

# MEASURING THE DIFFICULTY OF COMPUTER GAMES USING THE RASCH MODEL

MECH3890 Individual Engineering Project

***Measuring the difficulty of computer  
games using the Rasch model***

*Author: Pablo Trescoli Garcia 201075711*

*Supervisor: Brian Henson*

*Examiner: Mark Wilson*

*Date: 27<sup>th</sup> April 2017*

MECH3890 Individual Engineering Project

TITLE OF PROJECT

Measuring the difficulty of computer games using the Rasch model

PRESENTED BY

Pablo Trescoli Garcia

If The Project Is Industrially Linked Tick This Box  
And Provide Details Below

Company Name and Address:

This project report presents my own work and does not contain any unacknowledged work from any other sources.

Signed

date 27<sup>th</sup> April 2017

Page left intentionally blank

## Table of Contents

---

- Abstract	IV
- List of figures/tables	V
- Chapter 1	
1. Introduction	1
2. Aims	2
3. Objectives	2
4. Project layout	3
- Chapter 2	
Literature review	
2.1 Introduction	4
2.2 Literature review	5
2.2.1 Video game difficulty	5
2.2.2 Rasch model and its applications	6
2.2.2.1 Guttman scaling	8
2.2.3 Statistical parameters	9
2.2.4 Chosen videogames	11
2.2.5 Social factors	14
2.3 Summary	14
- Chapter 3	
Methodology	
3.1 Introduction	15
3.2 Recordings & gameplay	15
3.3 Gameplay performance conversion	16
3.4 Data processing for RUMM 2030	19

<b>- Chapter 4</b>	
<b>Results</b>	
4.1 Summary statistics	20
4.2 Individual person statistics	22
4.3 Individual Item statistics	23
4.4 Threshold maps	25
4.5 Unidimensional test	26
4.6 Summary & comparing games	28
<b>- Chapter 5</b>	
<b>Conclusion and discussion</b>	
5.1 Achievements	30
5.2 Discussion	30
5.3 Conclusion	32
5.4 Future work	33
<b>- References</b>	34
<b>- Appendix</b>	36

Page left intentionally blank

## **Abstract**

The purpose of this project was to use the Rasch model to develop a method that estimates videogame difficulty. A large number of people were recorded playing two different videogames so their gameplay footages could be analysed to measure their performance in the game by coming up with a set of quantitative parameters that could objectively measure respondents performance (items). The collected data was processed into software RUMM 2030 which generated the Rasch model fit from the data. The model is then analysed to evaluate if it is a good estimation of difficulty and if the results are reliable.

The 2 game results can be compared to conclude if one game fits better to the Rasch model and what parameter causes the better fit or higher reliability on the model.

The results were more reliable in terms of the person separation index for one of the games, which also proved to be the more unidimensional game of the two. Some of the parameters for both games were inconclusive but the general results were aiming in the right direction.

It is concluded that the Rasch model can be used to analyse the difficulty of videogames as long as the measured videogame has a strong unidimensionality and the items and their categorisation don't misfit heavily the model.

The project can also be interpreted as a starting point to what could be a method of analysing difficulty of computer/electronic tasks from the real world, which could be a useful resource to many technical professional fields.

## **List of figures/tables**

- Figure 1: Item category curve for polytomous item.
  - Figure 2: Picture 1 arcade old game
  - Figure 3: Picture alternative modern game
  - Figure 4: Picture Donkey Kong gameplay
  - Figure 5: Picture Galaxian gameplay
  - Figure 6: Data conversion for RUMM 2030
  - Figure 7: Summary statistics for Donkey Kong
  - Figure 8: Summary statistics for Galaxian
  - Figure 9: Individual person fit Donkey Kong
  - Figure 10: Individual person fit Galaxian
  - Figure 11: Individual Item fit Donkey Kong
  - Figure 12: Summary statistics for Donkey Kong after item removal
  - Figure 13: Individual item fit Galaxian
  - Figure 14: Threshold map Donkey Kong
  - Figure 15: Threshold map Galaxian
  - Figure 16: Unidimensional t-student test
- 
- Table 1: Person separation index classification
  - Table 2: Item and categorisation for Donkey Kong
  - Table 3: Item and categorisation for Galaxian



Page left intentionally blank

### Introduction

#### 1. Introduction

The Oxford dictionary defines difficult to be “A thing that is hard to accomplish, deal with, or understand.” [1] So how complex can the answer to the question “*How difficult is ...?*” be?

There are no units of measurement for difficulty, nor a scaling system that objectively describes how hard something is to accomplish, deal with, or to understand. And yet, difficulty is used as a rating for tasks, there are obvious scenarios of difficult or easy tasks, which sparks the possibility of measuring difficulty, or at least attempting to understand better the background and factors of the difficulty of a task.

Difficulty can be qualified as subjective, as it can be unique to each individual and thus the best way to analyse it are psychometric approaches. The Rasch measurement model is a psychometric approach that has been served for a variety of applications in the recent decade, especially in health sciences [2] [3]. The model verifies the standards for validity, reliability and responsiveness.

Using the model to evaluate the difficulty of a task could be an auxiliary tool to many fields and workplaces as it could help develop the productivity and the profile of the person who does the task.

To evaluate if the model works the Rasch analysis requires a latent trait [4] which is the main factor influencing the different probabilities of being successful at a task. This project brings forward a simplified scenario to evaluate the model with a “computer but not strictly professional” task: video games. Using video games as a starting point to evaluate the model simplifies the complexities of measuring performance and of setting up an experiment that every respondent has the same knowledge about, desirably none.

This project’s approach is to measure the difficulty of videogames using the Rasch model analysis.

The required basic background to understand the project concerns videogame difficulty, Rasch analysis and statistics.

The resources needed aren't numerous, any modern laptop that could install the video games and a program to record the gameplay, and a RUMM 2030 license, which is the program that implants the Rasch analysis. An important amount of time for this project concerns the collection of numerous respondents playing video games and their game footage being recorded.

## **2. Aims**

The aim of this project is to use the Rasch model as a potential measurement method to evaluate the difficulty of computer video games, which can be presented with visual tools such as graphs or tables, and can be validated with numerical statistical coefficients.

## **3. Objectives**

1. To research and choose video games that appear to be simple and easy to play so respondents can engage into the gameplay equally and without problems.
2. To obtain the resources in such way so that every player plays the same game in the same conditions as well as making the experiment fair and equal for every respondent.
3. To study and analyse how the Rasch model and RUMM 2030 process the data so the gameplays can be converted into adequate numerical data.
4. To study how each game's properties affect the suitability of the Rasch model to estimate their difficulty.
5. To evaluate the results obtained from the program and conclude whether the Rasch model is an appropriate and reliable measurement method for videogame difficulty.

#### 4. Project report layout

**Chapter 1** starts with a brief introduction to the purpose of the project, the potential of the method if it proves successful, and a main guideline to the objectives and aim.

**Chapter 2** provides a literature background which includes the primary theoretical concepts that help understanding the underlying blocks that build the project's theoretical foundations. It also provides background on the tools that will be used in the project such as the software, and what there is important to know about it.

**Chapter 3** will guide the reader through the methodology of the project, how each objective is being accomplished and how each step of the experiment is being carried out with enough transparency for readers to be able to make their own experiments basing themselves on this project.

**Chapter 4** presents the results after the data has been processed by the software. It explains the meaning of each result and gives some insight into the reliability of the method.

**Chapter 5** is the final chapter which summarises the achievements and how they cross over with the objectives, it discusses the overall method analysis's validity for the project, and finally concludes with a summary of results and an insight into the potential of the method given the results.

### Literature review

#### 2.1 Introduction

In order to understand the technical concepts behind videogame difficulty it is useful to study how gaming performance can be measured and which factors can affect performance. After all, as the objectives change in each games, so would the parameters to measure performance. Variables can't be extrapolated from one game to another.

There are past papers and studies that have attempted to measure videogame difficulty previously [5] [6], which can provide some insight into the challenge it raises, even if the method to estimate the difficulty is different.

The Rasch model and its software RUMM 2030, require deep understanding as it is important to understand how each graph display, graph trend or numerical value relates to the physical values of the player's performances. From the results the validity, reliability and responsiveness are evaluated which are three of the main conclusive parameters. Software and Rasch analysis knowledge combined are the key to finding the possible source of errors and to assessing possible solutions if it is a human error that is causing the problems.

There are many projects and papers about using the Rasch model in different applications such as medical, psychological or social experiments [2] [3] [7]. They can help build a solid background on how the data can be processed and how the results from the model can be related to the conclusions on the method.

Statistical parameters are of significant importance to the results of the model. Once the data is processed, understanding the distributions, scores, confidence intervals, variances, standard deviations and residuals amongst others is an essential to spot extreme respondents, anomalies or patterns that affect the reliability of the results.

Different tests are performed by the software (RUMM 2030) to study different aspects of the resulting data such as power of fit, unidimensional or invariance tests, which are strongly related to statistics.

## 2.2 Literature review

### 2.2.1 Videogame difficulty – References 5 and 6

*The following information is based on the cited references, which are studies on video game difficulty. Their literature reviews on videogame difficulty have been deemed useful to the purpose of research into the matter.*

As the Rasch model requires quantitative data (explained later in Rasch model literature), videogame performance has to be measured, and the performance should be a function of the difficulty of the game.

The studies “*Measuring Difficulty in Platform Video Games, F.J. Mourato, M. P. dos Santos, 2010*” and “*Measuring the level of difficulty in single player video games, 2011*” reflect on the nature of the videogames difficulty, and each develop a different method to compute the difficulty.

Difficulty doesn't have a clear definition as a measurable parameter, and its adjustment is based on subjectivisms. Difficulty is a primary component of any gameplay as the player exerts a certain effort into influencing the outcome of the game, and the effort is dependent on the difficulty of the game. When it is too challenging players become frustrated, and if it's too easy players become bored. So, finding a point in between will make the player stay in a *flow* status, hence a good game design should scale the game difficulty to maximise the player's enjoyment. [5]

Each videogame can have different difficulty systems such as variable artificial intelligence, multiple stages with different difficulty or dynamic difficulty adjustments so how the performance is measured would be different in each system.

