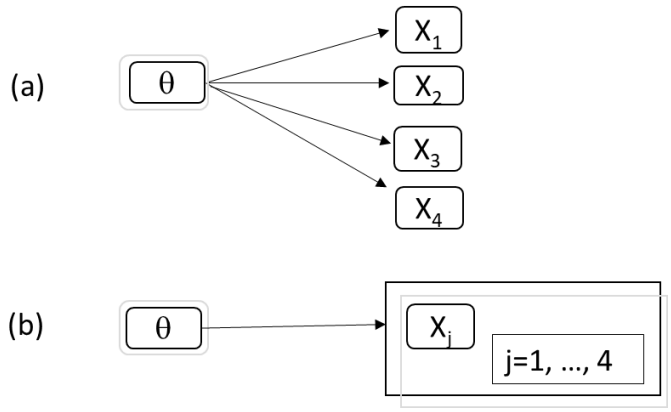


Figure 18 The Bayes IRT DAG model, with the prior, y, influencing the subsequent scores for individual tasks and plates used to express iterations (Levy, 2017, Bayesian psychometric modeling)

Carrying over of priors is at the heart of the last challenge in the ECgD criteria, that an iterative assessment model needs to be found to assess the iterative nature of gameplay (Mislevy, 2012). Carrying that prior over entails a range of value judgments on the role of old knowledge and new evidence. Often DAGs summarise repetitive patterns by using 'plates'. The second image (b) in Figure 18 shows the same process, but using a plate to indicate iteration. The ability, θ is used to weight the new data from the same four tasks, $x_1, \ldots, x_4$ with different people represented by j. The node for the prior θ is outside of the plate because it influences all the instances in the same manner in this particular model. This is still just a scoring only model, not a calibration model. The rounded rectangle shape of the nodes for x show that these are known values drawn from the pool of evidence. This scoring only model might suit, for example, something like the tally systems that are common in Cognitive Diagnostic Assessment models (Leighton and Gierl, 2011), but it is problematic for other types of assessment. All of the data are treated as backing evidence for the warrant that we are measuring fairly, when we know that some of this data are rebuttal evidence for the alternative argument. In other words, we have measured

something else by mistake.

This study aims to reveal the extent of rebuttal evidence in the data set, and this needs a different type of DAG. This will be discussed in the context of specific questions and concerns below.

4.2.1 When would scoring and calibration take place?

As discussed in the section 2.3 introduction in chapter 2, assessors tend to carry out a calibration phase to establish the value each tick should be assigned before they start to score children. The DAG for this needs to capture the unknown parameters of the value of the tick, as in Figure 19.

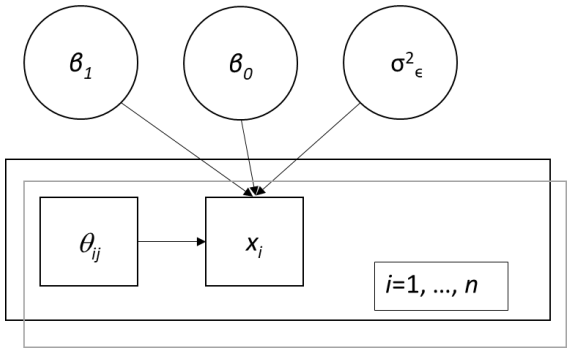


Figure 19 The values for *x* are considered unknown in a calibration and scoring model.

In Figure 19, the circles indicate values that inform an Item Characteristic Curve (both discussed in 2.3.2) for each task, and these need to be estimated from the data during that calibration phase (Levy and Mislevy, 2017). It produces the set of coefficients that make up the β_1 slope, β_0 intercept term, and σ^2_ϵ error variance for each item x_j so that the ability score θ of each child, j, on each item, i, can be calculated (Almond *et al.*, 2015).

The item difficulty values (x) are therefore unknown, and so too is the ability of each person (θ_j) until the estimation is carried out. With Bayes, there are two possible ways to approach this:

1. Two-stage calibration and scoring model

A two-stage calibration and scoring model is common in assessment practice. It carries out the calibration phase first, using learner performance data from a sample of test takers to estimate the values for *x*. At the end of the calibration phase, these values are 'anchored', in other words no further changes are made to the value assigned to each item *x*, even when new data are acquired. Values for *x* are entered as known parameters during the scoring phase to estimate new values for θ. This was the approach from the original Rasch model (1960).

2. Simultaneous calibration and scoring

Bayes allows an alternative approach to be taken, carrying out the calibration and scoring of test taker ability simultaneously (Levy and Mislevy, 2017). The values for *x* are not anchored, and can be continually updated in the light of new data. One of the possible affordances of using Bayesian mathematics is that it has the potential to calibrate and score simultaneously. The effect of simultaneous calibration and scoring on game scores is one of the features of games that this study will consider.

4.2.2 What will the outputs look like?

Although scores are very often reported to stakeholders as categorical variables (pass, merit, distinction), there are various ways of getting to this kind of output. Cohort referenced approaches set grade boundaries in relation to the other children in a particular cohort.

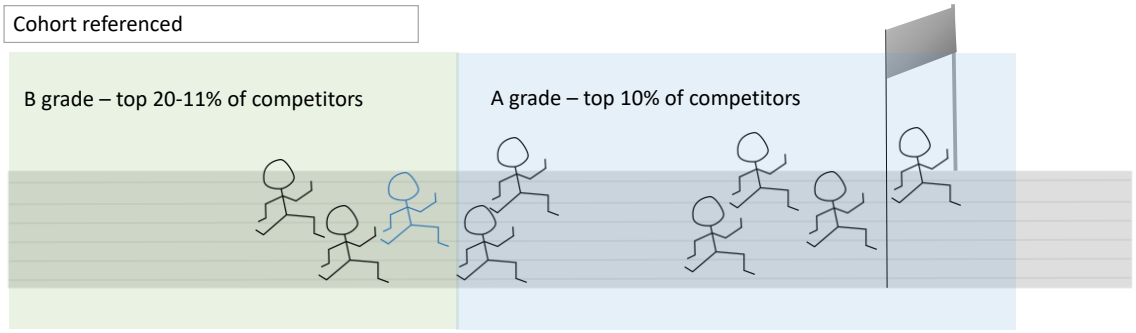


Figure 20 Cohort referencing puts children's performance into categories based on their relation to others taking the test

In this model, the top, say 10% of children will be given the top band, and information on how their performance relates to children in previous or subsequent years is ignored (Figure 20). That is not always the approach taken, though, and a criterion referenced or curriculum referenced approach is also an option. In these two approaches, children are rewarded scores if they have demonstrated a designated range of skills or knowledge in the test, and how their fellow candidates performed that year is ignored (Baird *et al.*, 2018). Even with categorical reporting, a continuous scale ability rating is usually generated from the performance data, and a senior member of the assessment team decides where to determine the grade boundaries for that particular delivery of the test (Mislevy, 2018).

Bayesian and other Machine Learning approaches often classify from the outset, with grade boundaries set a priori. Table 9 below shows Almond et al.'s (Almond *et al.*, 2015) suggested 5-band IRT scoring structure, with bands of logit values -2, -1, 0, 1, 2. It is assumed in the prior that 40% of the population will fall into the 0 band, 20% each into bands -1 and +1, and that the extreme bands, -2 and +2 are outlying behaviours, likely to contain just 10% each of the student population.

Table 7 Almond's categorisation model (Almond *et al.*, 2015)

θ	Prior	Conditional probability				
	θ	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**
		(β = -1.5)	**(β = -0.75)**	**(β = 0)**	**(β = 0.75)**	**(β = 1.5)**
-2	0.1	0.3775	0.2227	0.1192	0.0601	0.0293
-1	0.2	0.6225	0.4378	0.2689	0.1480	0.0759
0	0.4	0.8176	0.6792	0.5000	0.3208	0.1824
+1	0.2	0.9241	0.8520	0.7311	0.5622	0.3775
+2	0.1	0.9707	0.9399	0.8088	0.7773	0.6225

Table 9 shows how β values, or difficulty values, for each task might be used to estimate performance levels, and how the ability estimation would update in response. Looking at the distribution under the prior, there seems to be a fairly strong assumption in this model that ability in the population will be normally distributed. That may not be the case, and one of the arguments for using a logistical distribution in assessment is because skewed data are common (Wilson, 2004; Bond and Fox, 2015). A categorical classification may contain assumptions that the assessor would not share.

4.2.3 Differential performance

One of the main principles of assessment is that test-takers are assumed to have an equal probability of answering the question correctly, until we learn otherwise from their performance data (Wilson, 2004). This may not always hold true, and the assessor may have some prior knowledge that suggests that a question or task will be easier for boys than girls, or with children of different nationalities. For example, attitudes towards guessing in tests are largely shaped by the culture of education. In international comparison tests, such as the PISA exams, a nation of children instructed, 'If you don't know the answer, just guess.' tend to benefit from that strategy. In other cultures, children are conditioned from a

young age never to guess, as it distorts the true picture of their ability. It seems sensible to weight scores to take the impact of such cultural attitudes into account if the end goal is to create an international comparative study of attainment (Mislevy, 2018). Culture impacts in other ways. In an English test, children in Latin America, for example, might misunderstand a question that refers to 'America' with the intended meaning of the USA, as it is a very culturally sensitive term (Zheng and De Jong, 2011). The existence of an error term, which are often caused by socio-cultural factors that can be hard to identify a priori, is at the heart of Classical Test Theory (Lord, 1980). Differential Item Functioning (DIF) is the term referring to times when the question or task is experienced differently among different demographic groups. More recently, work has been carried out to make the identification of DIF through Bayesian approaches easier (Almond *et al.*, 2015).

The iterative nature of games that Mislevy (2014) drew attention to in ECgD maybe a DIF problem. In games, the characteristics that make an item function differently may not be affiliation to a particular demographic group, say male/female or Year 10/Year 11, but the structure of the tasks, including prior exposure to the games. Children who have played the same game several times already presumably are different from a child playing it for the first time. Children who have chosen a particular pathway through the game will have been exposed to different information.

The number of unique pathways through games may be caused by the different narrative structures through the game. Traditionally, tests have been delivered in linear format, as in pathway (a) in Figure 21 below.

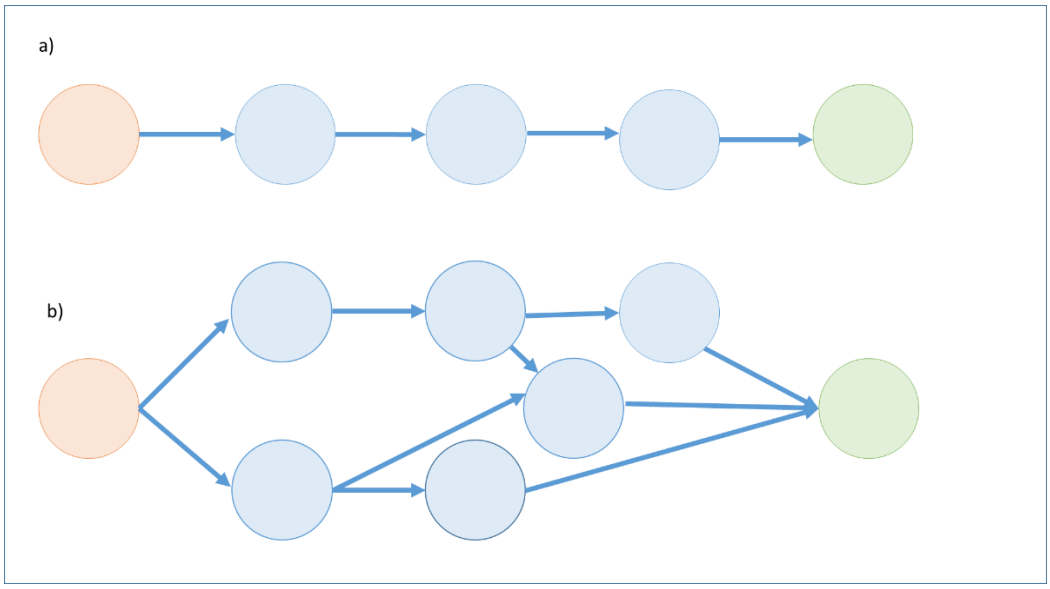


Figure 21 Linear designs of traditional tests in (a), compared to non-linear hypertext designs in games in (b)

Although the child might turn directly to, say, the last question, and then work backwards if they wish, under the Rasch model, the questions should all be discrete (Wilson, 2004, Bond, 2015), and therefore the child should not be affected. Games are characterised by branching patterns, illustrated by pathway (b) (Figure 21). Many games are engaging precisely because of the level of complexity and agency, or choice (Adams, 2010, Fullerton, 2014). This agency creates two possible issues for scoring games. The first is DIF. When children get back to a central point in the narrative, a task, say, that all players have to complete, have the challenges they were exposed to along the way made them different? Can we capture that? Should we capture that? The second issue is that it creates a huge amount of missing data, which will be elaborated on below.

4.2.4 Missing data

When the child chooses one pathway, we have no information on how they might have performed had they taken a different pathway. In essence, their performance on untrodden paths becomes missing data. In the simple branching pathway in Figure 21, for every 3 steps that the child completes to get to the green endpoint, we have three other steps with

no data. The data frames below in Figure 22 are taken from the data set under analysis in this study. You can see that the long format that is used by the games designers has no missing cells, but most analysis software requires a wide format to parse.

Long format

Child	Game	Score	Response Time
100230	BidmassBlaster	0	57.0
100230	BidmassBlaster	1	63.4
123045	Jet Stream Riders	3	35.6
167895	Flower Power	2	180.0
100230	BidmassBlaster	2	42.6
123045	Jet Stream Riders	3	32.7

Wide format

Child	Attempt	BidBla Score	BidBla RT	Jet SR Score	JetSR RT	FloPo Score	FloPo RT
100230	1	0	57.0				
100230	2	1	63.4				
100230	3	2	42.6				
123045	1			3	35.6		
123045	2			3	32.7		
123045	3						
167895	1					2	180.00
167895	2						
167895	3						

Figure 22 Data on performance in different tasks (game) displayed in long format and wide, showing the large amount of missing data in wide formats from the MangaHigh© data set

The 'game' variable, or the task that the children played, moves from row values to column heads, each with a score and response time value, and during this transformation, a very large amount of missing data is generated (Figure 22) because not all of the children completed all of the potential steps available to them. This is compounded further when they are able to have additional attempts at the same task.

Assessors need to interrogate assumptions around the impact of missing data. Some values may also be missing for technical reasons, where the computer failed to record a gameplay due to connectivity issues. Research from the field of clinical diagnosis assessment suggests that network interruptions are fairly common (Van den Broeck *et al.*, 2005). This type of missing data is not related to the target θ of ability, and so could be treated as Missing Completely at Random (MCAR) (Allison, 2001), although the best solution would be to investigate these anomalies and fix them. Van den Broeck *et al.* (2005) cautioned that irregular patterns in data tend to require manual inspection to identify them, and the size of gameplay data sets may complicate manual inspection.

Data that is missing due to test design requires a cautious approach, as stakeholders often

assume that a skipped question is synonymous with not being able to answer it. In technical terms, missing values are treated as zero, under the assumption that it is Missing Not At Random (MNAR). In other words, the reason the data was missing was related to the target variable of ability (Allison, 2001). In cases such as the example of DIF, discussed above, children's decision to leave a task blank is influenced by wider cultural norms (Mislevy, 2018), not just ability and it raises questions around fairness. In earlier studies, Mislevy and Wu (1996) argued that treating missing values as MNAR might also have an undesirable effect on the calibration of the results. They pointed out that in speeded tests, the test design forced some questions to be ignored, usually at the end of the test. It may be appropriate to treat those missing answers as 0 when scoring the child's ability, but during calibration of the difficulty of the question or task, it made the questions appear significantly more difficult than they were experienced (Mislevy and Wu, 1996). It was better to treat it as Missing At Random (MAR), or caused by other factors (Allison, 2001).

In games, the test design also forces some questions to be ignored. In treating all un-navigated paths as zero values, they may start to appear as more difficult, when in fact they were simply less popular. Mislevy and Wu (1996) suggest that when the data is absent because it was not administered, it is better to estimate and anchor difficulty estimates using only the full data set at the calibration stage. Ludlow and O'Leary (1999) recommended a similar approach, but extended this to the scoring phase, too. They suggested that in some cases it might be better to estimate the child's ability as a proportion of the number of tasks attempted.

There is still the question of whether one path was easier than the other, and whether ability was the reason that children chose the easier pathway. Ludlow and O'Leary (1999) recommended that scoring as incorrect only be carried out on data that was MNAR.

Translating that to a games scenario, it would seem to require that there was a clear optimal path, and the child's ability was partly or wholly responsible for not taking that optimal route. With games, it may be necessary to develop new terminology to reflect reasons for being missing that are unique to gaming scenarios, such as 'browsed but not played', or 'not yet discovered'. Once the participant is committed to a certain route, pathways that immediately become closed off may be best treated as 'not presented'.

However, many software packages for Bayesian modelling, such as WinBUGs, and many Machine Learning algorithms have fairly strict rules on data types. Variables can be integers (whole numbers), decimals, or in date or time form, but a blank space is a Not a Number (NaN) value. NaNs are not tolerated during the process of parsing data, and a substitute number has to be entered (Lunn *et al*, 2012). That means that choices around modelling missing data can be reduced to either a zero value, which biases the data (Mislevy and Wu, 1996), or imputing a value from nearest neighbour (Lunn *et al.*, 2012). Guidance on how to deal with this problem was not readily available in the literature and investigating technical solutions was out of the scope of this research, however the potential impact of forcing a value will be considered.

4.3 How should we conceptualize paradata in gameplay?

Gameplay results in a large amount of paradata. There is little consensus about what specifically can be designated paradata, but in general, it is data about how the data were collected, and is often automatically generated whether the client, in this case, the assessor, has asked for it or not (Kreuter and Casas-Cordero, 2010). Von Davier and Halpin (2013) pointed out that many published research projects on games have chosen to ignore the complexities of scoring from this micro data, using external pre- and post-test models to explain what was happening in the game instead. Several researchers in the field of GBA have drawn attention to the potential of paradata to reveal additional information about

ability (DiCerbo, 2014; Mislevy *et al.*, 2014; Shute and Ke, 2012). Paradata is, however, at a micro data level, such as time stamps or key strokes. Transitioning from micro-data to macro-understanding is not simple.

There is a case for consciously ignoring some paradata. Von Davier and Halpin (2013) argued that greater clarity over what will and will not be assessed seems to be necessary to deal with the mass of data produced by games. In trying to model collaboration, they observed firstly that very clear conceptualisations are needed. For example, they identified a number of potential different types of collaboration, before choosing one to measure. In the case of the games in this study, they all produced time stamps for every action, as do all actions in computing, but an overarching conceptualisation of speed was absent from the literature within the field of Games Based Assessment.

Von Davier and Halpin (2013) also observed that it is still acceptable to assume that some things are captured in the overall score, and should therefore not require a specific variable. These nuisance variables are not uncommon in assessment, and aspects such as crossings out are rarely incorporated into scores. In their case, they felt that skills such as leadership, influenced the final outcome, but did not require a specific variable. By ignoring some paradata, some of the complex interplay of dependencies in games can be removed, and it allows psychometrics to be brought back into the process of scoring some complex skills.

It is perhaps also telling that AI scoring methods that are currently in professional assessment usage seem to avoid trying to standardise performance at a micro-level. There is a long history of trying to assess complex skills such as proficiency in the English language, which itself is not even a stable entity (Mislevy, 2018). In essay scoring algorithms, the algorithm itself is rarely, if ever, subject to scrutiny. The algorithm might build a judgement using Natural Language Processing from a micro level up, but standardisation, for example, with the Pearson Test of English Academic, takes place at the

macro level of an interrater reliability analysis (Zheng, 2011). Interrater reliability is the measurement of the extent to which two classifications of a piece of work coincide (Gwet, 2014), in this case, the agents are a human and the AI (Zheng, 2011). The effort to include other variables therefore may not always be justified, and there is a precedent within assessment for taking this approach.

Returning to the issue of speed and time stamps, although it is anecdotally accepted that a skill like time management has a positive impact on children's test performance, how children manage their time within the upper and lower bounded limits is usually treated as inconsequential in the scoring process. Assessors rarely attempt to capture and measure this in-test, but instead assume that it is already captured by, for example, managing to answer all the questions. The variable 'response time' is usually held constant at a time deemed adequate, such as an hour to complete the test. In the wider field of games, though, speed is often treated as a central indicator of ability (Adams, 2010, Fullerton, 2014). Detailed time records are an example of paradata that are available to the assessor. Whether it is sensible to use these records needs to be investigated, which the next section will develop in more detail.

4.4 Competing proxies for ability

Most educational tests are completed under time conditions, and so response time is an aspect of the test experience, even if completion time is kept constant for all of the test takers. Often, things like successful time management are assumed to have been captured in the scores, even if they have not been modelled (Davier, 2013). In fact, many behaviours may influence performance, but are treated as nuisance variables, because of the challenges of capturing and explaining everything. However, in commercial games, response time is rarely treated as a constant. Players have varying completion times and response time is often the main criteria to compare performance by players in the same tasks (Fullerton,

2014). In other words, response time is often used as the main proxy for ability. Games analysts tend to use the manifest variable, or physical measure of time (El-Nasr, Drachen and Canossa, 2016). In other words, for gamers, time is already an accurately calibrated mechanical measurement process.

Some types of educational tests also award separate scores for accuracy and speed. All tests assess a proxy, such as accuracy of performance answering test questions, for another target measurement, such as a cognitive ability or useful world skills, such as communication. When using speed, rather than accuracy, as the proxy, there seem to be more questions around the validity of response time and what it actually measures, the nature of the relationship between speed and accuracy, and the reliability of scoring systems from response times.

Education and psychology share many assessment analytical approaches. There is a larger body of research around how response time could be used to measure a construct, particularly from psychology. The most common test assumes that rate of work is an important factor, and therefore imposes a constant maximum time limit to complete all of the tasks has been set. Often, this decision may well be influenced by the practical constraints of testing a large number of people at one time. Early theorists distinguished between a power test, which assesses on a scale of difficulty, and speeded tests, which assess on a scale of speed (Morrison, 1960). There was also evidence that the two different approaches did not measure the precise same thing (Davidson, 1945).

Under a strict definition, a power test gives students unlimited time to respond, removing time pressure entirely (Gulliksen, 2013). The 3PL Rasch model assumes, for example, a power approach, and speed is not factored into the equation. In practice, a true power test, with no time limit, is rare, and for practical purposes, most test design companies will introduce some constant maximum time limit. Lee and Ying preferred to use the definition

of a test delivered in power time as one that gives test takers the time that it would take to responds if time were unlimited (2015).

A speeded test focuses on speed as the scale of ability. In a speeded test, the questions or tasks are relatively easy, but the test takers are under pressure to respond quickly (Gulliksen, 2013). Fairly elementary questions are chosen because the main aim of speeded tests is to assess recall, rather than, say, a more complex skill such as problem solving. Individual exam bodies may have their own specific definition of the point when a test can be classified as speeded. Educational Testing Services (ETS), for example, define a speeded test in relation to the performance statistics from the sample taking the test. When a certain percentage of students complete the test, or a portion of the test in particular time, say only 80% get to the three quarters point in a certain time, it is considered speeded (Donlon, 1980).

There are other questions around the validity of time that are difficult to answer. For example, many tests target recall, but there is a point when even a straight forward recall task that might become a problem solving task, especially when people are given enough time. What is missing is a clear understanding of when a test crosses that line (1981). It has been argued that when response time is included as a variable, we are not measuring ability, but ability under extreme time limits (Beilock, 2004). These are not the same construct. In the field of psychology, Cattell-Horn-Carroll Theory introduced an iunfluential framework of test types, and some of those are classified as speed factors, such as processing speed tests, where participants classify objects, or cognitive speed tests, where they might match a letter to a picture (McGrew, 2005). However, even when speed is not named as a factor, it can be a feature of other types of task in their framework (Alfonso, 2005), and it is challenging to say what conclusions we might reach in educational assessment from their findings.

There is evidence that response time says something about ability. Looking at both speed and accuracy, Maris and van der Maas found that candidates completing tasks at their limen of level of ability, tended to take the longest to complete the task (2012). Those who were either much lower or much higher in terms of accuracy completed the tasks more quickly. This suggests that short response times are associated with guessing behaviour, as well as expert behaviour. In fact, the correlation between lucky guessing and fast response times has been so strong that Wright (2016) suggested that all rapid responses in a multiple choice task should be scored as wrong.

If speed and accuracy are both used in scoring, there are questions around the relationship between the two. Are they conditionally dependent, as Maris and Maas' work suggested loosely, or ar they conditionally independent? Theoretically, it would be possible to uncover evidence of whether response time and accuracy are conditionally dependent by regressing the results of estimations of both if we could estimate speed and difficulty. Once we start to estimate both, questions around the reliability of measurements become more important.

Thurstone (1937), an early theorist on assessment, felt that time was at the core of educational testing. He noted that 'speed' was often socially valuable, as well as accuracy, in both simple and complex tasks, and so he added a third dimension to the Item Characteristic Curve to represent expanding response time. He argued along the same lines of the power test theorists, that the probability of answering the question correctly increased in direct proportion to an increase in the time limit allowed. For Thurstone, this response space was cylindrical, a curve going back into a third dimension, with the effect of additional time on a person's ability to answer becoming smaller as time expands.

$$\Pr\ (a\ \geq N \mid T) = \Pr(t \leq T \mid N)$$

Equation 3 Van der Linden's summary of Thurstone's formula for including a variable for response time (T) in the Rasch model (Van Der Linden, 2009).

Equation 4 shows, for example, the probability assumptions behind a test of reading ability in very young children based on Thurstone's 1937 model. The probability of α (the number of misreadings) occurring in a text of N words, could be measured over a time period T.

Thurstone's use of the term speed is significant. Van der Linden (2009) pointed out that the assumption of a linear relationship between time and success rates that Thurstone takes, irrespective of the nature of the item itself, is problematic for two reasons. Firstly, speed is not the same as response time, and questions are not equally spaced in terms of the time taken to complete them. Under this treatment in Equation 4, a simple word such as 'the' or a more complex one like 'procrastination' would count as one word each, but they almost certainly do not require the same amount of time to read aloud. Even items with the same level of difficulty may need more time if more iterative steps are to be taken, though. Secondly, there is an assumption that we can identify a known optimal time for completing the task. As mentioned above, the fastest correct result may simply indicate lucky guessing.

The approaches taken in games analytics towards scoring from response time seem to share a similar a linear interpretation of the relationship between time and ability, rewarding the fastest finisher. Van der Linden (2009) suggested an alternative way of including response time in educational scoring would be to measure 'speed' instead. If speed, not the mechanical measure of time, is the target variable, it needs to be treated stochastically, as an unknown random variable which is a function of the child, the response time and the task. Van der Linden proposed taking steps to address his second

concern by identifying a range of optimal response times. The boundaries of a range within which a time θ ability estimate lies needed to be isolated before scoring (Van Der Linden, 2009).

Once such an estimate is obtained, it would, theoretically, be possible to regress speed and accuracy to make decisions on whether the two are conditionally dependent. Molennar et al (2015) suggested instead using a generalised linear factor model on the output of van der Linden's hierarchical model for speed and accuracy. They found that it allowed nonlinear relationships between speed and accuracy to be explored. These are some mathematical solutions to validation, but in terms of interpretation at the level of individual students, it may not be possible to say for certain what construct is being measured. Experts sometimes slow down, the time limits may act as a control on the level of difficulty of the task, test takers may experience a lapse of attention, or they make decisions to try, guess and then move on. A large number of behaviours are captured in response time variables. The potential of using van der Linden's stochastic approach to speed will be explored with response time in this study.

4.5 How can repeated, iterative play be modelled?

Under van der Linden's model, response time may not necessarily be paradata, but a proxy for ability and therefore a principle variable for measurement. Another example of paradata gathered would be the use of hints and clues (Schnepp and Rogers, 2015), or the role of iterative play. The challenge of creating an iterative model of assessment was discussed briefly above in the section on DIF. Finding a way to deal with iteration was one of the main criteria of ECgD (Mislevy *et al.*, 2012). It would seem logical that having played the game once, the child would have a better chance of a stronger performance second time round. Do we want to capture this in the assessment model? And if so, how?

There is very little literature on dealing with 'iteration' in games, but if it is instead conceptualized as 'familiarity', there are some published studies to turn to outside of the field of Games Based Assessment (GBA). In the field of mathematics assessment or tests of scientific reasoning, a problem is often introduced in the form of a story or a scenario, and there may be a number of questions or tasks based on their comprehension of that story. After the first question, the child already had some familiarity with the scenario, and the tasks presumably become easier. Some of the early uses of Bayes in assessment aimed at exploring the subtle differences in difficulty that familiarity introduced.

Almond *et al.* (2015) pointed out that when modelling 'familiarity', it may seem like a function of the child parameter, as it is the child who reads the scenario, becomes familiar with it and then performs the task with this new comprehension. However, he advocated adding it as a function of the item. This was because the distribution of being 'familiar' was stable in the tasks, but not among the population of children.

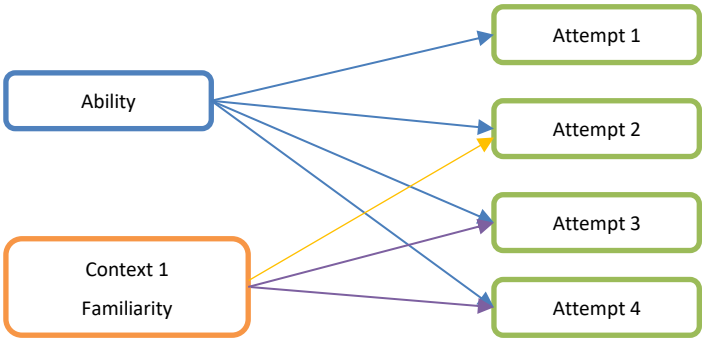


Figure 23 Simple DAG of modelling familiarity with the task, including it as a function of the task, not the child, adapted from Almond et al. (Almond *et al.*, 2015), page 173.

The DAG for introducing a familiarity weighting (Figure 23) shows how, using Almond *et al.*'s (2015) model, a context variable could, in theory, be introduced to reduce the weight of the evidence in the task slightly on subsequent attempts, when the child has had a chance to become familiar with the game. This allows the assessor to model the increased

facility of having a second or subsequent attempt, and could, in theory, also be used to model the use of hints, too.

The issue of iterative performance, however, may be more complex than the introduction of a familiarity variable suggests. The findings which will presented in this study show why this model is informative, but not optimal for gameplay.

4.6 Summary of methods of scoring from different fields

There is a rich body of research in the field of assessment that sheds some light on the conceptualisation of variables in gameplay data. These considerations were often missing from the literature on Games Based Assessment, with a few notable exceptions. At this stage, the parallels between findings in the wider field of assessment and behaviours in games are mostly speculative, but the research in this chapter provides some guidance on a more targeted methodological approach. It suggests that there are methods that could be used to model missing data, response time, iterations and inform whether a cohort or criterion referenced approach seem most suitable.

This chapter has focused on issues of conceptualisation, as often the choice of one model over another will be decided by non-mathematical considerations (Messick, 1987). Before going into the more specific detail of this study, which aims to test assumptions that the quantitative output of scoring games reflects performance, it is important to remember the wider picture. Validation of tests requires a range of approaches to identify not just the reliability of scores, but also other issues, such as whether there is adequate and appropriate coverage of a syllabus in any test. This investigation of new influences on scoring processes is not expected to result in a complete and satisfactory model, but it will draw on these insights into biases in scoring as a starting point.

This chapter was intended to add some depth and discussion to the methodological choices which will be laid out in the next chapter.

Chapter 5 Methodology

Previous chapters have discussed the way that gaming environments can perform a seemingly similar social function to traditional workbook tasks, homework sheets, or tests. However, when they are framed as a means to produce evidence for assessment analysis, games diverge greatly away from their paper-based siblings. The behaviours they elicit are different, and so the evidence of learning is different, and the way that we analyse it also needs to change. There are, therefore, many new challenges to solve, but this study in particular looks at the way that new variables in game play can be conceptualized, how these conceptualisations impact on modelling decisions, and in particular, whether there are ways to filter data from games to remove biases and improve the accuracy of measurement of to get a more intended, valid result.

The methodology for this chapter is pragmatic in many ways. The purpose was to observe what happened within the data set, rather than why it happened, and so a positivist design was used (Cohen, 2002). Positivist approaches can be better suited to describing the relations between behaviours (Vogt, 2012), and this is especially true when that data has been collated in survey form. During the application of the methods described below, there were many questions about why certain patterns of behaviour arose, but it was not the purpose if this research to investigate that further at this point. The study therefore is pragmatic, but also acknowledges that a large number of qualitative decisions and interpretations need to be made during a wider validation procedure (Messick, 1987).

This is a quantitative case study of a secondary data set. Although case studies are more typically associated with qualitative approaches, case study methodology is a contested terrain (Yin, 2003). With this case study, it was possible to build on existing theory around scoring games, outlined in the previous chapter, to simplify the design of the study. One of

the main features of case studies is that they are empirical inquiries carried out in a real-world context (Yin, 2011), and that is the context for this study.

A secondary data set was selected because the number of hours required to build an educational game meant that creating one from scratch was impractical. Sourcing a pre-existing telemetry set was a preferable option, although this also comes with restrictions on the type of data available. Initially, more than one suitable data set was acquired. After manual inspection, both data sets required a lengthy cleaning process, and the sheer size of gameplay data sets is a challenge in itself. For example, the dataset not eventually selected for this study was over 30GB in size and required a software kit to split it just to be able to open it on a regular desktop computer and inspect it. If more than one data set was to be analysed, it would have meant restricting the number of variables studied, given the available resources, and that, in turn, might have produced a more superficial analysis. In addition, the work carried out reviewing the literature concurred with the many other researchers that even since their findings (NRC, 2011; Eseryel, 2011; Hainey, 2012; Ifenthaler, 2014), optimal approaches and methods to study game data have not yet appeared.

For those reasons, only one data set was selected as the case. An obvious drawback of case studies is that the findings are often challenging to generalise to a wider population (Yin, 2003). Some methods of assessment, and in particular IRT informed approaches, often do transfer well across assessments with a strong qualitative differences (Bond, 2015; Wright, 1982). Mislevy *et al.* (2015) have cautioned that games appear to be an exception, though, and it seems necessary to adapt different assessment designs to different game mechanics. Despite these concerns, a case study still met the requirement of allowing an in-depth study of how games could be scored.

The research questions that this case study sought to answer are:

1. What counts as a valid attempt on task?
2. Does missing data impact the final score?
3. How can the game specific variables of response time and iterations be conceptualized and scored?
4. Within the game, how reliable do the results appear to be?

To begin to score games, it is necessary to devise a strategy for dealing with the many interfering variables, not least the element of choice and the impact this has on missing data. As response time and iterations transition from constants to variables, how they are conceptualized will impact on how they are measured. The final research question is speculative at this point, as there is very little in the published literature to suggest that games can be scored accurately.

This methodology chapter will begin by describing the data analysis strategy. This was carried out sequentially, and so attention will be drawn to where the specific research questions are addressed by each of the stages of analysis. The data strategy was broken down into 6 main stages and is outlined in the table below (Table 8). Although this was an iterative process in practice, with later stages informing and impacting on earlier stages, in this methodology chapter, it will be presented in the order below.

Table 8 Overview of the data analysis strategy

DATA ANALYSIS STRATEGY	
1 DATA COLLECTION	1.1 Identification of data requirements 1.2 Sourcing data sets 1.3 Play testing and gathering background information 1.4 Sourcing the software 1.5 Ethics clearance 1.6 Data acquisition
2 DATA PROCESSING	2.1 Data sampling 2.2 Re-structuring the rows and columns 2.3 Transforming data types
3 DATA CLEANING	3.1 Manual inspection of accuracy and quality 3.2 Identification of incomplete and duplicate cases 3.3 Identification of thresholds 3.4 Filtering and production of the data sets
4 EXPLORATORY PHASE	4.1 Production of descriptive statistics on the distribution of key variables 4.2 Outlier detection 4.3 Exploration of contextual information in the data set (age, period of play, patterns of play, reason for play) 4.4 Verification of derived values
5 MODELLING PHASE	5.1 Production of difficulty estimates for the different data sets for a) HighScore, b) mean response time, and c) the first five iterations 5.2 Production of ability estimates on students, and production of difficulty estimates from a data partitioned on high and LowAbility 5.3 Estimation of the degree of error and stability of results 5.4 Estimation of the impact of including dependent variables in the analysis model
6 DATA PRODUCTS	6.1 Production of graphs, charts and tables for the distribution of key variables

The rest of this chapter will outline how each of those 6 phases were addressed, the methodological and ontological questions around that, and the methods that will accommodate these concerns.

5.1 Data collection

5.1.1 Identification of the data requirements

Seven general requirements for the data set were drawn up that were compatible with the intended uses of the secondary data set:

1. It must have some elements of gamification (e.g. rapid play, competition, non-linear structures (Fullerton, 2014)).

2. It should assess skills that have a value in real-life contexts (i.e. not spell casting or magic).

3. There must be pre-existing substantive work on the skill targeted to avoid unnecessary complication, informed by the experiences of Von Davier and Halpin (2013).

4. A curriculum-based paradigm behind the assessment design would be desirable, in other words, the assessment should be aligned to statements of learning goals and the results will be criterion-referenced against those goals, not norm-referenced (Baird, 2018).

5. The game and the scoring within it should have gone through some kind of trialling or expert judgment to minimise the impact of poor test design.

6. The data set must provide insight into the key target variables of iteration, response time and either an element of choice or a random presentation of tasks.

7. It must be large enough (n >200 players) to evidence learning.

5.1.2 Sourcing the data sets

Initially open source data sets were consulted, but this quickly proved to be too optimistic.

At the time of investigation, 17th September 2017, no educational data bases were available on the Open Access Directory (OAD) (see http://oad.simmons.edu/oadwiki/About_OAD for more information about this data repository). A second site, DataWorld was consulted on the same day, and of the 509 data sets returned in a search for 'assessment', none referred to telemetry data (see https://data.world/ for more information on this repository). It is unfortunate that no data sets are currently publicly available for study on open source as scoring gameplay data is a complex challenge that will require the input of many researchers to overcome.

Sourcing a suitable secondary data set from proprietary sources began by snowballing requests among researchers who were working in the provision of games for possible access to data sets, beginning with a few published authors, and researchers in games design at the University of Southampton. Around 50 people were approached, and around 6 data sets were offered that failed to meet the criteria. An additional data set met the criteria, but permission to use it was declined at later stages of discussion. In the end, permission for two data sets which met all the criteria was given simultaneously, and so both were acquired. The process of searching for data sets stopped at this point.

The first data set was a series of Key Stage 2-3 / US Common Core Maths Games called MangaHigh, provided by Blue Duck Education. The second, QuizYourEnglish was provided by Cambridge University Press (CUP), and targeted English as a Foreign Language (EFL) skills, testing B1-B2 Level Proficiency in the Council of European Framework for Languages.

The way that these two data sets met the data requirements are summarised in Table 9.

Table 9 Satisfaction of the case study criteria

Criteria	Qualities of the data sets
1 Gamification features	Both games have game mechanics such as direct competition, speeded play, or manipulation of avatars along with leader boards, coin collection etc.
2 Valued skills	Both the primary / early secondary level Maths curriculum and the English as a Foreign Language (EFL) curriculum are of central economic importance in many countries, and are part of obligatory state-funded education.
3 Pre-existing substantive work	There is a large amount of pre-existing substantive work on testing Maths and EFL.
4 Curriculum-based	Both games are aligned to pre-existing international syllabi which have been extensively trialled and practised.
5 Game trialling	Both games were designed by professional teams with a range of computer and subject specialists inputting on the design process and have completed the initial trialling phases.
6 Target variables	Main target variables are present.
7 Size of the data set	MangaHigh has over 1 million users in their database, producing over a billion data points. The sample of gameplay extracted from the QuizYourEnglish dataset contained 170,678 unique gameplay cases.

Both provided detailed and useful data to investigate the research questions. As mentioned in the introduction, because of the extent of data processing and cleaning and the fact that there was a large amount of new information to process and understand, this thesis quickly became characterised by increasing complexity. On consideration, only the MangaHigh set was chosen for investigation in this case study. The games in MangaHigh were chosen because they had a variety of game mechanics and some features more unique to game delivery of assessments, such as the use of bot competitors and some opportunities for real time collaboration.

5.1.3 Play-testing and gathering background information

Play in MangaHigh takes place in rich online environments. The games are dynamic, with colourful interfaces, music, sound effects, animations and avatars. Successful outcomes are dependent on both maths ability and controlling the games mechanics. The game mechanics were chosen by each design team to best illustrate the mathematical function and so they vary. For example, knowledge of angles and triangles is tested by helping a spider to build its web with triangular shapes in the game *A Tangled Web*.

Figure 24 Game interface from *A Tangled Web*, showing a robot spider's web ©MangaHigh

As Figure 24 shows, the interface for *A Tangled Web* has a range of features such as stop clock, score board, help button, pause and restart functions. Keyboard and mouse skills are needed to complete the games.

Four broad types of games mechanics are used in the games:

5.1.3.1 Drill + animation

In these games the learner answers a quiz question and is rewarded for success with an engaging animation. There is often an additional gaming element, such as direct competition with other humans or an automated intelligent agent, or time-pressured play.

Figure 25 Screen shot of Sundae Times ©MangaHigh

Figure 25 shows the multiplayer game Sundae Times, where players type their answer to a question and are rewarded for correct answers with lively cheering, music and an additional scoop of ice-cream, which falls onto their ice-cream sundae with sound effects. The player with the tallest sundae 'wins'. Medals are awarded depending on the number of questions they have answered correctly. Some games in this genre are single player, but Sundae Times is a multiplayer with competitive elements.

5.1.3.2 Avatar or object manipulation

In these games the student manipulates objects in the game, such as an avatar or a line, as part of their response.

Figure 26 Screen shot of *Piñata Fever* ©MangaHigh

Figure 26 shows a screen shot from the game, Piñata Fever, which has a Latino festival

soundtrack, and the player drags the avatar along the number line to indicate the correct response. The avatar takes a swing and smashes the piñata if the answer is correct, with the whole crowd celebrating to mariachi music.

Figure 27 Screen shot of Flower Power ©MangaHigh

Figure 27 shows another game in this genre, Flower Power. This game tests ordering of fractions. The flowers need to be ordered from smallest to highest value by dragging the 'new shoot' to the correct place along the stem. Players can control the complexity of the challenge by harvesting the flowers. *Flower Power* has a very calming soundtrack, swooshing wind effects and smooth animations, reminiscent of the commercial game *Flower* (ThatGameCompany, published by Sony Entertainment). It also has elements of systems management as the child can choose how many flowers to manage at a time.

5.1.3.3 Platform game

Platform games are played across two-dimensional landscapes. The player manoeuvres a character across the screen, avoiding obstacles or locating and using objects that might assist them (Fullerton, 2014). There are often 'walkovers' which also give the player a boost or a reward.

Figure 28 Screenshot of Jet Stream Riders ©MangaHigh

Figure 28 shows a screenshot from the platform game, Jet Stream Riders. Players help balloon avatars cross the landscape in multi-player format, competing live against each other, or, if no-one else is online, against a robot competitor. In the case of this game, the robot is a pre-recorded performance that is replayed, not an intelligent agent that responds to input. Correct responses when the player is over the 'walkover', in this case, the jet stream are rewarded by propelling the balloon forward faster.

5.1.3.4 Shoot'em up

The aim in shoot'em up games is usually to shoot as many targets as possible. In the case of these games, the targets were robotic, and the pace was fairly slow.

Figure 29 Screenshot of Bidmas Blaster ©MangaHigh

In Bidmas Blaster (Figure 29), for example, the defensive shield of the advancing aliens

can be taken down by solving the equations. The aliens can then be shot down.

All of the gameplay environments are directed at the 7-16 age range, although the minimum age was estimated at around 4, and there was a consistent spread of ages up to an upper maximum of 18, which was artificially imposed by the research criteria. There were a number of much older users, up to the age of 50, but these were most likely to be teachers, and they were excluded from the sampling process.

In the MangaHigh games, learner outcomes are aligned to the key stage 2 and 3 outcomes and parts of the GCSE syllabus in the UK. In the USA, where the majority of its users are based, it is aligned to the Common Core Curriculum for maths grades 3 - High School. A table with all of the games and their genres and the areas of the mathematics syllabus that they cover can be found it Appendix B.

The interactions students can have within MangaHigh are quite short. In general, before any cleaning up of the data, game play was between 81 seconds (around a minute and a half) and 403 seconds (6 and a half minutes), depending on the game and the player. There is evidence of very short play (under 5 seconds per game) suggesting browsing behaviour, and although some games had a cut-off point, others evidenced very slow play (over 10 minutes per game) suggesting wandering off. Players cannot currently chat or communicate with one another in the game and collaboration, for example in the game Beavers Build It, is done by clicking on a suggested value to use.

MangaHigh is primarily sold as a homework practice tool, but it has an evaluative element, awarding a mixture of medals based on the percentage of correct answers out of the total possible (which reflect ability) and coins (which reflect a mixture of ability and perseverance). In this way, scoring addresses the dual purposes of providing feedback to

teachers on the degree of assimilation of the concepts and motivating the learner to continue. There are also medal tables and leader boards. Learners who have demonstrated adequate understanding of the concept are awarded a Bronze award. This can then be upgraded to Silver and Gold, which demonstrates increasing levels of mastery. The player can repeat games to accumulate more coins, but the developers have put considerable effort into making sure that the child cannot score a Bronze by repetition alone.

At the time of data extraction, membership of MangaHigh was only sold to institutions, and particularly in the United States, to purchasing teams of several schools from the same school district. A public version has since been launched. The data set involved real school students. All of the cases provided longitudinal data on their progress from the time they began to play, which varied, to the date of extraction of the data, on 30th January 2018.

The scoring in MangaHigh was at a grade band level for each game, and a more granular level of scoring was not available, which was a limitation of the data set. The average response time for each game was around a minute and a half, which is short for test conditions. No child had played all of the 21 games, and the data tended to go deeply into performance on a few areas of mathematics, rather than offer insight into performance over a whole curriculum.

No data was gathered on the gender of the participants, and all demographic information that was collected was self-reported, and therefore subject to error. Data for multilevel modelling, such as school or class associations, were not collected for reasons discussed in the ethics section below. In addition, nothing was known about how these children were playing, such as invested effort each time they played. A different research design would be needed to answer that question.

Soft skills such as collaboration in Beavers Build It, and Systems Management in Flower Power were present in the games, but the focus was on the mathematical skill. In general,

the problems that children were solving in the games were more complex and more procedural than they might encounter in, say, a paper-based quiz. Soft skills were not the sole focus of the assessment, but more an additional level of complexity.

5.1.4 Sourcing the software

Structured Query Language (SQL) was used to extract the relevant data tables, which was in string or numeric .csv format. Data were cleaned up in Excel, using Power Query. SPSS was used in the initial exploration of the data sets and for the production of descriptive statistics and hypothesis testing when comparing model results. Finding software for the main analysis was challenging. Bayes for IRT is a fairly specialised field. Software was available in STATA, and code has been made available for WinBUGs (Fox, 2010; Almond, 2015), but a way to manipulate missing data in the WinBUGs programme was not found after consulting with several experts. WinBUGs automatically imputes missing values, and it was out of the scope of this study to trial whether that imputation made a difference to the results. It appeared equally problematic in STATA. After considering costs as well, Winsteps FACETS, which uses a frequentist approach to IRT, was finally chosen. It allowed for control over how missing data is modelled, was able to produce Partial Credit scoring results and it allowed the incorporation of additional facets, which in this case was important to model the iterations. The software uses an iterative version of the PROX algorithm, written in Absoft Pro Fortran and MS Visual Basic 6.0, and uses Joint Maximum Likelihood Estimation (JMLE), to iterate the calculations until a pre-defined point at which convergence between estimates no longer improves (Lineacre, 1998).

5.1.5 Production of descriptive statistics

Initial distributions of the derived values for the mean score and mean response time were created as a starting point. In this early stage of the study, a pilot of 30 random cases selected from the final sample, which will be described below in 5.2.1, was carried out in advance of deciding what descriptive statistics to produce. It gave an impression of the shape of the data before producing descriptive statistics for the whole data set.

From the box plots for response time, it appeared that in gameplay, speed did appear to be a function of the game, the child and the response time as van der Linden posited (2009). Looking at the total response times for a pilot of 30 cases in the competitive times tables game, Sundae Times, there were also different patterns of behaviour in the different band groups, suggesting that it was also a function of the band score.

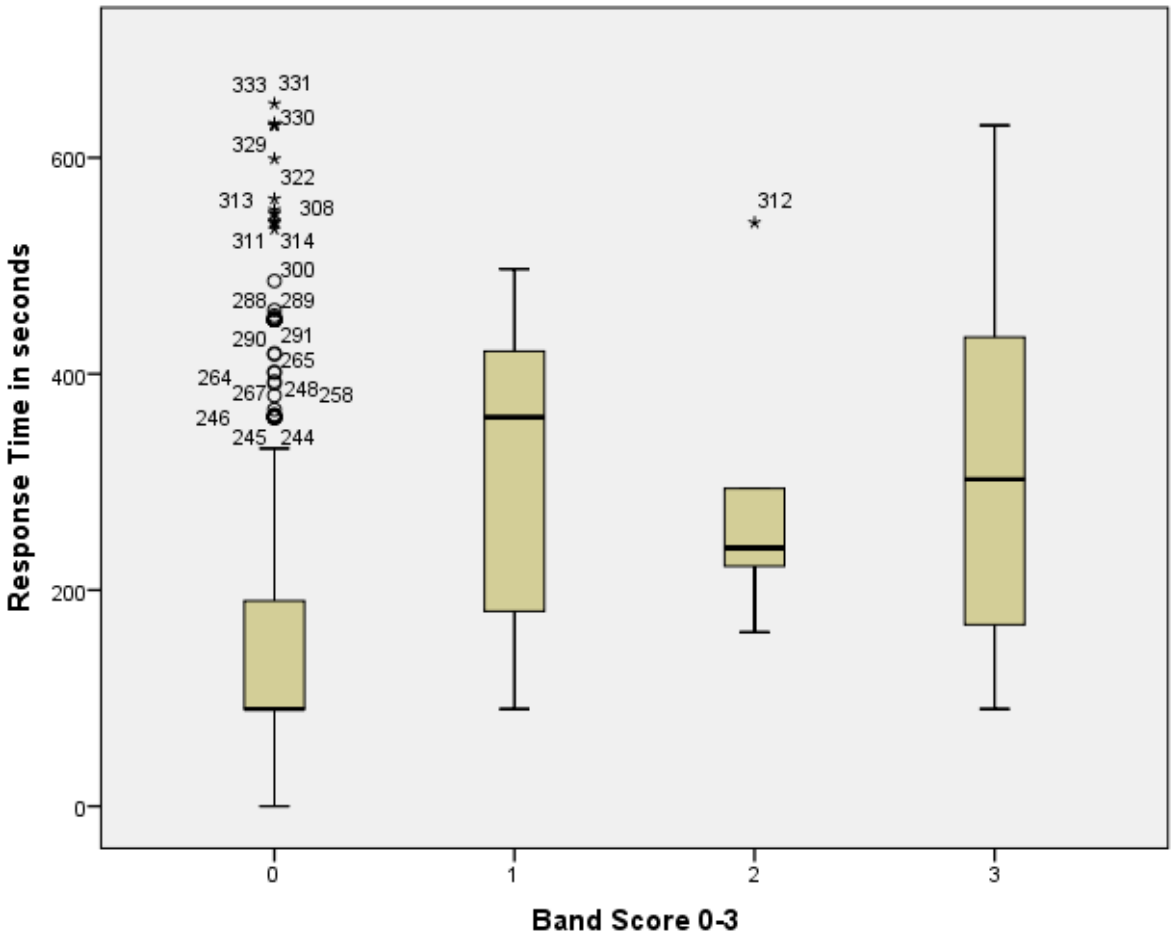


Figure 30 The distribution of response times for the game Sundae Times across the 4 bandings

Figure 30 shows the box plot for response times for SundaeTimes. Ignoring Band 0, the non-scoring band, a response time of 300 seconds would be relatively fast for a child in the bronze medal category, Band 1, relatively slow for the silver medal category, Band 2 and very close to the mean for the gold medal category, Band 3.

Another issue that emerged from the early pilot and production of descriptive statistics was that multiple attempts were very common. Some children played the same game over 500 times. It was initially anticipated that iteration would be dealt with through adding a familiarity weighting, as discussed in 4.5, but framing the treatment as a problem of which measure to use, particularly in terms of measures based on central tendency, seemed more appropriate from this early pilot stage.

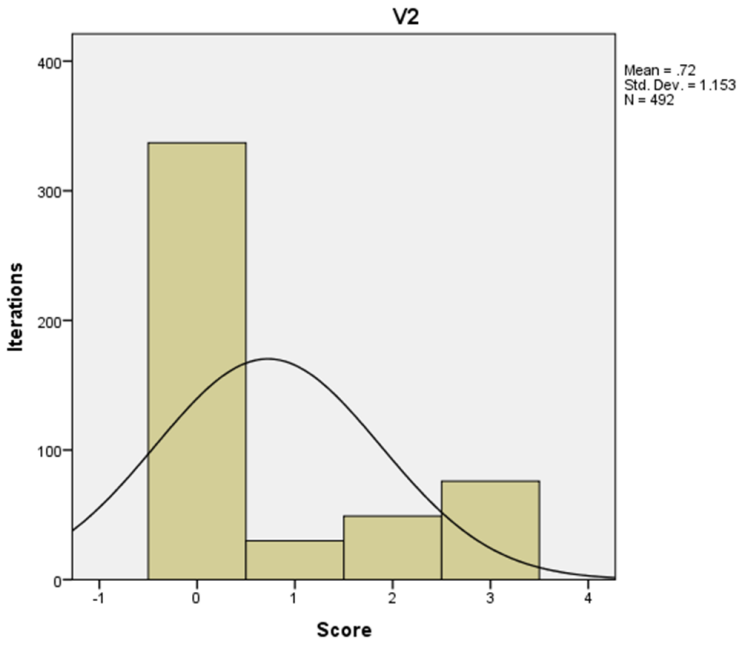


Figure 31 Distribution of scores for the same child over multiple iterations of the same game

Figure 31 shows a sample of one child playing the same game, a times tables platform game, 492 times. Options for a measure to score this child might include using the mode (Band 0), the location of the mean of .72 (Band 1), the location of the upper quartile limit of 1.873 (Band 2), the child's HighScore (Band 3), or the most recently recorded score (Band 2). Almond's (Almond, 2015) model of familiarity described in 4.5 was not

intended to work with around 500 levels of familiarity.

The descriptive statistics of the mean score for each game was generated. The mean response time for each game was calculated separately for each of the four bands. In addition, statistics were generated to show how many times each game had been played.

5.1.6 Ethics

Ethics approval for this project was received from the University of Southampton Ergo for the MangaHigh data set, ID 31046, on 18/12/2017. Ethics extends for the duration of this course of study, until November 2019. Please see Appendix C for the full ethics approval documentation.

From self-reported data, the learner sample contains mostly minors under the age of 12, some potentially as young as 4. Many aspects of education have increasingly fallen within the remit of information society services and rules around data use with minors is evolving. In the EU, children under the age of 13 are not considered to be old enough to provide their consent to data use under the GDPR (Article 8, 2018 reform of EU data protection rules. European Commission. May 2018). In the USA, following the failure of *InBloom* in 2014, hundreds of separate pieces of state-level legislation on child data use have been created to address parental concerns. Other children in the data set were in disparate regions of the world. Given the multiple rulings in force with this sample population, a very conservative approach was taken. No identifying data, such as the name the child uses in the game, or the school that they were studying in, was extracted.

Participant consent to use the data for educational research purposes was granted to the data owners, Blue Duck Education, by representatives of the school or purchasing district at the time that they agreed to the terms and conditions of play. The conflict of interests between fostering greater understanding through research and respecting the child's rights

is complex. Allowing school staff to give consent can be problematic as employed representatives have a tendency to move jobs. If a child or parent wishes to withdraw consent at a later date, the employee has to be traced, which may not be easy. It is also a matter for concern that the children in the sample had little choice over their data use. The games were sold as a homework package, and the children may actually have faced punishments if they had refused to participate. For this study, the use of data from GCSE results was taken as a precedent and guideline. Participation in producing GCSE results data for academic study is also a non-elective process, but steps are taken to reduce the risks that pseudo-anonymising the data will fail and expose the child to identification.

5.2 Data processing

The population data set for MangaHigh contained over a billion data points at the time of data collection. A sample was necessary to reduce the size of the extracted data set to a size that can be processed on a desktop computer. Research suggests that cluster sampling is the best approach when a known variable or criteria for cases exists (Ahmed, 2009). As many of the pre-existing assessment models work best on data sets where there is a minimum of 30 samples of learner performance (Bond, 2015; Wilson, 2004), a large enough data set was also a requirement. The top 95-100% contained records for the game developers and staff working for MangaHigh, who regularly playtest the games, and there were no data markers available to remove this subset of the population from the data set. Given that the criteria were known (Ahmed, 2009), and there were no precedent studies on sampling in this specific circumstance, the cluster sample took the 90-95% highest users, to meet the criteria for adequate data and exclude the developers. The sample produced a data set with 450,000 rows of data and 10 columns. Ten tables were extracted, and the variables were pre-determined by the information available in the log file codings (Table 10).

Table 10 Variables extracted in the data set

Name	Meaning	Type	Missing
UserRef	A unique identifier for each player	String	None
ObjectiveID	A unique identifier that maps to a Core Curriculum target. Different game levels may have more than one objective in each game.	Numeric	None
PlayedAtstart	The date of gameplay	Date	None
Achievement	A categorical variable with the four game award levels, None, Bronze, Silver and Gold	String	None
SecondsPlayed	An ordinal variable up to 3 decimal points indicating the duration of the game.	Numeric	None
Source	A categorical variable indicating whether the task was assigned by a teacher (recommendation), or the students' choice (freeplay)	String	None
Upgrade PointsEarned	A categorical variable ranging from 0-3 indicating an in-game rewards system	Numeric	None
UserCountry	A categorical variable indicating the student's country of origin	String	None
GameName	The name of the game as presented to users	String	None
GameTitle	A description of the Maths skills targeted by the game	String	None
GameDescription	A description of the game objectives	String	None
DateOfBirth	Self-reported date of birth.	Numeric	None

Tables that would allow multilevel modelling, the child's class grouping and educational institution affiliation, were not extracted for the reasons explained above in the ethics section, 5.1.6. Demographic details on the child's age and location geographically were extracted to provide context to the performances (Table 10). More information about how these variables were used in the study will be given in 5.2.3 below.

5.2.1 Data sampling

It quickly became clear that the 450,000 rows extracted needed considerable work to make it parse through the IRT software. Transforming the data, as laid out below in 5.2.2, was time consuming, and manual or detailed inspection of such a large document was highly

restricted. In addition, with almost half a million rows, the software crashed regularly because of the size of the file. Given the time constraints, the aim of the study, which was to describe in some detail what was happening in the data set, and the limitations of the software, the sample size needed to be reduced further. A case was taken to be the collection of performance data for all of the games played by one child.

The calibration process which estimates the discrimination powers of a test works with a sum of squared errors analysis between expected and observed behaviour (Bond, 2015), and is a form of Chi Square analysis, as discussed in section 2.3.2. Research on Chi Square tests suggests that the benefits to be gained from increasing the sample sizes drops off dramatically after a sample size of 200 cases (Hintze, 2001; Cumming, 2013). In fact, they argue that a sample of n=200 has around 98% power to detect an effect size, and so a random selection of 200 children from the 90-95% highest players list, using a Python random.sample () function on the UserRef variable. There were 32,213 individual games recorded in the data set for these 200 children.

5.2.2 Re-structuring the rows and columns

The SQL extracted tables presented the target data (the scores) in long format, shown in Figure 22 in section 4.2.4, and the following information (Table 11) in wide format was used to parse the data through the software. This produced a data set of 44 columns and 14,220 rows, with a potential of 298,620 cells before taking into account the extent of missingness. This data set was more stable.

Table 11 Wide format data structure

User Ref	Attempt	Game 1		Game 2		Game …n	
		Time	Band	Time	Band	Time	Band
134	1	180.00	Bronze	56.78	NULL	87.46	Silver
134	2	156.76	Silver	83.09	Gold		

The waves of data or attempts were added in the form of additional rows, with a column to

designate the number of the attempt (Table 11). FACETS was able to calculate a difficulty estimate on each attempt separately. This treatment is more commonly applied when, instead of multiple attempts, there are multiple judges (Lineacre, 1998). The function of the software can be helpful when cross moderation of assessments is used. Cross moderation compares multiple holistic judgements of performance (Baird, 2018). The same process of estimating different difficulty parameters for attempts has not been reported in any of the literature, but the necessary condition of having more than one score for a similar performance seemed to be satisfied by this analysis. After using response time as a context variable to decide if an attempt was valid, which will be discussed in detail below, score and time models were separated.

5.2.3 Transforming data types

The case or unit of measurement is one whole game, and the score recorded under 'Achievement' represents a partial score value of null (no medal awarded), bronze, silver and gold for each of the games played. To parse these through the software, the string values were changed to integer values, as shown in Table 12.

Table 12 Interpretation of the medal values

Medal	Assigned score	Interpretation
NULL	0	The game was played, but there was insufficient evidence of comprehension.
Bronze	1	The game was played, and sufficient comprehension was evidenced.
Silver	2	The game was played and superior comprehension was evidenced.
Gold	3	The game was played and mastery of the concept was evidenced.

The scoring bands were therefore 1-3 (Table 12), with increasing levels of mastery. The 'Upgrade Points' variable referred to coins and points collected for motivational reasons, rather than achievement, and so the decision was taken not to include this in the

performance analysis. The following minor changes to the tables shown in Table 10 in the introduction were also made:

1. UserRef is the person identifier. This had to be changed from a string to an integer variable to parse it. Also, the application of a random number code made it harder to link the data back to a specific child's records in the main database.

2. SecondsPlayed was the recorded raw response time. The response times were rounded to the nearest whole second where necessary to change the data type from decimal to integer in order to parse it.

3. PlayedAtStart recorded the start time and date for each game. By taking the very first date and last date in a child's records, this was used to create a new variable, LengthOfPlay, which roughly estimated the total number of months played.

4. DateOfBirth – Date of birth recorded the child's self-reported year of birth. It is possible that the children did not fully understand this concept if they were asked to complete it themselves, but it gave some context to the gameplay. Using PlayedAtStart and DateOfBirth, two new variables were created to roughly estimate the age of the child at their first recorded game, AgeAtStart, and at their last, AgeAtEnd, by extracting the year of birth from the year of play. This was a very rough estimate of age.

After processing the data, further work was carried out to clean the data, which will be explained below.

5.3 Data cleaning

5.3.1 Manual inspection of accuracy and quality

There were no missing values in the SQL extracted data set, but the data was manually

inspected for unusual patterns or signs of corruption and mis-codings, such as very short or very long play times. One issue was observed with the platform game, JetStreamRiders. Sometimes even silver medals were awarded for times as short as 2 seconds, suggesting that this was either a mis-reporting of the results, or that the player won on the basis of a competitor conceding. The games designers confirmed that this game was an anomaly among the group in that it was possible to win if all of the other competitors were human and all of them conceded. If the bot played, it did not concede.

The game Beavers Build It, the collaborative addition game where children choose the correct sized blocks to build a wall, only recorded response times of 300 seconds, with no exceptions, even in the 0 scoring band. This suggests that there was an error in recording the response times as, although this is an engaging and easy game, it seems unlikely that in 692 separate plays, none of the children opened the game and then closed it in less than 300 seconds. Beavers Build It was excluded from any further analysis of response time because time in that particular game was a constant, but it was included n the analysis of band scores. Finally, the game Wrecks Factor, which uses factorization of simple and hard quadratics to identify co-ordinates of a ship in distress, appeared to have a fixed completion time limit. All the scoring bands had the same mean response time, and all of the early finishers received a NULL score. It was therefore also excluded from specific analysis of time.

The initial 200 sample contained a number of people with ages that were not characteristic of the target population. Six gave a year of birth that would make their age higher than 25 years old; one gave a year that would make them more less than year old at the start; and a final student responded NULL to this question. These were rejected from the sample and eight more players were randomly sampled.

5.3.2 Identification of incomplete or duplicate cases

The data set was manually inspected to understand the extent of missing data and the nature of it. In addition, checks were run, and one duplicate set of data for one user in the sample was removed and replaced by the next selection from the random sample.

5.3.3 Identification of thresholds

Box plots were produced for the response time for each scoring band (0-3) within each game to be able to identify outlying behaviour. These were plotted and collated to get an overall impression of how the games were being played, and to identify extreme and outlying values, discussed in the next section.

5.3.4 Filtering and production of the data sets

In total, eight different data sets were created to allow the research questions to be explored. The filters for inclusion or exclusion were plausible response times and ability.

Table 13 Final data sets used in estimation

Data set	Description
Complete data set	All of the cells used in the final sample of 200 learners
Without Zeros	All of the score + time pairs where response time values greater than .5 from Set 1
OverTenSecondsOnly	All of the score + time pairs where response time values greater than 10.0 seconds were removed from Set 1
Main	All of the score + time pairs where extreme values at the upper boundary capped to the top end of Q3 in the Interquartile range from Over Ten Seconds.
MeanScore	The player's MeanScore and response time in each game from Capped upper boundary.
HighScore	The player's highest score and fastest response time from Capped Boundary.
FirstFiveAttempts	The FirstFiveAttempts at the task only were used.
DEMOGRAPHIC	A separate data set was produced with the variables for the child, their date of birth, date of their first play and country of origin

The production of the data sets in Table 13 was in order to make it possible to address the

research questions. Research Question 1 centred around what counts as a valid attempt on task, and response time appeared to be a contextual variable for ‘meaningful attempt’. In CompleteDataSet, a number of recorded gameplays lasted ‘0.05’ seconds. In fact, in the final sample data set, there were 42,663 such examples of games. The ‘0’ time score values, or those that rounded down to 0, were discarded to produce a second data set, called WithoutZeroTimes Data set (Table 13).

There were still many values in WithoutZeroTimes that were very short, and it was unclear at what point an attempt was no longer a meaningful attempt and would be best ignored. There was no guiding literature on removing these values in the filtering process. It seemed most likely that response times of less than 10 seconds, for example, evidenced browsing behaviour, or something else that was not related to the level of the challenge of the tasks. A third data set, OverTenSecondsOnly, was created with all values less than 10 seconds discarded as invalid responses.

After removing Outliers to produce the Main data set for analysis, which will be discussed in 5.4.1 below. The First Five Iterations, Mean and High scores data sets were produced to understand how iterations might be modelled, and the final data set, the Demographic data just contained information on the child for contextual information.

5.3.5 Verification of derived values

The MeanScore and HighScore data sets were produced in Excel, and they were manually inspected to check for anomalies in estimating the mean values, and high/fast scores for. A number of repeated values were found. PowerQuery was used to identify unexpected values.

5.4 Exploratory phase

Several of the researchers in the literature review made reference to the fact that a lot of

behaviours were emergent, in other words, they observed patterns of behaviour that had not been documented before (Vendlinski *et al.*, 2010; Davier and Halpin, 2013; Graf, 2014; DiCerbo, 2014; Mislevy *et al.*, 2014). For that reason, a range of methods were identified for the exploratory phase, to allow a full consideration of patterns in the data.

5.4.1 Outlier detection

The band score was a categorical variable and therefore there were no outlying values, but the continuous variable of response time did contain outliers in the OverTenSeconds data set. As already mentioned in section 5.3.4, and this idea will also be developed in more detail below in 5.5.1, short times that appeared implausible were removed. After that, there were no outliers at the lower boundary, but there were many at the upper boundaries.

Outliers can distort data analysis, particularly when reporting on mean values (Cumming, 2013), and unbounded response time variables can create issues interpreting the data. There is no one definition of an outlier, but in general, it is one that lies outside of the other values, and 'extreme' values can hold a lot of weight on the data (Cumming, 2013). There were substantive reasons to consider some values implausible. A response time of 10 minutes or more may not indicate effort and perseverance, but rather that the child has wandered away from the screen. However, as this information could only be obtained from a different research design, such as key stroke analysis or observation of play, and there was no substantive research in this field, a statistical approach to identifying and dealing with outliers was taken.

The point when a value becomes extreme is still a contested field several centuries into the debate, and a Tukey fence approach was taken, using the definition of extreme as greater than or equal to Q3 + 3(Q3-Q1) (Hawkins, 1980) to identify values from OverTenSecondsOnly. Rather than remove the cases completely, the decision was taken to cap the time at a value of Q3 + 3(Q3-Q2)-1. This allowed the cases to stay in the data set

but made it possible to explore the variable for response time within stricter parameters. This capped data set formed the Main data set. The research design used here could not confirm if this is an optimal point, but steps were taken to measure the impact of this decision, outlined in section 5.5.1 below.

5.4.2 Exploration of contextual information

Using the Main data set, a frequency tally was carried out to report on:

1. The popularity of each game measured by the total number of times that the game was played.

2. The intensity of play measured by the total number of times each child played any game.

The demographic data set was used with the variables of the child identifier, CountryOfOrigin, AgeAtStart; AgeAtEnd; and LengthOfPlay, with values for LengthOfPlay informed by the Complete data set. This was used to produce bar charts showing where the children had come from. Mean, mode and median and standard distribution value were obtained on their age at the start of play, at the end and the average length of play, as well as pie charts and bar charts as appropriate to explore these variables, and these will all be shown in section 6.1.

5.5 Modelling

5.5.1 Research question 1: What counts as a valid attempt on task?

There were three main steps to answer this first research question. In the first instance all cases where the response time was less than or equal to .5 seconds were identified and removed from the data set. The next decision was around what was a plausible minimum time. After playtesting the games, they all seemed to require a minimum time for the game

to launch and for play to start, around 10 seconds. However, a small number of games that had been played less than or equal to 0 seconds appeared in scoring bands. For that reason, the impact of removing these values was estimated by calculating the MeanScore and mean response times for WithoutZeroTimes and OverTenSecondsOnly. The sets of results were compared using a paired samples t-test to see if discarding these recordings had a statistically significant impact. A paired samples t-test was chosen as the mean of the same data set was measured twice, before and after the adjustments (Walliman, 2017). The hypothesis for this was that:

> H_0 = The exclusion of games with response time values of less than 10 seconds has no significant impact on the overall MeanScore.
>
> H_A = The exclusion of games with response time values of less than 10 seconds has a significant impact on the overall MeanScore.

This test was carried out to understand the impact of the decision to remove these values, if a means-based measure was to be used.

As there was little guiding literature available, it was also decided to test the impact of removing games which had recorded extreme values. These were considered to indicate wandering off behaviour, and because the sample size was considerably larger than 20, a one-tailed paired samples t-test was considered appropriate (Walliman, 2017).

The hypothesis was:

> H_0 = The capping of response time extreme values at the Q3 + 3(Q3-Q1) limit had no significant impact on the response times in any of the band scores for the games.
>
> H_A = The capping of response time extreme values at the Q3 + 3(Q3-Q1) limit had a significant impact on the response times in any of the band scores for the games.

5.5.2 Research Question 2: Does missing data impact the final score?

The extent of missing data in gameplay data sets is vast. It was mentioned above in section 5.3.1 that the CompleteDataSet contained 14,220 rows, and had the potential to produce data on 298,620 cases or games, but that would require every child to play every game the same number of times. Having removed untargeted behaviour and implausible values to produce the Main data set, only 64,090 cells recorded data on a game. The remaining 533,150 cells were blank. This means that only approximately 12% of the available cells actually contained information. This imbalance was caused largely by the fact that some children repeated the same game hundreds of times. Each additional attempt by any one child created an additional, and largely empty, row in the data. The software could not parse that extent of missing data.

One way around the problem of missing data was to reduce the number of zeros by taking just one value for each game, and so a Partial Credit model, described in section 2.3.1, was carried out on the highest score of each child. An additional way was to limit the number of possible attempts, and an additional Partial Credit model was carried out on a restricted number of attempts, which will be described below in section 5.5.3 on dealing with additional attempts.

Missing values were not replaced with a zero value, for the reasons discussed in section 4.2.4. Much of the data that was missing seemed to result from choice, and was forced by the games design, and that would suggest that missing values should be treated values as ignorable was taken as the default. Nonetheless, after piloting the methods with a small random sample of 30 children, there was some evidence that choice might be related to ability.

Some of the games offered ‘regular’ and ‘lite’ versions, which meant that the games mechanics were the same for both versions, but the mathematics in the lite version was

selected from parts of the mathematics syllabus that are generally presented earlier to children. In the pilot, the lite games were not consistently easier, and it was possible that the children were self-selecting into games based on their ability. If that was the case, the different sets of children may well have experienced the different levels of the games equally challenging, relative to their own ability. Isolating a group of children who had played every game was not an option. There was no child who had tried all of the games. Instead, the HighScore data set was partitioned in terms of ability of the child. After an initial Rasch processing of the HighScore data set in the FACETS software, using a Partial Credit Model, three groups were identified, a HighAbility group greater than or equal to +1.72 logits, a LowAbility group less than or equal to -.72 logits, and a middle ability group between those two values. No further action was taken with the middle ability group to allow for error in estimating their ability. This left a set of HighAbility children (n=80) and LowAbility children (n=79). The Rasch estimation of the difficulty of the task was repeated, estimated on each group to see if the order of difficulty of the tasks had changed, which is a common indicator of stability in Rasch modelling (Bond, 2015).

5.5.3 Research Question 3: How can the game-specific variables of response time and iterations be conceptualized and scored?

There was very little in the literature on how to deal with either of these variables specifically in a gameplay data set, and there were a number of confounding variables that influenced choices and patterns of behaviour regarding these variables, and both needed to be conceptualized as well as measured.

5.5.3.1 Response time

Box plots will be produced from the Main data set for visual comparison purposes. Box plots can show patterns of behaviour with continuous variables over large data sets (Walliman, 2017). With regard to the bot described in 5.1.3.3, no guidance was found on

how to deal with bot times. The bots will be identified from repeated data trails where in use (Jet Stream Riders and Sundae Times), and their performance will be used for comparison with human performances.

Although van der Linden (2009) provided models for how to conceptualise and include response time, there were a number of other factors to consider for gameplay. It was the intention to include a measure of response time as a proxy for ability, using that model but because response time is an unbounded continuous variable, the mathematics to carry out the estimation were challenging. After much experimenting at the pilot stage, and on the advice of a professor in the mathematics department at the university, it was decided not to use that approach. An alternative from psychology, estimating Inverse Efficiency Scores, by taking time divided by the score (Bruyer, 2011) was not possible either because the data was not available at a granular level. Another problem was that response time followed different patterns in each of the scoring bands.

A rough estimation of ability based on the shortest response time for each child in each game partitioned by the band scores was carried out by estimating an overall speed logit value for each child. This was carried out in excel and artificially imposed limits for each game were derived by taking the slowest time among the 200 children as the maximum. This was for indicative purposes only and is not intended to be a solution to estimating speed. It became clear that any more accurate estimation required a much more complex model mathematically, which was beyond the resources and scope of this study.

5.5.3.2 Additional attempts

Almond had posited that familiarity makes tasks easier (Almond, 2015), and there was descriptive evidence from other games that there was a difference in behaviour between first and second attempts (Shute, 2016). The variable for attempts was included in this data set, and the estimation process was carried out on each of the first five attempts separately

(Lineacre, 1998). The number of attempts was capped at 5 because beyond the fifth attempt, there was too much missing data to be able to parse it. This allows a comparison of difficulty levels and reliability indicators for the First Five Attempts, with a data set that was more than 50% complete.

5.5.4 Research Question 4: Within the game, how reliable do the results appear to be?

The assessment models produce an estimation of error of measurement for the individual items and individual children in the analyses provided above. Depending on the findings, it may be suitable to estimate SEM θ and CEM θ values for the overall test. The SEM θ is often a U-shaped curve that gives an indication of the accuracy of overall measurement, with greater precision around the middle (Bond, 2015), and is one of the requirements of, say, the American Educational Research Association (1999). SEM θ mostly assumes that that item parameters, in other words the difficulty of the games, are known values (Wilson, 2004), and it may be that there is not enough evidence from the games to suggest that it is appropriate to anchor them at this stage.

5.6 Conclusions from the methodology

The data set offered many opportunities, and during pilot investigative work, the methods used here were refined and altered many times. The methods here, while relatively simple, produced a large number of findings that required considerable unpacking. This reflected what many have observed previously, that scoring games is problematic, and the process of data cleaning was lengthy given the size of the data sets, and the fact that they were not collected for assessment purpose. However, patterns did emerge, and insights into how to improve that data collection and cleaning process were gained. The results will be presented in Chapter Six, and discussion of these results will be saved until Chapter Seven.

Chapter 6 Results

There were many aspects of the gameplay environment that had unanticipated impacts on the results. The data seemed relatively simple, consisting of only of the children's identifier codes and their response times and band scores in a total of 21 different games (section 5.1). Despite this, there were many challenges to the scoring process but also new opportunities appeared. As the literature review in Chapter Three had found, there was very little substantive evidence to draw on directly in the field of Games Based Analysis, which may be why the results were unexpected. A second reason maybe that the gameplay environments themselves were complex. The results, however, did point to some identifiable behaviours, such as guessing, and compounding variables that could be dealt with in the assessment design and data collection process.

The children in the sample playing the games were varied in terms of their age, ability, enthusiasm and preferences. The data was longitudinal, which is unusual in assessment, and the sporadic nature of gameplay made ability assessment difficult. Behaviours that were encouraged by the platform of delivery, such as browsing and wandering off without logging off were fairly common and had the potential to bias the results. There were human and bot competitors that affected performance. These all presented challenges.

The first section describes the findings from the demographic exploratory phase, to contextualise the results and behaviours observed. I will go on to report on the findings as they shed light on each research question. A discussion of how these results might be interpreted is saved for the following chapter, Chapter Seven, to allow for a fuller consideration.

6.1 What does the group playing MangaHigh look like?

There were many features of the children in the sample and their gameplay patterns and style that seemed quite specific to game environments.

6.1.1 Patterns of gameplay

The descriptive statistics for the popularity of each game was taken from the Complete data set using SPSS, before any cases were removed for validity reasons. One game, the platform game, Jet Stream Riders, dominated the records. Once children discovered it, they tended to repeat play it many times in a row and return to it months or sometimes years after their first attempt. It was also a game where they could challenge another person to play. Figure 32 shows the overall popularity of the games among this sample of children.