Some videogames have onscreen scores that can be used to monitor performance, other performance measurements can be how quick the players complete the stages, how many lives or attempts were needed to complete a stage or his/her consistency in terms of his average score after a certain amount of time playing. These are all parameters that reflect the challenge, effort and ability of the player.

### 2.2.2 Rasch model and possible application to videogames – References 4, 7, 8 and 9

*The main paper used to research this section is Reference 4 which is a study that updates the uses of the Rasch model, and provides a useful guide to the information of interest in any Rasch model analysis paper. The rest of references are complementary and used as evidence and to clear conceptual doubts.*

The Rasch model is an item response theory procedure that considers that a latent variable explains the response to the items. In this project the latent variable is the user's ability and the response level is the performance in the game.

The model is mainly used for dichotomous data (two possibilities of response, right or wrong, for each item), due to its spread use amongst questionnaires and tests. A polytomous form of the model can be generalised from the dichotomous model. The polytomous model will be used in this project, due to performance being classified by thresholds between more than two levels. The different levels in each item will be called *subcategories* in this project.

The items are parameters that can measure the performance in the game, their units can be time, score or any other measurable (numerical) unit. They should be independent, meaning the items shouldn't be relative with each other. The different subcategories in items are assumed to be a function of the latent variable, ability. For most items there will be 5 subcategories, higher ability players will be at higher subcategories unless the item has a reversed scaling. The number of items in the games will be 7 which is as many as possible whilst attempting to maintain the selected items independent.

The Rasch model calculates the probability of a specific response depending on the person and items being evaluated. The function that computes the probability is given by:

$$\Pr(P_{ni} = 1) = \frac{e^{An-Di}}{1 + e^{An-Di}}$$

**Where:**

$Pr(P_{ni} = 1)$  Is the probability of person  $n$  obtaining a performance high enough to fit into a certain subcategory for item  $i$ .

$A_n$  is the ability of the person in logit units, it is computed from the overall performance of the person based on the subcategory placement in the items.

$D_i$  is the difficulty of the item in logit units, it is computed from the distribution of persons along the subcategories of each item.

In a polytomous model this can be visually represented with a common Rasch model tool such as Item Category Curve (ICC), the following ICC is from one of the games in the experiment, it represents 1 item:

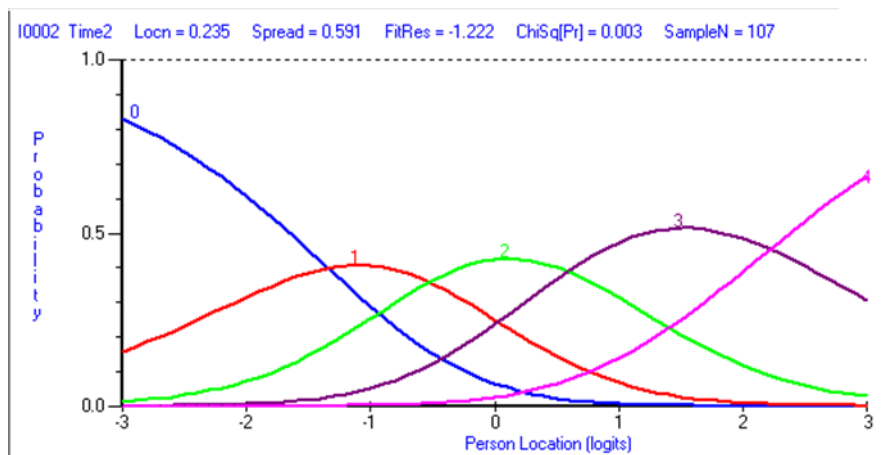


Figure 1, ICC for a polytomous model

Figure 1 is a visual representation of the probabilities each ability level stands with each subcategory. Lower subcategories are easier to attain, they are lower performances than higher subcategories within the item. When the person ability is low there is a much higher probability of attaining lower subcategories (0, 1 and slightly 2). At higher ability location there's a higher probability of attaining the higher performance subcategories (4, 3 and slightly 2). Each subcategory has a range of ability values for which they are the most likely subcategory the given respondent will fit in.

*E.g. for person ability -3 to -1.4 its subcategory 0, -1.4 to -0.6 its 1, -0.6 to 0.6 its 2, 0.6 to 2.3 its 3 and for 2.3 to 3 its 4.*

When an item has a subcategory that doesn't have a dominant range the item isn't functioning well, and limits the range of software tools that can be used to represent the item.



As long as the data obtained is discrete data and the trait can be qualified as quantitative, the model can be used, so there are many applications it can be used for. In this project gaming performance can be strongly qualified as discrete data. Ability is also a quantitative attribute.

The Rasch model is an apt method if the response depends solely on the latent trait; it has to be unidimensional. The level of performance should only depend on ability in the case of videogames which is the main assumption in the project, but after the analysis such assumption will be discussed.

This complicates the use of the method for general tasks instead of videogames, as unlike videogames, general task performance may not be a factor of just ability. Knowledge, age, or experience could affect in different ways the performance of people carrying out the tasks. This would reduce the range of tasks for which the Rasch model can be used. Other tasks would be required to be carried out by people with common properties such as same job, same knowledge about the task and same previous experience in the field as well as any other variables which could potentially affect the task performance (would require some study/research into the specific electronic task).

#### **2.2.2.1 Guttman scaling – Reference 14 and 4**

*Reference 14 has been used to extract the theoretical information and to understand the concepts for Guttman scaling. Its relation with the model is extracted from Reference 4.*

The reason behind the unidimensional requirement for the Rasch model, comes from the response patterns which are a probabilistic form of Guttman scaling.

The Guttman scaling orders the persons on a single continuum with a cumulative probabilistic function. In other words, if an answer to an item of high order is yes, it would be assumed that for every item with a lower order the answer would be yes too. This is better explained with a simple example:

Statement 1 (order 1): Do you know what tennis is?

Statement 2 (order 2): Do you enjoy tennis?

Statement 3 (order 3): Do you practice tennis?

The set of questions above can be considered unidimensional as someone who practices tennis is very likely to enjoy tennis, and to know what tennis is.

Guttman scaling assumes that positive (yes) answers to higher orders, such as order 3, which is statement 3: “Do you practice tennis?” assumes the answers to the lower order statements is positive too. A different scenario, if answer to statement 3 is negative (no), then either statement 1 and 2 are negative, or both are positive due to the Guttman scaling on a single continuum. This is backed up with the assumption of unidimensionality, if it were multidimensional, lower order statements could still be negated even if higher orders are affirmed.

In the project Guttman scaling would mean that if a respondent scores a good performance in a harder item, he/she is more likely to obtain good performances in lower difficulty tasks. This is a basis for the reliability of the model. If the performance is dependent on more factors other than ability (which would negate unidimensionality), the Guttman scaling will make the data unfit the Rasch model.

As a consequence a unidimensionality test has to be done and presented to evaluate the reliability of the results.

### **2.2.3 Statistical parameters – References 10, 11 and 12**

*Different references have been used to conceptualise about the statistical parameters and how they relate to the model. Especially with the Reference 11, which is provided by the University of Leeds and guides the user through the parameters of the program.*

The numerical results are mostly in the form of technical words from either statistics or Rasch model software. The parameters applied to Rasch model and RUMM 2030 to be understood are the following:

- Mean is the average value the program computes across all the components of the class, so “mean item location” is “average item difficulty” (RUMM 2030 set the item mean to 0 as an initial condition), and “mean person location” is the “mean person ability”. Mean for normal distribution is 0.
- Standard Deviation can be referred to as a measure of how disperse the data is, if the value is low the data is clustered close to the mean, standard deviation for normal distribution is 1.

- Residuals are the difference between the observed values and the estimated values from the sample. Each item and each person from the analysis have a residual value, the overall residual for item and person is the average residual for each type.
- Variance & chi-squared test are related, as the chi-squared test is based on the sample variance. The test will compare the observed data with the expected value and compute the chi-squared probability. It would be desirable to obtain chi-square probabilities higher than 0.05 or if multiple testing, 0.01 [11]. This would be regarded as the items are working as expected at grouped levels of mental state [11].
- Person-Separation index is an indicative of the power of the set of data to discriminate amongst the respondents. A higher index means a higher % of variance is not due to error, which allows the data to discriminate between groups of respondents as follows:

<b>Person Separation index</b>	<b>% variance not due to error/ due to error</b>	<b>No. of distinct groups</b>
0	0/100	1
0.5	50/50	1
0.7	70/30	2
0.8	80/20	3
0.9	90/10	4
0.94	94/6	5

*Table 1. Different Person index derivations. Reference 10*

This is also an indicator of how reliable the fit statistics are, the higher the index, the more reliable.

- Independent t-student is a test that determines how different 2 sets of data are from each other, it will be used to determine if the results from the videogame performances are multidimensional by comparing the difference between 2 subsets from the game (each subset with a different amount of items), which will be one of

the main factors of discussion when considering the reliability of the Rasch model for videogame difficulty estimation.

## 2.2.4 Games chosen – References 5 and 6

*The cited references have been used to take in account the factors that affect difficulty, simplicity and the unidimensionality, but the actual games are chosen based on an informed decision taking in account all the factors.*

There is a countless amount of videogames to be found online which can be downloaded and used in the experiment. Nevertheless there are properties these games have to acquire that make the experiment reliable, given the study conditions.

Games should be easy to play (not to complete), as explained in the game design previously. This project tries to measure performance, and if performance is dependent of more factors than ability alone, the results can start to lose reliability as the game stops being unidimensional. If the games have few control buttons and aren't visually demanding every player regardless of ability parts from the same starting point. Their performances are more strictly related to ability and not affected by a lack of understanding of the game due to its complexity.



Figure 2. Galaxian



Figure 3. Modern 3D game

The games should also be hardly known, as it would be desirable that no players have previous experience in playing the game as they would be in advantage

regarding the rest of players. Otherwise they would have an advantage and better performances non-related to ability alone but to knowledge, which could influence the results.

As a result, the amount of games that meet such properties can be reduced to a lower number: 2D arcade games from 1980s (Figure 2) meet perfectly the properties compared to modern 3D games (Figure 3) which are usually complex and somehow known to at least a minority of the respondents.