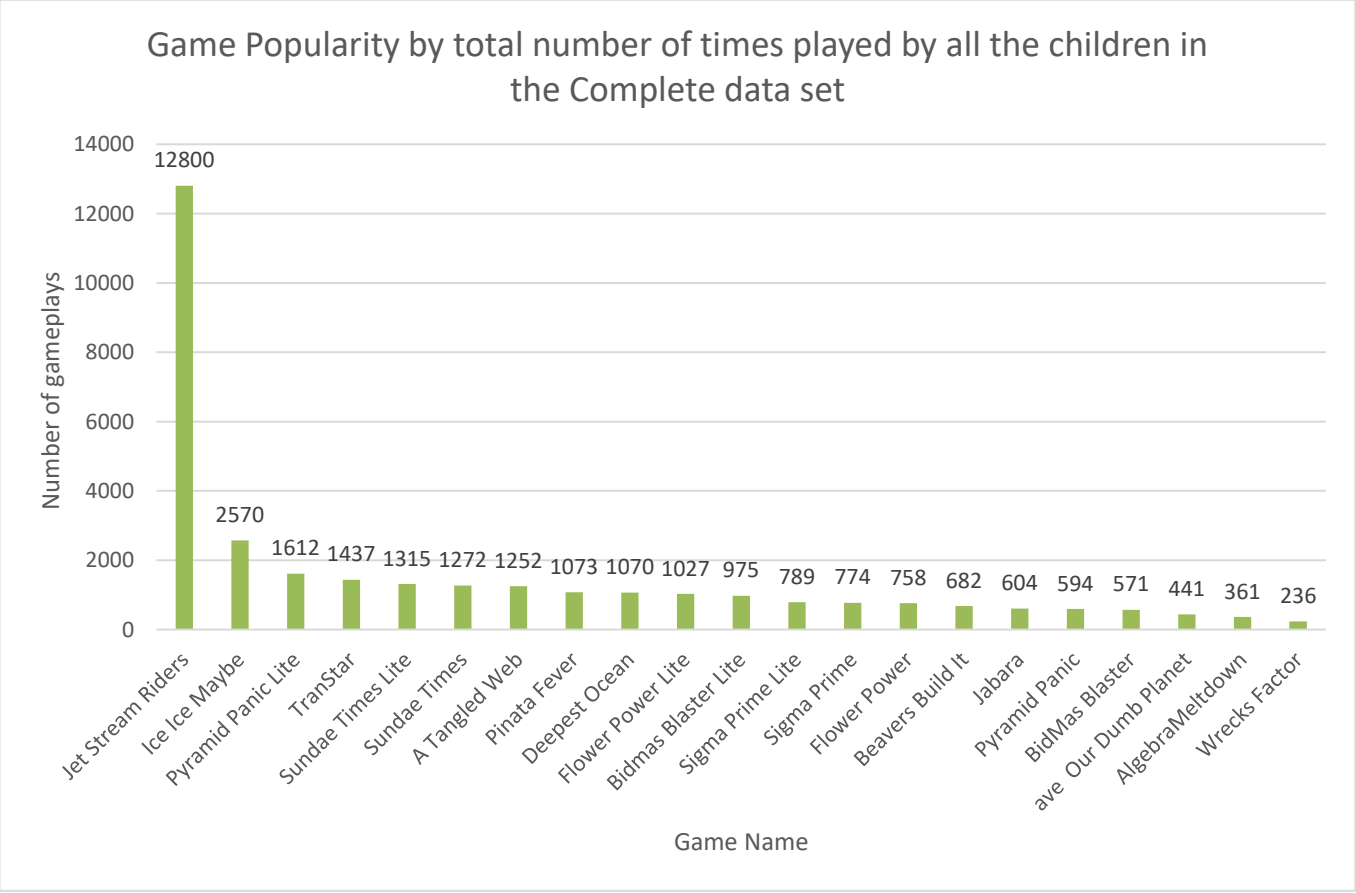


Figure 32 Distribution of the total number of game plays by game in the sample from MangaHigh

Jet Stream Riders was so popular, that 39.74% of the total available evidence in the data

set came from that one game, while 10 of the remaining 20 games contributed less than 20% of the cells with data between them (Figure 32). The 'lite' games, with slightly easier mathematics, were more popular than the regular versions with the same game mechanics for all five games offered at two levels (Bidmas Blaster, Flower Power, Pyramid Panic, Sigma Prime and Sundae Times).

The total number of plays was also the initial selection criteria for sampling, and so the group of children playing the games had already been filtered for similarities in terms of frequency of play. These children were sampled from the most active 90-95% players in the data set, which had a population of over 1 million registered players at the time of sampling. However, some children contributed considerably more evidence to the total than others.

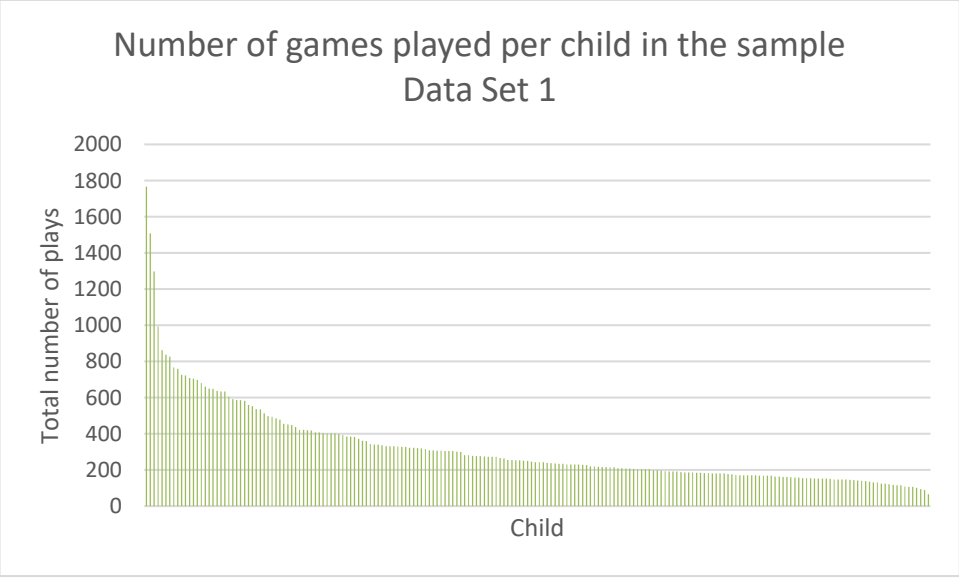


Figure 33 Distribution of the total number of game plays by child in the sample from MangaHigh

From the sample, 7 children had played more than 800 games in total (Figure 33). The most active child had played twice that number. Towards the lower end, gameplays of 100 times were observed. The very highest users, therefore, appear to contribute a disproportionate amount to the total volume of evidence collected.

6.1.2 Demographic information

After removing players from the initial sample if age appeared to be greater than 18 years, or those with no usable data in DateOfBirth, the children playing the games seemed to have a wide range of ages. Figure 34 shows the ages of the children at the time that they started playing the games in the first graph, from their self-reported year of birth, and the second shows their estimated ages on the date of data collection. Neither of these estimations were accurate, but gave a broad idea of the age range playing the games.

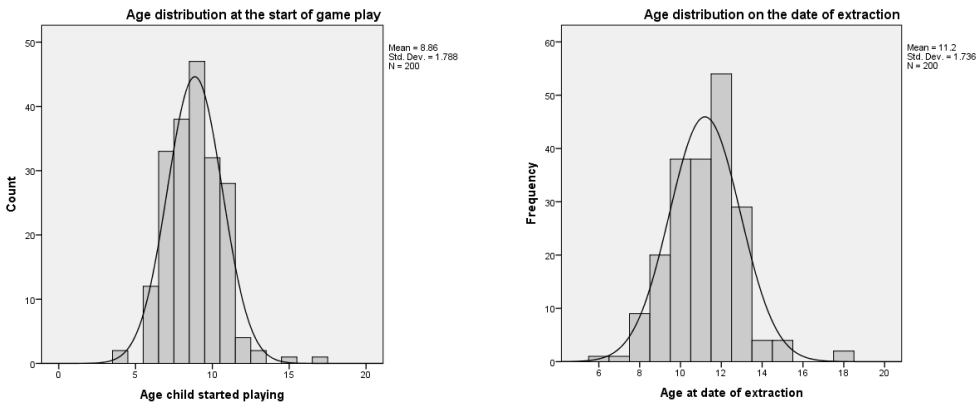


Figure 34 Distribution of ages of the children in the sample at the outset of play in MangaHigh over a period of time from 2011 – 2018 and at the time of extraction.

All ages were derived values and subject to an error of ± 11 months, or the possibility that the child's self-reported year of birth was wrong. Within that range of error, the ages were normally distributed with a mean age at the start of gameplay of 8.86 (Figure 35). The standard deviation was 1.78, and so there appeared to be a core age group, with around two thirds of the players within the age range of around 7 to 10.5 years at the time they started playing. There was still a fairly large number of children outside of that range, from potentially just 4 years old up to 17 at the time of signing up. There was a sharp drop off in the number of newcomers to the games after 11 years old, when children in many of the

countries represented in the sample (see below) change schools.

There was no drop out from the games among the children sampled, and the length of time recorded playing could be seen as a proxy for gaming experience. They all seemed to enjoy the game and were all still actively playing in the month before data collection. This was not a criterion for sampling. Figure 35 shows the number of years that the children had been playing MangaHigh games leading up to the time of data collection.

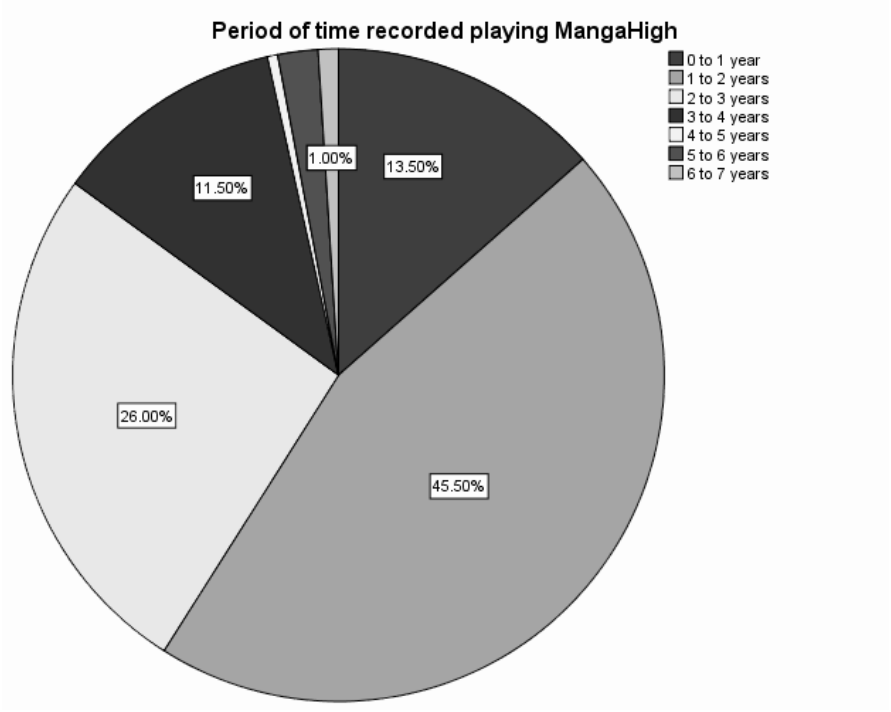


Figure 35 Length of time playing MangaHigh games in years among the children in the sample

The majority, seven out of ten of the children, had played the game sporadically for a period of between 1 and 3 years. On manual inspection, the very longest time periods were recorded from the youngest players at the outset of play, and they tended to follow a pattern of trying the games, abandoning them and returning to them more consistently when they were in the core age group of 7-10.5 years old.

The children in the sample came from 11 different countries. Figure 36 shows their countries of origin.

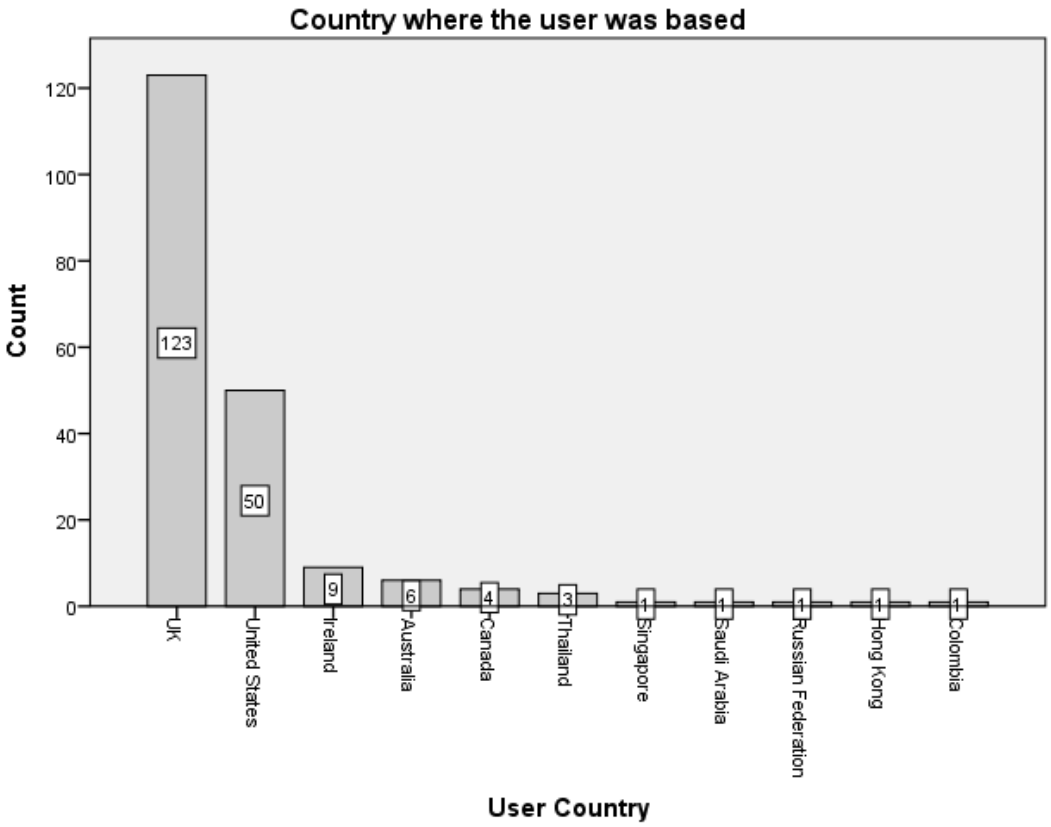


Figure 36 Country where children in the sample were based

Over half the sample were embedded in the school system in the UK (Figure 36). MangaHigh have stated that their core customer base is in the USA, and so this sample is also atypical in terms of where they are based. Instructions within the game are delivered in written English format at the outset of gameplay and countries where the English language has a recognised or historically important role in governance and education systems also dominated the sample. Outside of those countries, children potentially came from a diverse range of cultural and linguistic backgrounds, although it is possible that these children were in international schools. They were also playing multiplayer games in real time in different time zones.

Overall, the sample of children playing the games appears to be heterogeneous in terms of spread of ages, length of time in the 'course' of study and their country of origin.

6.2 Research Question 1: What counts as a valid attempt on task?

The data that appeared to evidence non-targeted behaviours was incrementally removed, using response time as a context variable to inform these decisions. As there was little guidance on how to decide what counts as a valid attempt at a game from the literature in Chapter 3, this section reports in some detail on the processes of removing cases. A lot of games were very short, and in deleting these, around 14% of the data in the CompleteData set was removed before dealing with outliers.

6.2.1 Deleting cases rounded down to 0

Games are state machines, and so they are usually programmed to record all of the states. This means that they keep records of all of the key strokes and input actions, and a conscious decision needs to be taken if these are wanted in the assessment data set. There were 2,783 games that rounded down to zero in the Complete data set out of a total of 32,045 recorded games. This meant that zero response time games accounted for 8.68% of the total cells with data from the Complete data set. The response times and their corresponding scores were deleted from the Complete data set in all cases where the response time less than or equal to .5 seconds, and this left 29,262 cases in the Without Zeros data set.

6.2.2 Deleting cases under 10 seconds

Of the remaining 29,262 cases, there were 1,793 cases with times of less than or equal to 10 seconds, accounting for 6.13% of the data set. Unlike the cases where response time less than or equal to .5 seconds, some of the cases between .5 and 10 seconds were in a scoring band (One, Two or Three). The decision to remove these cases had a statistically significant impact to the 95% confidence level on the mean response time in six of the

games. These were Bidmas Blaster ($p \leq .0005$), Jet Stream Riders ($p \leq .0005$), Piñata Fever ($p \leq .002$), Pyramid Panic ($p \leq .0005$), Pyramid Panic Lite ($p \leq .0005$), and Sigma Prime ($p=.001$). The null hypothesis, that deleting cases under 10 seconds would have no effect on the mean time scores, was rejected in these six games. The cases were removed to create OverTenSecondsOnly data set, which contained 27,469 games.

6.2.3 Dealing with outliers at the upper boundary

As discussed in the Methodology Chapter, some games had capped or controlled response times, but others stayed open for as long as the browser was left open. Some gameplays lasted as long as 95 hours consecutively, which was indicative that the child had wandered away from the screen. Figure 37 shows the maximum length of play for each of the games in the OverTenSecondsOnly data set.

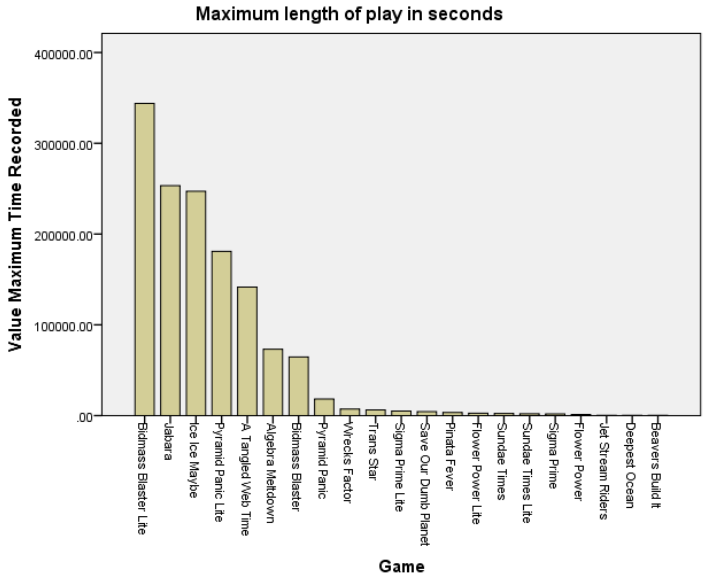


Figure 37 Maximum recorded length of game play for each game in the Complete data set from MangaHigh

Seven of the games recorded exceptionally long maximum response times (Figure 37) and this can be seen to have an impact on the mean response times for games, too, shown in Figure 38.

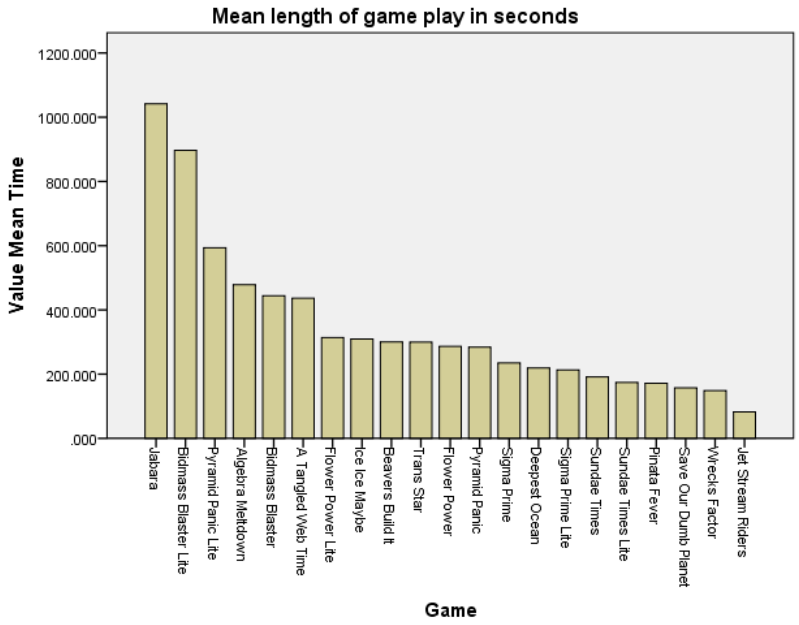


Figure 38 Mean response times for each game in the Main data set from MangaHigh

The mean play time in the games Jabara, a 'guide the avatar to the correct answer' game of algebra, and Bidmas Blaster, a shoot 'em up order of operations game, was still around 17 minutes and 15 minutes respectively (Figure 38). In reality those games could be successfully completed in around 3 minutes. It appeared that these long play times held a large amount of weight on the mean response time as a whole, and many games in the OverTenSecondsOnly data set had large standard deviations. For example, the mean response time for Jet Stream Riders, the times tables, addition and subtraction platform game, was 82.36 seconds with a standard deviation of 12,698 seconds. Bidmas Blaster had a mean of 897, with a standard deviation of 12,364 seconds. These standard deviations over 12,000 seconds amount to around 3.5 hours.

As described in the Chapter Five, in the absence of observation of the children at play or key stroke analysis, a purely statistical approach was taken to identifying the point when it became highly unlikely that the child was still playing. The definition of extreme values as greater than or equal to Q3 + 3(Q3-Q2) was used and extreme values were capped at the value of Q3 + 3(Q3-Q2)-1. This meant that there was no difference in size of

OverTenSecondsOnly and the Main data sets, but the spread of potential response times was greatly reduced, allowing patterns to more clearly show. Table 14 shows the frequency of statistically significant differences between OverTenSecondsOnly and the Main data sets, broken down into band scores.

Table 14 Impact of capping response times at the upper quartile limit on the mean response time

	Band 0	***Band 1***	***Band 2***	***Band 3***
Cases where deletion was statistically significant	6	1	1	0
Cases where deletion was not statistically significant	12	9	9	10
Cases where insufficient evidence was gathered	1	9	9	9

A one tailed paired samples t-test was conducted to compare the mean response time for each of the games in OverTenSecondsOnly and the Main data sets. There was not a significant difference in the mean response times for the games in the scoring bands (1, 2 and 3) of the majority of the games. The notable exception was the game Transtar, where players manipulate an avatar space ship to practice reflections and rotations. Taking statistical significance to the 95% confidence interval, 6 of the 8 response times that were found to be significantly different after capping were in the non-scoring band 0. There was a significant difference in the mean response time in the games:

Table 15 Report on the statistically significant differences found in the zero scoring band from a one-tailed paired samples t-test comparing results from the OverTenSeconds data set, and the Main data set

	Mean and Standard Deviation		*Difference*
	OverTenSeconds Band 0	**Main Band 0**	
FlowerPower Lite	**(M=298.94, SD=246.29)**	**(M=278.66, SD=163.24)**	**t(573)-1.70, p≤.0005**
JetStreamRiders	**(M=77.99, SD=55.51)**	**(M=77.90, SD=55.17)**	**t(6709)=-4.63, p≤.0005**
Pinata Fever	**(M=163.13, SD=182.38)**	**(M=145.68, SD=102.12)**	**t(861)=-4.14, p≤.0005**
SigmaPrime	**(M=145.66, SD=150.98)**	**(M=133.93, SD=106.28)**	**t(526)=-3.74, p≤.0005**
SundaeTimes	**(M=184.48, SD=179.81)**	**(M=169.37, SD=118.94)**	**t(1048)= -4.95, p≤.0005**
SundaeTimes Lite	**(M=172.81, SD=179.06)**	**(M=146.53, SD=93.72)**	**t(1052)= -7.285. p≤.0005**

These results (Table 15) suggest that there was a significant difference in the mean response time for OverTenSeconds and Main in band 0 in these games, although as these were all in the zero band group, and the number of degrees of freedom for Jet Stream Riders, Sundae Times and Sundae Times Lite were quite high, and this is more likely to produce statistically significant results (Walliman, 2017). The difference in the two means for Jet Stream Riders, for example, is .09 seconds, which is very small.

In the game Transtar, however, there was a significant difference in the mean response time for OverTenSeconds band 1 (M = 99.56, SD = 134.22) and Main band 1 (M = 82.29, SD = 85.78); t (-2.99), p ≤ .003. There was also a significant difference in the mean response time for that game for OverTenSeconds in band 2 (M = 121.28, SD = 199.40), and the Main data set in band 2 (M = 84.54, SD = 93.17); t (-3.71), p ≤ .0005. In that game,

once outliers were removed, the 0, 1 and 2 band all had a mean of around 84 seconds, whereas the 3 band still had a very high mean of 245.70 seconds. These results suggest that some children did in fact sometimes take longer to complete the game Transtar. Specifically, these results suggest that a different definition of a plausible time to play may be needed for that particular game to avoid punishing children who persist. The results for all of the games can be found summarised in Appendix D.

To sum up, response time served as a context variable to identify cases where there were reasons to believe that the data was unhelpful in informing about performance and in some cases this may have impacted on the results. The decision was taken to use the Main dataset for all subsequent analysis.

6.3 Research Question 2: Does missing data impact the final score?

As described in section 6.1.1, there was a very large amount of data missing in this study. This created two challenges, the first was to find a way to parse the data set, and the second was to investigate the assumption of Missing At Random, in other words the assumption that the reason for data being missing was not associated with the ability of the child (Allison, 2001). The next section describes the findings of investigations into this assumption.

6.3.1 Missing data and its relationship to ability

As described in section 5.5.2, a very large amount of data was missing from the Complete data set. Around 88% of all the cells were empty, and this was largely caused by children following unique pathways through the games. The pilot study had suggested that ability might be related to the children's choice of game, and so, following a Partial Credit Score estimation of the highest score for each child in the Main data set, it was partitioned into

three groups, a HighAbility, a LowAbility group and third set that were around the mean. The estimate of ability using a partial credit model was then repeated with the HighAbility (n=80, ability logit ≥+1.72) and LowAbility (n=79, ability logit ≤-.72) groups. No further action was taken with the group around the mean (n=41, ability logit = -.71 to +1.71) to ensure that there was a clear distinction in ability between the two data sets. The partitioning of the data had an overall impact on the order of difficulty in task rankings. The Wright Map below shows the difficulty in logits in the 'measr' column, the abbreviation of the game names, and the position of the four scoring bands (0-4) for the Whole HighScore data set, the HighAbility group extracted from the HighScore data set, and the LowAbility group extracted from the HighScore data set (Figure 39).

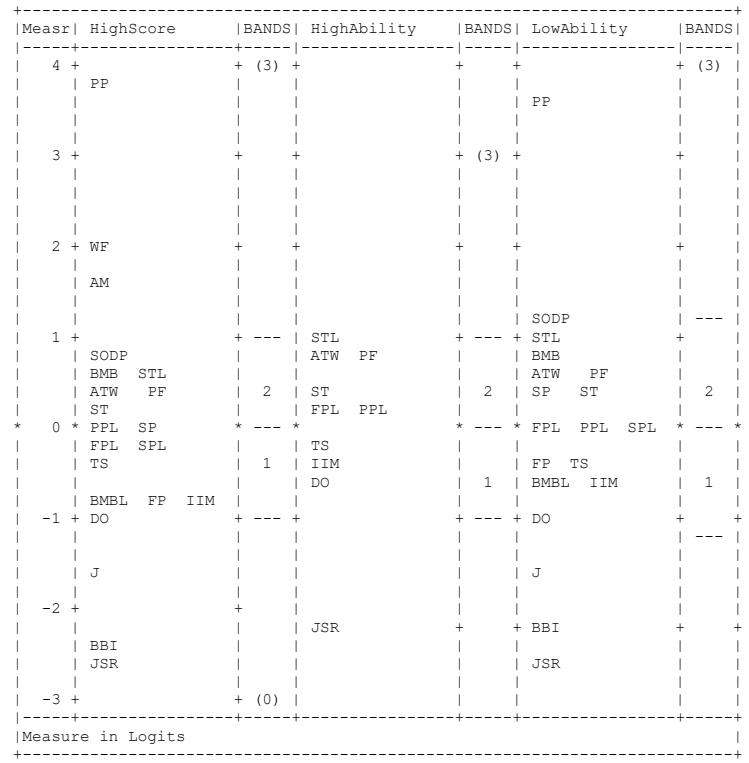


Figure 39 Wright map comparing the logit position for the whole data set, the HighAbility group only set and the LowAbility group only set

Higher positions vertically indicate a higher level of challenge in Figure 39, and the mean is represented by the 0 in the left hand column. The position of the cut off points of the band scores is relatively stable, and the order of difficulty of the tasks was also relatively

stable. Slight changes in the logit values are to be expected. The two groups played a fairly similar number of times in total (LowAbility = 1,196 plays; HighAbility = 1,169 games). The High Ability group, however, were highly specialised. They had played only 10 out of the 21 possible games, compared to the Low Ability group, who had tried 19 different games, only missing the two least popular (Wrecks Factor and Algebra Meltdown). The High Ability group avoided the four most difficult games (Pyramid Panic, Wrecks Factor, Algebra Meltdown and Save Our Dumb Planet) and the easiest ones (Beavers Build It and Jabara). The full Wright map with child ability estimates for the two divisions are in Appendix E.

6.3.2 Missing data and difficulty of the regular and lite games

One anomaly that came of the analysis of the partitioned data set was that there were some unexpected patterns there for the games that had regular and lite options. Five of the games, Pyramid Panic, Bidmas Blaster, Sundae Times, Sigma Prime and Flower Power offered the same game mechanics, but with a regular version that had more demanding mathematical content than the lite version, which focused on mathematical functions that are presented earlier in a child's schooling. The way that these games were experienced by the High and Low Ability groups differed in unexpected ways. Table 16 shows the results of the estimate of difficulty on all of the children's highest scores from the Main data set.

Table 16 Results of estimations on the HighScore with regular and lite versions using the whole HighScore data set, higher values in logits are associated with greater challenge

Game name	*Results for the regular game*			*Results for the lite game*		
	Logit values	**Infit**	**Outfit**	**Logit values**	**Infit**	**Outfit**
PyramidPanic	3.87	.92	.22	.01	.93	1.01
BidmasBlaster	.58	1.07	.93	-.79	.68	.74
SundaeTimes	.20	1.24	.94	.65	1.27	.79
SigmaPrime	.07	.96	.85	-.11	.89	.94
FlowerPower	-.81	1.26	1.28	-.19	1.13	.96

The logit values in Table 16 show the difficulty of the task, with higher numbers associated with higher levels of challenge, and two of the lite games, highlighted in red, were intended to be easier, but were in fact experienced as more challenging by the children in the sample. The result for Sundae Times and Flower Power both show a higher difficulty estimate for the lite game compared to the regular version. Looking at the error estimation in the table, the Infit (weighted around the child's threshold) and Outfit (over the entire data set) estimates of error suggest relatively stable results. Although there is no set threshold for error, values lower than 1.5 suggest fairly stable performance (Wilson, 2004). Higher values for outfit would indicate a greater degree of randomness in the results (Lineacre, 1998), but this is not the case.

One possible explanation was that the children had selected the appropriate level for themselves, and so the lower ability students would still find the lite tasks challenging, and the higher ability students might find the regular versions appropriate to their level. The results of how the games were experienced by the HighAbility group and LowAbility group are shown in Table 17 and 18, respectively.

Table 17 Results of estimations on the HighScore with regular and lite versions using only HighAbility children in the HighScore data set, higher values in logits are associated with greater challenge

Game name	*Results for the regular game*			*Results for the 'lite' game*		
	Logit values	Infit	Outfit	Logit values	Infit	Outfit
PyramidPanic	2.48	N/A	N/A	-.11	.68	.80
BidmasBlaster	.92	.78	.90	-.51	.48	.52
SundaeTimes	-.03	1.31	1.21	.70	.93	.58
SigmaPrime	.46	.70	.76	1.40	.81	.69
FlowerPower	-.63	.70	.66	.36	1.03	1.05

Table 18 Results of estimations on the HighScore with regular and lite versions using only LowAbility children in the HighScore data set, higher values in logits are associated with greater challenge

Game name	*Results for the regular game*			*Results for the 'lite' game*		
	Logit values	Infit	Outfit	Logit values	Infit	Outfit
PyramidPanic	3.69	N/A	N/A	.16	.61	.62
BidmasBlaster	.63	1.22	1.33	-.45	.48	.48
SundaeTimes	.36	1.37	1.29	.63	1.26	1.10
SigmaPrime	.03	.85	.82	-.08	.69	.68
FlowerPower	-.91	1.57	1.51	-.01	1.22	1.22

The results are relatively similar between the High and Low Ability groups. The infit and outfit statistics for Flower Power regular among the Low Ability group are quite high (infit 1.57 and outfit 1.51) but this might be acceptable for some testing scenarios (Wilson, 2004). With the HighAbility group, Sigma Prime also showed unexpected results, with the lite version proving more challenging (Table 18). Partitioning the data set on the grounds of ability did not have the effect or reversing the order of difficulty of Flower Power Lite and Flower Power, and Sundae Times Lite and Sundae Times. The lite games were

consistently experienced as more challenging than the regular version (Tables 17 and 18). It suggests that the level of challenge of Flower Power and Sundae Times is not aligned to the designers' intentions.

A second purpose of splitting the data set was that, under the assumptions of the Rasch model, the partitioning of the data into the two subsets should not have had an effect on the order of difficulty of this task (Wilson, 2004; Bond and Fox, 2015). Table 19, showing the order of just the lite games among the three groups, whole, high and low, shows some changes in the order.

Table 19 Impact on the order of the lite tasks when the analysis was run over the three different groups of children, the whole data set, the HighAbility and the LowAbility partitions

Whole Data set (n=200)		**HighAbility (n=80)**		**LowAbility (n=79)**	
Game name	**Logit**	**Game name**	**Logit**	**Game name**	**Logit**
SundaeTimes Lite	.65	SigmaPrime Lite	1.40	SundaeTimes Lite	.63
PyramidPanic Lite	.01	SundaeTimes Lite	.70	PyramidPanic Lite	.16
SigmaPrime Lite	-.11	FlowerPower Lite	.36	FlowerPower Lite	-.01
FlowerPower Lite	-.19	PyramidPanic Lite	-.11	SigmaPrime Lite	-.08
BidmasBlaster Lite	-.79	BidmasBlaster Lite	-.51	BidmasBlaster Lite	-.01

The position of Sigma Prime Lite, shown in blue, in particular, moves considerably as the data set is split. The High Ability group found Bidmas Blaster (-.51 logits) and Pyramid Panic Lite (-.11 logits) easier than Low Ability group (-.01 logits and .16 logits respectively). But they found Sigma Prime (1.40 logits) considerably harder than the Low Ability group (-.08 logits). They also found Flower Power Lite (.36) harder than the Low Ability group (-.01 logits). The results from Table 6.5 suggest that there might be a greater degree of unpredictability in performance than the reliability estimates alone imply.

6.4 Research Question 3: How can the game-specific variables of response time and iterations be conceptualized?

The two variables of response time and repeated attempts received very little mention in the published literature on GBA, and both variables proved challenging to model. Because they are such a strong feature of games, it may be worth including them in a measurement, even if it is to take steps to hold them constant, for example, by imposing a fixed time limit or delivering the games as a simulation, with controlled attempts. The results below will outline some of the issues found with these two variables.

6.4.1 Conceptualizing response time in the games

Response time was not experienced in the same way with the different games mechanics and games designs, and so there was no one way to conceptualise it. The box plots in the sections below were produced from the Main data set in SPSS. There were four categories of response time in the data set:

1. Set time

Beavers Build It was a multiplayer game, with players collaborating to build the highest wall. The game demonstrated an unusual pattern of response time in that all of the games were completed in 300 seconds.

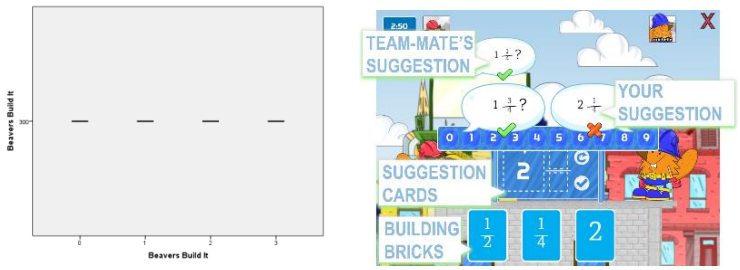


Figure 40 The distribution of response times for Beavers Build It using the Main data set, with game screen shot ©MangaHigh

There were no recorded early finishers in this game, no matter the band score (Figure 40).

2. Capped limit

Deepest Ocean had two set periods of gameplay, a basic time (up to 150 seconds), and then a bonus time period (up to 300 seconds) if they performed well. The distribution of the response times in the different band scores (0, 1, 2, 3) is shown in Figure 41.

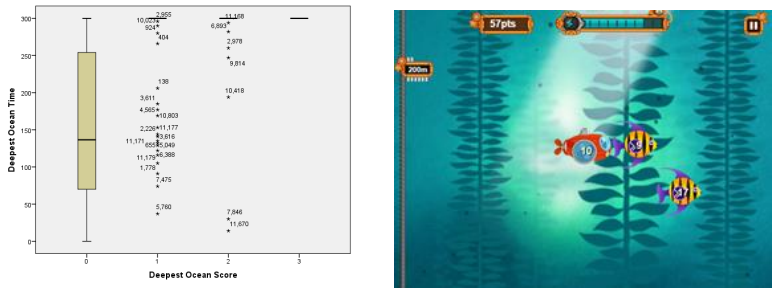


Figure 41 The distribution of response times from the Main data set for Deepest Ocean, with screen shot ©MangaHigh

Some children were able to get into a scoring band even if they left the game early before the first and second time limits ended (at 150 and 300 seconds respectively), but most continued to play until the end of the game. There were no outliers in this game at the upper boundary because of the definite time stop at 300 seconds (Figure 41).

3. Limits imposed by human and bot competitors

There was no data recorded on competitors, and so there was no specific variable to show when a child played against a human or a robot. Frequency counts for each potential response time were usually ‘1’, reflecting the fact that response times were mostly recorded to 3 decimal points, and so exact repeats were rare. However, a frequency count revealed some anomalies that were indicative of the pre-recorded robot performance, or a ‘bot’ time (Table 20). The mechanics required multi-player modes for these three games, and the repetition of such exact times suggests that the bot competitor was playing when other children were not available to play in real time.

Table 20 Frequency counts of bot times in games where children competed against pre-programmed performances

Game Name	Response time (seconds)	Frequency
Jet Steam Riders	61	1,299
SundaeTimes	90	403
	180	174
SundaeTimes Lite	90	455
	180	156

Both sets of game mechanics in these three games relied on competitive multiplayer modes of play. Screenshots from both games are shown Figure 42.

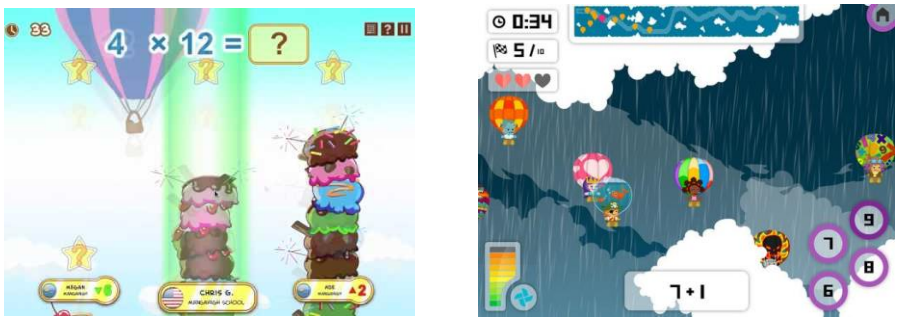


Figure 42 Screenshots from the multiplayer games Sundae Times and Jet Stream Riders ©MangaHigh

Sundae Times, shown on the left, had a fixed number of competitors, three. Jet Stream Riders, shown on the right, was a multiplayer platform game, and therefore a race-to-the-finish game, as the screen shot on the left in Figure 43 shows. There can be many players in any one session depending on the number of children online at the time, and in fact, the game can support up to 30 players racing simultaneously. The end of the game was determined when either the player or the other competitor finished first.

Sometimes only the bot was available, though, and the bot always followed a pre-set pattern, always completing in a specific time, 61 seconds. It essentially imposed a time limit on that game, but only sometimes. Table 21 shows the mean response times for each band, using the OverTenSeconds data set.

Table 21 Mean and minimum response times in seconds for Jet Stream Riders with children competing against real children and a bot

Time scores for humans and bots

	Mean	Bot time
Band 1	97	61
Band 2	95	61
Band 3	79	61

Overall, the data suggests that, although there were times when the child was able to beat the bot, in general, the bot won if it was in play (Table 21). It finished well under the mean winning finish times of 97, 95 and 79 seconds for Bands 1, 2 and 3. Although there was a bot also in play in both versions of Sundae Times, medals were allocated on the basis of the number of correct questions / scoops of ice cream, not finish times.

4. Open response time

Finally, in the rest of the games, the game stopped when either the child completed, conceded to a competitor, or when the child eventually came back to log off. The main issues with these types of game was discussed in more details in Section 5.2 on dealing with outliers at the upper boundary.

6.4.2 Response time as a proxy for ability and the game mechanics

Response times varied by game, as might be expected, but it also varied by band. Not all of the games provided data on all of the bands. Of those that did, a range of patterns appeared. The box plots below were produced on SPSS, using the OverTenSeconds data set, with response times of less than 10 seconds removed, and extreme values at the upper limit capped. Figure 44 below shows that a range of patterns across the bands were observed.

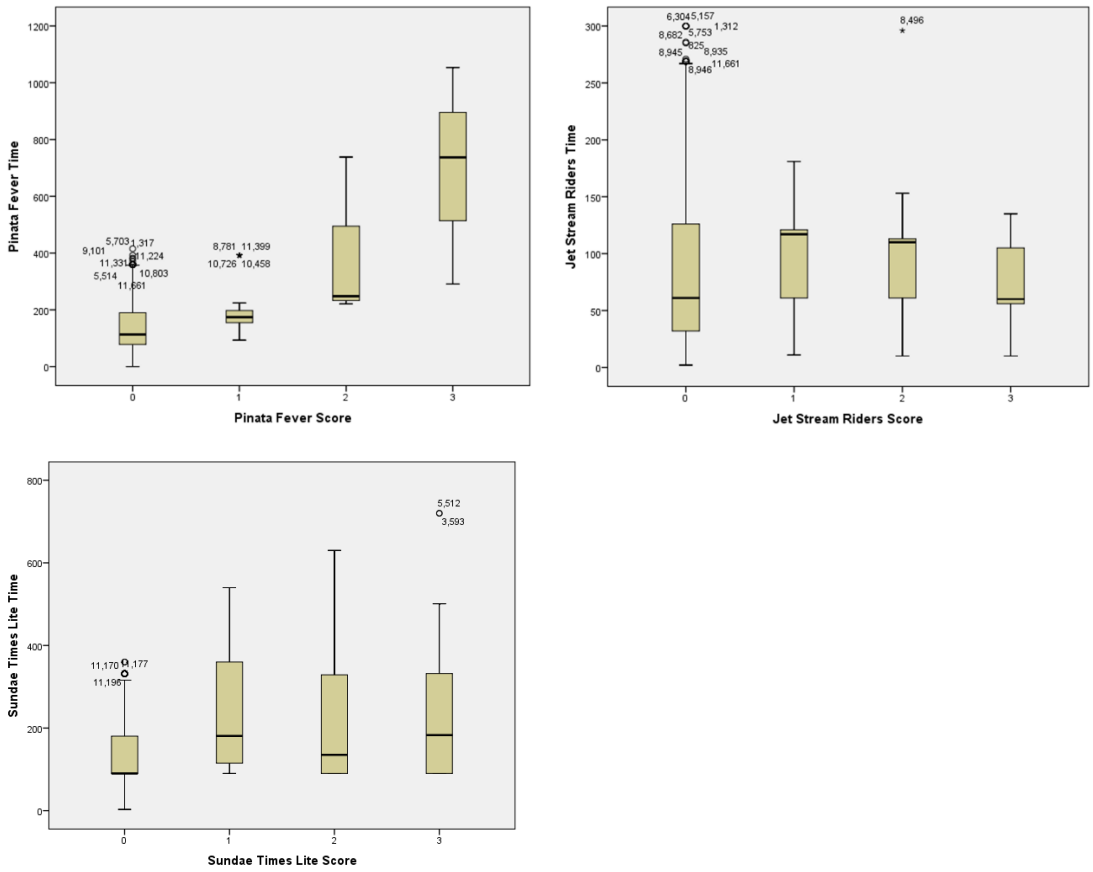


Figure 43 Box plots showing an ascending pattern (Piñata Fever), descending pattern (Jet Stream Riders) and varied (Sundae Times Lite) pattern

Ignoring the no-score bands, Piñata Fever was unusual in that there was no overlap between the response times for each band, apart from outliers (Figure 43, top left). The rising pattern of PiñataFever was typical of some of the games where the player had to either guide an avatar towards the correct answer or the shoot'em up games. It means that in order to score a higher result, the child had to persevere. This is in contrast to Jet Stream Riders (Figure 43, top right) which was the only platform game and the only one with a clear falling pattern, where speed was rewarded. Sundae Times Lite (Figure 43 bottom left) bottomed out slightly in band 2, although in general, the speed of finishing was more or less the same, and this game had a capped time limit.

Other games peaked in band 2. Table 22 below gives the full breakdown of the patterns of response time grouped by game mechanics.

Table 22 Patterns of response times grouped by game mechanics

Patterns of Response time in MangaHigh		
	Maths	Pattern
1. Guide the Avatar to the correct answer		
AlgebraMeltdown	Algebra	Peaked at band 2
FlowerPower	Ordering Decimals, Fractions and Percentages	Rising
IceIceMaybe	Fast estimation with basic number calculations	Rising
SaveOurDumbPlanet	Patterns, Algebra, Straight Line Graphs	Peaked at band 2
PiñataFever Guide	Add and subtract with negative numbers	Rising
PyramidPanic	Geometry	Rising
Transtar	Reflections, rotations, enlargements and translations	Rising
2. Maths with animation		
ATangledWeb	Angle	Peaked at band 2
Jabara	Algebraic simplification	Bottomed out at band 2
SundaeTimes	Times tables from x2 to x15	Peaked at band 2
SundaeTimes Lite	Basic mental maths	Bottomed out at band 2
3. Platform game		
JetStreamRiders	Times tables and numeracy	Falling
4. Shoot 'em up		
BidmasBlaster	Brackets, Indices, Division, Multiplication	Rising
SigmaPrime	Prime factorisation with multiplication and division	Rising

In general, in all but the platform game, Jet Stream Riders, stronger performances were associated with taking a little more time to complete the game (Table 22), which was in line with Thurstone's (1937) original view that the probability of doing well increases with longer response times. The game mechanics, though, appear to have more influence over the patterns of response times than the subject of the game. The three games where the correct answer was rewarded with an animation did not need games mechanics to play,

they were straightforward ‘type the correct answer’ type items, and they showed a more random pattern, with either a peak or a trough in band 2. The games that required an avatar to be moved, generally required more time to get the top band score. The two that peaked, Algebra Meltdown and Save Our Dumb Planet, are drag and drop multiple choice type tasks. The patterns of response time appear to be affected by the game mechanics, although few definite conclusions can be drawn. The box plots for all of the games can be found in Appendix F.

Estimating the speed of the players from their shortest response times in each of the bands showed a clustering of players around the mean, but some were significantly faster than the others. Using speed as a proxy for ability, as well as accuracy, Figure 44 shows a scatterplot of the overall logit estimations of each player based on their fastest performances in band 1.

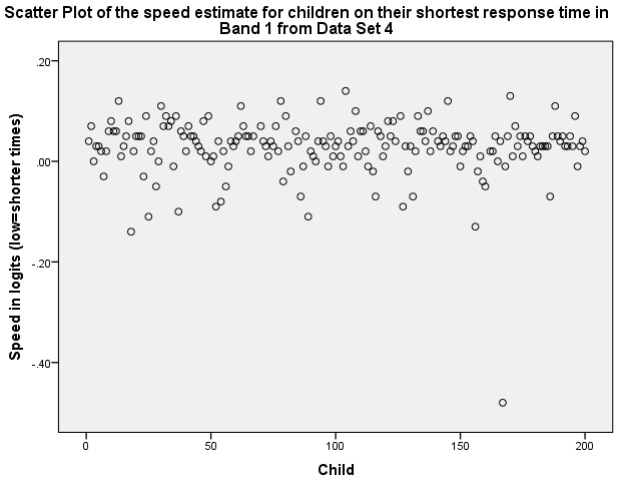


Figure 44 Scatterplot of the speed of children in Band 1 over all the games that they had played from the Main data set

Investigation of the fastest child at -.48 logits, showed that this child had only produced a score of band 1 in 3 games, but was consistently one of the quickest finishers among the players in those games. Overall this child had an ability level of that was slightly above average and fairly stable, at .9 (infit) and 1.12 (outfit). It suggests that perhaps some

children are fast. The slowest child at .4 logits for speed had produced band 1 performances in 5 games and was consistently one of the last to finish among the players in that game. This child had a slightly higher ability level, at 1.24, and also fairly stable overall at .21 (infit) and .19 (outfit). Bands 2 and 3 showed similar patterns. Again, it suggests that speed of response may be a variable that says something about the child. For a fuller breakdown, see Appendix G.

No further analysis of response time as an indicator of ability was made.

6.4.3 Conceptualizing and modelling attempts

The analysis of the First Five Attempts data set generally showed from the aggregation of the scores for all of the games, the games became easier the more they were played. Figure 45 shows a descending pattern, with a sharp drop between the first and second attempt, and then subsequent attempts become moderately easier.

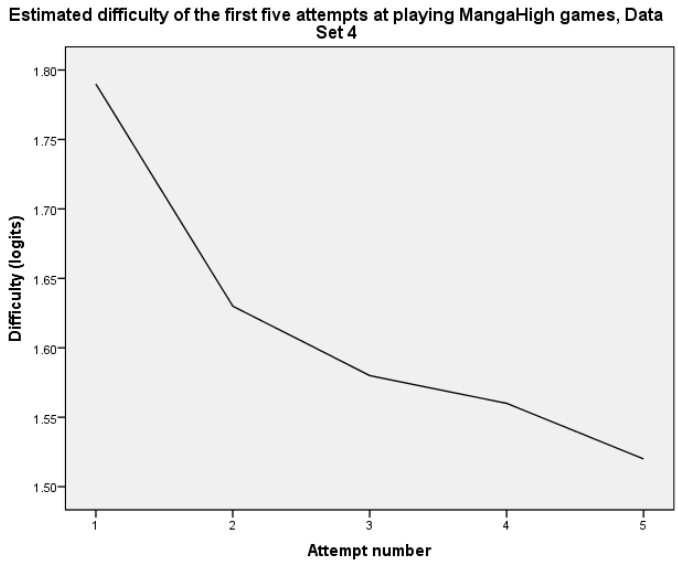


Figure 45 Difficulty values overall for the FirstFiveAttempts data set, representing the descending level of challenge with each additional attempt

The overall descending pattern in Figure 45, however, masks the way that the child might experience the game at an individual level. All the children had fluctuating patterns of

behaviour. In general, additional attempts increased the likelihood that the child would achieve the top score at some point, but they did not consistently maintain that level once achieved. Figure 46 shows an example of how the games were experienced by one child playing the same game.

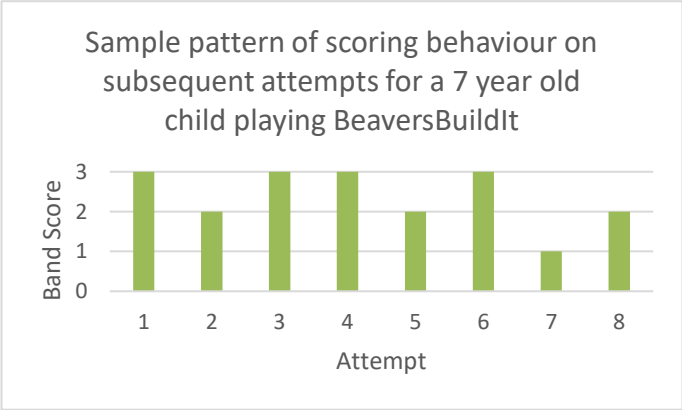


Figure 46 Scoring pattern from one child'7s records on the addition and subtraction game BeaversBuildIt

In this case the child actually achieved a band 3 score in the first attempt, and then slipped a little (Figure 46). There are multiple possible interpretations of this pattern. Beavers Build It is a collaborative game, and it is possible that collaborator ability was captured in the scores. Boredom, skills slip or even the fact that the child was asked to continue playing after losing interest, either by a teacher or a challenge from a friend, may be reflected in this pattern of play. This fluctuating pattern of scoring was observed across all of the games and all of the children and there were no typical patterns that could be observed by manual inspection among the children's game play.

Infit and outfit statistics were generated for the residuals of unexpected behaviour in the FirstFiveAttempts. These indicated a possible slight change in behaviour and are shown in Table 23 below.

Table 23 Results for the FirstFiveAttempts playing MangaHigh using the FirstFiveAttempts data set

Attempt	Infit	Outfit
1	1.06	.55
2	1.16	1.08
3	.95	.72
4	1.10	1.14
5	1.16	1.03

Attempts 1-2 show higher infit statistics than outfit (Table 23), which typically indicates a degree of specialisation (Wilson, 2004; Bond and Fox, 2015) either in the aspect of the syllabus, or perhaps in this case the ability of playing video games. By the 3rd attempt, these were lower, and after the 4th, they seem to stabilize more at levels that are generally acceptable for, say, a similar test of mathematics that might be delivered on paper in linear format.

Overall, there were indications that speed can be estimated, but requires specialist treatment that was outside of the scope of this study, and the reasons and implications of this will be discussed in greater detail in the next chapter. The results presented here on speed are only indicative of general patterns. In terms of additional attempts, although the statistics suggest that subsequent attempts at gameplay do become easier, as might perhaps be expected, how this is experienced on an individual level may be hard to quantify. This will also be discussed further in the chapter 7.

6.5 Research Question 4: Within the game, how reliable do the results appear to be?

There were many issues that arose through the process of analysis, and some of these that

are specific to the other research questions have been discussed above. Because so much of the behaviour in the games described above was observed for the first time in this study, and there is no substantive work to suggest how to interpret anomalies, the error measures reported above and in the subsequent sections are intended to be indicative only of the reliability of the game scoring process in this early study in the field.

6.5.1 Notions of invariance

The two main measures used to test invariance were the first five attempts from the Main data set, and the child's highest score from the Main data set, treating missing data as not presented, rather than 0 in each case. As already discussed, the whole data set contained too much missing data to be able to parse. The decision to keep missing values as not presented was made because the findings on whether ability was related to choice was inconclusive, but no child had played all of the games.

Taking the scores for the first five attempts did not produce noticeably distinct reliability measures compared to using just the child's highest score. For the first five attempts, 1,041 responses were used for the estimation, and there were 56 observations where the reliability was unexpected, above infit and outfit values of 2.0. Overall, the tasks showed average discrimination, with an overall model infit of 1.09 and outfit of .90. Using the high score data set, 200 responses were used for the estimation, reflecting that fact that just one score was escalated for each child. It resulted in 37 observations where the reliability was unexpected, above infit and outfit values of 2.0. The tasks also showed average discrimination, with an overall model infit of 1.03 and outfit of .91, which were very similar to the results from the first five attempts.

However, the individual task difficulty estimates were quite different between the first five attempts and the high score models. The order of difficulty of the games changed considerably, as shown in Table 24.

Table 24 The order of games from most to least difficult shows a dramatic difference in order between the two models.

Games in order from most to least difficult (logit difficulty value)	
FirstFiveAttempts	**HighScore**
BeaversBuildIt (1.65)	PyramidPanic (3.87)
Jabara (.95)	WrecksFactor (1.98)
PyramidPanic lite (.90)	AlgebraMeltdown (1.67)
AlgebraMeltdown (.41)	SaveOurDumbPlanet (.88)
SaveOurDumbPlanet (.25)	SundaeTimes lite (.65)
SundaeTimes lite (-.23)	BidmasBlaster (.58)
IceIceMaybe (.12)	PiñataFever (.41)
PiñataFever (-.02)	ATangledWeb (.26)
WrecksFactor (-.11)	SundaeTimes (.20)
FlowerPower lite (-.12)	SigmaPrime (.07)
FlowerPower (-.14)	PyramidPanic lite (.01)
SigmaPrime (-.12)	FlowerPower lite (-.19)
SundaeTimes (-.16)	SigmaPrime lite (-.11)
DeepestOcean (-.19)	Transtar (-.46)
JetStreamRiders (-.22)	IceIceMaybe (-.75)
SigmaPrime lite (.21)	BidmasBlaster lite (-.79)
PyramidPanic (-.33)	FlowerPower (-.81)
Transtar (-.37)	DeepestOcean (-1.05)
BidmasBlaster (-.39)	Jabara (-1.52)
ATangledWeb (-.70)	BeaversBuildIt (-2.46)
BidmasBlaster lite (-1.37)	JetStreamRiders (-2.58)

Beavers Build It, for example, was one of the most difficult for children who were playing for the first five times, but over the entire data set, it was the second easiest game. Beavers Build It targeted mathematics at the entry level point of the games. It begins with addition up to 10 and is likely to be the first game children experience, which may affect the pattern of behaviour. Jet Stream Riders, the most popular and the easiest game over the whole data set was also considerably more difficult (-2.58 logits and -.22 logits respectively) in the first few attempts. The children got a lot of practice at that game, and the chances of them getting a win from a competitor conceding almost certainly increased with more play. It may also be that, with practice, the tests of number bonds and times tables were no longer about problem solving, but tests of recall. Some other games became significantly harder

as the children became more experienced. Pyramid Panic, the game that appears to be very much higher than the level of the children in the high score model, was relatively easy in the first five attempts. Given that there were some quite radically different estimates of the overall difficulty of each game depending on which figures were taken, it was decided that there was no way of anchoring the scores, and therefore, reporting a SEM θ value seemed very premature. The logit, infit and outfit values and for all of the games in these two data set are shown in Appendix H.

The second model, using the high score only, was very slightly more reliable, according to the statistical output mentioned above, but it made considerably fewer demands computationally. For that reason, the high score model was used to produce the Item Characteristic Curve below for the overall test in Figure 47 from a Partial Score Rasch analysis.

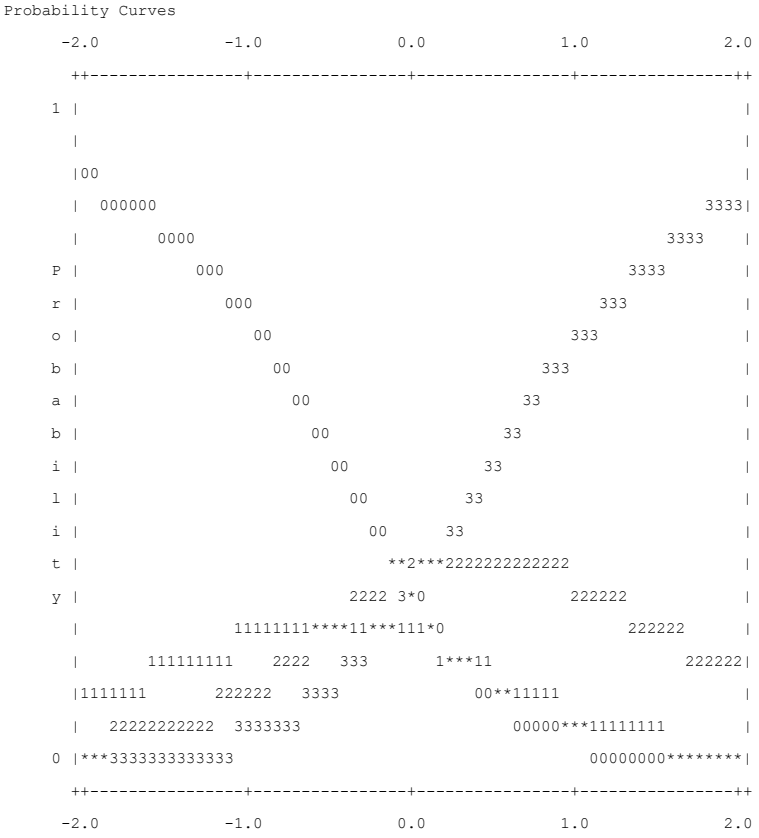


Figure 47 Item Characteristic Curve for the 4-band Partial Score model from high score data set

It shows reduced discrimination powers at Bands One and Two (Figure 47), with both curves producing a fairly flat line. The discrimination power at bands 0 and 3 was stronger. It should be noted, however, that there were significantly fewer performances in band 3 and therefore less evidence in that band to reach conclusions. There is also still the unresolved issue that, although the macro reports mentioned at the beginning of this section show a similarity, the individual performances shown in Table 24 show that results at the level of individual games are sensitive to modelling choices, and so this ICC is best treated as indicative only.

Looking at whether the games are suitable for the sample of children playing them, Figure 48 shows the Wright Map for the high score, including the child scores.

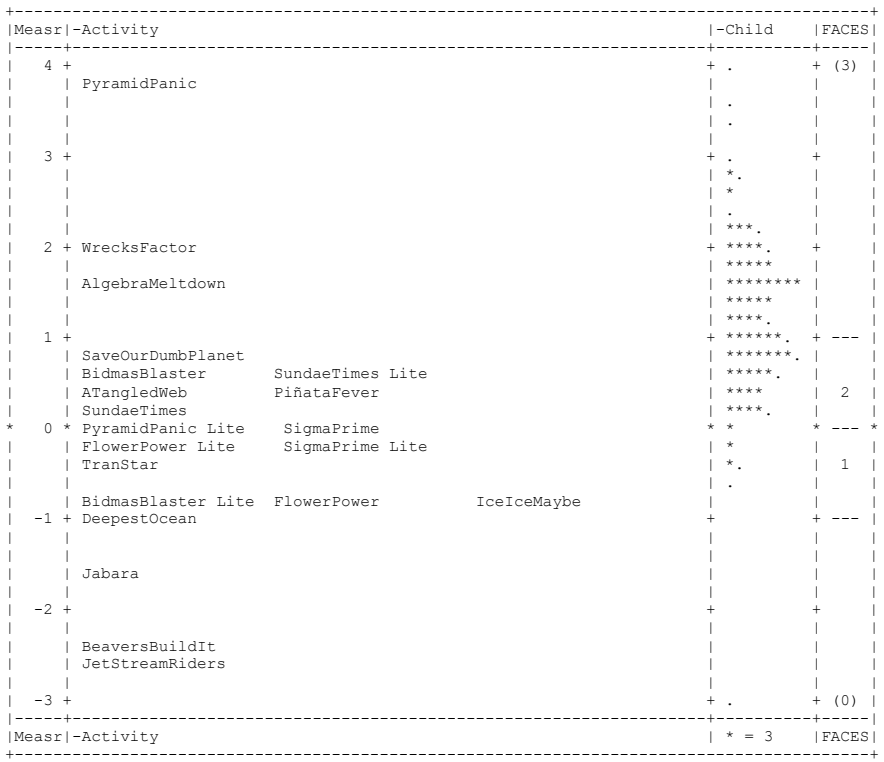


Figure 48 Wright Map for the high score data set

The games were slightly skewed to the difficult end, with a mean of 1.22 (Figure 48), although a few games (Pyramid Panic, Wrecks Factor and Algebra Meltdown) were largely responsible for that pattern, being significantly higher than the other games. The children were also very able, though, and were also skewed to the higher ability level, possibly because only their best performances were used in generating this map. In general, the majority of these children found the games fairly easy. The highest ability child had a logit score of +4.83, and the most difficult game, Pyramid Panic had a difficulty score of +3.87. The lowest ability child scored -2.74, but that child was an exception. All of the others scored greater than -1.0, and the easiest games, Beavers Build It and Jet Stream Riders had logit difficulty scores of -2.46 and -2.58 respectively.

6.6 Summary of the findings

There was a lot of information from this data set to consider. The group of people playing

the games were quite heterogeneous, and because of the element of choice over what to play, how long for and for how many times, the resulting data was also atypical in many ways. The results around whether ability was related to the reason for being missing showed that it is a complex question in games. The extent of missing data was huge and required some decisions over what represented a fair measure, the child's mean, high score or a score from a specific attempt. As van der Linden (2009) posited, response time did appear to need a stochastic treatment and was a function of the child, the time and the task, but there was also evidence here that it was a function of the band score, too. Response times were affected by the mechanics, and there were signs that plausible times needed to be established by a range of means to avoid punishing children who persevered with the games. There were human and robot competitors and collaborators who may have accounted for the fluctuating patterns of scoring or may have placed limits on the child's ability to score well. All of these behaviours summarised here have not been documented before. There is very little written on how to interpret the patterns and make decisions, and so the next chapter will consider the implications of these findings and either draw parallels or contrasts from the pre-existing literature to help understand what was happening.

Chapter 7 Discussion

The results produced insights into the behaviour of children in gaming environments that were interesting and unexpected. For some of the patterns that emerged from this research with this particular data set, there is a body literature from the broader field of assessment that provides guidance on how concepts might be reinterpreted to include these new variables. There were a large number of opportunities for better design choices to be made in creating task writer guidelines and a game data analysis strategy for future work in this field.

There were also behaviours that were completely new, and with no supporting substantive theory. This research was designed to be a quantitative case study, but the decision over whether data supports or refutes the claim that we are measuring ability, and nothing else, has to be embedded in a qualitative environment of gameplay and data collection. There were times when the achievement was to identify areas where more information about the context is needed, and so this study does not resolve all of the concerns. It does, however, suggest the next steps that alternative research designs could target for study.

This chapter will start by looking at who was playing the game, and how they played, as this information is important in a longitudinal study with a heterogeneous sample and imbalanced data sets. There were hints in the results that children in certain global locations might be disadvantaged by the games design at times, and the reasons for reaching this conclusion will be discussed, with suggested solutions. There was also the issue that the children themselves had changed considerably over the period of data collection, growing up over the years, and that introduced several concerns not yet discussed from the literature.

The question of what to filter out was not discussed in the studies in the literature review in

chapter three, and the value and limitations of the approach taken in this study will be discussed. The conceptualisation of missing data may need to vary at the different stages of assessment analysis, and this will be considered below. The attempts made in this study to rule out selection on the basis of ability were inconclusive, and the solution may not lie in a robust definition, but a slightly different game design to restrict the amount of missing data.

One key finding of this study was the complexity of using response time as a second, or only, proxy for ability. 'Speed' was affected by a large number of compounding variables such as the games design and game mechanics, the rules of play, and the presence of human and bot competitors, and is challenging mathematically to model. Going forward, the bots which complicated this assessment process may actually provide a good solution to measuring ability in games. It will also be argued that a novelty parameter, rather than a familiarity parameter, can improve the quality of the assessment design.

All in all, at every stage, this research suggested that a number of conscious decisions around the inclusion of new variables in the assessment process could reduce the noise in the data.

7.1 Demographic information

Initially, demographic information and descriptive statistics were gathered to provide general background information, and to suggest whether the findings here might be generalizable to a wider population. It is a limitation of case studies that the findings might be very specific to one particular group (Yin, 2003), and the cluster sampling used here meant that the results were not generalizable. Nonetheless, there were some features from the demographic review that raised interesting questions.

7.1.1 Imbalanced evidence

The number of play sessions per game and per child revealed evidence of possible bias in the results. In general, assessors are looking for a balanced structure in the pool of evidence from which to draw conclusions, and Item Response Theory works best with complete data sets (Wilson, 2004). Some individual games and some individual children contributed a disproportionate amount of evidence to the total pool of data. The imbalance in these games may not be a trivial concern, the fact that Jet Stream Riders contributed around 40% of the total evidence, for example, is particularly problematic. Aggregating data from the whole set to provide an overall impression of maths ability would been inappropriate, because it would focus too much on a narrow aspect of the mathematics curriculum, which is undesirable (Carmines, 1979).

The most issues appeared in the game Jet Stream Riders, the very popular platform game, and it was unclear whether this was because the game mechanics made accurate assessment challenging, or because the game's popularity simply created more opportunities for evidence of untargeted behaviour to emerge. The games developers have said that maths was not as centrally integrated into the gameplay of Jet Stream Riders as it was with the other games, suggesting the former. Of particular concern, though, was the fact that the fastest performances in a scoring band in this game were sometimes as low as 2 seconds. Assessment must be plausible to key stakeholders (Messick, 1987), and it seems unlikely that many stakeholders would accept that such short times provide enough an evidence learning. It may reflect the fact that in competitive games, a win can be achieved when a competitor concedes, however early on in the game, and dealing with and incorporating competitor actions will be returned to below. There is the possibility though that there are a large number of issues with games, which a larger study might reveal.

The number of times that the children in this sample played was not representative of the

overall population of MangaHigh players, as they were cluster sampled from the 90-95% most frequent player group, to capture frequent players but avoid the developers performances. It may be worth developing ECgD (Mislevy, 2015) into a set of more rigorous standards, such as those of the IEEE on learning design, to make the requirements of assessment designers clear. Data markers for the different types of player, become necessary if the designer understands that a calibration phase must be carried out.

Even though frequent play was a criteria for selection, the marked difference in the number of times that each child in the sample returned to the games may cause problems with dynamic Bayesian models. It appears to be possible to dynamically update the difficulty of the tasks and score the ability of the student simultaneously using Bayes (Almond, 2015), but there were three highly active individuals in this data set, contributing well over 1000 pieces of evidence each. If each new piece of evidence altered the difficulty estimate, these three children may have a disproportionate influence over difficulty estimates, which in turn would affect ability estimates. It seems sensible to anchor values (Wilson, 2004, Bond, 2015) in a pre-test calibration phase (Mislevy, 2003) using just one value per game per child. The challenge is to identify which value might be best, or to find a way to control creeping estimates.

There was further evidence from this study that gave support to the concern that an unbounded number of repetitions may be highly problematic. There was a difference between the difficulty estimate of JetStreamRiders on the FirstFiveAttempts only (-.22 logits), compared to the HighScore model (-2.58 logits), suggesting that this particular game became much easier with repetition. At 2.36 logits, it was the largest change in difficulty among the 21 games, but even so, four out the five games that were experienced as the most difficult at the level of the HighScore (PyramidPanic, WrecksFactor, AlgebraMeltdown, SaveOurDumbPlanet, BidmasBlaster) were also the five games that

were the least popular, with only 594, 236, 361, 441 and 571 plays each respectively among the 200 children. In general, those same games did not appear to be the most challenging on the FirstFiveAttempts.

It suggests that games capture performance data, as the children experiment, and not learning data from skills that have been acquired, discussed in section 2.2.8 (Soderstrom, 2015). Complete and total mastery may be the aim of knowledge based progress tests of skills like these, of mathematical skill such as time tables. Progress tests measure the extent to which a particular lesson was understood, and they can be quite granular (Black, 2009). IRT is more commonly deployed on proficiency tests of a broad skill base (Bond, 2015), and so there are questions around the most suitable way to standardise scores for the type of assessment (Baird, 2018) that remain unanswered. Although a distinction is rarely directly stated in the literature, proficiency measurements of complex skills seem to be the goal of the Games Based Assessment design community.

In general, the evidence here suggests that with so many iterations, the difficulty estimate and the child ability estimate will have a tendency to creep upwards unless only one measurement per child, such as a first attempt or their high score, is selected. Delivering the game as a simulation, rather than a game should solve the issue. Some in other fields, who have proposed the use of elo ratings from chess analytics with psychometric measurements (Regan and Biswas, 2013), which has features to control this upwards drifting of scores, and this may also be an option worth exploring.

Attempts were not made to see if performance correlated to age, and gender differences were not brought into the scope of the study, which, perhaps, might be of interest to educators.

7.1.2 Demographics

There was a wide range of ages of children playing the games in MangaHigh. In schoolbook publishing, the market is generally segmented by age. Books which target young children have bright colour washes and cartoons, and a very different visual feel from books for older teenagers. These games seemed to attract every segment of the school market, despite having many of those features of books for young children.

Figure 49 Screenshot of BeaversBuildIt contains bright colour washes and cartoon characters ©MangaHigh

Some of the games have 'babyish' features, in particular in terms of visual appeal (Figure 49), but they still managed to appeal to older users. In fact, although 18 years was the upper limit to be included in the final sample, a number of adults appeared in the top 90-95% of users. This may be significant as educational games can require a significant investment of time and money to build. Development teams, software licences, data and equipment are expensive and grow exponentially with the complexity of the game. Rockstar games, for example, invested $265 million in the development and marketing of *Grand Theft Auto 5* at the time of being built in 2013 (cited in Wikipedia, accessed September 20th 2019). Games Based Assessment will likely be more expensive to develop than paper-delivered assessments. The fact that the games in MangaHigh clearly appealed to a wide demographic and geographic base, too, may be an important commercial

consideration in the future development of games for assessment.

Overall, the sample population was relatively young, though. The mean age at the start of game play was around 8 and a half years, with 68% of the children between 7 and 10.5 years old. In general, there has been progress in identifying developmental red flags with very young age groups, but providing overall measurements of ability can be fairly problematic (Shepard, 1994). Young children need to be socialised to learn the rules of being a test taker, and assessment methods need to be sensitive to the age of the children. Young children also need time to develop their hand eye co-ordination and the fine motor skills, and will typically not become fluent readers until 9-15 years old (Ioannou-Georgiou, 2003). All of these can affect assessment processes. The children in the MangaHigh data set were reliably getting the answers right or wrong, and the games seemed to measure either one latent ability, or two that co-occur frequently (Wilson, 2004; Bond and Fox, 2015), but another research design would be needed to say whether that latent ability was definitely mathematics (Messick, 1987). Nonetheless, if assessing this age group becomes politically necessary, the overall reliability statistics of this data set seemed promising enough to justify further investigation that games might be compatible with this age group.

The geographical spread of the children could also be significant. Although the children can most likely work out the rules intuitively once they start playing, and this is often the goal of games design (Adams, 2010, Fullerton, 2014), there are instructions at the beginning in English. Very few children in the sample came from countries where English does not have a privileged or national language status. Presumably a non-English speaking child may have had a disadvantage, particularly in the first few attempts while they work out the game mechanics. There are several other issues around cultural assumptions that may not travel well (Mislevy, 2018), such as the heavy use of anthropomorphism for avatars.

In terms of geographical spread, the impact of time zones is also a cause for concern, especially when some children were geographically remote from the majority of the group. As discussed in Section 6.4.1, some games introduced a bot competitor, and when the bot played, it tended to win. From a play perspective, this means that children are still able to play dynamically when no one else is online. There are no indications on screen or any data markers in the telemetry set that the competitor is or is not human, and in fact the presences of bot performances in this study was only determined from the repetitive response time patterns.

It seems logical that children who are in time zones that are quite distant from the majority of users, such as the lone children in Singapore, Hong Kong or Thailand, will be less likely to be online at the same time as other players, and therefore more likely to have to play against the bot. The child in the Russian Federation may be anywhere from UTC +02:00 to UTC +12:00 hours, and there are also fairly large time differences across the United States. From an assessment perspective, this is problematic, as the bot appeared to be a challenging competitor. It never concedes early, and in these games, it always finished under the mean response time. This means that geographic location may well be a proxy for a confounding variable in those types of games because it affected who the child was likely to play against. Again, guidelines for games designers might require data on competitors to be recorded so that this can be included in analysis.

The length of play variable showed that this was a longitudinal data set that appeared to track the progress of some of the children from 4 to 11 years old. This seems to be further evidence that this data set contains performance data (Soderstrom and Bjork, 2015). On the other hand, it may be that in virtual games ability grouping surpasses the notion that all children grouped into the same year should follow the same curriculum, regardless of how well they coped with their lessons in the previous years. Next steps analysis was out of the

scope of this study, but through manual inspection, children often chose a game to repeat in the same play session, suggesting that the child was in a state of flow (Cscikszentmihaly, 1975). If the state of flow is also linked to the level of educational challenge, as in Vygotsky's (1987) theory of the zone of proximal development, it may be that the online environment groups children according to ability, rather than age cohorts. There was a wide age range playing the same games.

The relationship between age and ability expectations is a broad question, and one that games designers may not need to take a position on. Age has featured as a common variable in Differential Item Functioning (DIF) (Almond, 2015). Because of the longitudinal nature of the data set, it contained a large amount of historical data, and it was unclear how long earlier scores for children should be kept on their records. If the model of assessment gave equal weighting to all evidence, or was based on a Bayesian model that allowed a prior to be carried forward (Almond, 2015), some very old data may still find its way into the assessment. Assessors may need to take a substantive position on how long results are still valid (Newton, 2014). This could be done by weighting the Bayesian prior to have considerably less impact than new data (Lunn, 2012, Almond, 2015), but there were also questions around whether the most recent score is representative of ability, which will be discussed further in 7.4.3. An alternative, and possibly preferable, approach would be to remove records from the scoring model after an agreed period of time.

As well as being problematic on an individual level, very long periods of play also introduce the issue of cohort improvements. Year on year, students at the same stage of study get better at taking tests, and grades go up. This 'grade inflation' can often simply reflect improvements in teaching, better resources year on year, or growing familiarity with the test structure and requirements (Kohn, 2002; Flynn, 2007). About every 10 years, to accommodate this, many high stakes tests of national importance either have to make

qualitative changes to the syllabus, or have to adjust grade boundaries, as was the case with the 2018 GCSE revision in the UK. At seven years, the time periods in this sampled data set may be large enough to be affected by this phenomenon. Given how the data is stored, isolating annual cohorts from the sample would have been a challenge. Age could only be guessed from self-reported data, and so this phenomenon could not justifiably be tracked based on the available evidence. It would be desirable to factor a means to measure it into the design of the data collection process as good practice.

All in all, the demographic and contextual information extracted revealed features that give evidence to the argument that the data needs some analysis and treatment before sending it to a scoring model if the target is to isolate ability. A number of interfering variables should be considered and ruled out of the measurement process, where possible.

7.2 Research Question 1: What counts as a valid attempt on task?

There were many considerations around what was and was not a valid attempt at a task. While it may be challenging to identify times when children were trying their best, there were some response times that made it highly unlikely or impossible that this was the case. Both very short and very long response times were indicators of non-targeted behaviours.

It was easy to justify the deletion of cases that rounded to zero. Some of the children are very young, possibly with poor motor control and keyboard skills, and browsing or accidentally clicking on the wrong link is not evidence of a failed attempt. The impact of deleting those records did not appear to affect the mean difficulty, but if the measure was an overall mean, the scores of some individual children who browsed a lot went up after removing them.

A clear definition, backed up by substantive evidence of what counts as a failed attempt should be agreed, and there are a range of methods from the field of Human Computer

Interaction (HCI) that may be suitable for this (Cairns, 2008). All such decisions also need to be reported back to other decision-makers that come later, such as teachers, but also have to bear in mind the role of construct validity in assessment (Messick, 1987). Messick, a highly influential figure in assessment, has convincingly argued for a unified judgment on empirical evidence, based on statistical and also wider criteria (Messick, 1994) and teachers may themselves have an idea of the minimum length of time the child should be playing before suggesting that the child has been given adequate time to demonstrate mastery. In addition, one child managed to score a silver medal after 6 seconds of play, which appears to be what educators might consider 'a false win', where a player wins if one or more competitors concede. A position on 'false wins' would ideally be taken at an early stage.

Returning to the idea of the Flynn effect (Flynn, 2007), even when reasonable minimum time periods are defined at the start, it is possible that children get faster and better than previous cohorts as time passes, and those reasonable minimum time limits need to be adjusted. To avoid punishing the most able children who do manage to score well in very short periods of time, it seems desirable to flag up unexpectedly short response times with scoring behaviour every time, so that the Flynn effect can be monitored.

At the upper boundary of response time, it was implausible that a child played a mathematics game continuously for 95 hours. It seems to be miscoding of the data, where the child wandered away from the game but it stayed open. Assessment designers are used to anticipating how their tests might be experienced in socially diverse contexts (Mislevy, 2018), but the child going off for dinner or to catch the school bus halfway through the test are not familiar contexts for assessment. The point when the evidence crosses the boundary from being unlikely to being implausible (Hawkins, 1980) needs to be identified. In the less popular games, like Jabara and Bidmas Blaster, wandering away seemed to hold

considerable weight on the estimation of the mean response time, but even in games where there were enough plays recorded to balance out these extreme values, such as Jet Stream Riders, the size of the standard deviation (mean = 82.36 seconds, sd = 12,698.64 seconds) indicates that there is still a problem around pooling all of the data for speed estimates. The question remains around where that point might lie. The fact that the mean scores, as well as the mean response time, of the game Transtar were affected by capping at the upper boundary suggests that plausible times need to be estimated for each game and also at each grade boundary to avoid punishing persistence in challenging games.