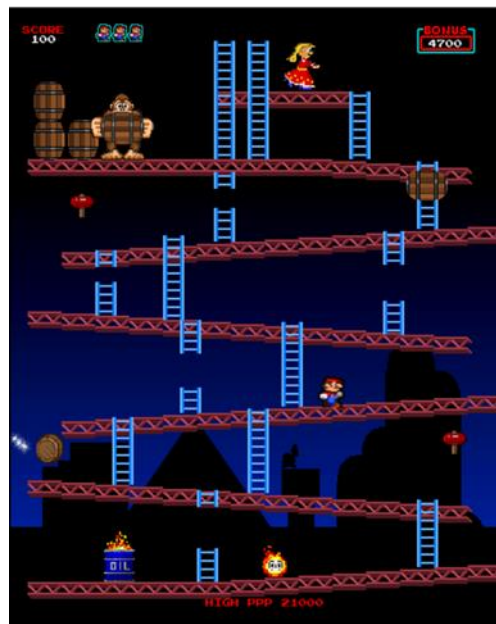
The selection of games for this project are called Donkey Kong and Galaxian.

## 1. Donkey Kong

---

Donkey Kong is a game from the 1980s which consists in the following:

The player controls Mario figure, they can use arrows and space keys to move and jump. Each stage has the same purpose, climbing from the floor to the highest point of the stage and saving the princess by walking into her which will take the player to the next stage. There is a maximum of 4 stages.



*Figure 4. Donkey Kong*

Mario should climb ladders to get to higher levels but he can be killed in many ways. In the first stage barrels roll down the different floors without a fixed route, so their paths are supposedly unpredictable to the player. There's also fireballs enemies that roam around and will also kill Mario if he is touched by them.

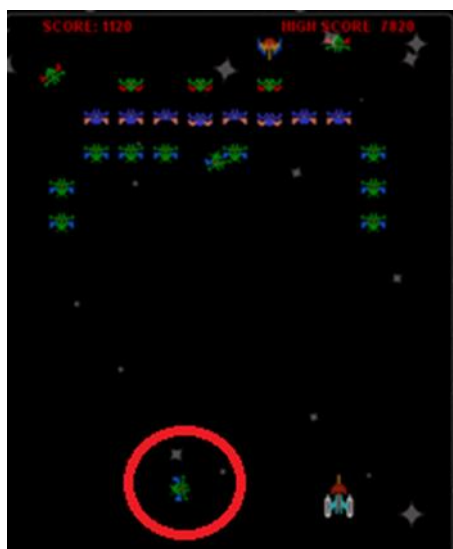
Players can choose how to deal with the enemies, they can avoid them by taking different routes, they can jump them, or they can defeat them by picking a hammer that grants Mario power to defeat them for a limited amount of time.

Different stages have different enemies and different structures but the player's choices are always the same three. If Mario dies, he starts again at stage 1, if he advances to stage 2 and is killed he will restart at stage 1 too.

## 2. Galaxian

---

Galaxian is also a 1980s game, but Galaxian is even simpler to play. Players take the form of a space ship encountering an alien fleet which attacks him. The player fights back and tries to shoot down as many aliens as he can.



*Figure 5. Galaxian screenshot.*

The controls are simple, arrows to move, space to shoot. When the whole fleet is destroyed, it reappears, but the scores and lives aren't reset. The alien fleet frequently break their formation to attack from close range to the player, which makes their movements unpredictable and challenging to the player, as when they are still it is fairly easy to complete the game.

Players have 3 lives to sum up the best score. Score is obtained by defeating alien spaceships. If they are hit while they are out of formation to attack you (circled enemy in figure 5), they will score double points.

### **2.2.5 Human factors**

In this project the specific ability to perform better at the videogames can be comprised of mental agility, reactivity time or other traits that are triggered when the player actively tries to complete the game.

Subjects that find the game too hard, too easy or who got frustrated after the first tries could corrupt the results. It is difficult to verify from watching the gameplays but as an external spectator, it wasn't hard to appreciate that a minority of players lost interest if by example, their first 3 tries at the game were poor. This relates to the previously mentioned game design parameter of difficulty, as it can make players get frustrated if the game seems too hard to them.

Some other subjects would also corrupt the results when they would kill themselves mid-game to reset the number of lives and have a better go with a fresher start.

## **2.3 Literature Summary**

The Rasch model will be considered theoretically to be a good method to evaluate the difficulty of videogames as these have been cautiously chosen to be as unidimensional as possible considering the range of videogames available. Unidimensionality is a requirement for the probabilistic form of the Rasch model, Guttman scaling, to fit the data to the model.

A range of post analysis tools such as tables, graphs, and statistical parameters will be used to represent the concept of difficulty in video games. The results will be discussed and finally, it will be concluded how well the difficulty of the videogames is being represented as well as the degree of precision of the estimated difficulty model.

The project will involve carrying out the experiment with 2 different games, to obtain a wider range of results and parameters that can be compared between the games, and finally evaluate and conclude the reliability of the method with videogames. If it is deemed reliable, the possibility to use the method for other computer tasks can be more thoroughly studied.

### Methodology

#### 3.1 Introduction

This is not the first project that attempts to develop a method to compute the difficulty of videogames, and so, it is an iterative process as newer projects try to learn from previous projects and innovate the method or change factors to a certain extent. Previous projects have kept innovating by changing the computer video games, each producing a different conclusion, which in the long term didn't provide any improvements on the reliability of the method.

This project attempts to compare 2 games and analyse the reasons which make the video games reliable. At the expense of a higher recording, analysing and processing time, it could shed some light on how and why different games are differently suited to the Rasch model.

The following chapter explains how the experiment was carried out and in which conditions, how the gameplay was analysed to obtain discrete data, and how this data was and processed by RUMM 2030.

#### 3.2 Recordings & gameplay

Every player received the same trait and game conditions. The recording software didn't affect the gameplay in any way, every player knew their gameplay was being recorded and that they could stop playing if they wished so.

Beforehand respondents are explained the controls of the game, and that they have freedom to ask any doubts they may have as they play.

Respondents played for 10-15 minutes each game. Later on the recording could be trimmed for every video so every respondent has the same gameplay footage length to be analysed.



### 3.3 Gameplay performance conversion

As explained previously, the inputs to the RUMM 2030 and to the Rasch model are a series of items and their subcategories.

These items and subcategories are extracted from the gameplay footage from each respondent. The items represent a measurement of performance for a certain aspect of the game. The following table summarises the items and subcategories for game 1:

#### Donkey Kong:

Donkey Kong	Subcategories				
	0	1	2	3	4
Item 1	0-60	61-120	121-180	181-240	>240
Item 2	15-30	31-46	47-62	63-78	>78
Item 3	0-20	21-40	41-60	61-80	81-100
Item 4	0-800	801-1400	1401-2000	2001-2600	>2600
Item 5	0-2	3-5	6-8	9-11	>11
Item 6	0-2	3-5	6-8	9-11	>11
Item 7	0-2	3-4	5-6	7-8	>9

*Table 2. Donkey Kong Item and Categorisation*

#### Where:

Item 1 is time in seconds it takes the player to complete for the first time the first stage of the game, if players have high ability, they should complete the stage earlier as their learning curve is steeper. Lower times indicate higher ability, so this item has reverse scaling.

Item 2 is the maximum time in seconds the player can stay alive for. One of the parameters that determines survival is ability to not be killed by the incoming enemies. The higher the survival time the higher the ability, so normal scaling.

Item 3 is the maximum height, which is a simile to how far the players advance in the game in the given time. Each stage is 25m and the score is the max height achieved, e.g. 2 first stages completed and part of stage 3 would be somewhere between 50-70m depending on how much they advance before being eliminated in the third stage. Normal scaling.

Item 4 is the total score which is an overall performance score as it is a function of enemies defeated, jumps, items collected and stages completed. Normal scaling.

Item 5 is the total number of enemies the player defeats, as explained previously in the literature. Normal scaling.

Item 6 is the total number of times the player died during the gameplay time, which measures the consistency with which the player is performing if they die many times they are either inconsistent or consistently low which would mean a lower player ability. The lower the number of deaths, the higher the ability, so this item has reverse scaling.

Item 7 is the total number of successful jumps the player does, as jumping the enemies requires a degree of ability. Normal scaling.

**Galaxian:**

Galaxian	Subcategories				
	0	1	2	3	4
Item 1	0-50	50-100	100-150	150-200	>200
Item 2	0-2	3-4	5-6	7-8	>9
Item 3	501-1300	1301-2100	2101-2900	2901-3700	3701-4500
Item 4	0-400	401-800	801-1200	1201-1600	>1600
Item 5	1-3	4-5	6-7	8-9	>9
Item 6	0-100	101-200	201-300	301-400	>400
Item 7	30-60	61-90	91-120	121-150	>150

*Table 3 Galaxian Item and Categorisation*

**Where:**

Item 1 is the maximum survival time in seconds the player managed to stay alive for as there is a constant input of attacks which requires skill to avoid and thus stay alive. The maximum time of survival corresponded to the maximum time within the gameplay time. Normal scaling.

Item 2 is the total number of times the player died during the gameplay time, as explained in the Donkey Kong game. Reversed scaling.

Item 3 is the maximum score which is an overall performance score as it is a function of enemies defeated, staying alive, being fast and stages completed. Normal scaling.

Item 4 is the minimum score which is a measure of the worst game of the player and a measure of his/her consistency. Normal scaling.

Item 5 is the total number of enemy bosses the player defeats. Bosses are located at the back of the fleet (can be seen in figure at the back of the fleet) and players are aware they are worth more points as well as a threat (they shoot at the player's spaceship) so higher performance players should get rid of them as quickly as possible, increasing the number of total bosses defeated. Normal scaling.


Item 6 is the time to complete stage, which is the time it takes for the players to complete the first stage of defeating the whole fleet starting from when the recording started. In the case the player wouldn't accomplish this, their highest score would be used as that time, but it would be penalised as a function of the number of remaining aliens standing from the full fleet. Better players should learn quicker the game mechanics and acquire lower times in this item. Reverse scaling.