There is a question over whether these values would have been better deleted, as is more typical with outliers (Hawkins, 1980), if they are so strongly associated with non-scoring behaviour. More research is needed to know whether walking away from the game was related to the variable of ability, and thereby to know whether to treat the task as not presented, as we do with other behaviour not related to ability (Mislevy, 1996; Ludlow, 1999). An alternative would be to impose time caps within the game, which may also resolve a number of issues discussed under response time below in section 7.4.1, or to shut down when no new activity from input devices are recorded. Power tests have proven impractical in most other test settings (Lee, 2015) and that may be true of games, too.

This study could only conclude that both short times and wandering off appear to evidence confounding variables and they should be taken into account when designing the rules of play.

7.3 Research Question 2: Does missing data impact the final score?

Overall, the full dataset had very large amounts of missing data. Once the invalid attempts discussed in 7.2 were removed, the ratio was around 18:1 missing to useful data, which was problematic. DiCerbo (2014) made the analogy that games contain an ocean of data,

with a huge volume of information, compared to the desert environment of a standard test, with perhaps just 40 responses recorded. In many ways, that is true. The whole original data set had over a billion data points on over a million children. Additional data, though, does not necessarily equate to more precise measurement, especially if some of that data is repetitive and redundant. If children were playing the same game up to 500 times, not all of the useful data in the 18:1 ratio above necessarily added anything new to our understanding of how the child was performing. DiCerbo (2016) did concede that we need to find a way to make use of all of those additional data points.

One limitation of the size and scale of the data set was that, at over a billion data points, only higher level information could be stored for any length of time. The information on medal performances was not at a granular level, such as each step they took as they completed the task, which presumably would be more useful for calibrating the task (Mislevy, 2015). The power law distribution of the graph of the most popular games (Figure 34) is typical of data driven by human choice and common in online analytics (Shirky, 2003), but it does not suit assessment purposes. It just meant that more and more evidence was accumulated over a limited spread of tasks, raising the concerns over construct validity (Messick, 1994) raised above in 7.1.2. All in all, more data does not mean better evidence.

More specifically, the hypertext structure of the MangaHigh website encouraged a power law distribution, and therefore a large amount of missing data.

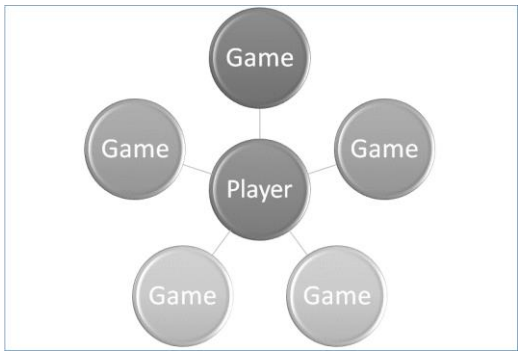


Figure 50 Structure of choice in MangaHigh

The games in MangaHigh placed the player in a central position at all times, and that person could choose any of the 21 games in any order (Figure 50). It creates a very high number of unique pathways which is problematic (Mislevy, 2014), and in fact it is highly unlikely that there were any identical patterns of play among the children, given the possible number of combinations.

Other common hypertext structures (Fullerton, 2014) might be better suited to assessment purposes by virtue that they elicit fewer alternative pathways.

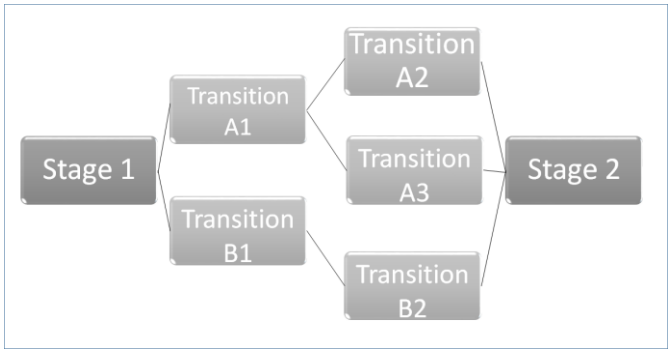


Figure 51 Alternative structure of choice that might be more conducive to assessment purposes

The *Choose Your Own Adventure* style games can lead players down temporary pathways, but bring them back to a central route at key points, shown in Figure 52 (Fullerton, 2014). From an assessment perspective, as well as reducing the amount of missing data, these key points in the pathway offer a chance for seeds, which are quality checks on consistent scoring (Baird, 2018) to help in the standardization process. It may also facilitate the coding, as individual lines of code can grow exponentially when very complex choice

patterns are made available (Adams, 2010), which may be an issue with procedural skills. Alternatively, a system of encouraging children to choose certain pathways to play new games with rewards or unlocking features might produce breadth as well as depth of data.

The main problem with the data that is missing is that it is challenging to say whether or not it is associated with the target variable of ability, and this is something that needs to be resolved (Allison, 2001). Not clicking on an optional game link in a hypertext environment does not seem to be the same behaviour as, say, avoiding question 4 in a linear sequence of 1-10 questions in a 10 point test. It seems to have more in common with the concept of 'not reached' tasks (Mislevy and Wu, 1996).

However, when one of those choices is labelled 'lite' and another 'regular', then it seems entirely plausible that the child chose to play one of those two options on the basis of that label, and therefore the target variable, their ability, or their self-perception of ability, impacted that choice. If that is not the reason for data being missing, it may be better to use Mislevy and Wu's (1996) solution, which was to treat items at the end of the test, or in the MangaHigh case 'games not yet played', as Missing At Random (MAR) when estimating difficulty and therefore ignorable. They argued these should be included in the ability estimation as failed attempts. One piece of evidence suggested this may not be appropriate.

The estimations of difficulty for some of the lite and regular games were hard to reconcile with the intentions of the games designers, as, according to the data set using the high score from the Main data, Sundae Times Lite and Flower Power Lite were harder than their regular versions. The infit and outfit for the regular and lite versions of SundaeTimes and FlowerPower are a little high, which suggests 'cruising' behaviour, where able children make mistakes on tasks below their level of ability because their efforts are focused on the more challenging tasks that are at their level (Lineacre, 1998), but the evidence of that is far from conclusive. Children may well play the 'lite' version first, make mistakes, and

then play the regular version once they have more mastery, or there could be a mistake in the games design. Next steps analysis from Machine Learning approaches might produce a data trail to pass better judgment on the other behaviours in this list (Anjewierden, 2012).

It could also be that the able children did not play the lite version at all, and only the lower ability children contributed evidence. If there were indeed, few children playing both games, it would be problematic (Bond, 2015). Partitioning the highest scores from the Main data set into high and low ability groupings was intended to investigate whether the lite and regular games followed a more expected pattern among a more homogenous sample. The ordering of the tasks was relatively stable across the three divisions and so notions of exchangeability held (Wilson, 2004), but that did not resolve the problem that the lite game should have been easier. It suggests that self-selection or ability was not the issue, though, and therefore perhaps the reason for missing could be MAR (Allison, 2001), but more information is needed.

There was one clear difference between the high and low ability group's behaviours, and that was that the high ability group had played a narrow range of games, almost half that of the low ability group. It may be that such narrow specialisation or repetition gave them an advantage. In which case, they benefitted from missing data being treated as ignorable. The infit and outfit statistics suggest that the higher ability group produced more stable results than the lower ability children. Table 25 puts the results for the groups together for comparative purposes.

Table 25 Overall performance for five games in MangaHigh, using the whole sample, HighAbility and LowAbility groups using the HighScore data set for five games that offered lite and regular versions

Calibration of the different sample divisions in MangaHigh in ten games						
	Whole data set		HighAbility		LowAbility	
	Infit	*Outfit*	*Infit*	*Outfit*	*Infit*	*Outfit*
Regular	1.09	.84	.87	.88	1.25	1.24

Lite	.98	.89	.79	.73	.85	.82

Overall, the high ability students produced consistently more stable results than the lower ability children (Table 25). This may be evidence of the stable quality that Soderstrom and Bjork (2015) associated with learning data, in other words, the point where the skill transitioned from being acquired, to a skill that is established. Stability of performance may therefore correlate more closely to ability in games than has previously been thought in the wider field of assessment, but it is not possible to reach such conclusions from this research design.

7.4 Research Question 3: How can the game specific variables of response time and iterations be conceptualized?

The two variables of response time and iteration are often held constant in assessments. Just the physical measurement of time or the number of additional attempts did not seem to capture the full complexity of these variables. From the results, there were several confounding variables that influenced both of these features. Giving a child complete control over their response time and additional attempts may not be the best option, for the reasons discussed in 7.2.1 and which will be expanded on below.

7.4.1 Response times

Response time seems to be a challenging variable to use an alternative proxy for ability. Looking at the interquartile range of the response times for each band in each game, using the whole Main data set, a simple linear model that uses the physical recorded time, or treating time as a manifest variable, does not seem appropriate. Even after removing invalid attempts, there seems to be a probability of responding in a particular period that depends on the game, as posited by van der Linden (Van Der Linden, 2009). This data also

suggested that speed in games is also dependent on the band score and game mechanics.

There were times when speed was an indicator of ability, typified by the race to the finish type of game. The most obvious example was the platform game, which had more in common with speeded tests of recall (Gulliksen, 2013), in this case Times Tables recall, than with problem solving. That kind of game was suited to incorporating a θ value for speed, and it also seemed justifiable to condition the speed estimate on band grouping, to reflect an accumulation of accurate responses, as well as speed. It had much in common with speeded tests. There is, however, a problem with this approach.

The use of a mixture of bots and other human competitors, mentioned in section 7.2, makes these games more exciting to play, but makes them problematic to evaluate as the child's score reflects both their own ability and their competitor's ability, best illustrated when one competitor concedes to the other in less than 10 seconds. An apparent solution would be to condition the score on an additional variable, a competitor ability ranking, which could be similar to the severity weighting of human judges (Gwet, 2014). A win against a strong competitor, such as the bot, might be weighted more favourably than a win against a weak competitor.

With this particular data set, the games designers had not chosen to record details of competitors, and so this was not an option. Even if they had, though, this solution has a flawed assumption behind it that competitor ability ranking is a stable entity. Seven-year-old children may not meet that criteria, and certainly the sample results illustrated in Figure 47 provides strong evidence that there is a great deal of fluctuation in their performance. Human competitor ability is not stable. If fair results really do matter, another solution would be to impose a constant competitor ability ranking by only using the bot, or a range of bots with stable ability, for evidence that is to be escalated to a scoring model. Seeds allow standardisation (Baird, 2018) and given that Jet Stream Riders, for example, had up

to 30 competitors at a time, a few bots of various levels could be seeded into every game play.

PiñataFever was an example of a game that rewarded higher bands only to those who persevered and played for longer. Grit and perseverance have been associated with school success (Laursen, 2015) and so these are positive qualities to foster. There are many questions about the value of generating θ values for speed with this type of game, though. If the number of tasks is held constant, then it seems sensible to generate a speed estimate (Thurstone, 1937). That may still be problematic, though, for the same reasons as the concern mentioned in section 2.3.1. If the programme randomly generates a set of sums, which would seem necessary to make the game less predictable and therefore more engaging and encourage flow (Csikszentmihalyi, 1997), the children could get a very lucky or unlucky set of questions. Only estimating individual speed estimates for each sum would overcome this problem (Van Der Linden, 2009).

Another option is Computer Adaptive Testing (CAT), which may solve this. With CAT, each question in a test is pre-tested with a sample group of test takers, to establish difficulty values (Van Der Linden, 2009; Van der Linden and Glas, 2000), or in this case they would be speed values (Van Der Linden, 2009). These values are anchored, or fixed, a priori to the live use of the test. An algorithm gradually refines the range of questions that the child is presented with in response to their answers. If they get the question correct, a slightly harder question will be presented. If they get it wrong, they see a slightly easier task. From the systematic literature review, Göbel *et al.* (Göbel *et al.*, 2010) proposed a model of adaptive computer games for learning that may enlighten how this could work computationally, although he did not cover evaluative elements. A similar approach as CAT could, in theory, be used to generate a range of tasks with known speed parameters, but it would be a complicated model. There would essentially be two competing proxies

for ability in such a model. Does the computer select a task to present according to its difficulty or speed? There may be a mathematical way to balance or estimate from both sets of values, but there would probably also need to be a strong assessment reason to justify the additional computational complexity. It was unclear from this type of game in this data set whether speed was such a strong indicator of ability that it was worth these extra costs.

Other games, such as Jabara, produced patterns of game play that were hard to classify. The mean of the response time for each of the scoring bands in Jabara was fairly similar, but there was a dip in band 2, and the standard deviations varied considerably by band. Others, such as ATangledWeb peaked in band 2. This was more in line with the older predictions of Maris (2012) that longer speeds are associated with test takers operating at the limen of their ability. But that is in contrast to the pattern in Jabara, where longer times were associated with novices and experts. Conclusions on games seem to need a different research approach, such as observation or key stroke analysis, to understand how response time operated in these games.

Much in line with the research from education on including a speed scale, as well as an accuracy scale, the analysis suggested that speed may measure many different behaviours. There are many possible conceptualizations, but in all of these, the mechanical measure of time alone does not appear to be adequate as a proxy for ability.

7.4.2 Producing speed estimates

If response time is to be used as a latent variable for ability, a number of factors need to be taken into account. A logit estimation was carried out on the shortest response times for each child in each game, separating results out for each of the 3 scoring bands. The results suggest that the games were actually quite similar in terms of overall speed challenge. For example, in band 1, 9 out of the 21 games had a speed value of 0 logits, and the overall

range was only from -0.4 to 0.2. The order of some of the games did change for the different band scores, though. It may be desirable to condition scores on bands. Games that appeared problematic to estimate a speed from, just looking at the box plots for mean response time, did not seem so problematic according to this estimation. PiñataFever, for example, produced a difficulty estimate of zero across all three bands. However, there were limitations in the mathematics used to process these results, and these Wright Maps are an imperfect measure. Estimating continuous variables proved very challenging, and these problems will be discussed further in the limitations in Chapter Eight. As a result, no more analysis was carried out on response times, and very little should be concluded from the current findings on the logit estimations.

7.4.3 Additional attempts

Using data from the First Five Attempts only, there was a marked decrease in the overall difficulty of the games from the first to the second attempt, and then marginal decreases after that. In theory, this change in difficulty can be quantified and used to weight scores accordingly, but that approach has several problems. To start with, using the available software, Facets, the statistics could only be generated for the data set as a whole, not at the level of individual games, as it treated the 'attempt' parameter as though it were assessing the impact of a lenient or sever judge (Lineacre, 1998). It seems inappropriate to use a generic 'additional attempt' parameter value across all of the games, as presumably the games became easier at different rates.

Another problem is more practical. Going back to Almond *et al*.'s (2015) familiarity variable discussed in 4.5, it was a dichotomous variable and the weighting values seemed to be constants (see pages 167-170). In Games Based Assessment, the number of attempts each child made causes problems because it is a continuous variable. It would be impossible to model an unbounded number of subsequent attempts. Every time a child

goes over the existing maximum, for example, they become the first and only player to play for the 462nd time, and there would be no data to create a weighting value from. It seems sensible to identify a point where the diminishing returns of additional play become minimal, and instead add a 'novelty' weighting to earlier attempts. In the case of these games, the child might be given additional novelty credit for scores obtained on the 1st, 2nd and 3rd attempt, which appear harder, but by the 4th and 5th, the benefits of repeating seem to have become minimal, and the straight score could be used.

A novelty weighting instead of a familiarity weighting is not without problems, though. If the scores are to be reported on a continuous scale and used for high stakes tests, then there is a risk that a child could go over the maximum limit if they also receive a bonus weighting for early success. This happened in the French Baccalaureate in 2017, when a girl scored 21.29 out of a maximum of 20 as a result of using a similar weighting approach to behaviour considered more challenging (Baird *et al.*, 2018). Alternatively, the impact of additional attempts could be controlled by either delivering the games as a simulation or limiting the number of attempts that a child can make.

Although a descending pattern of difficulty appeared overall, on an individual level, a far more random pattern was observed. The patterns observed in line graphs from individual children playing individual games (Figure 31) show that the grade seems to depend on modelling choice: Whether to use the mean, the mode, the high score or the most recent score. Even if the assessment design team have substantive grounds to decide that one of these options is fair, there are further questions that need to be addressed. If the scoring model uses a mean, for example, two children with the same mean of 2.2 could have got there by very different routes, shown in Table 26.

Table 26 Observed scoring patterns for two children over 5 attempts

	Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5

Child A	3	3	3	1	1
Child B	1	2	2	3	3

If Child A from table 26 was rewarded with a novelty weighting for getting higher scores on the apparently more difficult first and second attempts, on aggregate their results would be higher than Child B. This does not obviously reflect the difference between them at the end. A recency weighting is fairly easy in Bayesian IRT approaches (Almond *et al.*, 2015), as the weight of the prior can be reduced and the weight of the new data strengthened, but there were few signs that the most recent score was typical of the overall performance of the child.

From this data set, the infit and outfit values suggested that the first attempt contains poorer quality evidence than subsequent attempts, and so it may be better not to use the first trial run in the case of these games. ECgD highlighted that iterations were important (Almond, 2015) but there is still a real need for a body of literature on this aspect of games.

7.5 Research Question 4: Within the game, how reliable do the results appear to be?

The main conclusion on this final question reflects many of the concerns already discussed. Although the final results suggest some degree of stability, there were many patterns in the data set that challenged the notions of validity and reliability. Messick's (1987) principles of taking a wide range of statistical and qualitative information into account when creating a validation argument are well-respected in the field of assessment. It is clear that there delimiting the problem space of scoring from gameplay data by introducing robust assessment design will be helpful. Many of the issues raised in the Literature Review in chapter 3 seemed to come from the problem of making the most of a pre-existing data set. The games in this study were designed to be used as a homework tool, with some

evaluative purposes, and without doubt, a more targeted assessment design could reduce some of the problems.

From this data set, there seemed to be more potential to use the child's high score to produce stable estimates, rather than a mean. There is a political and social precedent for using the high score. When children take school leaving exams and get good grades, they do not go back again and take the test a month later to see if their skills have slipped. It is difficult to say whether the fluctuating pattern evidenced skills slip or one of the other behaviours, but it may have. The US Navy, however, has started to offer periodic re-tests of skills (Swartout *et al.*, 2016) as part of the permanent records on training and qualifications of staff. The US military has a historic record of innovation that has become adopted in more general education (Smith, 2010), as mentioned briefly in section 2.4, and skills slip may become more significant in future.

The high score may be preferable for another reason. It was challenging to say what the zero scoring band measured at times. Untargeted behaviour such as the browsing and wandering off were already discussed in 7.2, and it was possible that strong competitors, like the bot described in 7.1.2 had forced a no-score result. A risk of stealth assessment (Shute, 2011) is that if the players were not explicitly aware that every performance was being recorded and would be used to judge them, it is very possible that they did not invest effort in every attempt, also resulting in a zero score. For those reasons, a mean estimate that captures the zero bands might not be preferable.

One aspect of these games that is particularly difficult to reconcile, though, is the changes in rankings of the difficulty of the games between the First Five Attempts model and the high score model. The game Beavers Build It, for example, goes from being the most challenging games to the second easiest, which may reflect the fact that it is intended to be the first game that children play in MangaHigh. Next steps analysis (Anjewierden, 2012)

might confirm that players followed this pattern. It adds further weight to the idea that there is a lot performance data in this data set (Soderstrom, 2015) as it is plausible that the apparently challenging level of this game in the first five attempts just evidences a process of familiarization with games in general.

An alternative explanation could also be that Beavers Build It is also a collaborative game. It is possible that there were three latent skills in this game: Mathematics, gameplay skills and collaboration. Because of the way that the data was recorded, the soft skills were captured in the overall scores, but not isolated for individual measurement. However, in all of the types of games that introduced a soft skill, the game became easier with repeated play. These were:

1. Collaboration in Beavers Build It
2. Competition in Jet Stream Riders and Sundae Times
3. Managing a system in Flower Power

If there are, in fact, three latent skills in all of these games, it is possible that the children enjoyed the original types of challenge, and were motivated to play them more often (Csikszentmihalyi, 1997) and as a result of repeated play, the mathematics, game mechanics and soft skills all became easier through practice. Where there was possibly more than one latent skill being evidenced, Almond et al. (2015) suggested that there may be three different types of inter-relation:

1. **A conjunctive distribution, in other words all three skills were necessary to perform the tasks well**
2. **A disjunctive distribution, where there were alternative ways to solve the task and not all of the skills were in use**

3. **A compensatory distribution, in other words the presence of one skill was enough to compensate for the lack of another skill (Almond *et al.*, 2015).**

Looking at the way that the games are structured, it seems most likely, with the exception of Jet Stream Riders, that some form of conjunctive distribution with one of the skills being mathematics was in play. It does not look as though it would be possible to score without being able to answer the mathematics questions. This returns to Messick's (1987) contention that reliability does not equate to validity and a different research design would be needed to pin point the range and interrelation of skills that are evidenced in each game.

A causal relationship could not be stated confidently, but it seems unlikely that the games became popular because they were easy, but rather that a particular game became easier because they were popular. That is because games that were both popular and easy by the time the children had achieved their high score were not experienced as the easiest games in the children's first five attempts at the game. Ericsson (2008) has argued that deliberate practice is the key to success, although for these children, it does not appear to be the grinding often associated with repetitive, boring tasks in games (Fullerton, 2014), but more a sense that they have achieved flow (Csikszentmihalyi, 1997). The number of times that they played through their own choice was higher than the number of times that they were challenged by another child or by their teacher, in other words, something, probably not the teacher, compelled them to go back and have additional attempts.

A final point worth noting is that the results here were skewed, and that any approach that seeks to categorise children into band groupings might need to take that into account, depending on the assessment paradigm in used.

As the discussion chapter as a whole has shown, there are many aspects of scoring educational games that have not entered the field of academic debate yet. A decision was taken not to estimate a SEM θ value for the overall test because the evidence did not point

to the fact that maths skills alone were captured in the score. It seems very early to suggest that any estimated degrees of confidence in the scoring mechanisms are meaningful.

7.6 Conclusions from the discussion

This discussion chapter has shown that there are limited resources in the literature that directly shed light on interpreting GBA results. For researchers approaching GBA from a strictly computer programming background, and not familiar with the authors cited here, the challenges might be even greater. There were many insights into opportunities to refine the data collection process, and in turn improve the quality of scores and say with more confidence what is being measured. This chapter also highlights the fact that work is needed both in refining statistical models, which is where Machine Learning and AI practitioners excel, but also in understanding the many other aspects to the games that statistical approaches cannot shed light on. As an overall point that perhaps touches on all of the previous issues, there was evidence that children have not been socialised into playing the role of ‘game player under assessed conditions’. Children appear to be making up their own rules in gaming environments, which may well be the popular appeal of these games, but that lack of socialization also seems to impact on the validity of the score, and until that changes, steps need to be taken to account for this type of behaviour.

Chapter 8 Conclusions

8.1 Introduction

The possibility of using gameplay data for assessment purposes is exciting, because even simple games often require skills that are highly valued socially. These games practised and gave feedback on performance in the late primary and early secondary mathematics curriculum, but they still required collaboration and managing different aspects simultaneously. Educational games can produce very large performance data sets. Computational power and machine learning (ML) algorithms have developed in recent years to find patterns in data sets that are too large to be inspected manually by humans. The hope of Game Based Assessment is that gameplay data collection and new analysis tools will mean that approaches to measuring skills in games can be aligned to assessment processes. In turn, we may be able to measure modern work place skills through games.

High stakes assessment demands high standards of accountability, and measuring from gameplay data is a complex process. The first challenge for assessors crossing into gameplay analytics is the structure of the data, which has already received some attention in the literature. Gameplay data contains conditional dependencies, which are not aligned to current methods to extract proficiency scores (Mislevy, 2014, Almond, 2015), however, some machine learning algorithms work with some types of conditionally dependent data (Mislevy, 2014, Almond, 2015). This study has found other problems, in addition to questions around the independence of responses. These are specifically, the complexity of dealing with a large amount of historic data; the need to separate data that captures learning processes from final assessment data; the effect of not filtering untargeted behaviour, such as browsing, from the evidence set; and finally, the effect of choice in producing highly imbalanced data sets. These features of the structure of data are

challenging to work with from an assessment perspective, and affected scoring when generating an overall proficiency score.

In the data, there was a task variable, with a score and a response time, and other factors were held constant in this study. It is not usual for assessors to hold one or two variables constant, even though we know that they may be experienced differently by test takers. An example would be a constant response time, when we know that test takers will have different approaches to time management. The hope is that better time management will be reflected in more correct responses, and captured in the overall score, and so a constant total time is used. From the results of this study, more environmental variables could be captured and added to the data sets for proficiency assessment. These included:

- Rules around response time;
- the measure of response time as a proxy for ability;
- the familiarity of the child with the game environment;
- the ability of direct competitors;
- the ability of direct collaborators.

The different game genres and mechanics appeared to underpin some of these distinctions.

In terms of solving these issues, this study found that there are some approaches that could, for example, weight for novelty of early attempts or robot competitor or collaborator ability. For other variables, it is necessary to return to some fundamental issues in the design stage. For example, it was hard to reach conclusions on the relationship between response time and ability, even though response time is a valued metric used in games analytics. Measuring on a scale of speed is complicated in games, as it is in other forms of educational and psychological assessment.

The objective of proficiency testing is important to establish in a discussion on Games Based Assessment, as it is not automatically a goal of games designers, who favour progress testing. Progress tests are longitudinal examinations that sample at regular intervals and are often used for diagnostic purposes. They typically allow educators to pass judgement on the acquisition of a narrow construct, such as a knowledge recall task like times tables, or spot problems. Proficiency tests, on the other hand, aim to rank children along a scale, such as 65% or Grades A-F, and be able to compare results in similar tests with some confidence. They are more often used to assess complex cognitive skills which cannot be measured directly (Mislevy, 2018), such as the overall skill of numeracy. The structural issues and missing variables in gameplay data are not problematic when results are treated as a progress test, as they are in the MangaHigh games.

The distinction between progress testing and proficiency testing seemed important from the findings of this study. The IEEE[6] guidelines for learning technology architecture are influential in the field of educational technology development. Although not always stated directly, the goal of progress or diagnostic test reporting underpin the advice in those guidelines. A series of binary choices (Right/Wrong) are scored with a 0/1 value, and the challenge for computer scientists is how to store and retrieve that information reliably, and find associations in the data set that might facilitate learning. This case study, for example, used games scored in line with their standards. Assessors working with similar right/wrong binary choice tasks tend to prefer to treat the value of that correct choice as an unknown, to be estimated from the data. That is the moment when the structure of the data and the limited number of variables begin to be problematic for assessors.

[6] IEEE Standard #1484.1-2003 Learning Technology Systems Architecture (LTSA) ©IEEE

The solution to proficiency scoring of valuable skills from games does not currently seem to reside in either discipline, games design or assessment design, but at an intersection of the two, which is where this study positions itself.

8.2 The challenge of scoring from games, and the literature review

In the more general field of games analytics, a large body of literature exists on the technical process of building a system to process and store data from one game to the next. The purpose is usually to produce analytics that will help to refine the game play experience or to motivate the player to persevere (El-Nasr, 2016). Designing a system that stores and retrieves important information about the child's learning is not a simple process, and this is discussed in more detail in 2.4.1. There are guidelines for programmers on how to collect and store data, but not what data is necessary.

The more specific field of Games Based Assessment tackles the problem of scoring in an ethically acceptable way, as well as a computationally efficient way. There is considerably less literature available in this field. Despite establishing broad inclusion criteria for the literature review, no relevant articles were more than 10 years old at the time of the review, reflecting that this is a nascent field, and only 41 texts were found, mostly from software engineering journals or conference papers.

Case studies of scoring systems dominate the field of Games Based Assessment (GBA). Studies tend to focus on finding a mathematical function that will best turn evidence collected in the games into a meaningful score, and this is an area where many of the authors specialise. This was discussed in detail in 3.3. The use of a mathematical function to represent performance is common practice in education and assessment. At the simplest level, teachers who write their own classroom quizzes will often record evidence as a

series of binary ticks, and then convert the total into a percentage, to create a score that is more easily interpreted. Many professionally designed assessments go further, and add an estimate of more precise values of individual ticks from performance data, rather than just use 0 or 1, and this process has created a large body of research, introduced briefly in section 2.3. They may also weight some of the evidence, or possibly remove some badly performing aspects of the test, in order to reduce untargeted bias in the scoring process. The research to refine mathematical functions that can optimise gameplay data is an essential part of the process, but should not be the only part. The findings here show that greater control over the data input into the model, and assessment design seem necessary stages in fair scoring.

Machine Learning algorithms appear to have much in common with the psychometric paradigm in assessment. They both take a stochastic approach to scoring, and rely heavily on the mathematics of probability. Games produce conditionally dependent data, and it is important to factor that into the choice of mathematical model. This was first pointed out by Mislevy, and Almond, and it is discussed in more detail in 3.3.2. Item Response Theory (IRT), the suite of statistical analysis techniques favoured by assessors who use a psychometric paradigm, and was discussed in 2.3.2. Traditionally, frequentist mathematics are used for IRT, and that requires an assessment design that produces discrete items and conditionally independent data, and to that end, many test papers have become meticulously engineered environments. Even when a standardised test assesses complex skills, such as linguistic competency, questions in a reading test will be carefully designed and worded to elicit the discrete items and conditionally independent data that is favoured by assessment analytics. Gameplay environments are engineered with a different set of principles, and the resulting data does not have that structure.

Some machine learning (ML) algorithms can handle conditionally dependent data.

Bayesian solutions dominated the literature, because of this requirement. Bayesian probability also addresses a second challenge of game analytics. Gameplay does not often take place in one play session, and so scores need to be carried over from one day, week or month to the next. The prior in Bayesian approaches has been found to be ideally suited to both tasks: carrying over previous scores and making judgments that are conditionally dependent on earlier decisions the test taker makes (de Klerk, 2014, Almond, 2015). Assessors working with game data showed a preference for Bayesian Knowledge Nets in scoring (de Klerk, Eggen and Veldkamp, 2014; de Klerk, Veldkamp and Eggen, 2015). One drawback of Bayesian Knowledge Nets is that some authors described problems computationally when the complexity in the game grows (Mislevy *et al.*, 2014; Frezzo *et al.*, 2009; Shute *et al.*, 2016). Alternative suggestions include the use of a rule-based solution that uses formal logic modelling to better understand the construct, suggested by Vendlinski among others (2010). The case studies from the literature review all reported limited success but offered insights into some potential interfering variables, as described in 3.3.

Instead of framing the problem as solely about processing the data in a mathematically optimal way, it is also possible to see fair scoring as an issue with the quality of the data that is collected, and the environment that it is collected in. The over and under-inclusion of data going into any model is problematic. Some authors described the problem from a cognitive domain modelling perspective, described in 3.3.2. An assessor would see this as the fundamental challenge of all assessment, not a possibility worth considering. There is a very wide and rich literature on how to get better data from tests. This includes research on interfering variables, which often has the aim of influencing policy and practice. For example, this research might inform guidelines on what is and is not allowed in the test. Interfering variables are usually controlled by an experienced assessment design team, usually by removing them before the test reaches children. Observations of this process are

discussed at some length in Mislevy's book, Sociocognitive Foundations of Educational Measurement (2018). It is worth bearing in mind that assessments need to exist in a social context, where several social groups place limits on what design teams can do.

While some factors that unfairly influence scores should be removed from gameplay data, the focus of this study was whether some of those environmental aspects can and should be incorporated into the scoring process. The second scoping of the literature for methods found research in this field. The first issue was varying response times. In some assessments, speed has been treated as a proxy for ability, but this is more common in psychology tests, or tests of very simple recall tasks. These theories are discussed at some length in 4.4. In educational assessment, Thurstone theorised that response time acted as a context variable, and that additional time made tasks easier (1937). Van der Linden (2009) refined this with a stochastic model of response time. He considered the overall shortest time was not necessarily the optimal performance, as some questions need less time to complete. For that reason, speed has to be conditioned on individual tasks, and this is particularly important when children complete different tasks in the same game.

Other contextual variables, not just response time, have been incorporated into scoring models. These included different socio-cultural attitudes to guessing, in Mislevy and Wu's work (1996), evidenced by missing responses, which in turn led to a broader discussion on missing data in assessment; and the introduction of a variable for familiarity with the test environment, from Almond (2015). These studies all suggested that scores can be weighted, or the value of correct responses be adapted, to represent these changing circumstances. The methods for this study were heavily influenced by findings from these key authors.

All in all, there is literature to support researchers working on the use of games based assessments but it is not easy to find at the moment, as cross citation between computer

science and assessment research seemed limited in Games Based Assessment.

8.3 Methods

A secondary data set was acquired by snowballing requests, and once two potential data sets had been offered, from around 50 organisations or individuals approached, the process stopped. Sourcing data to study gameplay performance is not easy. Data from the MangaHigh maths website were chosen because the construct being measured, primary and early secondary maths, has a well-established literature, and the website introduced very rich gaming environments into that familiar process. There was one platform game and two shoot 'em up games. The other nineteen games used the games mechanics of drill and animation, or avatar or object manipulation. Both require either an answer to be entered or an object to be dragged to the correct response. Correct answers are rewarded with a lively animation. Several games also introduced other skills such as real time competition against another child or a robot competitor, real time collaboration with other children or a robot, or simple project management. These features created a range of gameplay environments to explore.

The total data set (n>1million) was collected from children aged 10-13 years in real school settings. Scores were stored as aggregated band scores (0, 1, 2 and 3). A case was taken to be the collection of performance data for all of the games played by one child. An initial cluster sampling based on known criteria (Ahmed, 2009) took the top 90-94% most active players, to ensure adequate data on each child was available. At over 400,000 rows and 10 columns, the resulting data set was still an unmanageable size for the depth of study intended. Random sampling reduced the data set to 200 cases, and 32,213 games, which allowed for some manual inspection. A sample of 200 children was chosen because research suggested that methods based on a Chi Square approach would not lose too much power to detect effect size with that number (Hintze, 2001; Cumming, 2013b). For future

studies, it would be worth bearing in mind that gameplay data sets may be imbalanced. In some games and band scores, there was not enough data, and a larger sample might be required.

Although the games have an evaluative element, it is important to remember that the makers of MangaHigh did not intend them to be used for proficiency assessment. The prime purpose of the games is to motivate children to engage in classroom maths at home, and the gameplay data suggested that the children enjoyed the games, with some children playing the same game up to 470 times. With perhaps the exception of the platform game, the results suggest that the use of these game to provide general progress test feedback of a specific mathematical concept, taking only the high score into consideration, reliably represents acquisition of that particular targeted mathematical construct. This is the approach MangaHigh takes. Under this research design, it is not possible to say that a child with a low score has not acquired the mathematics, as keyboard skills and other skills were also required to complete the games. This study used the data set to explore the potential of adding proficiency scoring to the existing games, to see if a more general proficiency estimate of the skill of numeracy could be inferred from the performance data. If it can, there is a possibility that the approach could be used with other complex skills.

The research questions for this thesis were:

1. What counts as a valid attempt on task?

2. Does missing data impact on the final score?

3. How can the game specific variables of response time and iterations be conceptualized and scored?

4. Within the game, how reliable do the results appear to be?

The first question arose because estimating the value of a correct response in IRT, whether through Bayesian or Frequentist approaches, is often carried out on whole data sets (Raykov, 2011, Bond, 2015). An exploratory investigation suggested that some response times were implausibly short, less than 1 second, or implausibly long, lasting several hours. Response time was therefore used as a context variable to indicate whether the attempt on task seemed valid. On play testing, ten seconds seemed to be a reasonable minimum time required to go through the launch process and start playing the games, and most games could be completed in under 3 minutes. These short and long times therefore seemed to evidence either browsing behaviour or wandering off. In order to know if such results occurred frequently enough to affect measures based on mean response times, such as when calibrating difficulty, three hypotheses were tested using a one tailed paired samples t-test:

- The first removed times of less than or equal to 1 second.
- The second removed times of less than or equal to 10 seconds.
- The final one removed extreme response time values at the upper boundary.

Very long response times were challenging because, with no upper time limit, the games were essentially delivered as power tests. A power test with no fixed time limit has always proven problematic to implement in practice. In addition, in games, response time can be treated as a significant aspect of the gameplay experience, even when the game has unlimited response times. An upper limit resolves some of the issues discussed below, but there are ethical questions around imposing limits after the assessment has been delivered. For future studies, if a limit is to be imposed, it might be better to refer to Lee and Ying's definition of 'power time' as the time it would reasonably take to answer a question if time were unlimited (2015). Determining a reasonable time from observation of children at play was not an option in this study, and so a statistical "Tukey fence" approach was taken

instead. Under this model, extreme response times were defined as anything above Q3 + 3(Q3-Q1) (Hawkins, 1980). The null hypotheses in each test were that removing cases with very short or very long RT would have no effect on the calibration of the mean response time or mean score.

The second research question on missing data seemed central. The narrative structures of games force a large amount of choice (Fullerton, 2014), and that in turn leads to very imbalanced data sets. When those data sets, often stored in long format in the game database, are analysed in a data frame that requires wide format, it results in a large amount of missing data. In this data set, around 88% of the cells were empty before any cleaning was undertaken. Missing data matters in assessment because sometimes it can evidence the absence of the target construct, but it can also sometimes be forced by the assessment design, and contain no information on the child's ability. From assessment, the work of Mislevy and Wu (1996), and Ludlow and O'leary (1999) suggested that automatically treating missing data as zero can distort calibration and scoring. The assumption that an unanswered question evidences a lack of ability requires investigation, particularly in estimating the difficulty of tasks. This also an area where assessment principles clash with ML approaches, as some ML algorithms impute values that are missing.

The extent of missing data, as well as ambiguities around the reason for data to be missing, made analysis problematic. One solution is to gather less data, for example, just one value per child per game. The bar chart in 5.1.5 showed that one child could be judged to be in all 4 bands, 0, 1, 2 or 3, depending on whether the mean, mode, high score or upper quartile limit was chosen to represent ability. This is not an optimal choice in assessment. Instead, scores from the children's performance on their first five attempts only were taken, and this produced a data set that was around 50% complete. A frequentist partial credit scoring model was run on that data to estimate the difficulty value of those early

attempts (Bond and Fox, 2015). This model was chosen because, among the alternatives found, it allowed missing data to be treated as not presented rather than as zero or an imputed value (Linacre, 2006). As most children persisted in some games many times, their best performance in each of the 21 games was also extracted, and a partial credit scoring model was run on that, too. The initial assumption was that missing data were Missing at Random (MAR) and not related to the target variable of ability (Allison, 2001) because the narrative structure forced that situation. However, there appeared to also be some elements of missing data that were associated with ability.

The 'lite' games, which had the same game play environment and mechanics as the regular games but easier mathematics, were not experienced by the children who played them as 'easy'. It may have been that the children were self-selecting into an appropriate level, and so data were missing due to their ability. In that case, and the assumptions of MAR could not apply (Allison, 2001). To investigate this, the dataset was partitioned into a high (n=80 cases) and low (n=79 cases) ability group, with a small border line group removed from the analysis. A partial credit scored model was repeated on the two separate ability groups to see if there was more consistency between the games designers' intentions and the way that they were experienced. An alternative explanation may have been that both games were played by all of the children, and perhaps the 'lite' games were played while they were younger, but a different research design would be needed to investigate that possibility, and that was not pursued here. This is discussed in sections 4.2.3 and 5.5.2 more fully.