Item 7 is the time to clear the first stage/fleet. Not to be confused with item 6, it is the (time since the game started when total fleet is destroyed – time at the start of the game where fleet was totally destroyed), same penalisation was applied as in item 6 for players who didn't manage to destroy the full fleet during their gameplays. Reverse scaling.

### 3.4 Data processing for RUMM 2030

Once all the data has been obtained in its respective item and categories column the Excel file is modified to enable RUMM 2030 reading the data:

Person ID	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
1	1	0	0	0	0	3	1
2	1	1	1	1	1	2	1
3	1	2	1	1	0	2	0
4	3	1	2	1	0	2	2
5	4	1	2	2	1	2	1
6	1	3	2	3	2	1	1
7	1	2	1	1	1	2	1
8	1	3	1	2	3	1	0
9	2	1	0	0	1	3	1
10	3	1	0	2	1	3	3
11	2	3	2	2	2	2	1
12	4	1	2	2	0	3	2
13	4	2	1	1	1	2	0
14	2	2	2	1	0	2	1
15	3	0	0	1	3	1	1
16	1	2	1	2	3	1	2



1	1	0	0	0	3	1	
2	1	1	1	1	2	1	
3	1	2	1	1	0	2	0
4	3	1	2	1	0	2	2
5	4	1	2	2	1	2	1
6	1	3	2	3	2	1	1
7	1	2	1	1	1	2	1
8	1	3	1	2	3	1	0
9	2	1	0	0	1	3	1
10	3	1	0	2	1	3	3
11	2	3	2	2	2	2	1
12	4	1	2	2	0	3	2
13	4	2	1	1	1	2	0
14	2	2	2	1	0	2	1
15	3	0	0	1	3	1	1
16	1	2	1	2	3	1	2

Figure 6. Data conversion for RUMM 2030

The first column corresponding to person ID has been set to a width of 3 characters as the number of digits goes up to 3, for the rest of columns it is set to 1 for the same reasoning. The column titles have also been deleted. This file is then saved with a .PRN extension.

This file is then loaded into RUMM 2030, where some basic instructions concerning how the data should be read are entered and then the next window will ask the user to edit or modify any of the items properties such as their names, and more importantly if they are reversed.

Reversed items are those which higher subcategories don't mean higher ability, but lower. An example could be, Item 4 which is total score will classify high scores, and thus high ability players to a higher subcategory, but on the other hand, item 6 which represents the number of deaths classifies the higher number of deaths into higher subcategories. Higher number of deaths signifies a lower player ability so the item punctuation should be reversed.

Once this has been done the analysis is ready to be done.

## Chapter 4

### Results

The following results are presented in order of parameters with both games:

#### 4.1 Summary statistics

Game 1 – Donkey Kong

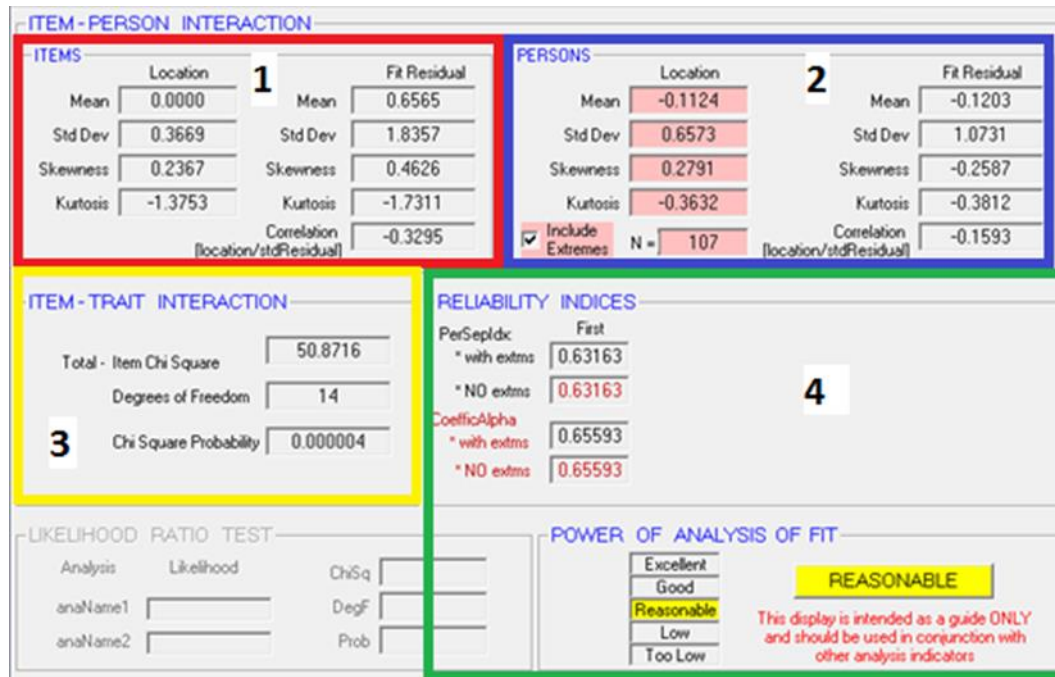


Figure 7. Summary statistics for Donkey Kong.

Items – Persons (RED table numbered 1 and BLUE table numbered 2)

In table 1 the left set of values display the statistics for the items, how the items in average interacted across the persons with different ability levels. RUMM 2030 locates the mean at 0 logits for items always.

In table 2 the left set of values display the statistics of how the persons interacted across the different items. Location of persons is -0.1124 logits with a standard deviation of 0.6573, which is still lower than 0 so response group was of slightly lower ability level than difficulty level asserted by items and their categories, range of logits is -4 to 4.

From the same rectangles and tables the fit residual statistics are obtained, these are used to describe the data fit to the Rasch model as they assess the difference between how each person/item should have fit into the item's categories/person and how they actually fitted.

Fit residuals are approximated to a standardized normal distribution which always has mean 0 and standard deviation (SD) of 1. Items fit residual mean and SD are (0.6565, 1.88357) which doesn't approximate strongly to normal distribution. Persons fit residual are (-0.1203, 1.0731) which is very similar to normal distribution. This shows that for game 1 the people responses residual appear to approximate the normal distribution, but the item residual doesn't, so items may not be the best fit to the model.

The power of analysis of fit (marked 4 in the green rectangle) represents an indicative of how much we can rely on our fit data, as explained previously a value between 0.6 and 0.7 means there is a 30% to 20% variance due to an error in the fit. In game 1, this value is labelled as "reasonable" but it's not desirable as the index should be 0.7 to discriminate in the data 2 different groups.

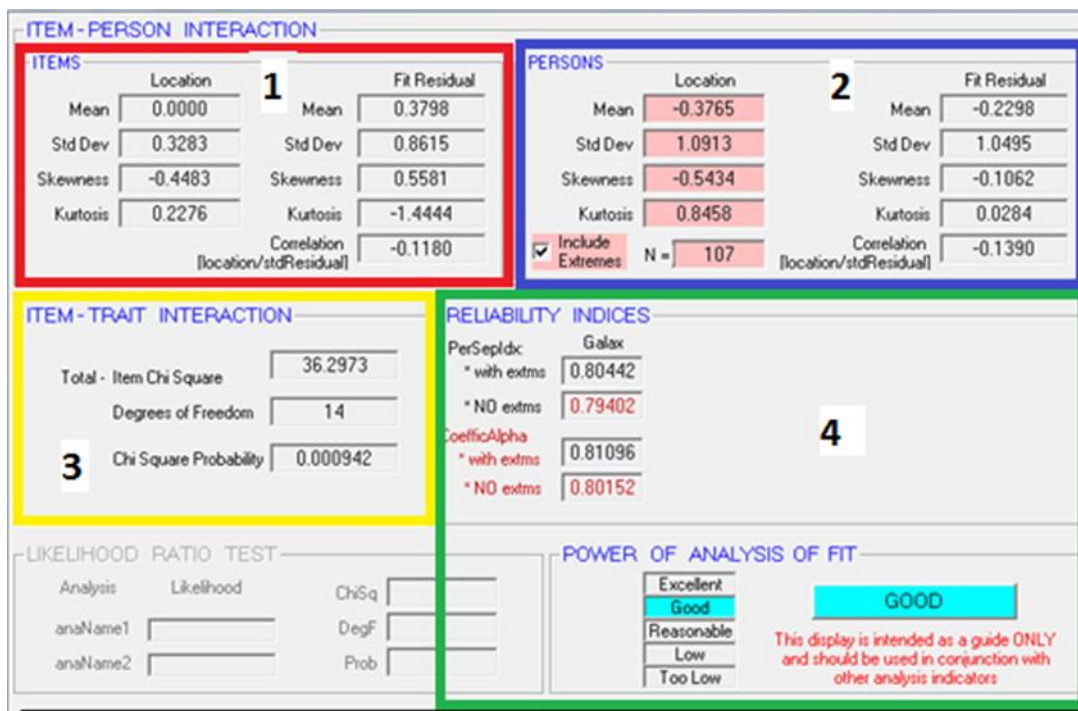


Figure 8. Summary statistics for Galaxian

Items – Persons (RED table numbered 1 and BLUE table numbered 2)

Location of persons is -0.3765 logits with a standard deviation of 1.0913, which is still lower than 0 so response group was of lower ability level than difficulty level asserted by items and their categories, range of logits is -3 to 3.

Items fit residual mean and SD are (0.3789, 0.8615) which has a significant degree of normal distribution. Persons fit residual are (-0.2298, 1.0495) which is very similar to normal distribution. This shows that for game 2 the people responses and the item (to a certain degree) residual appear to approximate the normal distribution so they will fit better the model.

For game 2 the person separation index is 0.8 so there is a 20% variance due to error. This value is labelled as “good”, which is an indicative of the power to discriminate different classes of respondents based on response. Higher than 0.8 person index means the data can be discriminated in 3 different groups, which means a better reliability than the minimum of 2.

## 4.2 Individual person fit

Individual person fit is used to find detailed stats about each respondent, as well as identifying any misfits, which are considered respondents that don't fall on a fit residual range of +- 2.5, which in this case in game 1 there is 1 person, with ID 35 and fit residual -2.758. This person is believed to not follow the same behaviour than the natural behaviour (as the rest of respondents).

INDIVIDUAL PERSON-FIT for Analysis Name FIRST - Item-Person ResFit [Descend Order]: PersEstm Weighted Maximum Likelihood method

recID	TotExp Sc	MaxSc	Items	Extm	Location	SE	FitResid	DegFree	Data Pts	PersID
46	10	28	7		-0.575	0.385	-1.045	5.7	7	46
65	18	28	7		0.518	0.397	-1.069	5.7	7	65
52	10	28	7		-0.575	0.385	-1.234	5.7	7	52
17	15	28	7		0.094	0.374	-1.266	5.7	7	17
45	13	28	7		-0.170	0.371	-1.270	5.7	7	45
92	15	28	7		0.094	0.374	-1.538	5.7	7	92
91	15	28	7		0.094	0.374	-1.620	5.7	7	91
7	11	28	7		-0.436	0.378	-1.646	5.7	7	7
2	10	28	7		-0.575	0.385	-1.648	5.7	7	2
98	13	28	7		-0.170	0.371	-1.691	5.7	7	98
16	16	28	7		0.229	0.379	-1.713	5.7	7	16
76	22	28	7		1.247	0.473	-1.775	5.7	7	76
34	16	28	7		0.229	0.379	-1.808	5.7	7	34
51	12	28	7		-0.302	0.373	-2.053	5.7	7	51
40	15	28	7		0.094	0.374	-2.138	5.7	7	40
50	13	28	7		-0.170	0.371	-2.175	5.7	7	50
83	17	28	7		0.370	0.387	-2.329	5.7	7	83
43	15	28	7		0.094	0.374	-2.377	5.7	7	43
54	22	28	7		1.247	0.473	-2.384	5.7	7	54
35	15	28	7		0.094	0.374	-2.758	5.7	7	35

Mean: -0.112, SE: 0.657, FitResid: -0.120, SE: 0.432, DegFree: 1.073, Data Pts: 1073

Selection: 1, Exclude Extreme Persons:  Extm Pers Criterion: 0.220, Separation Index: 0.63163, Coefficient Alpha: 0.65593, Mean Error Variance: 0.159, Est. True Variance: 0.273

Sort Persons by: Fit Resid Order [Desc], File Text Format: Fixed, Tab Delimit

Figure 9. Game 1 Individual Person Fit

For game 2 there are 2 people, and one person who is an extreme value (obtained lowest score in every item and a logit value of -4.127). These player's unexpected performances could be a cause of the human factors considered previously in the literature.

recID	Tot/Exp Sc	MaxSc	Items	Extm	Location	SE	FitResid	DegFree	Data Pts	PersID
30	16	27	7		0.425	0.427	-1.192	5.8	7	30
76	10	27	7		-0.658	0.438	-1.238	5.8	7	76
43	16	27	7		0.425	0.427	-1.302	5.8	7	43
103	8	27	7		-1.043	0.454	-1.310	5.8	7	103
53	17	27	7		0.603	0.431	-1.346	5.8	7	53
101	15	27	7		0.249	0.425	-1.403	5.8	7	101
60	11	27	7		-0.472	0.433	-1.407	5.8	7	60
12	11	27	7		-0.472	0.433	-1.407	5.8	7	12
40	15	27	7		0.249	0.425	-1.559	5.8	7	40
9	14	27	7		0.071	0.428	-1.806	5.8	7	9
64	22	27	7		1.630	0.523	-1.848	5.8	7	64
16	22	27	7		1.630	0.523	-1.848	5.8	7	16
28	11	27	7		-0.472	0.433	-1.903	5.8	7	28
22	8	27	7		-1.043	0.454	-2.091	5.8	7	22
27	8	27	7		-1.043	0.454	-2.091	5.8	7	27
34	21	27	7		1.384	0.490	-2.179	5.8	7	34
36	11	27	7		-0.472	0.433	-2.434	5.8	7	36
5	17	27	7		0.603	0.431	-2.628	5.8	7	5
72	13	27	7		-0.108	0.427	-3.109	5.8	7	72
3	0.190	27	7	extm	-4.127	1.275	...			3

Figure 10. Game 2 Individual Person Fit

### 4.3 Individual item fit

Tells how each individual item fits the Rasch model, again, we can locate items that aren't fitting to the model, due to:

- Significant chi sq. probabilities such as those above 5% (0.05 probability) may mean the items are not fitting the model. However, RUMM 2030 highlights the items where the chi square probability is below the Bonferroni adjustment which for 7 items is 0.001429 which is a much conservative probability.
- Fit residuals where the divergence between the estimated value and the actual value for every person and a given item, follow the standardized normal distribution. If they don't fall on the range of +/-2.5 fit residual they can be considered misfitting. These items may be classified as redundant as they don't provide any new information to the model. As a consequence it can affect the person fitting evaluation and person to model fitting.

For game 1 there are 2 items to appear to misfit into the model, both of them in terms of the fit residual, and one of them its probability is below Bonferroni adjustment. These 2 items are number of successful jumps, and enemies defeated.



As explained in the Donkey Kong game literature, players can choose how to advance through their enemies: defeating them, avoiding them, or jumping them.

This led to the model-items misfit as very good players who jumped enemies would have bad scores in the defeating enemies' item, or players who avoided enemies at all cost would have poor scores in both items, despite being higher ability players.

INDIVIDUAL ITEM-FIT for Analysis Name FIRST - Chi Square Probability Order

	Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob	F-stat	DF-1	DF-2	Prob
6	6	I0006	Poly	-0.433	0.115	0.577	87.86	1.618	2	0.445242	...	...	...	...
1	1	I0001	Poly	-0.333	0.098	0.353	87.86	4.173	2	0.124092	...	...	...	...
3	3	I0003	Poly	0.555	0.110	-0.717	87.86	4.611	2	0.099722	...	...	...	...
4	4	I0004	Poly	0.049	0.106	-0.753	87.86	5.989	2	0.050051	...	...	...	...
5	5	I0005	Poly	0.229	0.091	3.161	87.86	9.399	2	0.009100	...	...	...	...
2	2	I0002	Poly	0.235	0.109	-1.222	87.86	11.762	2	0.002793	...	...	...	...
7	7	I0007	Poly	-0.302	0.088	3.195	87.86	13.319	2	0.001283	...	...	...	...

Figure 11. Individual item fit game 1

The analysis was repeated removing items 5 and 7 which improved the power of analysis of fit, as the misfit led to decreasing the reliability of the data fitting into the model.

SUMMARY STATISTICS for Analysis Name GAME1

**ITEM - PERSON INTERACTION**

ITEMS				PERSONS				
	Location		Fit Residual		Location		Fit Residual	
Mean	0.0000	Mean	0.1423	Mean	-0.1609	Mean	-0.3036	
Std Dev	0.5360	Std Dev	2.0936	Std Dev	1.1745	Std Dev	1.0681	
Skewness	0.4176	Skewness	0.3789	Skewness	0.5766	Skewness	-0.2541	
Kurtosis	-1.0648	Kurtosis	-1.7647	Kurtosis	0.2115	Kurtosis	-0.4799	
	Correlation [location/stdResidual]		-0.8864	<input checked="" type="checkbox"/> Include Extremes	N = 107	Correlation [location/stdResidual]		-0.1001

**ITEM - TRAIT INTERACTION**

Total - Item Chi Square: 15.8528  
 Degrees of Freedom: 10  
 Chi Square Probability: 0.103926

**RELIABILITY INDICES**

PerSepIdx: Game1  
 \* with extms: 0.76413  
 \* NO extms: 0.76413  
 CoefficAlpha  
 \* with extms: 0.77814  
 \* NO extms: 0.77814

**LIKELIHOOD RATIO TEST**

Analysis Likelihood ChiSq  
 anaName1 \_\_\_\_\_ DegF \_\_\_\_\_  
 anaName2 \_\_\_\_\_ Prob \_\_\_\_\_

**POWER OF ANALYSIS OF FIT**

Excellent  
**Good**  
 Reasonable  
 Low  
 Too Low

This display is intended as a guide ONLY and should be used in conjunction with other analysis indicators

< Display Control      File Text Format      Fixed Tab Delimit      Save

Figure 12. Summary statistics of game 1 after item removal

This result from the item removal highlights the importance of item selection in the performance of the task being monitored. Items used to measure the performance of the players should be related to ability, if they aren't they can lead to model misfits. After the item removal the power of fit person index has increased above the minimum of 0.7. [11]

For game 2 there are no items appearing to misfit into the model.

INDIVIDUAL ITEM-FIT for Analysis Name GALAX - Serial Order														
	Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob	F-stat	DF-1	DF-2	Prob
1	1	I0001	Poly	0.321	0.124	0.008	87.14	3.663	2	0.160202	...	...	...	...
2	2	I0002	Poly	0.056	0.124	0.041	87.14	3.171	2	0.204828	...	...	...	...
3	3	I0003	Poly	-0.165	0.107	-0.608	87.14	8.699	2	0.012913	...	...	...	...
4	4	I0004	Poly	0.426	0.108	-0.174	87.14	7.807	2	0.020171	...	...	...	...
5	5	I0005	Poly	0.057	0.132	1.809	87.14	2.491	2	0.287733	...	...	...	...
6	6	I0006	Poly	-0.556	0.110	0.273	87.14	3.951	2	0.138659	...	...	...	...
7	7	I0007	Poly	-0.140	0.119	1.309	87.14	6.514	2	0.038500	...	...	...	...

Figure 13. Game 2 individual item fit

#### 4.4 Threshold maps

Threshold maps are a useful tool to visually represent the most likely category each ability level player would fit in. By drawing a vertical line at the desired logit unit (x axis) we can infer the most likely score at the different categories.

Threshold maps have been used as opposed to category probability curves as they condense information from every item into one bar graph.

It can also be used to perceive irregular item distribution as the spacing between subcategories is clustered at certain points.