The first two research questions uncovered areas where there are conflicts of interest between assessment and games design, which creates structural issues with the data. The third research question looked at interfering variables or context variables, depending on how they are used. A body of research was found in the literature review covering varying

response times, and this is discussed in section 4.4. An early finding from the pilot showed that the value of speed needed to be conditioned on the task, but it also appeared to be necessary to condition speed on the band score. In some games, there was no overlap in optimal speeds for any of the bands. This was explored further by producing box plots for the different band scores in the 21 games. A basic estimation was carried out, using a natural logarithmic transformation of the response time in the different band scores for each task. This was to explore the nature of response time in games, not to extract a score.

The second context variable explored in Research Question 3 was the effect of familiarity, or additional attempts. Almond et al had dealt with a similar but distinct concept in Bayesian IRT (2015). They observed that when children had already explored some aspects of the question in earlier sections of the test, it was fairer to weight the difficulty value of a subsequent, related performance. An important finding from their work was that familiarity remained stable in the items, but not in the children. For that reason, they recommended applying the familiarity weighting to the anchored value of the difficulty of the item before scoring. To calculate the value of that familiarity weighting, the IRT partial credit score model on the first five attempts allowed a facet for 'attempt' to be introduced. This produced a numerical estimation of the drop in difficulty each time the child repeated the game.

All of the research methods above produced statistical estimates of the degree of randomness or error in each model, and so the fourth research question aimed to bring together this output and reach some conclusions. There are certain reliability statistics that are required of some standardised tests to be accredited by assessment governing bodies (AERA, 1999). Producing these statistics does not complete the validation process (Messick, 1987), but it is a starting point. This final research aimed to look at those figures to see if one model performed well statistically, but would not reach any other conclusions.

8.4 Results and recommendations from the research questions

Proficiency analysis of the four main variables, a child's id, the game they played, a score, and a time and date stamp, was not easy because of interfering factors.

The findings from research question 1 show that an assessor who based difficulty calibration on the whole data set would find some data that does not evidence ability included in the data set. After defining browsing behaviour as any attempt that ends before play starts, browsing accounted for approximately 14.30% of all of recorded play. Removing all performances that ended in under 10 seconds affected the mean response time to a statistically significant level ($p \leq .001$) in 6 of the games, but only in the 0 scoring band, or the fail band. That means that the null hypothesis was not rejected in any of the difficulty estimates in the scoring bands, when looking at the difficulty of the task. Some individual children browsed a lot, though, and the overall mean ability response time for specific children was negatively affected to a significant degree by keeping this data.

Wandering off had a strong effect on the shape of the data. For example, the shoot 'em up game had a mean of 82.3 seconds, and a standard deviation of 12,698 seconds, or over three and a half hours. After capping very long times, in general, only the zero scoring band was significantly affected ($p \leq .001$) in 6 out of the 21 games, suggesting that long times did not appear to evidence extended effort or success. However, one game where players manipulate a space ship through rotations and reflections, there was an exception and the null hypothesis also had to be rejected in scoring bands one and two. It suggests that with some game mechanics, persistence may be important, or that the task is complex enough to go from being a recall task, to something more like problem solving, which demands time. The recommendations for assessors are therefore that:

- They should investigate reasonable response times;
- they should explore long times before setting rules about their removal.

Research question 2 revealed much wider problems with the structure of gameplay data. None of the children had attempted all of the games, and the reason why data is missing is complex. The level of choice means that missing paths appear to be 'not presented', in other words, the missing data are essentially forced by the narrative structure of the games, and not related to the child's ability. The reason data is missing matters because it affects how we treat it. As in the pilot, the full analysis showed 2 out of the 5 lite games as more difficulty than the regular games, using the full data set. One explanation might be that the lite tasks were only completed by the weaker students, who found them at their level, and therefore quite demanding. Perhaps if the higher level students had played the 'lite' version, they would have found it easier. This is discussed in more detail in 6.3.2. This possibility was explored.

The methods used to explore this unexpected pattern did not work because the division of the children into high and low ability groups did not work. On inspection, the 'high ability' children were specialised, and not necessarily of high ability. Overall, both high and low ability groups had played more or less the same number of times, but the 'high scoring' group only produced data for 10 out of the 21 games. The children who appeared in the 'low scoring' group produced data for 19 out of the 21 games. It was not possible, therefore, to say that one group was, in fact, more able. Obtaining more complete data was not possible under this game narrative design. This, however, could be controlled in the assessment design process through a different narrative design, suggested in section 7.3.

The findings therefore showed that there were more issues with the structure of the data than just the problem of conditional dependency. The contribution of this research shows that consideration needs to be given to:

- the overinclusion of data affecting estimates
- the extent and reasons for data to be missing
- the imbalances in the tasks performed

The next research question tackled the transient environments of gameplay.

For the first part of research question 3, using speed as a proxy for ability proved very challenging. The findings show that a number of features of the technically engineered environment of games are not aligned to the goal of accurate scoring on a scale of speed. That is despite the value placed on speed in the games community. There was no single construct of time in the games. The four models of time found were:

- unbounded response times;
- games with a capped upper time limit;
- games with real time variable human competitor performances and constant robot competitors. Both created a time limit when, and only when, they finished first;
- response time as a constant.

With unbounded response times and a capped upper time limit, neither the selection, nor necessarily the number of tasks presented, were constant, and nor was the response time. The two options to assess on a scale of speed, a speeded test or a two-construct model of speed and difficulty, are challenging if neither task selection nor time is constant at any point in the calibration and scoring process. The games with direct competitors were interesting because they sometimes had a variable upper time limit when the competitor won, but sometimes the child being assessed won, and so they were not restricted by the competitor. The data therefore shows if Child A beat Child B, but any speed estimates

from that kind of game were conditionally dependent on the speed and ability of the direct competitor. In theory, if details of competitor ids were stored, it would be possible to weight a speed performance against a competitor ability. In practice, it did not seem worthwhile as there was little evidence that child ability was stable from one game to another, and not all game performances necessarily represented extended effort.

It seems that future games based assessments that intend to use a scale of speed, rather than accuracy, will need informed decisions to be made in the games design phase about how speed will be conceptualised. The box plots of the response times for each game and band score proved useful in understanding how speed was experienced by players. There seems to be an association between the different game mechanics and patterns of speed, and future work might look at patterns in response time. For example:

- In the platform game, accuracy on task was a means to get to the end of the game more quickly, and so a falling pattern of response time was observed as the band score ranking went higher.

- In the tasks which required the child to manipulate an avatar, and the two shoot ‘em up games, it was easier to get higher scores on an accuracy scale with longer play times, and that may be consistent with Thurstone’s early assumptions that the likelihood of a correct question response increases with more time.

- The games that needed the child to perform maths and rewarded them with an animation generally had the fastest speeds among the most and least able students, in Bands 1 and 3. This is a pattern observed in the field of psychology by Maris (2012), where people working close to their threshold need longer to finish. Maris reasoned that high ability test subjects find answers quickly, and weaker ones guess and move on quickly.

The hardest game to explain is where the Band 2 group were the fastest.

All in all, there was no single concept of speed in evidence. Therefore there was no clear way to estimate ability on a scale of speed. More work is needed to see if the findings here, which suggest games mechanics determine speed patterns, are generalizable or if speed patterns can be anticipated in a design phase.

There was, however, one finding from the study of speed that might facilitate assessment. A robot played when no human was available in real time and it ran in constant time. The robot was a formidable opponent in the platform game, as it always finished in 61 seconds, which was well below the overall scoring mean (97 seconds, 95 seconds and 79 seconds in Bands 1, 2 and 3 respectively). In other words, when the bot was in play, it almost always won and produced a predictable cap on the response time. Some children were in very remote time zones globally from the majority of players and could be disadvantaged by this, because they were less likely to play another human who is closer to average ability. Therefore, those children were much less likely to win, and appear less able. It seems worth investigating the possibility that the use of several bots of different levels of ability might create a known constant ‘seeded’ performance necessary to weight scores for competitor ability. More research is needed to explore this possibility.

In terms of the second part of research question 3, there were mixed findings with familiarity. Some children were very familiar with the games, having returned voluntarily to the same game up to 470 times. Looking at just the first five attempts, the games became easier with more practice, and it was possible to isolate a value for the reduction in difficulty, from an overall difficulty of +1.78 logits at the first attempt to +1.5 logits by the fifth attempt. Interestingly, this was not a dramatic drop, suggesting that the correct solution was not immediately obvious, once failed solutions had been discovered. The infit and outfit statistics stabilised more by the third attempt, which also suggests that there is a

randomness in the first couple of attempts. A logical explanation is that this represents a phase while the child works out the rules of play, and it might be better to discard very early attempts as a trial run.

A value for the drop in difficulty had been identified, and it seemed justified by the data to include that familiarity weighting in the scoring model. Almond's work was the starting point for this, but following his model, a familiarity weighting would be needed on all subsequent attempts (2015), not just the first five attempts. This was not an option, because there were no limits on the number of subsequent attempts. The difficulty weightings were estimated with means-based mathematics, and it was impossible to calculate the familiarity weighting each time just one or two children went over the current maximum number of attempts. The alternative suggested by this data set would be to introduce a novelty weighting instead of a familiarity weighting. In other words, children could be given additional credit for success in early attempts but after a certain point, an unweighted estimate better reflects what we can realistically and reliably measure. This novelty weighting seemed well matched to the construct and easy to implement technically, but in these games the impact of novelty was fairly small. In other words, in these games, the problem space that the children were working in was not particularly delimited each time they played. In other games, this might not be the case, or there may be a more restricted number of additional attempts, akin to a multiple choice task. The novelty weighting was also only calculated on the first 5 attempts in this study, but by the time children achieved their high score the overall difficulty of the games had clearly dropped further. The high score, however, could come in any attempt, from the first to the 470th, and so there was no obvious way to model this.

Another suggestion from the results is that guessing parameters may help to identify a moment when performance seems stabilised, as an alternative to either choosing a specific

performance measure to use, or weighting iterative attempts for familiarity. Children's performance could fluctuate, as shown in Figure 46. The results of the partial credit scoring model on the 'high ability' or specialised group produced an infit value of 0.79 and outfit of 0.73. This result was slightly more stable than the results for the 'lower ability' or less specialised group, at 0.85 and 0.82 respectively. With some types of tasks where problem spaces are not delimited much in additional attempts, it may be more sensible to identify a moment when learning and experimenting appear to be replaced by informed intentions, rather than weight difficulty values for familiarity. This was also consistent with the findings from the partial credit scoring model on the first five attempts, which showed more stable results with later attempts. This was not explored any further in this study.

Addressing research question 4, it is still early stages in the field of Games Based Assessment for proficiency testing. The choice of when to measure was crucial in determining the difficulty of the task and the measure of the child, and an optimal moment could not be identified from this research design. The results of the division of the data set into 'high' and 'low' scoring groups also suggested that the measurement tool used in this study was sample dependent, which is problematic if it is necessary to compare performance in one paper to performance across different test delivery formats, and that is usually a requirement of tools used in high stakes proficiency testing.

8.5 Other recommendations and limitations

This section will look at recommendations for policy, practitioners, and at the limitations of this study that have not yet been discussed in the main methods and results and recommendations for researchers sections.

The first is the issue of the children's ages, and particularly the fact that they were growing

up over the period of data collection. For some, there was a very large gap, up to seven years, in the time between their first play and the date of extraction. There was evidence of growth and change over this period, which an educator would hope for. For example, the collaborative addition game that tested number bonds up to 10 was designed to be an entry game, and it was the children's first introduction to the website. It was experienced as the most difficult game when children played it for the first five times (at a difficulty of +1.65), but with practice, many of the children in the sample achieved the highest band possible, and it became the easiest of all of the games in that data set (-2.58). This may also indicate the development of other skills during game play, such as keyboard skills, or completing an online task collaboratively, both of which were required for success, but may be unfamiliar at the outset. No attempts were made to suggest the nature of the relationship between these different skills. It is encouraging to see an educational environment where children are allowed to pace their learning themselves, and complete tasks when they felt ready. However, stake holders, particularly educators, might want age to be included as an environmental factor as a central issue of fairness. These limitations were discussed in detail in section 7.1.2.

A second concern may be curriculum coverage and the construct assessed. Games go deep into certain, key areas, but perhaps do not offer breadth of curriculum. It would be worth investigating whether the child is best served educationally by playing the same game at the expense of spending time on a broader coverage of the curriculum. There may not be one answer to that question. With foundational skills, such as times tables recall, which features in several games here, it seems justified to invest time in this one skill. Other aspects of a curriculum may not be core or may not be suited to recall, and therefore intensive focus may become problematic.

A third observation from the literature review was that assessment design paradigms were

almost always absent from discussions, and very few studies reported including an assessor on the team. Subject specialists were preferred as a source of guidance about how to assess. In this data set, as in other games, mixed models of assessment paradigms were in play. The challenges of reconciling both cohort referencing and criterion referencing in the same assessment design are discussed in detail in section 4.2.2. A stated assessment paradigm for games matters because it affects the way that the results are validated, and in particular, the compromises educational communities may be prepared to make between things like validity, reliability and the authenticity of the task or its relation to real world scenarios. If existing paradigms are unsuited to games, and from this study that may well be the case, a set of principles over what to prioritise seems necessary.

Several limitations of this study have already been discussed. One that perhaps is worth pointing out is that the gameplay data in this study was clearly quantitative survey data. Some authors in the Game Based assessment field felt they had collected observation data, and so findings here are not transferable to those games. Nor are the methods used here recommended to score performance. They were chosen to explore the data structure and content only. It is also worth bearing in mind that the MangaHigh games were not intended to be used as a proficiency test. They serve the purposes they were designed for: to encourage learning, and produce progress data on targeted constructs for teachers.

There are other more general concerns that were not discussed. For machine learning researchers, who are most likely to take up the challenge of scoring models, issues around the transparency of algorithms used in assessment are not a trivial matter. There is also a broader discussion that needs to take place in the educational community on the reproductive nature of machine learning algorithms. Ultimately machine learning algorithms learn what is normal from historic data, and that approach may not be appropriate for a future facing assessment system. Some algorithms can be re-trained as

new data sets become available, but others carry with them a large amount of history. In short, scoring from games with an algorithm raises many questions, and is one reason why they are avoided in this study.

For educators, there is work needed on validating the constructs being measured. There were at least three constructs involved in these games, ability in maths, keyboard skills and sometimes another construct, such as collaboration, working under pressure and management skills, depending on the game. It was unclear whether they had a conjunctive relation. Factor analysis or a clustering algorithm might uncover some patterns in a more targeted assessment design. Methods from the field of human computer interaction might also shed light on whether all three skills were necessary to complete the task, or if ability in one compensated for weaknesses in the other.

8.6 Conclusions

At the moment, it is difficult to look at game data without seeing the potential for new types of test. It seems that machine learning algorithms are most likely to be the solution to modelling the scoring process. I hope this research has shown that the data collection process, and in fact, the way that the problem is initially framed, are equally important phases in constructing a solution to scoring on a calibrated scale from games.

Returning to the constructs that draw many people into the field of Games Based Assessment, complex modern work skills, there has been some development in this field. Work has begun on the wide spread promotion of complex skills in schools in Wales. All children in full time education in Wales aged 14-18 are now assessed on a range of complex work skills, such as planning and organising, project management and collaboration. The Welsh Skills Challenge Certificate is a recognised university entrance

qualification in the UK. I was invited to consult on the revision of the assessment criteria and exam structure as a result of my work in this research. Fragmented and imbalanced data sets, created by a large amount of choice, are not unique to gameplay environments. There may be possibilities to learn from the experience in Wales where these skills are already being assessed in high stakes environments.

Games seem to work for progress test assessment of simple constructs. The next steps in turning them into proficiency tests of complex skills may be a compromise that works with the strengths of both fields. For example, games designers might build stronger elements of assessment design into their games. Existing proficiency tests might introduce some of the features of complex skills that games encourage, but in a controlled way. There is currently no obvious middle ground where the concerns of both groups can be wholly satisfied. In addition, many people need to understand and participate in debates around changes to educational assessment. Decisions about what to test are made politically. The output of tests is used by school teachers, employers, parents and a range of other non-specialists. The most central person in assessment is the test taker, often a minor, who will find their life opportunities changed by test results. The technical challenges of fair scoring are just the beginning of a long process.

Appendix A : Literature review inclusion criteria

A.1 Terms searched:

Key word	Synonyms
Games-based	Serious Games, SEG, Game
Assessment	Testing

A.2 Abbreviations of data bases sourced

WoS	Web of Science
ERIC	Educational Resources Information Center
GooG	Grey search through Google Scholar
OGr	Grey search through Open Grey (opengrey.eu)
OpenDoar	Directory of Open Access Repositories

A.3 Coding key for reasons for rejecting the text

Med	Text refers to assessment used in medicine or psychological assessment
Tech	Text refers to technical assessment
Non-ass	Text offered little about assessment in the game
Non-game	Text was not about game-delivered assessment
Non-source	Text could not be sourced

As well as the sources mentioned above and below, a few people working in the field were contacted directly for findings that were not published, through email and a follow up email two weeks later. Di Cerbo was willing to share some unpublished findings at a conference.

Appendix B : Literature review texts

B.1 Texts Rejected at the Abstract Stage

Authors	Title	Source and date searched	Reason
Abad, Enrique, 2015	Participatory learning and knowledge assessment with a game based method relying on student generated questions	OpenDoar 3/3/2017	Non ass The assessment was external to the game
Alexandrowicz, Rainer W, 2011	Statistical and practical significance of the likelihood ratio test of the linear logistic test model versus the Rasch model	WoS 22/2/2017 Educational research and evaluation	Non game
All, Anissa, 2015	Validating a standardized procedure for effectiveness assessment: learning English vocabulary through gameplay	OpenDoar 3/3/2017 65th ICA annual conference	Non game
Angel del, Blanco, 2012	A framework for simplifying educator tasks related to the integration of games in the learning flow	ERIC 03/3/2017 Journal of educational technology and society	Non-Ass
Annoni, Paola, 2013	Measuring the impact of the web: Rasch modelling for survey evaluation	WoS 22/2/2017 Journal of applied statistics	Non-game
Augustin, T, 2011	Individualized skill assessment in digital learning games: Basic definitions and mathematical formalism	ERIC 03/3/2017 UEEE Transactions on learning technologies	Non-Ass
Baumert, Anna, 2014	Economic Games A Performance-Based Assessment of Fairness and Altruism	WoS 22/2/2017 European Journal for Psychological Assessment	Med Psychological assessment
Boeker, Martin, 2013	Game-Based e-learning is more effective than a conventional instructional method: a randomized controlled trial with third year medical students	OpenDoar 3/3/2017 PloS one	Non-Ass
Bolivar Baron, Holman, 2014	Graph Isomorphism in Fuzzy Cognitive Maps for Monitoring of Game Based Learning	WoS 22/2/2017 Proceedings of the 2014 9th Iberian Conference on information systems and technologies	Tech Text refers to the use of crowd sourced tagging

Bressler, M D	A mixed methods assessment of student flow experiences during a mobile augmented reality science game	ERIC 03/3/2017 Journal of computer assisted learning	Med Psychological assessment
Capuano, Nicola. 2015	Adaptive Serious Games for Emergency Evacuation Training	WoS 22/2/2017 2015 International conference on intelligent networking and collaborative systems IEEEE	Non Source
Chang, Yizhe, 2012	Overcoming the limitations of current online laboratory systems using game-based virtual learning environments	WoS 22/2/2017 Proceedings of the Asme International Mechanical Engineering Congress and Exposition	Non Ass
Charlier, Nathalie, 2011	Game-based assessment of first aid and resuscitation skills	WoS 22/2/2017 Proceedings of the 3rd European Conference on Games Based Learning	Med
Charlier, Nathalie, 2013	Game-Based Learning in Health Sciences, 2013	WoS 22/2/2017 Resuscitation	Med
Chaudy, Yaelle, 2013	Specifications and design of a generalized assessment engine for GBL Applications	ERIC 03/3/2017 No source specified	Non source
Chaudy, Yaelle, 2014	An assessment engine: educators as editors of their serious games assessment	ERIC 03/3/2017 No source specified	Non source
Clegg, Benjamin A, 2015	Effective mitigation of anchoring bias, projection bias and representativeness bias from serious game-based training	WoS 22/2/2917 6th international conference on applied human factors and ergonomics	Non-game Although it was about game based training, it was not in an online context
Cowley, Ben, 2014	Learning When Serious: Psychophysiological Evaluation of a Technology-Enhanced Learning Game	WoS 22/2/2017 Educational Technology and Society	Med Psychological assessment
Craig, Ashley B, 2015	Differences Between Japanese and US Children's Performance on 'Zoo U': A Game-Based Social SKills Assessment	WoS 22/2/2017 Games for Health Journal	Med Social skills were assessed
Crawford, B C, 1995	Assessing student learning outcomes in teaching organizational communication	ERIC 03/3/2017 No source	Non-source

Cutumisu, M, 2014	A game-based assessment of students' choices to seek feedback and to revise	WoS 22/2/2017 Proceedings of the 11th international conference on cognition and exploratory learning in the digital age	Med Study skills were assessed, not cognitive performance
Cutumisu, M, 2016	The effect of choosing versus receiving feedback on college students' performance	WoS 22/2/2017 13th international conference on cognition and exploratory learning in the digital age	Med Study skills were assessed, not cognitive performance
Denham, Andre R, 2016	Integrating game-based learning initiative: increasing the usage of game based learning within K-12 classrooms through professional learning groups	ERIC 3/3/2017 Techtrends	Non-ass
Dev, Pavarti, 2011	Training Tools: Game-Based Assessment/Quiz Template	OpenDoar Not given	Tech
DiCerbo, Kristen E, 2014	Game-Based Assessment of Persistence	WoS 22/2/2014 Educational Technology & Society	Med Psychological assessment, not cognitive performance
Donghwa, Jeong, 2010	TaG-games: Tangible Geometric Games for Assessing Cognitive Problem-Solving Skills and Fine Motor Proficiency	WoS 22/2/2017 2010 IEEE International conference notes	Med and Non-Ass Physical motor skills were assessed and the skills were tested externally to the game
Enfield, Jacob, 2012	Innovation diffusion: Assessment of strategies within the diffusion simulation	ERIC 3/3/2017 Simulation and gaming	Non-ass The assessment was of game-playing strategies
Eseryel, Deniz, 2014	An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning	WoS 22/2/2017 Educational technology and society	Med Psychological assessment
Fotouhi-Ghazvini, F, 2011	Using a conversational framework in mobile game based learning – assessment and evaluation	ERIC Enhancing learning through technology: Education Unplugged: Mobile technologies and the web 2.0	Non Ass The assessment was external to the game
Forsyth, C, 2013	Didactic Galactic: Types of Knowledge Learned in a Serious Game	WoS 22/2/2017 Artificial intelligence in education conference notes	Non-Ass The text contained a theoretical justification of using games to develop skills.
Gamito, Pedro, 2015	Assessing cognitive functions with VR-Based Serious Games that reproduce daily	WoS 22/2/2017	Med The study look at the psychological impact of

	life: Pilot Testing for normative values	Tests for improving patients rehabilitation research techniques	using VR with flight simulation games
Grace, Lindsay, 2015	Designing Microgames for Assessment: A case study in rapid prototype iteration	WoS 22/2/2017 12th Advances in Computer entertainment technology conference	Non-ass Study on the design of games, not scoring
Gutierrez, David, 2014	Assessment of Secondary school students' game performance related to tactical contexts	WoS 22/2/2017 Journal of human kinetics	Med Study of non-cognitive outcomes
Hildmann, Hanno, 2009	A critical reflection on the potential of mobile device based tools to assist in the professional evaluation and assessment of observable aspects of learning or game playing	WoS 22/2/2017 Proceedings of the 3rd European Conference on Games Based Learning	Non ass The study proposes assessing non-cognitive skills using apps in a highly theoretical way
Hsin-Hung, Yu. 2015	Jingnan campaigncopy – using game-based assessment with the mechanism of strategy games for history teaching: system development and learning evaluation	WoS 22/2/2017 2015 IIAI 4th international congress on advanced applied informatics	Non ass Assessment takes place externally to the game, or with the raw scores from the game used to imply historical learning
Huang, Kuang-Min, 2009	Developing the 3D Adventure Game-Based Assessment System with Wii Remote Interaction	WoS 22/2/2017 Advances in Web Based Learning	Non ass The study looks at the technical process to generate raw scores from a Wii game with a cognitive slant.
Hummel, Hans, 2014	Validation of game scenarios for the assessment of professional competence: Development of a serious game for system managers in training	OpenDoar 3/3/2017 Not given	Non game
Hwang, G J, 2014	Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach	WoS 22/2/2017 Etr&D Educational Technology Research and Development	Non ass The assessment methods proposed are entirely external to the game.
Irblich, D, 2002	Game-based interrogation techniques – Interaction diagnostics for assessment, counselling and research	WoS 22/2/2017 Praxis der Kinderpsychologie und Kinderpsychiatrie	Non-source The abstract implies medical assessment
Jaffal, Yasser, 2015	Employing game analytics techniques in the psychometric measurement of game-based assessments with dynamic content	WoS 22/2/2017 Journal of E-learning and Knowledge Society	Non-source The abstract describes using play metrics and evidence centred design.
Jing-Tao, Tang, 2009	Constructing the 2D adventure game-based assessment	WoS 22/2/2017	Non ass The study looks at the programming and design

		Advances in Web Based Learning – ICWL conference notes	considerations of mapping the skills to syllabi, but not the scoring.
Jogi, J, 2015	On-line puzzle game based assessment and training for ADHD	WoS 22/2/2017 European Psychiatry	Med
Ketamo, Harri, 2014	Replacing PISA with Global Game Based Assessment	WoS 22/2/2017 Proceedings of the 8th European Conference on Games Based Learning	Non ass Theoretical discussion, not a consideration of scoring issues
Ketelhut, Diane Jass, 2014	Design and Gender in Immersive Learning Environments	WoS 22/2/2017 Proceedings of the 8th European Conference on Games Based Learning	Med A consideration of social inequalities in online interactions
Ketelhut, Diane Jass, 2016	Blending Formal and Informal Learning Environments: The Case of SAVE Science	WoS 22/2/2017 Proceedings of the 10th European Conference on Games Based Learning	Non ass Although the game is mentioned to be incorporated into the learning model, the assessment is external
Khalil, Mohammad, 2016	What massive open online courses (MOOC) stakeholders can learn from learning analytics	OpenDoar 3/3/2017	Non game
Kumar, R, 2013	OAMS – a game-based online formative knowledge assessment system using concept map	WoS 22/2/2017 Journal of theoretical and applied information technology	Non ass The assessment was external to the game paradata and user-generated
Lai, Ting-Ling, 2015	A case study of the feedback design in a game-based learning for low achieving students	WoS 22/2/2017 Proceedings of the International conference E-Learning	Non ass The assessment was external to the game paradata and user generated.
Lot, Marhani, 2016	Game based learning as a platform for formative assessment in principles of account	ERIC 3/3/2017 International information institute	Tech
Mahroeian, Hamidreza, 2013	An analysis of web-based formative assessment systems used in e-learning environment	WoS 22/2/2017 2013 IEE 13th International conference on Advanced Learning Technologies	Non ass The assessment was human generated.
Mather, Richard, 2015	Multivariate gradient analysis for evaluating and visualising a learning system platform for computer programming	ERIC 3/3/2017 IAFOR Journal of education	Tech

McAlpine, Mhairi, 2010	Using Game Based Technology in Formal Assessment of Learning	WoS 22/2/2017 Proceedings of the 4th European Conference on Game Based Learning	Tech
Meyer, Bente, 2013	Game-based language learning for pre-school children: a design perspective	ERIC 3/3/2017 Electronic journal of game based learning	Non ass
Miller, W L, 2014	Unifying computer based assessment across conceptual instruction problem solving and digital games	WoS 22/2/2017 Technology, Education and Society	Tech
Moccozet, L, 2013	Gamification-based assessment of group work	2013 International conference on Interactive Collaborative learning	Non ass The assessment was external to the game
Moshfeghi, Y, 2016	A Game-Theory Approach for Effective Crowdsource-Based Relevance Assessment	WoS 22/2/2017 Acm Transactions on Intelligent Systems and Technology	Non-game
Mustika, M, 2015	A mobile phone camera text recognition game as an alternative assessment in vocabulary instruction for learning Indonesian as a foreign language classroom	WoS 22/2/2017 2015 IEEE 15th instruction for learning Indonesian as a foreign language classroom	Tech
Oulhaci, M'hammed Ali	A multi-agent system for learner assessment in serious games: Application to learning processes in Crisis Management	WoS 22/2/2017 2013 Ieee Seventh International Conference on Research Challenges in Information Science	Med This was a technical paper on business tracking behaviour patterns in information-seeking behaviours
Perry, J C	Improving the match between ability and challenge: Toward a framework for automatic level adaptation in game-based assessment and training	WoS 22/2/2017 IEEE International conference on rehabilitation robotics	Med
Pitchford, Nicola J, 2015	Development of early mathematical skills with a tablet intervention: a randomized control trial in Malawi	WoS 22/2/2017 Frontiers in psychology	Non-ass The assessment was external to the game
Simic, G, 2015	Assessment based on serious gaming interactive questions	ERIC 3/3/2017 Journal of computer assisted learning	Non ass The assessment was external to the game
Smits, Jarka, 2011	Game-Based Assessment and the Effect on Test Anxiety: A Case study	WoS 22/2/2017 Proceedings of the 5th European Conference on Games Based Learning	Med The study proposed measuring the psychological impact of the test

Stefan, Antoniu, 2016	Approaching assessment in educational games	ERIC 3/3/2017 No-source given	Non-source
Tobias, Sigmund, 2011	Computer games and instruction	ERIC 3/3/2017	Non ass
Tang, Jing-Yao, 2009	Constructing the 2D Adventure Game-Based Assessment System	WoS 22/2/2017 Advances in Web Based Learning	Tech
Thompson, W, 2016	Improving Game Based Learning Through Formative Assessment and Iterative Development	WoS 22/2/2017 Proceedings of the 10th European Conference on Games based learning	Non-Ass The assessment was not integrated into the game
Tong, Tiffany, 2016	Test-retest reliability of a serious game for delirium screening in the emergency department	WoS 22/2/2017 Frontiers in Aging Neuroscience	Med
Tsai, Fu-Hsing, 2015	The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment	WoS 22/2/2017 Computers and education	Non-Ass The assessment was of the effectiveness of genres of games
Ventura, Matthew, 2013	The relationship between video game use and a performance-based measure of persistence	WoS 22/2/2017 Computers and Education	Med This study was on psychological assessment
Ventura, Matthew, 2013	The Validity of a game-based assessment of persistence	WoS 22/2/2017 Computers in Human Behaviour	Med The study was on psychological assessment and validity was understood as correlation with other external assessments
Weakland, Natalie Lynn, 2013	Implementation of educational games-based instruction for improving sight word recognition	OpenDoar 3/3/2017 Not given	Med
Wiloth, Stefanie, 2016	Validation of a computerized game-based assessment strategy to measure training effects on motor-cognitive functions in people with dementia	WoS 22/2/2017 Jmir Serious Games	Med
Wittman, W, 2004	Complex computer based decision games: Challenge and promise for personnel assessment	WoS 22/2/2017 International Journal of Psychology	Med Psychological assessment
Zapata Rivera, Luis Felipe, 2015	Game based assessment for radio frequency circuits courses in engineering	WoS 22/2/2017 Frontiers in education conference	Tech

Zhang, B Y, 2011	Network Security Situation Assessment based on stochastic game model	Advanced intelligent computing	Non ass

B.1.1 Texts Accepted

Authors	Title	Source and data searched	Summary
Allen, Laz, 2009	The implementation of Team Based Assessment in Serious Games	WoS, 22/2/2017 Proceedings of the Ieee Virtual Worlds for Serious Applications	The text looked at the technical process of scoring more than one person simultaneously.
Almond, R G, 2015	Tips and Tricks for Building Team Based	WoS 22/2/2017 Advanced methodologies for Bayesian Networks, Second international workshop	The study contains advice on implementing Bayes Nets in serious educational games
Bertling, Maria, 2015	Measuring Argumentation Skills with Game-Based Assessment: Evidence for Incremental Validity and Learning	WoS 22/2/2017 Artificial intelligence in education	The text looked at qualitative judgments of argumentation.
Charlier, Nathalie, 2009	Game-Based Assessment, can Games Themselves act as Assessment Mechanisms? A Case Study	WoS 22/2/2017 Proceedings of the 3rd European conference on Games Based Learning	The study looked at whether a critique of the weaknesses in the cognitive underpinnings of the game Spore could be used to test understanding
Chin, Doris B, 2016	Got Game? A Choice-Based Learning Assessment of Data Literacy and Visualization Skills	WoS 22/2/2017 Technology Knowledge and Learning	The study looked at assessing the quality of representations through peer crossmoderation.
de Klerk, Sebastian, 2015	Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example	WoS 22/2/2017 Computers and education	This was a summary of Bayes techniques and concluded that they were best suited to simulation studies.
DiCerbo, 2016	Inference in game-based assessment	WoS Using Games and Simulations for Teaching and Assessment	The study looked at scoring complex games and found challenges with using work products alone because they give limited information.
Conrad, Shawn, 2014	A framework for structuring learning assessment in a massively multiplayer online educational game: Experiment centered design	OpenDoar 3/3/2017	The study looked at using peer cross moderation with a MOOC gamified challenge.

Authors	Title	Source and data searched	Summary
Cutumiso, Maria, 2014	A Game-Based Assessment of Students' choices to seek feedback and to revise	OpenDoar 3/3/2017	Although this assesses non-cognitive skills, it offers some insight into the social use of the game scores.
de Klerk, Sebastiaan, 2014	A blending of computer-based assessment and performance-based assessment Multimedia-Based Performance Assessment (MBPA). The introduction of a new method of assessment in Dutch Vocational Education and Training (VET)	OpenDoar 3/3/2017	A Bayes based approach to assessment incorporating scoring within the game.
de Klerk, Sebastiaan, 2015	Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example	OpenDoar 3/3/2017 Computer and education	A systematic literature review into the use of Bayes that concludes that more work is needed to implement the technology more fully.
Eseryel, Deniz, 2013	Validation study of a method for assessing complex ill-structured problem solving by using causal relationships	ERIC 3/3 2017	A study of complex problem solving looking at the identification of patterns of behaviour
Frezzo, Dennis C, 2009	Psychometric and Evidentiary Approaches to Simulation Assessment in Packet Tracer Software	WoS 22/2/2017 Icna: 2009 Fifth Iternational Conference on Networking and Services	The study laid out the principles of evidence centred design and the use of Bayes nets in a complex procedural problem solving game, where learner patterns were compared for similarities with the behaviour patterns of experts conducting the same task.
Graf, Edith Aurora, 2014	Connecting lines of research on task model variables, automatic item generation and learning progressions in game based assessment	ERIC 3/3/2017	The study lays out the challenges of measuring emergent behaviour and generating items through adaptive testing
Hainey, Thomas, 2012	Assessment integration in games-based learning: a preliminary review of the literature	ERIC 3/3/2017	A literature review of the state of research in 2012, largely concluding that more work was needed
Halverson, R, 2014	Game-based assessment: an integrated model for capturing evidence of learning in play	WoS 22/2/2017 International journal of learning technology	A technical study of how to record paradata during game play.
Hooshyar, D, 2016	A solution-based intelligent system integrated with an online game based formative assessment: development and evaluation	WoS 22/2/2017 Etr&D – Educational Technology Research and Development	A mostly technical account of using in game data to provide feedback to the learner on areas to improve or revisit.

Authors	Title	Source and data searched	Summary
Ifenthaler, D, 2014	Challenges for education in a connected world: Digital Learning, Data Rich Environments, and Computer-Based Assessment – Introduction to the Inaugural Special Issue of Technology Knowledge and Learning	WoS 22/2/2017 Technology, Knowledge and Learning	A review of the literature on a number of challenges in digital learning, including the conclusion that more work is needed to develop games-based assessments.
Kearney, P, 2007	Cognitive assessment of game-based learning	WoS 22/2/2017 British Journal of Educational Technology	A general introduction to the issue of game based assessment
Kim, Yoon Jeon, 2015	The interplay of game elements with psychometric qualities, learning, and enjoyment in games based assessment	WoS 22/2/2017 Computers and education	A justification for the use of games from a motivational perspective, with the results of the game regressed onto external gold standard tests.
Kim, Yoon Jeon, 2016	Applying evidence centred design for the development of game-based assessments in physics playground	WoS 22/2/2017 International Journal of Testing	An account of how the game was designed to evidence aspects of the syllabus in a physics game
Lamb, Richard L, 2014	Cognitive Diagnostic like approaches using neural-network analysis of serious educational video games	WoS 22/2/2017 Computers and Education	The study looked at the theoretical role of cognitive diagnostic assessment, and proposed a neural-network analysis approach to feedback
Leighton, Jacqueline, 2016	First among equals: Hybridization of cognitive diagnostic assessment and evidence centered game design	ERIC 3/3/2017 International journal of testing	The study looked at the theory of evidence centred game design and how it was applied in a game.
Mavridis, Apostolos, 2015	Gamified assessment supported by a dynamic 3D collaborative game	ERIC 3/3/2017 International journal of game based learning	A largely technical description of how latent variables were assessed
McAlpine, M, 2010	Using games-based technology in formal assessment of learning	WoS 22/2/2017 Proceedings of the 4th European Conference on Games Based Learning	This was a theoretical summary and reached few clear conclusions.
Mitgutsch, Konstantin, 2012	Purposeful by design? A serious game design assessment framework	OpenDoar 3/3/2017	A summary of how syllabi can be mapped to games design tasks
Nelson, B C, 2010	Exploring cognitive load in immersive educational games: The SAVE Science project	WoS 22/2/2017 International journal of gaming and computer	This reported on a identifying conditions for success to help students manage cognitive overload in virtual environments.

Authors	Title	Source and data searched	Summary
		mediated simulations	
Nelson, Brian C, 2014	Visual signalling in virtual world-based assessments: The SAVE Science project	WoS 22/2/2017 Information sciences	This looked at ways of using visual signalling to reduce cognitive load.
Shute, Valerie J, 2008	Monitoring and fostering learning through games and embedded assessments research report, ETS	ERIC 3/3/2017 ETS Research Report	A summary of the challenges to introducing games based learning gathered from trialling with the US testing body, ETS.
Shute, V J, 2011	Stealth Assessment in Computer Based Games to Support Learning	WoS 22/2/2017 Proceedings of the 5th European Conference on Games Based Learning	A practical proposal to harvest data without learning awareness, and a theoretical justification of this approach
Shute, Valerie J, 2016	Measuring problem-solving skills via stealth assessment in an engaging game	WoS 22/2/2017 Computers in human behaviour	This study proposed the harvesting of data on game play as an unobtrusive means of gathering formative assessment on progress
Shute and Ke, 2012	Games, learning and Assessment	WoS 22/2/2017	A study which used user choices to assign a creativity value to choices based on the frequency of use by game players.
Snow, E L, 2016	Taking control: Stealth Assessment of Deterministic Behaviors Within a Game-Based System	WoS 22/2/2017 International journal of artificial intelligence in education	This sought to use log data to classify students in a game into either a group that showed signs of deterministic or one that showed signs of random behaviour, using self-explanations and random walk analysis.
Steiner, Christina, 2015	Making sense of Game-Based User Data: Learning Analytics in Applied Games	OpenDoar 3/3/2017 International Association for Development of the Information Society	Reported on a software to score skills in games developed from learner analytics.
Tong, Tiffany, 2014	Designing a Game-Based Cognitive Assessment for a Tablet	OpenDoar 3/3/2017 Not given	A suggestion for mapping syllabi content to assessment design
Usart, M, 2014	Entrepreneurship competence assessment through a game based learning MOOC	WoS 22/2/2017 Games and learning alliance	Used a final questionnaire to assess performance in a game based assessment of entrepreneur skills. The assessment was of the course quality, not the students.
Vendlinksi, Terry P, 2010	Developing High Quality Assessments that align with instructional video games	ERIC 3/3/2017 CRESST Report	The CREST report was an early document that summarised some of the problems with assessing rational numbers use in a games

Authors	Title	Source and data searched	Summary
			platform, Puppetman. The assessment was largely to determine next steps, and used pre-test post test evaluation of the children.
Wood, Lincoln, 2013	The role of gamification and game-based learning in authentic assessment within virtual environments	OpenDoar 3/3/2017 Not given	Used game based mechanisms to improve assessment, particularly in terms of authenticity. Assessment was largely understood as feedback on performance.
Zapata Rivera, D, 2009	Combining learning and assessment in assessment based gaming environments	ERIC 3/3/2017 Interactive Technology and Smart education	This was a technical paper on creating adaptive learning environments.
Zapata Rivera, D, 2010	Adaptive, Assessment Based Educational games	WoS 22/2/2017 Intelligent Tutoring systems	This was a technical paper on creating adaptive learning environments.
Zourou, K, 2014	Assessment in Game-Based Learning: Foundations, innovations and perspectives	WoS 22/2/2017 Language learning and technology	This was a review of the book.