Threshold map for game 1 (without item removal to explain irregularities):

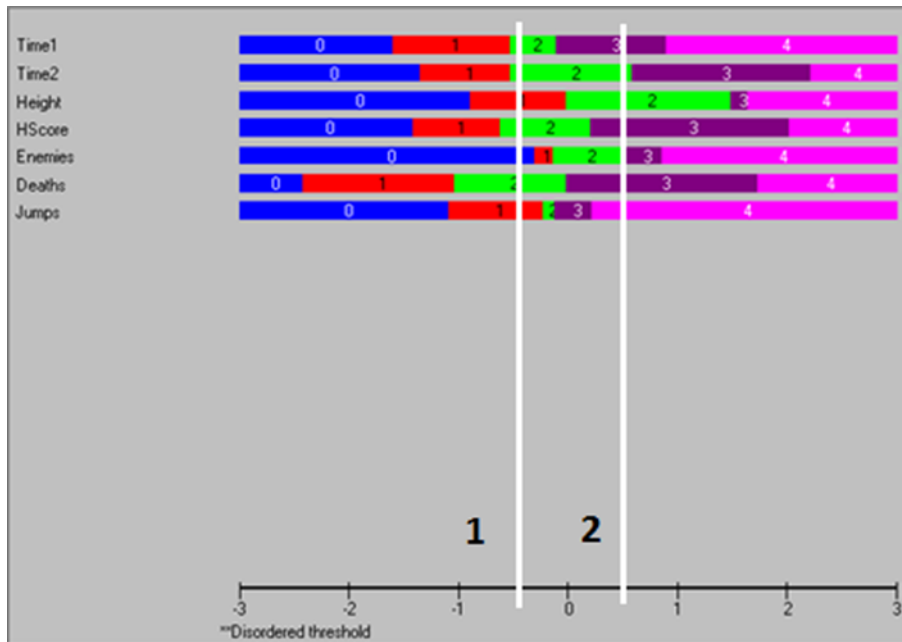


Figure 14. Game 1 threshold map without item removal

Lines 1 and 2 have been added to help perceive the irregularities in the items 5 and 7 (the same that showed misfit to the model).

In line 1 an ability level of (-0.45) out of [-3, +3] was picked. This threshold map describes that a player with that ability level will most likely be in the subcategory 1 or 2 for every item, except for item 5 where it would still be at 0, which results misfitting.

In line 2 an ability level of (+0.5) out of [-3, +3] was picked. This threshold map describes that a player with that ability level will most likely be in the subcategory 2 or 3 for every item, except for item 5 where it would already be at 4, which results misfitting.

Threshold map for game 2:

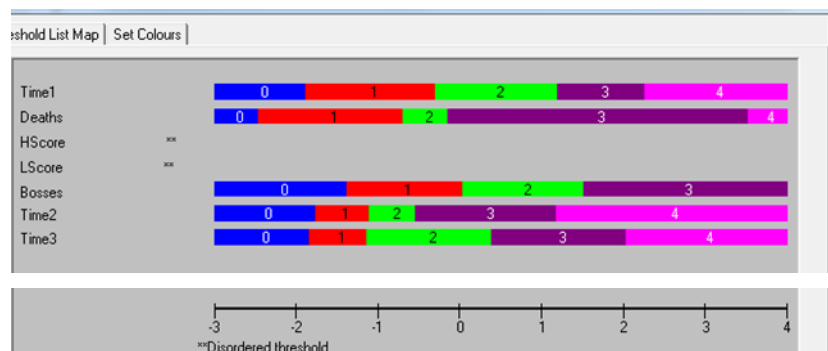


Figure 15. Game 2 threshold map

Most items seem to have regular spacing between them which proves fitting to the model. Items 3 and 4 (H.score and L.score) have a \*\*\* symbol because in both items there was a subcategory which was never the most likely for a given ability level player to fit in, (explained in 2.4 Rasch model literature).

#### 4.5 Test of unidimensionality

As mentioned previously a significant importance has been given to the factors affecting the game performance. Not only for the validity of the results of this project but to compare with possible future tasks which their execution could depend on numerous factors.

To do so RUMM 2030 can perform a test proposed by Smith EV. [9]. It examines the correlation between items and the first residual factor to define 2 subsets of items; one positive correlated items, and the other one the negatively correlated, which are then used to make separate person estimates.

These estimates are then tested with an independent t-test in which the residuals outside the boundaries -1.96 to +1.96 shouldn't exceed 5%. [9]

Summary Table of t-test analyses for this Subtest pair

Test	Subset Pair	No. < 5%	No. < 1%	PerC < 5%	PerC < 1%	Total
1	pos; neg	10	4	9.35%	3.74%	107

Figure 16 Donkey Kong t-test analysis

Summary Table of t-test analyses for this Subtest pair

Test	Subset Pair	No. < 5%	No. < 1%	PerC < 5%	PerC < 1%	Total
1	neg; pos	8	5	7.48%	4.67%	107

Figure 17 Galaxian t-test analysis

Donkey Kong game has a 9.35% of person estimates outside the range [-1.96, +1.96] which means the game isn't strongly unidimensional as the % shouldn't be greater than 5%.

Galaxian game has a 7.48% of person estimates outside the range [-1.96, +1.96] which means the game isn't totally unidimensional as the % shouldn't be greater than 5%, but it's a better result than Donkey Kong and much closer to 5%.

These results for the t-student test indicate a higher unidimensionality for game 2. If the gameplay is studied it would make sense. Donkey Kong was a game with a clear objective but with different ways of achieving it (different routes or enemies) which introduces factors besides ability such as tactical thinking or finding strategies.

Game 2 on the other side, had a clear objective and a clear route to do so, there were no other possibilities to complete the game than to have high ability. Tactics and strategies like in Donkey Kong wouldn't help significantly the player to achieve the goal.

Human factors such as boredom, frustration or non-willingness to aim for the game goal could influence the correlation values. In an ideal situation where the respondents were morally responsible (as expected in a realistic scenario) there would be no data distortion due to such human factors. Consequently unidimensionality would be higher, so in this case the % of values outside the accepted boundaries may be slightly lower than their real values.

#### **4.6 Summary of results for game 1 and game 2**

The Rasch model is being used to evaluate the difficulty of the videogames, and the model requires to evaluate in-game measured parameters and to classify the scorings into discrete polytomous data.

This data is processed into fitting the model, and RUMM 2030 will inform of how well each item is fitting into the created model, so choosing which items and quantifying their subcategories is an important factor.

If the task being measured isn't unidimensional, using Smith's proposed method should assess the multidimensionality of the games.

In Donkey Kong there is a clearer case of multidimensionality which could be caused by the nature of the game as explained in its correspondent section. This means that the results of the analysis can't be directly extrapolated to estimate the difficulty of the game, but it can be used as an auxiliary guide, due to:

- Person separation index being  $>0.7$  after removing misfitting items indicates the separation of respondents into at least 2 different groups meaning there are significant trends between ability and difficulty.
- Chi square after item removal is higher than 5% so there isn't a significant deviation between expected and actual values.
- Item-persons interaction for persons does show a similar behaviour to normal distribution so the interaction for persons is being well evaluated/modelled with the Rasch analysis.
- Items are mostly fitting into the Rasch model and category curves are sensible and logical. The threshold map also displays proportional bars between the items.

In Galaxian there is a lower degree of multidimensionality, but still significant. This means that the results of the analysis can't be used as a direct indicator but an auxiliary guide due to:

- Person separation index being  $>0.8$  indicates the discrimination amongst respondents into at least 3 different groups meaning there are meaningful trends between ability and difficulty.
- Chi square after item removal is lower than 5% so there could be some deviation between expected and actual values.
- Item-persons interaction for persons does show a similar behaviour to normal distribution so the interaction for persons is being well evaluated/modelled.
- Items are mostly fitting into the Rasch model and category curves are sensible and logical. The threshold map also displays proportional bars between the items, with the exception of the items that didn't satisfy the criteria to be displayed.

### Conclusion

#### 5.1 Achievements

The experiment set-up met the proposed objectives: Selecting two videogames which respondents weren't familiarised with but at the same time were simple to play. No respondent had any issues understanding the game objectives, mechanics and controls, which led to an equally fair performance for every player. Every player played the same game versions, in the same laptop, and for the same amount of time.

The data was processed into RUMM 2030 and the fit between the data and the Rasch model was evaluated with both games. The tools provided by the software allowed to make technical comparisons both qualitatively with maps or graphs, and quantitatively with the numerous statistical parameters between the games. Links were made between the meaning of the results and the game mechanics that helped understand better how to measure more efficiently the performances of the respondents.

Reliability of the study was assessed obtaining different reliabilities for each game, providing significant arguments concerning the use of the method for videogames.

#### 5.2 Discussion

Game 2, Galaxian, proved to be a better game than game 1, Donkey Kong, in terms of the reliability of using the Rasch model measurement analysis to determine its difficulty.

Difficulty is a concept that for some fields and situations, can be represented as a number. However, for a different situation such as videogames, it can't be summarised into a single number, but into a series of graphs, maps and tables which can help understand the difficulty level of the game, but not to quantify it. In this project the output was a performance estimator based on probability and ability.

There is a source of uncertainty as there are certain criteria the model has which the videogames may not strictly be in accordance with. After all, Rasch model is generally used for questionnaires and psychological responses from patients, where item categorisation is characterised by different physical values for each possible test

answer. Not by an ordered quantitative parameter of performance like in the videogames. An ability trait may not be treated like a psychological trait.

The set of results from this specific project should attempt to answer the question **“how hard is this game?”** It does so by presenting the graphs, maps and charts that illustrate how ability relates with probability of performance, and at the same time how reliable they are.

Carrying out the same experiment with 2 different games provided 2 different sets of results, each with a different reliability result. The analysis on game 2 provided a better reliability and thus a better estimate on the difficulty of the game. The causes could be: a better fit to the model, an item and category threshold selection which suits better the performance measurements or a higher unidimensional character. These were the three main differences between the game results.

Nevertheless, both games failed to meet the unidimensionality test, meaning the performances were influenced by more than one trait. Poor Item selection and categorisation, as well as response dependency could affect the unidimensionality test.

Discussing if the answer to the question **“how hard is this game?”** Is adequate would mean analysing if the Rasch model as a whole is a realistic approach to analysing videogame difficulty. As analysed in the literature and due to the range of appliances and “manoeuvrability” of the model it has been accepted to be so, but after the results does it still seem as the best approach?