Appendix C Game name, mathematics and mechanics

Game name	Mathematics	Game details*
Genre: Guide the avatar to the correct answer		
Algebra Meltdown	Algebra	Feed the nuclear generator the correct atoms
Deepest Ocean	Inequality signs	Guide the submarine to catch fish of a specified size
Flower Power and Flower Power Lite	Ordering Decimals, Fractions and Percentages	Multiplayer competitive – Move the buds up and down the stem in size order
Ice Ice Maybe	Fast estimation with basic number calculations	Position floating icebergs to save the penguins from the hungry killer whales - discontinued
Save Our Dumb Planet	Algebra	Destroy meteors by calculating trajectories
Piñata Fever	Add and subtract with negative numbers	Move the avatar to the correct position to jump and hit the piñata
Pyramid Panic and Pyramid Panic Lite	Patterns, Algebra, Straight Line Graphs	Calculate rope lengths to move mummies through a tomb
Transtar	Reflections, rotations, enlargements and translations	Guide Transtar through gates by clicking on lines of reflection or orbs
Genre: Maths with animation		
A Tangled Web	Angles	Guide the spider up a clock to rescue his family by calculating angles on webs
Beavers Build It	Addition	Multiplayer collaborative – Build the highest wall possible sharing information about each player's bricks
Jabara	Algebraic simplification	Isolate the variable in the minimum number of moves
Sundae Times and Sundae Times Lite	Times tables from x2 to x 15 and basic mental maths	Multiplayer competitive - build the tallest ice cream by answering mental maths questions
Wrecks Factor	Quadratics	Factor quadratic equations to stop ships falling into the Bermuda Rectangle

Genre: Platform game		
Jet Stream Riders	Times tables and numeracy	Multiplayer competitive - Answer the sums to make the balloon rise over obstacles or find the jet stream
Genre: Shoot 'em up		
Bidmas Blaster	Order of operations (brackets, indices, division, multiplication, addition, subtraction)	Each advancing robot has a formula on its head and typing the answer disables the robot's shields so it can be blasted.
Sigma Prime	Prime factorisation with multiplication and division	Alien space ships appear with a number value that needs to be decomposed to defeat it by selecting the right prime factor.

Appendix D Ethics

D.1 Ethics Application for Secondary Data set MangaHigh

Version 2 November 27th 2017

Ethics Application Form for SECONDARY DATA ANALYSIS

Please consult the guidance at the end of this form before completing and submitting your application.

1. Name(s): Clare Walsh

2. Current Position: PhD Researcher

3. Contact Details:

Division:

Email: cew2g15@soton.ac.uk

Phone: 07922180918

4. Is your research being conducted as part of an education qualification?

   Yes ☒ No ☐

5. If Yes, please give the name of your supervisor:

   Dr Christian Bokhove and Dr Su White

6. Title of your research project / study:

   PhD Web Science - Can educational games be scored fairly?

7. Briefly describe the rationale, aims, design and research questions of your research

Please indicate clearly whether you are applying for ethics approval for a specific piece of research, or for overarching ethics approval to use certain datasets for a range of research activities. Approval for the latter will only cover the datasets specified here, for a maximum of 3 years and then subject to renewal.

The study is a secondary data analysis of logfile data from the database of MangaHigh, an online Maths game site, which is owned by Blue Duck Education. MangaHigh have granted permission to access their data. Major test providers are interested in using gaming data to inform formative assessment decisions, and potentially higher stakes tests, but as yet, no-one has been able to produce a validation argument around gaming data. There is very little written from an assessment perspective. A validation argument informs judgments on the value of assessment scores by producing reliability estimates and other analysis of the data. In this study I intend to use the data to answer the following research questions:

1) What can be escalated to the final score (key strokes, time stamps, number of attempts etc.)?

2) How can we escalate that (as dichotomous pass/fail judgments, partial 4/5 scores, categorised, delimited problem spaces, or as a separate parameter)?

3) How do we adapt current models of analysis of the internal consistency of score data to a hypertext environment of conditionally dependent, varying states?

4) How do we control for interfering variables (such as interface, keystroke controls, poor design) in the test construction model?

The secondary data set will use historical data on user patterns collected by MangaHigh.

8. Describe the data you wish to analyse

Please give details of the title of the dataset, nature of data subjects (e.g. individuals or organisations), thematic focus and country/countries covered. Indicate whether the data are qualitative or quantitative, survey data, administrative data or other types of data. Identify the source from where you will be obtaining the data (including a web address where appropriate).

The data base contains over a billion data points, and over a million users are currently registered for the game. The games are aimed at children of primary and secondary school age, and cover the Key Stages 1 and 2 Maths syllabus. The majority of users come to the game through school subscription, and so it is possible I may have access to the names of individuals, teachers and educational institutions. The tables in the data set need to be specifically queried by SQL requests, and no personally identifying information, such as user name or the name of the school or class will be requested. Users and schools are assigned a unique number and these will be used instead. Users self-report their age, and this will be accessed, but date of birth is not collected by the site. No attempt will be made to access data stored in these three data tables, or to track down the identity of the person(s) registered. This information is irrelevant to the current study and poses obvious data protection risks. The rest of the data, which I intend to use, contains a user id, age in years, information on the learner activities during the game, their status, points total, medal total, responses, and times stamps, as well as information on the games they have played.

9. What are the terms and conditions around the use of the data? Did data subjects give consent for their data to be re-used? If not, on what basis is re-use of the data justified?

Please state what (if any) conditions the data archive imposes (e.g. registration, signing of confidentiality agreement, specific training etc.). In many cases the data controller will have given explicit permission for data re-use. Please explain how you justify the use of data if approval and consents for the original data collection and re-use are not in place. This may be the case where, for example, the original data collection predated requirements for ethics review or occurred in a jurisdiction where explicit consent and approval are not required.

MangaHigh gives permission to access the database through their VPN and password login. The VPN will be installed on my laptop at their site in London. Users have agreed that data could be used for research purposes when signing up.

10. Do you intend to use personal data (https://ico.org.uk/media/1549/determining_what_is_personal_data_quick_reference_guide.pdf) or sensitive personal data (http://www.legislation.gov.uk/ukpga/1998/29/section/2) as defined by the Data Protection Act (even if the data are publicly available)?

Yes ☐ No ☒

If YES, please specify what personal data will be included and why.

11. Do you intend to link two or more datasets?

Data linkage refers to merging of information from two or more sources of data to consolidate facts concerning an individual or an event that are not available in any separate record. Please note that for the purposes of research ethics we are not interested in the merging of different waves of a particular survey, or the merging of data from different countries for the same survey.

Yes ☒ No ☐

If YES, please give details of which datasets will be linked and for what purposes.

The data described above is stored in different tables, and these will be brought together. Personal or identifying data will not be accessed.

12. How will you store and manage the data before and during the analysis? What will happen with the data at the end of the project?

Please consult the University of Southampton's Research Data Management Policy (*http://library.soton.ac.uk/researchdata/storage* and *http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html*), and indicate how you will abide by it.

Any data downloaded will be stored on a password protected computer. The anonymised datasets will be kept in the researchers' archives. Hard copies of any of the data (in paper or digital will be kept in a locked drawer. Access to the database will be password protected and I will stay logged out and not save the password on my laptop when not in use. I will also inform MangaHigh immediately if I think that the security has been breached in any way.

13. How will you minimise the risk that data subjects (individuals or organisations) could be identified in your presentation of results?

Please consider whether disclosive ID codes have been used (e.g. date of birth) and whether it is theoretically possible to identify individuals by combining characteristics (e.g. widow in Hampshire with 14 children) or by combining datasets. How will you protect individuals' anonymity in your analysis and dissemination?

No identifying data will be downloaded from the data set and no attempts will be made to identify individuals.

14. What other ethical risks are raised by your research, and how do you intend to manage these?

Issues may arise due to the nature of the research you intend to undertake and/or the subject matter of the data. Examples include: data or analysis that are culturally or socially sensitive; data relating to criminal activity, including terrorism, and security sensitive issues.

Not applicable

15. Please outline any other information that you feel may be relevant to this submission.

For example, will you be using the services or facilities of ONS, ADRN, or HSCIC and/or are you obtaining ethical review from NRES (through IRAS) or other? Please confirm whether the data being used are already in the public domain.

Not applicable

16. Please indicate if you, your supervisor or a member of the study team/research group are a data controller and/or data processor in relation to the data you intend to use as defined by the Data Protection Act, and confirm that you/they understand your/their respective responsibilities **https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/**).

I believe my supervisor understands my responsibilities and I will keep him updated on my actions.

D.2 Risk Assessment Form for Assessing Ethical and Research Risks

- Please see Guidance Notes at the end of this document.
- *Students*: Please make sure you have discussed this form with your supervisor!

Researcher's name:	Clare Walsh
In case of students:	
Supervisor's name:	Dr Christian Bokhove and Dr Su White
Degree course:	PhD Web Science

Part 1 – Research activities
What do you intend to do? (*Please provide a brief description of your study and details of your proposed methods*.) Access a database of log file data from an educational mathematics game, and run analysis over the scoring of game play.
Will your research involve collection of information from other people? (*If yes, please provide a description of your proposed sample*.) MangaHigh, a company owned by Blue Duck Education, have given permission to use the data in their records for this purpose. In order to access the data set, I will need to travel to their office in central London to have their VPN installed on my computer.
If relevant, what locations are involved? (*Please specify which country/region/place you will be working in, and details of where data collection activities will take place (e.g. public or private space*).)

I will be working at my office in the education department, or in my house. I will need to travel to central London. The MangaHigh Offices are in a communal block with several start-up companies, and very public and open. The research project requires a one-off visit to have software installed on my computer.
Will you be working alone or with others in the data collection process? **Alone**
Part 2 – Potential risks to YOU as the researcher
Please specify potential <u>safety issues</u> arising from your proposed research activity. (*Give consideration to aspects such as lone working, risky locations, risks associated with travel; please assess the likelihood and severity of risks.) If you have already completed a departmental H&S risk assessment, this may be attached to cover these aspects.* I will mostly be working on data analysis from my desk at home, or in the education department. I will need to make one trip to the MangaHigh offices in London.
What precautions will you take to minimise these risks?
Please specify potential <u>distress or harm to YOU</u> arising from your proposed research activity. (*Give consideration to the possibility that you may be adversely affected by something your participants share with you. This may include information of a distressing, sensitive or illegal nature*.) I will need to ensure that I take sufficient breaks from screen work and from sitting down while I am analysing the data. I will also need to take sensible precautions when travelling to the office.
What precautions will you take to minimise these risks?

When I go to the MangaHigh offices, I will be travelling during normal office hours, and will inform my supervisor of my proposed timetable and where I will be. I will be visiting a shared office space in a publically open building, with both MangaHigh staff and other start-ups.

Otherwise, I will be working at my desk. I have attended the Health and Safety training that the university has provided and will be taking suitable breaks from screen work, and using a chair, desk and screen at the appropriate height.

Part 3 – Potential risks to YOUR RESEARCH PARTICIPANTS

Please consider potential safety risks to participants from taking part in your proposed research activity? (*Give consideration to aspects such as location of the research, risks associated with travel, strain from participation, and assess the likelihood and severity of risks.) If you have already completed a departmental H&S risk assessment, this may be attached to cover these aspects.*

As stated above, travelling may involve some risk.

What precautions will you take and/or suggest to your participants to minimise these risks?

As stated above, I will make sure I travel in day time hours, inform my supervisor and someone at home of my whereabouts, and will be in public spaces at all times.

Please specify potential harm or distress that might affect your participants as a result of taking part in your research. (*Give consideration to aspects such as emotional distress, anxiety, unmet expectations, unintentional disclosure of participants' identity, and assess the likelihood and severity of risks*.)

The data set is collected from minors. No personally identifying information will be collected, and no attempt will be made to identify any participant. No personally sensitive information will be involved. The analysis will be on the patterns of key strokes, time stamps, and scoring of maths games.
What precautions will you take and/or suggest to your participants to minimise these risks? No personally identifying data will be collected. The laptop that will have the VPN with access to the data set will be pass word protected, as well as an additional pass word needed to access the dataset. This will be kept in my home or in a locked bag.
Part 4 – Potential wider risks
Does your planned research pose any additional risks as a result of the sensitivity of the research and/or the nature of the population(s) or location(s) being studied? *(Give considerations to aspects such as impact on the reputation of your discipline or institution; impact on relations between researchers and participants, or between population sub-groups; social, religious, ethnic, political or other sensitivities; potential misuse of findings for illegal, discriminatory or harmful purposes; potential harm to the environment; impacts on culture or cultural heritage.)* The standards for respectful digital monitoring will undoubtedly tighten in the next decade as awareness spreads more, and agreement can be reached.
What precautions will you take to minimise these risks? This project will anticipate stronger best practice in the future, and keep the focus on the game and its role in assessment, not on the participants.

CONTINUED BELOW …

Part 5 – International Travel
If your activity involves international travel you must meet the Faculty's requirements for Business Travel which are intended to: 1. Inform managers/supervisors of the travel plans of staff and students and identify whether risk assessment is required. 2. Provide contact information to staff and students whilst travelling (insurance contact details, University contact in case of emergency etc.) Full details are provided in the Faculty H&S Handbook in the **Business Travel** section. Selecting **Business Travel** from the Contents list will take you straight to the relevant section.

Departmental H&S risk assessment attached (for Part 2/3)	**YES / NO**	**(Delete as applicable)**
Business Travel and Risk Filter Form attached (Part 5)	**YES / NO**	**(Delete as applicable)**

Version April 2016

D.3 Ethics Application Form for Secondary Data set QuizYourEnglish (not used in the main analysis)

Please consult the guidance at the end of this form before completing and submitting your application.

1. **Name(s):** Clare Walsh

2. **Current Position:** PhD Researcher

3. **Contact Details:**

   Division:

   Email: cew2g15@soton.ac.uk

   Phone: 07922180918

4. **Is your research being conducted as part of an education qualification?**

   Yes ☐**X** **No** ☐

5. **If Yes, please give the name of your supervisor:**

   Dr Christian Bokhove and Dr Su White

6. **Title of your research project / study:**

   PhD Web Science - Can educational games be scored fairly?

7. **Briefly describe the rationale, aims, design and research questions of your research**

   Please indicate clearly whether you are applying for ethics approval for a specific piece of research, or for overarching ethics approval to use certain datasets for a range of research activities. Approval for the latter will only cover the datasets specified here, for a maximum of 3 years and then subject to renewal.

   The study is a secondary data analysis of logfile data from the online game 'QuizYourEnglish', which is owned by Cambridge University Press. Cambridge University Press have granted permission to use the data under their terms and conditions. Major test providers are interested in using gaming data to inform formative assessment decisions, and potentially higher stakes tests, but as yet, no-one has been able to produce a validation argument around gaming data. There is very little written from an assessment perspective. The idea is to inform judgments on the value of gaming data for formative

assessment purposes by trying to get a reliability estimate on the data. In this study I intend to use the data to answer the following research questions:
1) What can be escalated to the final score (key strokes, time stamps, number of attempts etc.)?
2) How can we escalate that (as dichotomous pass/fail judgments, partial 4/5 scores, categorised, delimited problem spaces, or as a separate parameter)?

3) How do we adapt current models of analysis of the internal consistency of score data to a hypertext environment of conditionally dependent, varying states?

4) How do we control for interfering variables (such as interface, keystroke controls, poor design) in the test construction model?

The data was collected from gameplay over 2017.

8. Describe the data you wish to analyse

Please give details of the title of the dataset, nature of data subjects (e.g. individuals or organisations), thematic focus and country/countries covered. Indicate whether the data are qualitative or quantitative, survey data, administrative data or other types of data. Identify the source from where you will be obtaining the data (including a web address where appropriate).

There are two sets of anonymised data. The first is a summary of 184,218 users' actions in the game. Their nationality is recorded, but no information on their age, but the game is more likely to appeal to older learners (over 16). The ratio of children to adults is estimated on this application. The first set consists of a number of fields on the id number of the user, nationality, topic and question that learners were asked in the quiz, their status, their responses, the id number of their competitors, times stamps, coin counts, and their updated score counts. The second set contains the same information but in a raw data file from telemetry set. The set has already been anonymised by CUP, and the id number cannot be linked to individual people. No attempts will and can be made to track down the identity of the person(s).

9. What are the terms and conditions around the use of the data? Did data subjects give consent for their data to be re-used? If not, on what basis is re-use of the data justified?

Please state what (if any) conditions the data archive imposes (e.g. registration, signing of confidentiality agreement, specific training etc.). In many cases the data controller will have given explicit permission for data re-use. Please explain how you justify the use of data if approval and consents for the original data collection and re-use are not in place. This may be the case where, for example, the original data collection predated requirements for ethics review or occurred in a jurisdiction where explicit consent and approval are not required.

Cambridge University Press gives permission to use the anonymised data set. Users have agreed that data could be used for research purposes when signing up. CUP have additionally requested the right to reply, should any negative findings emerge from the analysis, and the right be informed of any major publications using the data.

10. Do you intend to use personal data (https://ico.org.uk/media/1549/determining_what_is_personal_data_quick_reference_guide.pdf) or sensitive personal data (http://www.legislation.gov.uk/ukpga/1998/29/section/2) as defined by the Data Protection Act (even if the data are publicly available)?

Yes ☐ **No** X☐

If YES, please specify what personal data will be included and why.

11. Do you intend to link two or more datasets?

Data linkage refers to merging of information from two or more sources of data to consolidate facts concerning an individual or an event that are not available in any separate record. Please note that for the purposes of research ethics we are not interested in the merging of different waves of a particular survey, or the merging of data from different countries for the same survey.

Yes X☐ **No** ☐

If YES, please give details of which datasets will be linked and for what purposes.

Possibly. As stated above, the dataset comes in two versions, and I may benefit from using information from each one. Both data sets come with several tables. The first is a summary of the data contained in the second, and no new data external data will be linked. Neither can be linked to actual people or used to identify person(s).

12. How will you store and manage the data before and during the analysis? What will happen with the data at the end of the project?

Please consult the University of Southampton's Research Data Management Policy (http://library.soton.ac.uk/researchdata/storage and http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html), and indicate how you will abide by it.

The data will be stored on a password protected computer. The anonymised datasets will be kept in the researchers' archives. Hard copies of any of the data (in paper or digital will be kept in a locked drawer.

13. How will you minimise the risk that data subjects (individuals or organisations) could be identified in your presentation of results?

Please consider whether disclosive ID codes have been used (e.g. date of birth) and whether it is theoretically possible to identify individuals by combining characteristics (e.g. widow in Hampshire with 14 children) or by combining datasets. How will you protect individuals' anonymity in your analysis and dissemination?

No identifying data is included in the data set and no attempts will be made to identify individuals.

14. What other ethical risks are raised by your research, and how do you intend to manage these?

Issues may arise due to the nature of the research you intend to undertake and/or the subject matter of the data. Examples include: data or analysis that are culturally or socially sensitive; data relating to criminal activity, including terrorism, and security sensitive issues.

Not applicable

15. Please outline any other information that you feel may be relevant to this submission.

For example, will you be using the services or facilities of ONS, ADRN, or HSCIC and/or are you obtaining ethical review from NRES (through IRAS) or other? Please confirm whether the data being used are already in the public domain.

Not applicable

16. **Please indicate if you, your supervisor or a member of the study team/research group are a data controller and/or data processor in relation to the data you intend to use as defined by the Data Protection Act, and confirm that you/they understand your/their respective responsibilities https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/).**

I believe my supervisor understands my responsibilities and I will keep him updated on my actions.

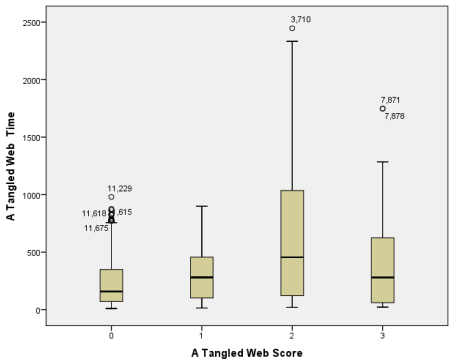
Appendix EEffects of deleting less than 10 seconds

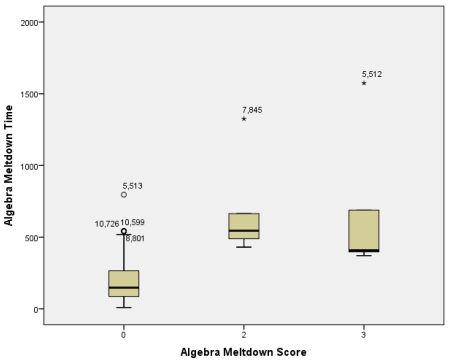
				Without Zeros	Over Ten Seconds Only	
Game	**Number of cases**	**IQR**	**Extreme cap**	**Mean**	**Mean**	**Paired t-test sig.**
ATangledWeb						
Band 0	993	278	774	428	234	-
Band 1	47	379	N/A	324	324	-
Band 2	58	924	2444	668	629	-
Band 3	43	567	1747	431	430	-
AlgebraMeltdown						
Band 0	275	182	542	472	194	.290
Band 1	0					
Band 2	5	164	N/A	690	690	.374
Band 3	5	748	N/A	688	688	.374
BidmasBlaster						
Band 0	444	266	787	437	274	.256
Band 1	32	401	1458	509	509	.325
Band 2	11	474	N/A	654	654	-
Band 3	2		N/A	245	245	-
BidmasBlaster Lite						
Band 0	576	218	666	1140	258	.158
Band 1	164	181	676	354	334	.011
Band 2	96	275	1042	423	385	.052
Band 3	12	406	N/A	432	217	.339
FlowerPower						
Band 0	574	120	487	251	249	.091
Band 1	46	141	N/A	380	380	-
Band 2	20	153	781	452	452	.330
Band 3	37	211	1097	630	626	.317
FlowerPower Lite						
Band 0	830	190	656	299	278	.000
Band 1	57	1321	748	405	395	.322
Band 2	22	179	836	467	463	.329
Band 3	20	443	N/A	570	568	.330
IceIceMaybe						

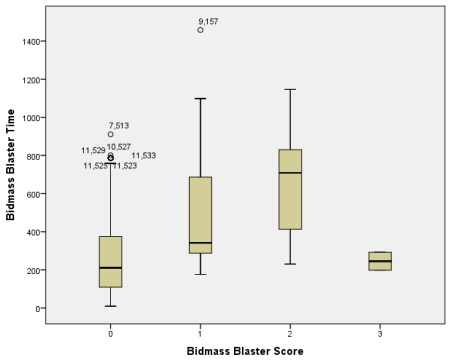
				Without Zeros	Over Ten Seconds Only	
Game	**Number of cases**	**IQR**	**Extreme cap**	**Mean**	**Mean**	**Paired t-test sig.**
Band 0	1530	82	253	343	102	.158
Band 1	441	67	232	123	101	.013
Band 2	265	58	225	443	114	.291
Band 3	81	86	298	141	122	.298
Jabara						
Band 0	326	199	564	1614	176	.118
Band 1	4	716	N/A	451	342	-
Band 2	92	224	713	217	207	.055
Band 3	107	223	613	190	186	.140
JetStreamRiders						
Band 0	6709	94	271	78	78	.000
Band 1	1217	61	N/A	97	97	
Band 2	1330	52	300	95	95	.732
Band 3	1296	49	N/A	79	79	
Pinata Fever						
Band 0	863	113	360	163	146	.000
Band 1	4	44	392	214	190	.012
Band 2	3	389	N/A	364	364	-
Band 3	1		N/A	694	694	
PyramidPanic						
Band 0	461	208	592	283	193	.042
Band 1	2		N/A	323	323	-
Band 2						
Band 3						
PyramidPanic Lite						
Band 0	1225	245	691	611	224	.063
Band 1	141	308	1052	467	457	.078
Band 2	20	456	N/A	527	527	-
Band 3	6	247	1559	360	340	-
SaveOurDumbPlanet						
Band 0	321	118	352	149	127	.098
Band 1	7	368	N/A	186	186	
Band 2	7	542	1013	319	319	.356
Band 3	5	196	461	289	288	.374
SigmaPrime						

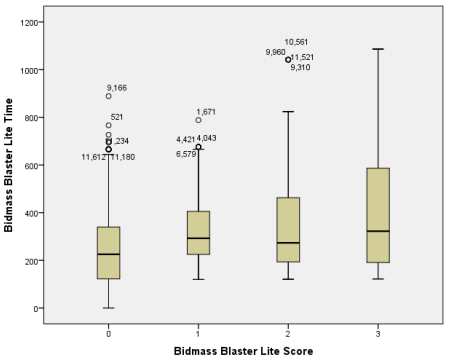
				Without Zeros	Over Ten Seconds Only	
Game	**Number of cases**	**IQR**	**Extreme cap**	**Mean**	**Mean**	**Paired t-test sig.**
Band 0	528	137	402	146	134	.000
Band 1	75	399	N/A	454	453	.321
Band 2	41	414	N/A	714	714	-
Band 3	10	920	N/A	1235	1235	-
SigmaPrime Lite						
Band 0	610	141	413	161	141	.025
Band 1	57	386	4962	542	469	.322
Band 2	22	396	N/A	663	663	-
Band 3	6	605	N/A	979	979	-
SundaeTimes						
Band 0	1049	135	448	184	169	.000
Band 1	27	338	1000	311	304	.321
Band 2	15	271	800	333	304	.316
Band 3	22	278	N/A	289	389	-
SundaeTimes Lite						
Band 0	1056	91	332	173	146	.000
Band 1	27	270	N/A	253	253	-
Band 2	12	240	N/A	227	227	-
Band 3	19	270	N/A	267	265	.148
TranStar						
Band 0	862	88	244	361	83	.057
Band 1	132	89	271	100	81	.003
Band 2	174	100	276	92	86	.000
Band 3	11	460	600	340	245	.041
WrecksFactor						
Band 0	160	125	414	133	90	.311
Band 1	10	2	N/A	231	231	-
Band 2	6	35	372	262	261	.363
Band 3	6	1	238	260	239	.347

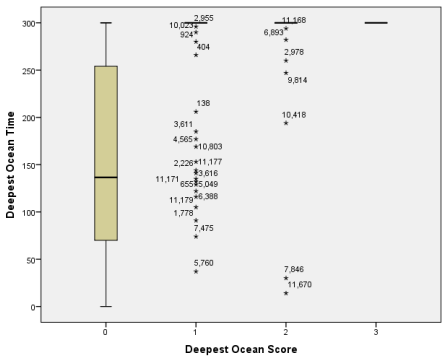
Appendix F Box Plots for response time

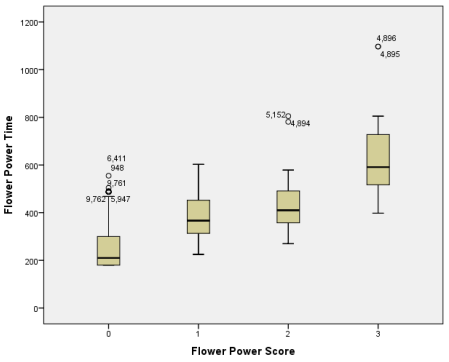


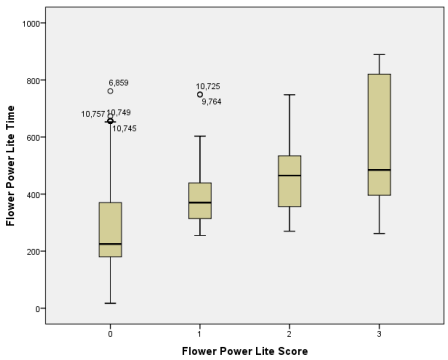


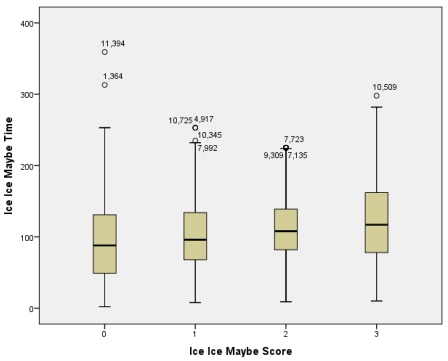


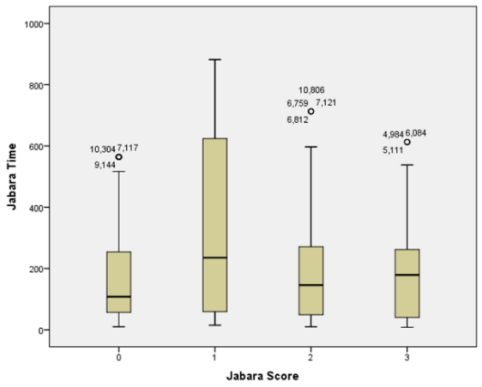


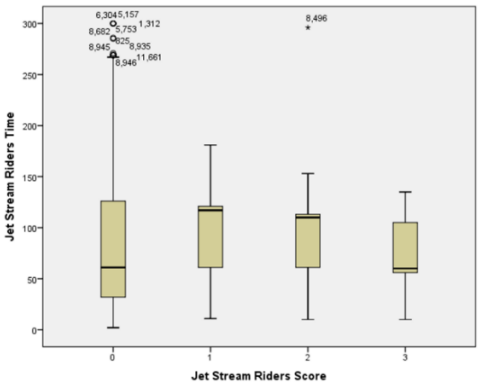


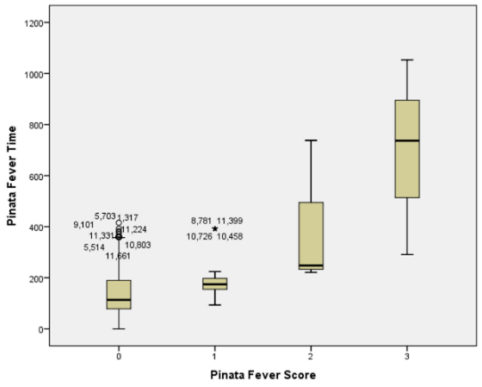


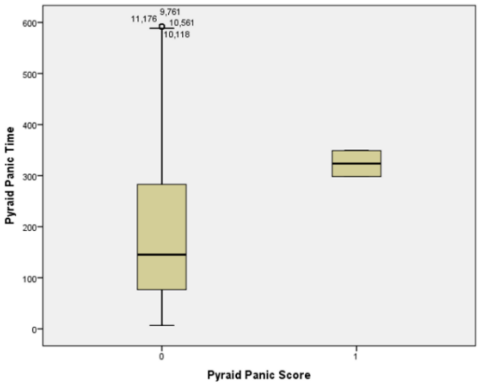


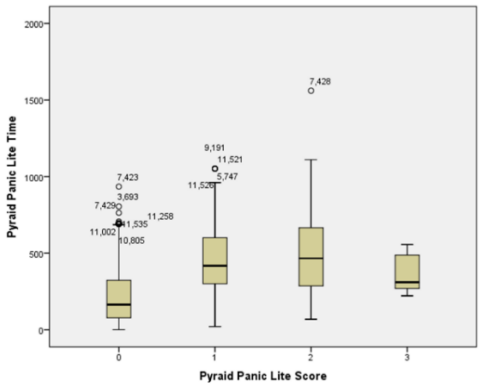


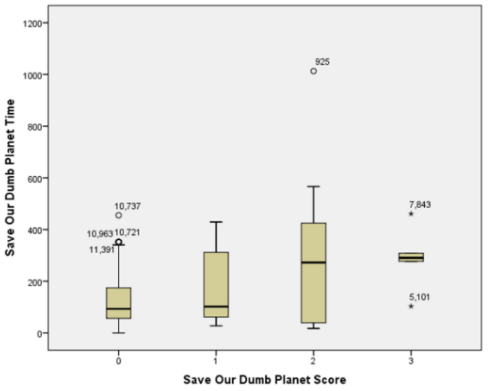

Jabara Time
Jabara Score

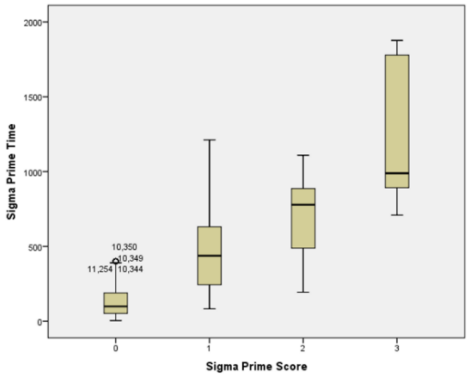

Jet Stream Riders Time
Jet Stream Riders Score

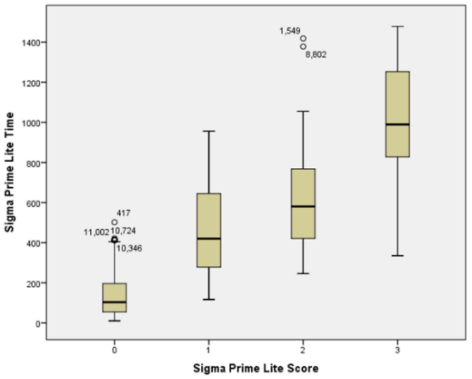

Pinata Fever Time
Pinata Fever Score

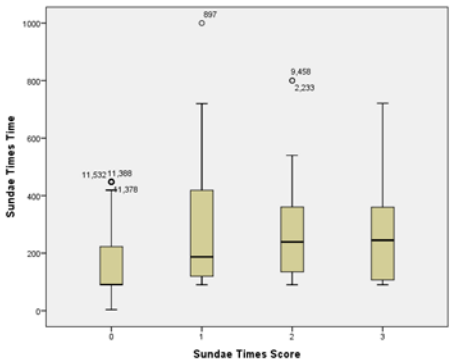

Pyraid Panic Time
Pyraid Panic Score

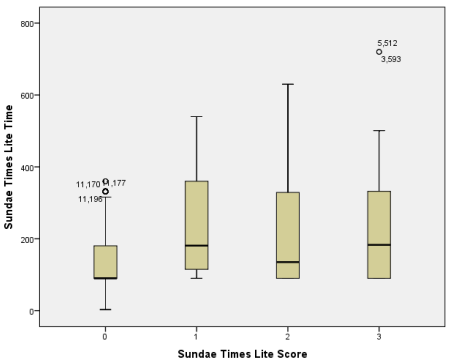

Pyraid Panic Lite Time
Pyraid Panic Lite Score

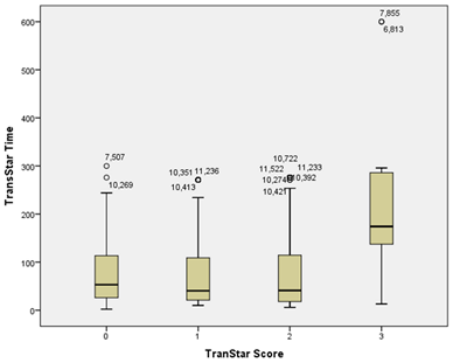

Save Our Dumb Planet Time
Save Our Dumb Planet Score


Sigma Prime Time
Sigma Prime Score


Sigma Prime Lite Time
Sigma Prime Lite Score

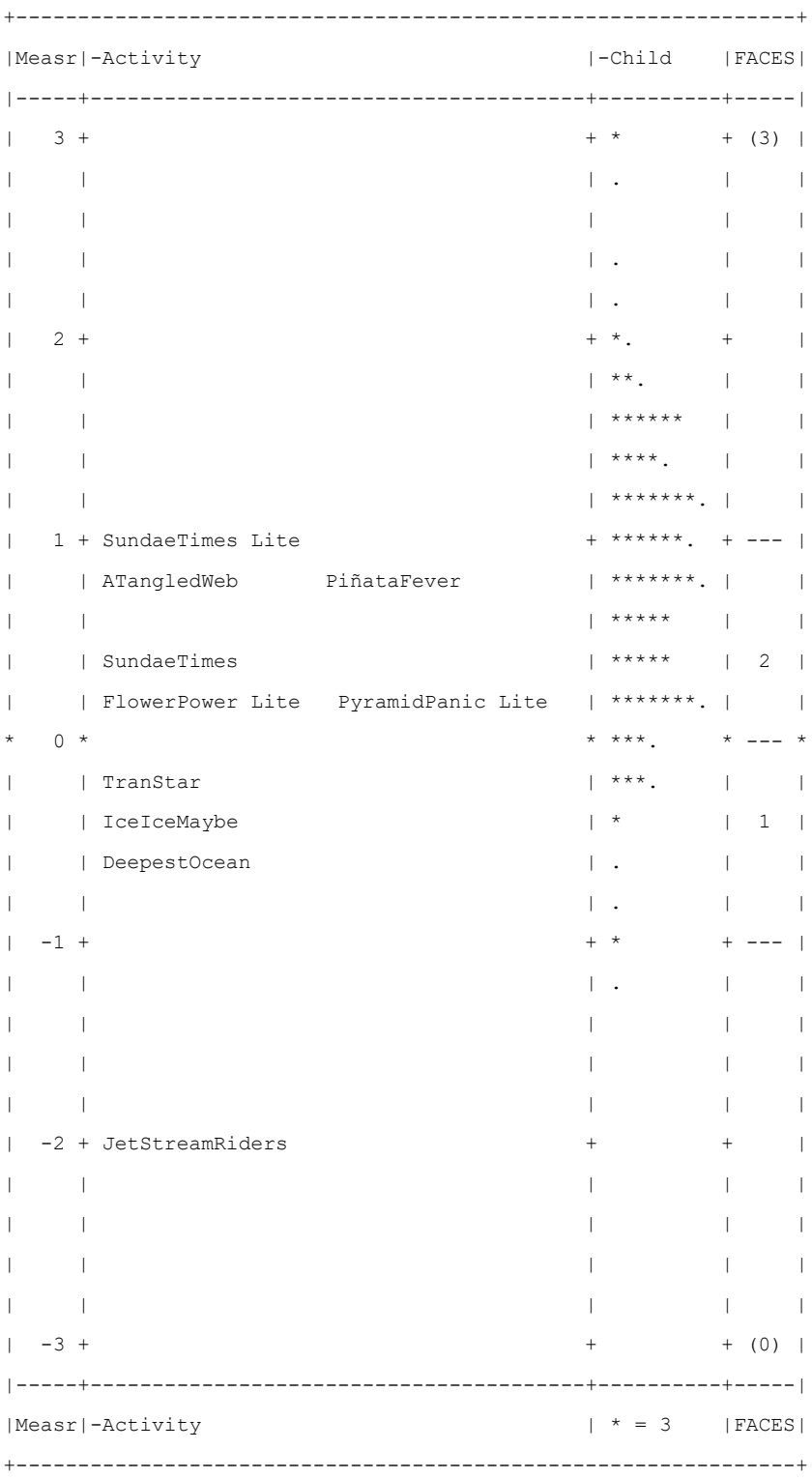

Sundae Times Time
Sundae Times Score
897
9,458
2,233
11,532
11,388
11,378


Sundae Times Lite Time
Sundae Times Lite Score
5,512
3,593
11,170
11,177
11,195


TransStar Time
TranStar Score
7,855
6,813
7,507
10,269
10,351
11,236
10,413
10,722
11,522
11,233
10,276
10,392
10,421