Item selection and the item categorisation thresholds should be specified with more technicality as the analysis and the results would greatly vary as a function of the item selection and categorisation thresholds.

The idealised “perfect for Rasch model” scenario if applied to video games would be: *“Using a set of items which all depend solely on one latent trait, but all those items are independent of each other”*. To make this happen, would make the Rasch model fully appropriate for the project. To do so, items and their categorisation have to be independent, and the video games have to meet the unidimensionality criteria; which in this project neither game did, even being cautiously chosen to be unidimensional.



### 5.3 Conclusion

After analysing the data from 107 recordings for 2 different games the results proved that the model could generate a set of relevant detailed quantitative and qualitative parameters regarding how ability and performance varied with the difficulty of the game. Game 2 (Galaxian) was discussed to be more reliable due to a higher Person Separation index and to a lower multidimensionality, which were two of the main factors affecting reliability.

The information extracted from the results points in the right direction and is logical, but nevertheless the results can't be considered 100% reliable as both games displayed a higher degree of multidimensionality than the maximum accepted by the model. This means that in these 2 games the model can be used as an auxiliary guide, but certainly not as an absolute difficulty estimator.

The project's results led to conclude that the Rasch model can be used to evaluate the difficulty of videogames as long as the selected items and their categorisation are independent. The task being developed should also be unidimensional for the results to fit to the Rasch model. Such condition isn't easy to attain with videogames, as seen in the results but there are chances that a better item selection and categorisation could improve the unidimensionality test. The model still provided a significantly good guide to the relation between performance, ability and difficulty, besides not meeting the unidimensionality requirement.

## 5.4 Future Work

As stated initially, analysing videogames with the Rasch model could be a starting point to analysing other computer/electronically executed tasks, but unfortunately the results from the current project raised the bar in terms of the difficulty of advancing to other tasks than videogames.

There is a wide gap between videogames and other computer tasks which has to be covered with further research such as breaking down the factors that affect performance in computer tasks, or perhaps developing further the Rasch model so it has the capability of handling multidimensional tasks, which raises the question: is there a better model that does address multidimensionality?

Regarding the videogame difficulty estimation, an improvement on the targeting of the items and their categorisation as well as their independency would lead to a much more reliable method of computing the difficulty. Combined with the right multidimensionality analysis it could provide more reliable results on this method. The project could be replicated considering the proposed changes, and if concluded to be reliable, then experiment with a computer or electronic task.

## References

---

1. Oxford Dictionary definition of “difficult” available online from:  
<https://en.oxforddictionaries.com/definition/difficulty>
2. Wolfe F, van der Heijde DM, Larsen A. Assessing radiographic status of rheumatoid arthritis: introduction of a short erosion scale. *J Rheumatol* 2000; **27**: 2090–9.  
Available online from: <https://www.ncbi.nlm.nih.gov/pubmed/10990218>
3. opec JA, Sayre EC, Davis AM, Badley EM, Abrahamowicz M, Sherlock L, et al. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. *Health Qual Life Outcomes* 2006; **4**: 33.  
Available online from: <https://www.ncbi.nlm.nih.gov/pubmed/16749932>
4. A. Tennant, PG. Conaghan, 2007 “The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?” Pages 1358-1362  
Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18050173>
5. M. V. Aponte, G. Levieux, S. Natkin, 2011, “Measuring the level of difficulty in single player video games” Available from:  
<http://www.minesweeper.info/articles/ScalingDifficultyLevelVideoGames.pdf>
6. F. J. Mourato, M. P. dos Santos, 2011, “Measuring difficulty in platform videogames” Available from:  
<https://comum.rcaap.pt/bitstream/10400.26/6087/1/Measuring%20Difficulty%20in%20Platform%20Videogames.pdf>
7. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* 2007; **46**: 1–18. Available online from:  
<https://www.ncbi.nlm.nih.gov/pubmed/17472198>
8. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* 2004; **7** Suppl 1: S22–6. Available online from:  
[http://www.valueinhealthjournal.com/article/S1098-3015\(10\)60232-X/pdf](http://www.valueinhealthjournal.com/article/S1098-3015(10)60232-X/pdf)

9. Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas* 2002; **3**: 205–31 Available from:  
<https://www.ncbi.nlm.nih.gov/pubmed/12011501>
10. Cook, R. Dennis; Weisberg, Sanford. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall. Retrieved from the University of Minnesota Digital Conservancy, available from  
<http://hdl.handle.net/11299/37076>
11. Setiati D. Introductory Rasch analysis using RUMM2030. (Hand out). Psychometric Laboratory for health Sciences. University of Leeds, 2016
12. Weisstein, Eric W. "Chi-Squared Test" "Students t-distribution" available from:  
<http://mathworld.wolfram.com/>
13. Fair tests: A do it yourself guide  
[http://undsci.berkeley.edu/article/fair\\_tests\\_01](http://undsci.berkeley.edu/article/fair_tests_01) University of California, Berkeley
14. Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological Methodology*, San Francisco, Jossey-Bass. (Chapter 2, pp. 33–80.) available online from:  
[http://hbanaszak.mjr.uw.edu.pl/TempTxt/Andrich\\_1985\\_An%20ElaborationOfGuttmanScalingWithRaschModels.pdf](http://hbanaszak.mjr.uw.edu.pl/TempTxt/Andrich_1985_An%20ElaborationOfGuttmanScalingWithRaschModels.pdf)

## Appendices

---

### Meeting log

#### MECH3890 Individual Engineering Project Supervisor Meeting Log

Date of Meeting	Summary of Discussion	Objectives for next meeting	Supervisors Initials
19/10/2016	Introduction project	Read past projects	
2/11/2016	Scoping document	Scoping document	
9/11/2016	Past projects review	Innovate project	
16/11/2016	Presenting new idea	Researching idea/Gantt chart	
23/11/2016	Literature review/idea appliance	Develop experiment	
30/11/2016	Literature review/idea appliance	Develop experiment	
7/12/2016	Practical considerations of experiment	Try out software for experiment/set up experiment	
14/12/2016	Evaluating experiment	Apply changes + start recordings	
11/01/2017	Rectify game selection	Remodel experiment changing games	
25/01/2017	Modified experiment discussion	Restart recordings	
08/02/2017	Practical considerations with recordings	Continue recordings	
15/02/2017	Report plan	Start report writing + continue recordings	
22/02/2017	Report plan	report writing + continue recordings	
8/03/2017	Report structure	report writing + continue recordings	
15/03/2017	Software review	Research software functioning	
22/03/2017	Introduction to software analysis tools	-	
23/03/2017	Guiding through software + data first impressions	Analyse data	

Game 1

### Class intervals

Item	CI1	CI2	CI3
I0001	36	37	34
I0002	36	37	34
I0003	36	37	34
I0004	36	37	34
I0005	36	37	34
I0006	36	37	34
I0007	36	37	34

### Summary statistics

SUMMARY STATISTICS for Analysis Name FIRST

**-ITEM - PERSON INTERACTION-**

ITEMS		PERSONS	
	Location		Fit Residual
Mean	0.0000	Mean	0.6565
Std Dev	0.3669	Mean	-0.1124
Skewness	0.2367	Std Dev	0.6573
Kurtosis	-1.3753	Skewness	0.2791
		Kurtosis	-0.3632
		<input checked="" type="checkbox"/> Include Extremes	N = 107
	Correlation [location/stdResidual]		Correlation [location/stdResidual]
			-0.1593

**-ITEM - TRAIT INTERACTION-**

Total - Item Chi Square: 50.8716  
 Degrees of Freedom: 14  
 Chi Square Probability: 0.000004

**RELIABILITY INDICES**

PerSepIdx: First  
 \* with extms: 0.63163  
 \* NO extms: 0.63163  
 CoeffAlpha  
 \* with extms: 0.65593  
 \* NO extms: 0.65593

**-LIKELIHOOD RATIO TEST-**

Analysis Likelihood ChiSq  
 anaName1  
 anaName2  
 DegF  
 Prob

**POWER OF ANALYSIS OF FIT**

Excellent  
 Good  
**Reasonable**  
 Low  
 Too Low

This display is intended as a guide ONLY and should be used in conjunction with other analysis indicators

File Text Format  
 Fixed  Tab Delimit

### Individual Person fit

recID	Tot/Exp Sc	MaxSc	Items	Extm	Location	SE	FitResid	DegFree	Data Pts	PersID
42	19	28	7		0.677	0.411	2.169	5.7	7	42
37	9	28	7		-0.721	0.396	1.978	5.7	7	37
30	22	28	7		1.247	0.473	1.859	5.7	7	30
90	22	28	7		1.247	0.473	1.856	5.7	7	90
25	6	28	7		-1.234	0.452	1.649	5.7	7	25
85	7	28	7		-1.047	0.428	1.641	5.7	7	85
24	8	28	7		-0.877	0.410	1.552	5.7	7	24
23	9	28	7		-0.721	0.396	1.502	5.7	7	23
107	11	28	7		-0.436	0.378	1.303	5.7	7	107
58	8	28	7		-0.877	0.410	1.279	5.7	7	58
49	6	28	7		-1.234	0.452	1.193	5.7	7	49
33	9	28	7		-0.721	0.396	1.162	5.7	7	33
15	9	28	7		-0.721	0.396	1.084	5.7	7	15
95	11	28	7		-0.436	0.378	1.025	5.7	7	95
81	11	28	7		-0.436	0.378	0.998	5.7	7	81
106	10	28	7		-0.575	0.385	0.972	5.7	7	106
70	16	28	7		0.229	0.379	0.937	5.7	7	70
64	15	28	7		0.094	0.374	0.899	5.7	7	64
80	18	28	7		0.518	0.397	0.861	5.7	7	80
104	15	28	7		0.094	0.374	0.829	5.7	7	104
60	9	28	7		-0.721	0.396	0.805	5.7	7	60
61	12	28	7		-0.302	0.373	0.764	5.7	7	61
97	6	28	7		-1.234	0.452	0.758	5.7	7	97
84	15	28	7		0.094	0.374	0.673	5.7	7	84
96	18	28	7		0.518	0.397	0.670	5.7	7	96
47	11	28	7		-0.436	0.378	0.635	5.7	7	47
1	5	28	7		-1.448	0.484	0.625	5.7	7	1
18	19	28	7		0.677	0.411	0.601	5.7	7	18
36	18	28	7		0.518	0.397	0.577	5.7	7	36
8	15	28	7		0.094	0.374	0.536	5.7	7	8
72	9	28	7		-0.721	0.396	0.515	5.7	7	72
78	24	28	7		1.752	0.547	0.487	5.7	7	78
59	15	28	7		0.094	0.374	0.450	5.7	7	59
63	9	28	7		-0.721	0.396	0.376	5.7	7	63
28	20	28	7		0.849	0.427	0.370	5.7	7	28
93	15	28	7		0.094	0.374	0.345	5.7	7	93
20	10	28	7		-0.575	0.385	0.317	5.7	7	20
71	7	28	7		-1.047	0.428	0.305	5.7	7	71
53	10	28	7		-0.575	0.385	0.291	5.7	7	53