Appendix G Wright map for the High and Low Divisions

G.1 HighAbility

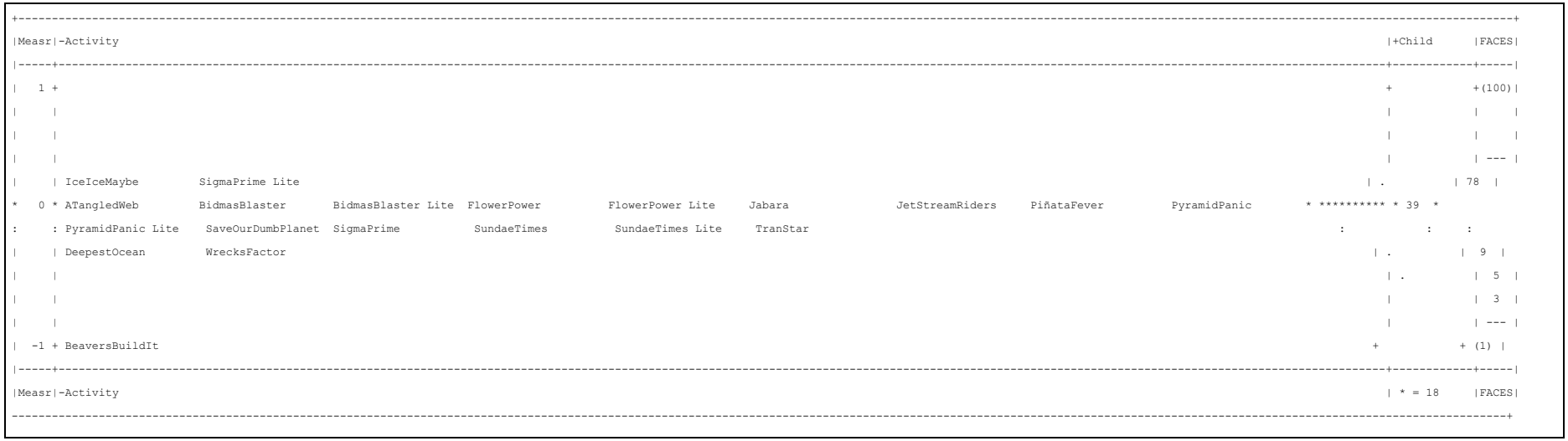


G.2 LowAbility

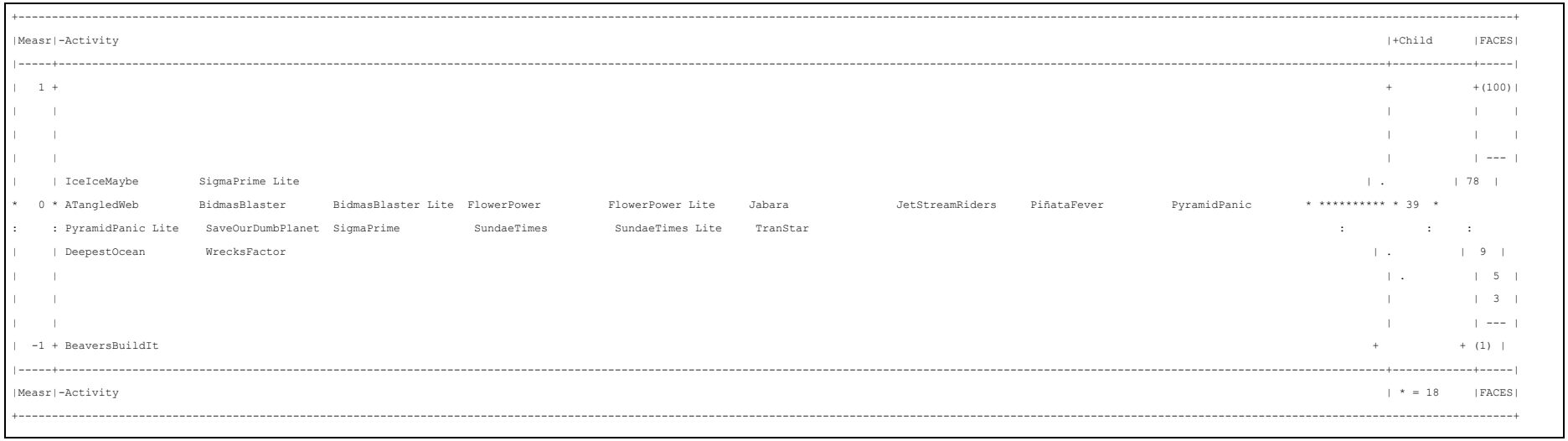
```
+-------------------------------------------------------------------------------------------+
|Measr|-Activity                                                          |-Child    |FACES|
|-----+-------------------------------------------------------------------+----------+-----|
|   4 +                                                                   +          + (3) |
|     |                                                                   |          |     |
|     | PyramidPanic                                                      |          |     |
|     |                                                                   | .        |     |
|     |                                                                   | .        |     |
|   3 +                                                                   +          +     |
|     |                                                                   | .        |     |
|     |                                                                   | .        |     |
|     |                                                                   |          |     |
|     |                                                                   |          |     |
|   2 +                                                                   + ***      +     |
|     |                                                                   | ****     |     |
|     |                                                                   | *****.   |     |
|     |                                                                   | ******* |     |
|     | SaveOurDumbPlanet                                                 | ****.    | --- |
|   1 + SundaeTimes Lite                                                  + ***.     +     |
|     | BidmasBlaster                                                     | ******.  |     |
|     | ATangledWeb         PiñataFever                                   | *****.   |     |
|     | SigmaPrime           SundaeTimes                                  | ****.    |  2  |
|     |                                                                   | ***.     |     |
*   0 * FlowerPower Lite     PyramidPanic Lite    SigmaPrime Lite        * ***      * --- *
|     |                                                                   | .        |     |
|     | FlowerPower          TranStar                                     | *.       |     |
|     | BidmasBlaster Lite  IceIceMaybe                                   |          |  1  |
|     |                                                                   | .        |     |
|  -1 + DeepestOcean                                                      +          +     |
|     |                                                                   |          | --- |
|     | Jabara                                                            |          |     |
|     |                                                                   |          |     |
|     |                                                                   |          |     |
|  -2 + BeaversBuildIt                                                    +          +     |
|     |                                                                   |          |     |
|     | JetStreamRiders                                                   |          |     |
|     |                                                                   |          |     |
|     |                                                                   |          |     |
|  -3 +                                                                   +          + (0) |
|-----+-------------------------------------------------------------------+----------+-----|
|Measr|-Activity                                                          | * = 2    |FACES|
+-------------------------------------------------------------------------------------------+
```

Appendix H Wright Maps for time by Band Division

H.1 Band 1 minimum speed



H.2 Band 2 minimum speed



H.3 Band 3 minimum speed

Appendix I Fit for HighScore and FirstFiveAttempts

	HighScore model				FirstFiveAttempts model			
	Pt Bis	**Difficulty (logits)**	**Infit**	**Outfit**	**Pt Bis**	**Difficulty**	**Infit**	**Outfit**
AlgebraMeltdown	.33	1.67	1.83	.41	.19	.41	.69	.54
ATangledWeb	.36	.26	1.01	.79	.44	-.70	1.53	1.59
BeaversBuildIt	.33	-2.46	1.14	1.32	.22	1.65	1.04	.10
BidmasBlaster	.18	.58	1.07	.93	.45	-.39	1.01	1.10
BidmasBlaster lite	.38	-.79	.68	.74	.57	-1.37	1.87	1.73
DeepestOcean	.44	-1.05	.75	.75	.59	-.19	.84	.80
FlowerPower	.41	-.81	1.26	1.28	.59	-.14	1.12	.81
FlowerPower lite	.51	-.19	1.13	.96	.49	-.12	.64	.45
IceIceMaybe	.36	-.75	1.06	1.09	.23	.12	.73	.89
Jabara	.15	-1.52	1.40	1.48	.11	.95	1.51	1.64
JetStreamRiders	.28	-2.58	1.42	1.36	.41	-.22	.92	.47
Pinata Fever	.40	.41	.67	.62	.30	-.02	.69	.33
PyramidPanic	.22	3.87	.92	.22	.48	-.33	1.64	1.85
PyramidPanic lite	.17	.01	.93	1.01	.30	.90	1.02	.41
SaveOurDumbPlanet	.31	.88	1.22	.81	.44	.25	1.46	.68
SigmaPrime	.35	.07	.96	.85	.54	-.12	.58	.49
SigmaPrime lite	.40	-.11	.89	.94	33	-.23	.91	.84
SundaeTimes	.42	.20	1.24	.94	.69	-.16	1.99	2.43
SundaeTimes lite	.47	.65	1.27	.79	.53	.21	.82	1.17
Transtar	.33	-.46	.97	1.08	.37	-.37	.72	.57
WrecksFactor	.26	1.98	.92	.22	.32	-.11	.98	.64

Appendix J

The data set used in this study consisted of:

A sample was taken of the student records for the 90-95% highest users of the MangaHigh data set (n>1 million students) at the time of extraction in January 2018.

This resulted in a data set of 427,783 rows, each row represented one individual game play.

The tables extracted were :

Name	Meaning	Type	Missing
UserRef	A unique identifier for each player	String	None
ObjectiveID	A unique identifier that maps to a Core Curriculum target. Different game levels may have more than one objective in each game.	Numeric	None
PlayedAtstart	The date of gameplay	Date	None
Achievement	A categorical variable with the four game award levels, None, Bronze, Silver and Gold	String	None
SecondsPlayed	An ordinal variable up to 3 decimal points indicating the duration of the game.	Numeric	None
Source	A categorical variable indicating whether the task was assigned by a teacher (recommendation), or the students' choice (freeplay)	String	None
Upgrade PointsEarned	A categorical variable ranging from 0-3 indicating an in-game rewards system	Numeric	None
UserCountry	A categorical variable indicating the student's country of origin	String	None
GameName	The name of the game as presented to users	String	None
GameTitle	A description of the Maths skills targeted by the game	String	None
GameDescription	A description of the game objectives	String	None
DateOfBirth	Self-reported date of birth.	Numeric	None

Glossary of Terms

Term	Definition
Anchoring values	To fix a difficulty value for a particular item when the analysis process is complete
Artificial neural networks	An artificial neural network is an interconnected way of organising computational processes so that they mirror the way signals travel between neurons in the brain. It is a form of machine intelligence.
Bayes probability net	A Bayes network is a directed graph that models the relationship between objects, using probability estimates (e.g. modelling the relationship between a test item and the concept it is believed to target). It is a form of machine intelligence.
Calibration	The meaning of calibration depends on the context. In Item Response theory, calibration is the process of describing the numerical parameters that describe the characteristics of each item. It can also mean checking to make sure that standards are being applied equally.
Classical test theory	Classical test theory, or true score theory, is the idea that a person's observed score is the sum of a true score (the person's real ability) and an error value.
Clustering	A way of grouping objects so that they are more similar to each other than they are to others not in the group
Clustering algorithm	A mathematical process to identify how objects can be grouped into a cluster
Cognitive competencies	Cognitive competences are our ability to carry out brain-based tasks from simple to complex.

Computer adaptive testing	Computer Adaptive testing adapts the difficulty of test items presented in a test to the test-takers' ability, which is often estimated dynamically during the test.
Concept	An abstract idea of a latent variable that we wish to measure (e.g. problem solving)
Concept model	A representation of how people know, understand or simulate a concept.
Concurrent validity	Concurrent validity of a test is how well it performs measured against another test that targets the same concept.
Conditional error of measurement	A conditional error of measurement is similar to SEM, but takes into account variation in reliability at the higher and lower boundaries of a scale.
Dichotomous score	Data that has only two values (right/wrong, pass/fail)
Dimension	In Bayes nets, a dimension model is a believed connection between nodes and edges.
Directed acyclic graph	A set of objects that are connected together, where the relationship between them goes one-way. E.g. a person (object) takes (relationship) a test (object). The order of the objects cannot be reversed.
Discrete	Treated as individual, separate or distinct. In testing, it is the principle that one item should not depend on or be influenced by the response to another. In programming, discrete means that the number of variables included are likely to be finite.
Discriminate	In assessment, the power of a test to separate high-achievers from low-achievers.
Domain model	In a Bayes net, the domain model is an abstraction of believed knowledge-object relationships (e.g., that item X evidences creativity, but not knowledge retrieval)
Dynamics	The behaviour in a game (e.g. the controls, visuals, audio etc)

Embedded assessment	Embedded assessments are activities that are completed as part of the learning process or lesson, but are also used to provide assessment data about a learning outcome.
Error estimate	The difference between the observed response and an expected response
Facet	In factor analysis, a grouping of items that show statistical evidence of measuring the same concept.
Factor analysis	A statistical method to describe variability among observed, correlated variables, particularly with the aim of reducing redundant variables from a measurement tool.
Formal logic	A set of rules for making deductions that seem self-evident. It is often used to describe the symbols used to represent those rules.
Formative assessment	A range of assessment activities during the learning process with the aim of informing subsequent activities to encourage progress.
Fuzzy cognitive maps	Fuzzy cognitive maps are a formal way of representing social scientific knowledge and modelling decision-making or causal knowledge in complex systems.
Graph	In computer terms, a graph is an abstract representation of a finite set of vertices or nodes (such as a test item, skill or concept) and edges, or relationships (such as ‘demonstrates’ ‘impacts on’).
Guessing behaviour	In testing guessing behaviour covers a range of practices such as random guessing, ruling out a number of options and then guessing, using flaws in the test construction to aid guessing or other non-targeted means to get the right answer.
Heat map	A heat map is a graphical representation on a matrix where colours are used to represent intensity of activity, such as failure or success.

Hypertext	A non-linear organisation of ideas, with cross-referencing between related ideas
In-fit	In-fit statistics are residual based statistics that indicate the extent of misfit between observed and expected results, with a weighting applied to observations around the point of interest
Interface	In gaming, the computer interface is what the player sees or interacts with directly.
Interference	Error caused by the influence of non-targeted variables on results
Item	A test question, including the question wording and stimulation materials, options if it is a multiple choice question, and the rules for scoring the question
Item invariance	When an item's difficulty level remains stable across different populations of test takers.
Item response theory	Item Response Theory is a test theory that assumes that the probability of a correct response is a mathematical function of a person's ability parameter and one or more parameters that characterize an item.
Iteration	A repetition In gaming, it would be repeating a task. In IRT, it is the repetition of a statistical process until a specified convergence is achieved.
Latency	In computing, latency is a delay between the instruction to begin the transfer of data and the actual completion of that transfer.
Limen	In testing, the limen is the level where the failure to answer an item correctly indicates a person's likely failure to answer the remaining questions above that threshold in difficulty (e.g. a 50-50 probability of a correct response, or an 80-20 probability if evidence of mastery is required)

Linear narrative form	In gaming, a linear narrative dictates the order of game play to the user, with no choice of next steps.
Logarithmic transformation	A process used to convert nominal raw score data into a log odds ratio on a common interval scale around a mean of 0.
Mechanics	The rules of a video game
Multi-dimensional	A multi-dimensional measurement has more than one line of measurement, for example self worth.
Next Steps Analysis	Next steps analysis looks at patterns of behaviour to identify in one action is always followed by another.
Natural Language Processing	Natural language processes is an AI technique that identifies word patterns in sentences to extract meaning.
Non-linear narrative form	In gaming, a non-linear narrative gives the player a choice of what to do next.
Out-fit	An out-fit statistic is a residual based statistic that indicates the extent of misfit between observed and expected results, unweighted
Paradata	In gaming, it is data about how the data was collected, such as a time stamp.
Parameter	A number that describes a measurable characteristic, such as the difficulty or the discrimination power of task.
Partial credit	A partial credit is a polytomous score awarded to a response, such as 3 out of 5.
Person invariance	When a person's ability estimate remains stable across different measurement tools.
PISA	PISA is a triennial survey that aims to evaluate global education systems, by testing 15-year-old students in problem-solving tasks.
Priming	The effect when one stimulus affects another

Psychometrics	The study and use of statistical operations associated with tests of psychological characteristics, mental abilities, educational, occupational or knowledge skills.
Rasch analysis	A form of Item Response Theory that assumes that a probability measurement can be obtained by measuring only one characteristic of an item, its difficulty.
Raw scores	Scores in their original state, or that have not been statistically manipulated to be comparable with scores in other tests. It is generally expressed as the sum of the number of correct responses or the sum of ratings, and is generally not used as an expression of test taker ability.
Reliability	In assessment in general, it is the evidence collected to support the argument that test scores will be consistent on two or more occasions of testing, assuming that the test taker's knowledge has not changed between tests. It is generally understood as an argument, not a statistic. Statistically, a reliability value for a test is the ratio of sample or test variance to the total observed variance.
Scale	Statistically transformed scores to make them comparable to other scores.
Semantic ontologies	In computing, an ontology is a formal way of naming and defining the types, properties and interrelation of entities. A Semantic Ontology embeds these into computer code made available on the Web.
Serious educational game	A game which has education as the primary purpose.
Standard error of measurement ($\theta\theta$)	Standard error of measurement is an indication of the variation we would expect in an observed score, taking into account the error of measurement in any given test. It is expressed as standard deviations on a normal curve. For example, for a test with an SEM of ±2, and a score of 25/40, we can be 95% certain (2 standard deviations) that the student's true scores lies between 21 and 29 points (2 x 2 = ±4).

Standards	In the World Wide Web, standards are the internationally agreed rules and guidelines among the international community to promote consistency in design code. This agreement allows web pages to be experienced consistently in many different contexts.
Substantive material	In education, substantive literature would be the findings based on the study of a phenomenon in the real world.
Summative assessment	Summative assessments measure a student's learning at the end of an instructional period and are used in comparison to some standard or benchmark.
Threshold	In testing, the limen is the level where the failure to answer an item correctly indicates a person's likely failure to answer the remaining questions above that threshold in difficulty (e.g. a 50-50 probability of a correct response, or an 80-20 probability if evidence of mastery is required).
Telemetry data	Data collected during the process of gameplay
Test fatigue	In assessment, test fatigue is the increased influence of tiredness on test results as the test progresses.
Timss	Regular international comparative assessment of children in maths and science to measure the effectiveness of educational systems
Transference	In education, transference is the ability to learn a skill in one context and apply it in another.
Uni-dimensional	A uni-dimensional measure has just one line of measure, such as height, or is a multi-dimensional measure that has been forced to behave uni- dimensionally, such as creating a test of self-worth by asking people to place themselves on a scale from low to high.
Validity	The evidence collected to support the claim that a tool measures the concept that it claims to measure. Validity is

	generally understood as an argument, not a statistic. It is a feature of the test and the use of the test.
Zone of Proximal Development	The Zone of Proximal Development was Lev Vygotsky's theory that learning is optimised when the level of difficulty of a task is slightly higher than the current ability of the person.

Bibliography

Adams, E. and Rollings, A. (2010) *Fundamentals of game design.* 2nd ed. edn. Berkeley, Calif.: New Riders ; London : Pearson Education [distributor].
AERA (1999) *American Educational Research Association Standards for educational and psychological testing.* Amer Educational Research Assn.
Ahmed, S. (2009) 'Methods in sample surveys', *Johns Hopkins Bloomberg School of Public*.
Allison, P. D. (2001) *Missing data.* Sage publications.
Almond, R. G. (2015) *Tips and Tricks for Building Bayesian Networks for Scoring Game-Based Assessments. Advanced Methodologies for Bayesian Networks. Second International Workshop, AMBN 2015. Proceedings: LNCS 9505*.
Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D. and Williamson, D. M. (2015) *Bayesian networks in educational assessment.* Springer.
Ananiadou, K. and Claro, M. (2009) '21st century skills and competences for new millennium learners in OECD countries'.
Andrich, D. (1988) *Rasch models for measurement.* Sage.
Augustin, T., Hockemeyer, C., Kickmeier-Rust, M. and Albert, D. (2011) 'Individualized Skill Assessment in Digital Learning Games: Basic Definitions and Mathematical Formalism', *IEEE Transactions on Learning Technologies,* 4(2), pp. 138-148.
Baird, J., Isaacs, T., Opposs, D. and Gray, L. (2018) *Examination standards : how measures and meanings differ around the world.*
Baron, H. B., Crespo, R. G., Espada, J. P. and Martinez, O. S. (2015) 'Assessment of learning in environments interactive through fuzzy cognitive maps', *Soft Computing,* 19(4), pp. 1037-1050.
Baron, H. B., Salinas, S. C. and Crespo, R. G. (2014) 'An Approach to Assessment of Video Game-based Learning using Structural Equation Model', in Rocha, A., Fonseca, D., Redondo, E., Reis, L.P. and Cota, M.P. (eds.) *Proceedings of the 2014 9th Iberian Conference on Information Systems and Technologies Iberian Conference on Information Systems and Technologies*.
Baumeister, R. F. (2013) 'Writing a literature review', *The Portable Mentor*: Springer, pp. 119-132.
Baumeister, R. F. and Leary, M. R. (1997) 'Writing narrative literature reviews', *Review of general psychology,* 1(3), pp. 311.
Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S. and Bavelier, D. (2018) 'Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills', *Psychological bulletin,* 144(1), pp. 77.
Begg, C. B., Cooper, H. and Hedges, L. (1994) 'Publication bias', *The handbook of research synthesis,* 25, pp. 299-409.
Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N. and Weitzner, D. J. (2006) 'A framework for web science', *Foundations and trends in Web Science,* 1(1), pp. 1-130.
Bolivar Baron, H., Castillo Salinas, S. and Gonzalez Crespo, R. (2014) 'An Approach to Assessment of Video Game-based Learning using Structural Equation Model', in Rocha, A., Fonseca, D., Redondo, E., Reis, L.P. and Cota, M.P. (eds.) *Proceedings of the 2014 9th Iberian Conference on Information Systems and Technologies Iberian Conference on Information Systems and Technologies*.
Bolivar Baron, H., Gonzalez Crespo, R., Pascual Espada, J. and Sanjuan Martinez, O. (2015) 'Assessment of learning in environments interactive through fuzzy cognitive maps', *Soft Computing,* 19(4), pp. 1037-1050.
Bolivar Baron, H., Martinez Rojas, M., Trujillo Diaz, J. and Velasquez Contreras, A. (2014) 'Graph Isomorphism in Fuzzy Cognitive Maps for Monitoring of Game-based Learning', *2014 International Conference on Intelligent Networking and Collaborative Systems (Incos)*, pp. 304-310.
Bond, T. and Fox, C. M. (2015) *Applying the Rasch model: Fundamental measurement in the human sciences.* Routledge.

Bressler, M. D. and Bodzin, M. A. (2013) 'A Mixed Methods Assessment of Students' Flow Experiences during a Mobile Augmented Reality Science Game', *Journal of Computer Assisted Learning,* 29(6), pp. 505-517.
Camara, W. J., Dorans, N. J., Morgan, R. and Myford, C. (2000) 'Advanced placement: Access not exclusion', *education policy analysis archives,* 8, pp. 40.
Chin, D. B., Blair, K. P. and Schwartz, D. L. (2016) 'Got Game? A Choice-Based Learning Assessment of Data Literacy and Visualization Skills', *Technology Knowledge and Learning,* 21(2), pp. 195-210.
Csciksz entmihaly, M. 1975. Beyond boredom and anxiety: Experiencing Flow in Work and Play. Josey–Bass inc, Publishers.
Csikszentmihalyi, M. (1997) *Finding flow: The psychology of engagement with everyday life.* Basic Books.
Cumming, G. (2013a) 'The new statistics why and how', *Psychological science*, pp. 0956797613504966.
Cumming, G. (2013b) *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.
Davier, A. A. and Halpin, P. F. (2013) 'Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations', *ETS Research Report Series,* 2013(2).
de Klerk, S., Eggen, T. J. and Veldkamp, B. P. (2014) 'A blending of computer-based assessment and performance-based assessment: Multimedia-Based Performance Assessment (MBPA). The introduction of a new method of assessment in Dutch Vocational Education and Training (VET)', *Cadmo*.
de Klerk, S., Veldkamp, B. P. and Eggen, T. (2015) 'Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example', *Computers & Education,* 85, pp. 23-34.
DiCerbo, K. E. (2014) 'Game-Based Assessment of Persistence', *Educational Technology & Society,* 17(1), pp. 17-28.
DiCerbo, K. E., Mislevy, R. J. and Behrens, J. T. (2016) *INFERENCE IN GAME-BASED ASSESSMENT. Using Games and Simulations for Teaching and Assessment: Key Issues*.
El-Nasr, M. S., Drachen, A. and Canossa, A. (2016) *Game analytics.* Springer.
EPPI (accessed January 2017) *Evidence for Policy and Practice Information and Co-ordinating Centre: What Works*. Available at: http://eppi.ioe.ac.uk.
Eseryel, D., Ifenthaler, D. and Ge, X. (2011) 'Alternative assessment strategies for complex problem solving in game-based learning environments', *Multiple perspectives on problem solving and learning in the digital age*: Springer, pp. 159-178.
Eseryel, D., Ifenthaler, D. and Ge, X. (2013) 'Validation Study of a Method for Assessing Complex Ill-Structured Problem Solving by Using Causal Representations', *Educational Technology Research and Development,* 61(3), pp. 443-463.
FAS (2006) *Summit on Educational Games*. Available at: https://fas.org/programs/ltp/policy_and_publications/summit/Summit%20on%20Educational%20Games.pdf.
Ferzli, M., Pigford, K. and Black, B. (2015) 'DEVELOPMENT OF A GAMIFIED LEARNING OBJECT (GLO) WITH CORRELATED CLASSROOM ACTIVITIES TO ENHANCE STUDENT UNDERSTANDING OF EVOLUTION', in Chova, L.G., Martinez, A.L. and Torres, I.C. (eds.) *Iceri2015: 8th International Conference of Education, Research and Innovation ICERI Proceedings*, pp. 2898-2906.
Finfgeld-Connett, D. and Johnson, E. D. (2013) 'Literature search strategies for conducting knowledge-building and theory-generating qualitative systematic reviews', *Journal of Advanced Nursing,* 69(1), pp. 194-204.
Flynn, J. R. (2007) *What is intelligence?: Beyond the Flynn effect.* Cambridge University Press.
Frezzo, D. C., Behrens, J. T., Mislevy, R. J., West, P., DiCerbo, K. E. and Ieee (2009) *Psychometric and Evidentiary Approaches to Simulation Assessment in Packet Tracer Software. Icns: 2009 Fifth International Conference on Networking and Services*.
Fullan, M., Langworthy, M. and Barber, M. (2014) 'A rich seam', *How New Pedagogies Find Deep Learning. Online: http://npdl. thumbtack. co. nz/wpcontent/uploads/2015/08/A-Rich-Seam. pdf*.

Fullerton, T. (2014) *Game design workshop: a playcentric approach to creating innovative games.* CRC press.
Gee, J. P. (2005) 'Learning by design: Good video games as learning machines', *E-learning and Digital Media,* 2(1), pp. 5-16.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian data analysis.* Chapman and Hall/CRC.
Giddens, A. 2009. Sociology (th edition). Cambridge: Polity Press.
Gosper, M. and McNeill, M. (2012) 'Implementing game-based learning: The MAPLET framework as a guide to learner-centred design and assessment', *Assessment in Game-Based Learning*: Springer, pp. 217-233.
Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C. and Germany, M.-L. (2018) 'Challenges of assessing collaborative problem solving', *Assessment and teaching of 21st century skills*: Springer, pp. 75-91.
Graf, E. A. (2014) 'Connecting Lines of Research on Task Model Variables, Automatic Item Generation, and Learning Progressions in Game-Based Assessment', *Measurement: Interdisciplinary Research and Perspectives,* 12(1), pp. 5.
Gwet, K. L. (2014) *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC.
Göbel, S., Wendel, V., Ritter, C. and Steinmetz, R. 'Personalized, adaptive digital educational games using narrative game-based learning objects'. *International Conference on Technologies for E-Learning and Digital Entertainment*: Springer, 438-445.
Hainey, T., Connolly, T., Baxter, G., Boyle, L. and Beeby, R. 'Assessment Integration in Games-Based Learning: A Preliminary Review of the Literature'. Reading, 2012/10//

Oct 2012: Academic Conferences International Limited, 174-XII.
Halford, S., Pope, C. and Carr, L. (2010) 'A manifesto for Web Science'.
Harvill, L. M. (1991) 'Standard error of measurement', *Educational Measurement: issues and practice,* 10(2), pp. 33-41.
Hawkins, D. M. (1980) *Identification of outliers.* Springer.
Hintze, J. (2001) 'NCSS and Pass', *Number cruncher statistical systems. Kaysville, Utah*.
Ifenthaler, D., Adcock, A. B., Erlandson, B. E., Gosper, M., Greiff, S. and Pirnay-Dummer, P. (2014) 'Challenges for Education in a Connected World: Digital Learning, Data Rich Environments, and Computer-Based Assessment-Introduction to the Inaugural Special Issue of Technology, Knowledge and Learning', *Technology, Knowledge and Learning,* 19(1-2), pp. 121-6.
Ioannou-Georgiou, S. (2003) *Assessing young learners.* Oxford University Press.
Kim, Y. J., Almond, R. G. and Shute, V. J. (2016) 'Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground', *International Journal of Testing,* 16(2), pp. 142-163.
Kohn, A. (2002) 'The dangerous myth of grade inflation', *The Chronicle of Higher Education,* 49(11), pp. B7.
Koster, R. (2005) *A theory of fun for game design.* Scottsdale, Ariz.: Paraglyph.
Kreuter, F. and Casas-Cordero, C. (2010) '4. Paradata'.
Lamb, R. L., Annetta, L., Vallett, D. B. and Sadler, T. D. (2014) 'Cognitive diagnostic like approaches using neural-network analysis of serious educational videogames', *Computers & Education,* 70, pp. 92-104.
Lee, Yi-Hsuan. (2015) A mixture cure-tate model for responses and response times in time limit tests. *Psychometrika*
Leighton, J. P. and Chu, M.-W. (2016) 'First among Equals: Hybridization of Cognitive Diagnostic Assessment and Evidence-Centered Game Design', *International Journal of Testing,* 16(2), pp. 164-180.
Leighton, J. P. and Gierl, M. J. (2011) *The learning sciences in educational assessment: The role of cognitive models.* Cambridge University Press.

Leonard, T. and Hsu, J. S. (2001) *Bayesian methods: an analysis for statisticians and interdisciplinary researchers.* Cambridge University Press.
Levy, R. and Mislevy, R. J. (2017) *Bayesian psychometric modeling.* Chapman and Hall/CRC.
Levy, Y. and Ellis, T. J. (2006) 'A systems approach to conduct an effective literature review in support of information systems research', *Informing Science: International Journal of an Emerging Transdiscipline,* 9(1), pp. 181-212.
Linacre, J. M. (2006) 'WINSTEPS Rasch measurement computer program', *Chicago: WINSTEPS. com*.
Lineacre, J. and Wright, B. 1998. A user's Guide to BIGSTEPS/WINSTEPS Rasch Model Computer Program. Chicago: MESA Press (University of Chicago).
Lord, F. M. (1980) *Applications of item response theory to practical testing problems.* Routledge.
Luckin, R. (2018) *Enhancing Learning and Teaching with Technology: What the Research Says.* ERIC.
Ludlow, L. H. and O'leary, M. (1999) 'Scoring omitted and not-reached items: Practical data analysis implications', *Educational and Psychological Measurement,* 59(4), pp. 615-630.
Lunn, D., Jackson, C., Best, N., Spiegelhalter, D. and Thomas, A. (2012) *The BUGS book: A practical introduction to Bayesian analysis.* Chapman and Hall/CRC.
MacKenzie, D. and Wajcman, J. (1999) *The social shaping of technology.* Open university press.
Mead, C. (2013) *War play: Video games and the future of armed conflict.* Houghton Mifflin Harcourt.
Maris, Gunter. (2012) Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*
Messick, S. (1987) 'Validity', *ETS Research Report Series,* 1987(2).
Mislevy, R. J., Almond, R. G. and Lukas, J. F. (2003) 'A brief introduction to evidence-centered design', *ETS Research Report Series,* 2003(1), pp. i-29.
Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Frezzo, D. C. and West, P. (2012) 'Three things game designers need to know about assessment', *Assessment in game-based learning*: Springer, pp. 59-81.
Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., John, M. and Drasgow, F. (2015) 'Psychometrics and game-based assessment', *Technology and Testing: Improving Educational and Psychological Measurement*, pp. 23-48.
Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K. and John, M. (2014) 'Psychometric considerations in game-based assessment', *GlassLab Report*.
Mislevy, R. J. and Wu, P. K. (1996) 'Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing', *ETS Research Report Series,* 1996(2).
Mislevy, R. J. a. (2018) *Socio-cognitive foundations of educational measurement.*
Mitgutsch, K. and Alvarado, N. 'Purposeful by design?: a serious game design assessment framework'. *Proceedings of the International Conference on the foundations of digital games*: ACM, 121-128.
National Research Council . Committee on Science Learning: Computer Games, S. a. E., Honey, M. and Hilton, M. L. (2011) *Learning science through computer games and simulations.* Washington, D.C.: National Academies Press.
Nelson, T. (1982) 'Literary machines'.
Nind, M. (2006) *Conducting systematic review in education: a reflexive narrative.*
Ofqual, archives, N. (2018) *Entries for GCSE, AS and A level, Summer 2018 exam series.* Coventry: Office for Qualifications and Examinations Regulation.
Okoli, C. and Schabram, K. (2010) 'A guide to conducting a systematic literature review of information systems research', *Sprouts Work. Pap. Inf. Syst,* 10(26).
Oudshoorn, N. E. and Pinch, T. (2003) *How users matter: The co-construction of users and technologies.* MIT press.
P21 (2002) *P21 Framework for 21st Century Learning*. Available at: http://www.p21.org/our-work/p21-framework (Accessed: 13/01/2017.
Pan, Y.-C. (2009) 'A review of washback and its pedagogical implications'.

Petticrew, M. and Roberts, H. (2006) 'How to appraise the studies: an introduction to assessing study quality', *Systematic reviews in the social sciences: A practical guide*, pp. 125-163.
Rasch, G. (1960) 'Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests'.
Raykov, T. and Marcoulides, G. A. (2011) *Introduction to psychometric theory.* Routledge.
Reese, D. D., Seward, R. J., Tabachnick, B. G., Hitt, B. A., Harrison, A. and Mcfarland, L. (2012) 'Timed Report measures learning: Game-based embedded assessment', *Assessment in Game-Based Learning*: Springer, pp. 145-172.
Reese, D. D. and Tabachnick, B. G. (2010) *The Moment of Learning: Quantitative Analysis of Exemplar Gameplay Supports CyGaMEs Approach to Embedded Assessment*: Society for Research on Educational Effectiveness. 2040 Sheridan Road, Evanston, IL 60208. Available at: https://search.proquest.com/docview/822505032?accountid=13963.
Regan, K. W. and Biswas, T. 'Psychometric modeling of decision making via game play'. *2013 IEEE Conference on Computational Inteligence in Games (CIG)*: IEEE, 1-8.
Reingold, J. (2015) *Everybody hates Pearson*. Fortune. Available at: http://fortune.com/2015/01/21/everybody-hates-pearson/ (Accessed: 21 Jan 2015.
Repko, A. F. (2008) *Interdisciplinary research: Process and theory.* Sage.
Riesenberg, L. A. and Justice, E. M. (2014) 'Conducting a successful systematic review of the literature, part 1', *Nursing,* 44(4), pp. 13-17.
Rodway, C., Tham, S.-G., Ibrahim, S., Turnbull, P., Windfuhr, K., Shaw, J., Kapur, N. and Appleby, L. (2016) 'Suicide in children and young people in England: a consecutive case series', *The Lancet Psychiatry,* 3(8), pp. 751-759.
Sailer, M. and Homner, L. 2019. The Gamification of Learning: a Meta-analysis. Springer.
Sala, G., Tatlidil, K. S. and Gobet, F. (2017) 'Video Game Training Does Not Enhance Cognitive Ability: A Comprehensive Meta-Analytic Investigation'.
Sanni, A. and Mohammad, M. (2015) 'Computer based testing (CBT): An assessment of student perception of JAMB UTME in Nigeria', *Computer,* 6(2).
Schnepp, J. and Rogers, C. (2015) 'Just Give Me a Hint! An Alternative Testing Approach for Simultaneous Assessment and Learning', in Uskov, V.L., Howlett, R.J. and Jain, L.C. (eds.) *Smart Education and Smart e-Learning Smart Innovation Systems and Technologies*, pp. 141-150.
Scientists, F. o. A. (2006) *Summit on Educational Games*. Available at: https://fas.org/programs/ltp/policy_and_publications/summit/Summit%20on%20Educational%20Games.pdf.
Sellar, S. and Lingard, B. (2014) 'The OECD and the expansion of PISA: New global modes of governance in education', *British Educational Research Journal,* 40(6), pp. 917-936.
Shute, V. J. (2011) 'Stealth Assessment in Computer-Based Games to Support Learning', *Computer Games and Instruction*, pp. 503-524.
Shute, V. J. and Ke, F. (2012) 'Games, learning, and assessment', *Assessment in game-based learning*: Springer, pp. 43-58.
Shute, V. J., Wang, L. B., Greiff, S., Zhao, W. N. and Moore, G. (2016) 'Measuring problem solving skills via stealth assessment in an engaging video game', *Computers in Human Behavior,* 63, pp. 106-117.
Soderstrom, N. C. and Bjork, R. A. (2015) 'Learning versus performance: An integrative review', *Perspectives on Psychological Science,* 10(2), pp. 176-199.
Stobart, G. *Testing times : the uses and abuses of assessment.*
Swartout, W. R., Nye, B. D., Hartholt, A., Reilly, A., Graesser, A. C., VanLehn, K., Wetzel, J., Liewer, M., Morbini, F. and Morgan, B. 'Designing a personal assistant for life-long learning (PAL3)'. *The Twenty-Ninth International Flairs Conference*.
Thurstone, L. L. (1937) 'Ability, motivation, and speed', *Psychometrika,* 2(4), pp. 249-254.
Tinati, R., Carr, L., Halford, S. and Pope, C. (2014) '(Re) integrating the Web: beyond 'socio-technical''.

Van den Broeck, J., Cunningham, S. A., Eeckels, R. and Herbst, K. (2005) 'Data cleaning: detecting, diagnosing, and editing data abnormalities', *PLoS medicine,* 2(10), pp. e267.
Van Der Linden, W. J. (2009) 'Conceptual issues in response-time modeling', *Journal of Educational Measurement,* 46(3), pp. 247-272.
Van der Linden, W. J. and Glas, C. A. (2000) *Computerized adaptive testing: Theory and practice.* Springer.
Vendlinski, T. P., Delacruz, G. C., Buschang, R. E., Chung, G. K. and Baker, E. L. (2010) *Developing High-Quality Assessments that Align with Instructional Video Games. CRESST Report 774*: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). 300 Charles E Young Drive N, GSE&IS Building 3rd Floor, Mailbox 951522, Los Angeles, CA 90095-1522. Available at: https://search.proquest.com/docview/822507322?accountid=13963.
Vygotsky, L. (1987) 'Zone of proximal development', *Mind in society: The development of higher psychological processes,* 5291, pp. 157.
Walliman, N. (2017) *Research methods: The basics.* Routledge.
Walls, E., Santer, M., Wills, G. and Vass, J. (2015) 'The Dreams Plan: A Blupoint Strategy for e-Education Provision in South Africa', *The Electronic Journal of Information Systems in Developing Countries,* 70(1), pp. 1-24.
Wilson, M. (2004) *Constructing measures: An item response modeling approach.* Routledge.
Wright, B. D. and Masters, G. N. (1982) *Rating scale analysis.* MESA press.
Wright, D. B. (2016) 'Treating all rapid responses as errors (TARRE) improves estimates of ability (slightly)', *Psychological Test and Assessment Modeling,* 58(1), pp. 15-31.
Yin, R. (2003) *Designing case studies.*
Zheng, Y. and De Jong, J. (2011) 'Research note: Establishing construct and concurrent validity of Pearson Test of English Academic', *Retrieved on September,* 29, pp. 2016.