87	18	28	7	0.518	0.397	0.280	5.7	7	87
12	8	28	7	-0.877	0.410	0.279	5.7	7	12
57	7	28	7	-1.047	0.428	0.269	5.7	7	57
77	17	28	7	0.370	0.387	0.230	5.7	7	77
102	17	28	7	0.370	0.387	0.227	5.7	7	102
67	12	28	7	-0.302	0.373	0.224	5.7	7	67
69	17	28	7	0.370	0.387	0.213	5.7	7	69
32	17	28	7	0.370	0.387	0.201	5.7	7	32
21	17	28	7	0.370	0.387	0.190	5.7	7	21
101	14	28	7	-0.039	0.372	0.151	5.7	7	101
73	8	28	7	-0.877	0.410	0.144	5.7	7	73
94	15	28	7	0.094	0.374	0.142	5.7	7	94
100	15	28	7	0.094	0.374	0.110	5.7	7	100
105	8	28	7	-0.877	0.410	-0.028	5.7	7	105
89	16	28	7	0.229	0.379	-0.101	5.7	7	89
48	16	28	7	0.229	0.379	-0.108	5.7	7	48
41	15	28	7	0.094	0.374	-0.144	5.7	7	41
66	20	28	7	0.849	0.427	-0.161	5.7	7	66
31	12	28	7	-0.302	0.373	-0.195	5.7	7	31
29	19	28	7	0.677	0.411	-0.196	5.7	7	29
10	9	28	7	-0.721	0.396	-0.208	5.7	7	10
26	11	28	7	-0.436	0.378	-0.221	5.7	7	26
3	9	28	7	-0.721	0.396	-0.222	5.7	7	3
56	19	28	7	0.677	0.411	-0.265	5.7	7	56
19	8	28	7	-0.877	0.410	-0.308	5.7	7	19
5	9	28	7	-0.721	0.396	-0.321	5.7	7	5
13	7	28	7	-1.047	0.428	-0.352	5.7	7	13
6	17	28	7	0.370	0.387	-0.362	5.7	7	6
99	15	28	7	0.094	0.374	-0.441	5.7	7	99
82	19	28	7	0.677	0.411	-0.469	5.7	7	82
88	19	28	7	0.677	0.411	-0.498	5.7	7	88
22	15	28	7	0.094	0.374	-0.501	5.7	7	22
27	18	28	7	0.518	0.397	-0.507	5.7	7	27
79	15	28	7	0.094	0.374	-0.524	5.7	7	79
103	11	28	7	-0.436	0.378	-0.577	5.7	7	103
38	14	28	7	-0.039	0.372	-0.585	5.7	7	38
9	6	28	7	-1.234	0.452	-0.585	5.7	7	9
39	21	28	7	1.038	0.448	-0.628	5.7	7	39



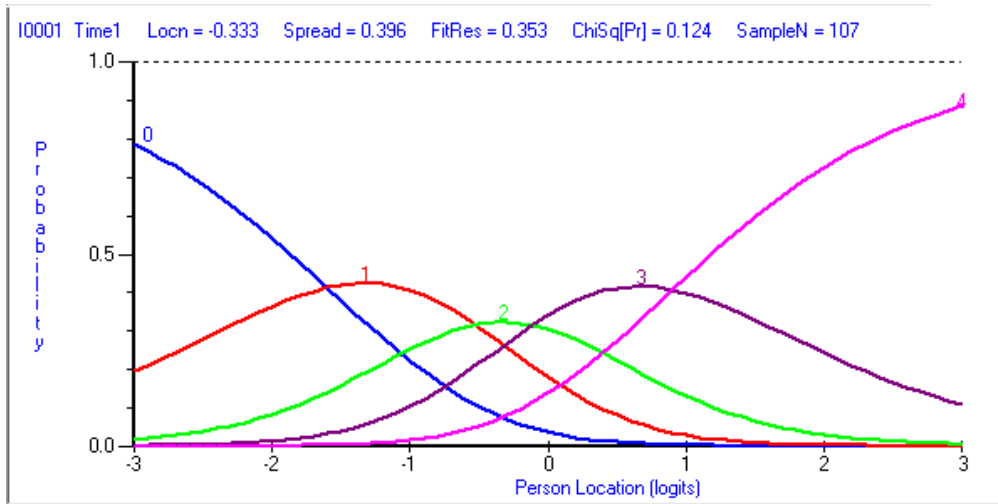
86	17	28	7	0.370	0.387	-0.709	5.7	7	86
68	9	28	7	-0.721	0.396	-0.717	5.7	7	68
55	12	28	7	-0.302	0.373	-0.748	5.7	7	55
75	20	28	7	0.849	0.427	-0.798	5.7	7	75
4	9	28	7	-0.721	0.396	-0.808	5.7	7	4
44	13	28	7	-0.170	0.371	-0.825	5.7	7	44
11	14	28	7	-0.039	0.372	-0.897	5.7	7	11
62	10	28	7	-0.575	0.385	-0.931	5.7	7	62
14	10	28	7	-0.575	0.385	-0.931	5.7	7	14
74	8	28	7	-0.877	0.410	-0.975	5.7	7	74
46	10	28	7	-0.575	0.385	-1.045	5.7	7	46
65	18	28	7	0.518	0.397	-1.069	5.7	7	65
52	10	28	7	-0.575	0.385	-1.234	5.7	7	52
17	15	28	7	0.094	0.374	-1.266	5.7	7	17
45	13	28	7	-0.170	0.371	-1.270	5.7	7	45
92	15	28	7	0.094	0.374	-1.538	5.7	7	92
91	15	28	7	0.094	0.374	-1.620	5.7	7	91
7	11	28	7	-0.436	0.378	-1.646	5.7	7	7
2	10	28	7	-0.575	0.385	-1.648	5.7	7	2
98	13	28	7	-0.170	0.371	-1.691	5.7	7	98
16	16	28	7	0.229	0.379	-1.713	5.7	7	16
76	22	28	7	1.247	0.473	-1.775	5.7	7	76
34	16	28	7	0.229	0.379	-1.808	5.7	7	34
51	12	28	7	-0.302	0.373	-2.053	5.7	7	51
40	15	28	7	0.094	0.374	-2.138	5.7	7	40
50	13	28	7	-0.170	0.371	-2.175	5.7	7	50
83	17	28	7	0.370	0.387	-2.329	5.7	7	83
43	15	28	7	0.094	0.374	-2.377	5.7	7	43
54	22	28	7	1.247	0.473	-2.384	5.7	7	54
35	15	28	7	0.094	0.374	-2.758	5.7	7	35

### Individual Item fit

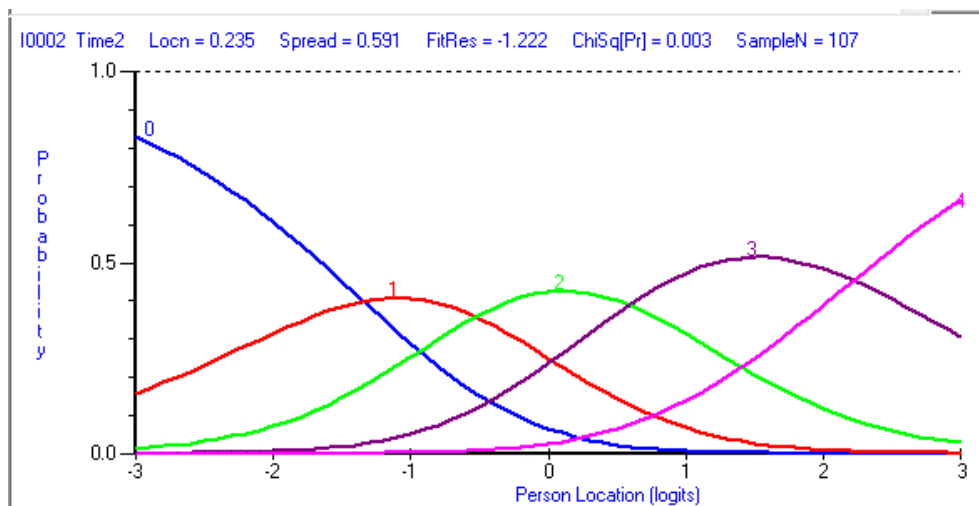
	Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob	F-stat	DF-1	DF-2	Prob
6	6	I0006	Poly	-0.433	0.115	0.577	87.86	1.618	2	0.445242	...	...	...	...
1	1	I0001	Poly	-0.333	0.098	0.353	87.86	4.173	2	0.124092	...	...	...	...
3	3	I0003	Poly	0.555	0.110	-0.717	87.86	4.611	2	0.099722	...	...	...	...
4	4	I0004	Poly	0.049	0.106	-0.753	87.86	5.989	2	0.050051	...	...	...	...
5	5	I0005	Poly	0.229	0.091	3.161	87.86	9.399	2	0.009100	...	...	...	...
2	2	I0002	Poly	0.235	0.109	-1.222	87.86	11.762	2	0.002793	...	...	...	...
7	7	I0007	Poly	-0.302	0.088	3.195	87.86	13.319	2	0.001283	...	...	...	...

### Item category curve

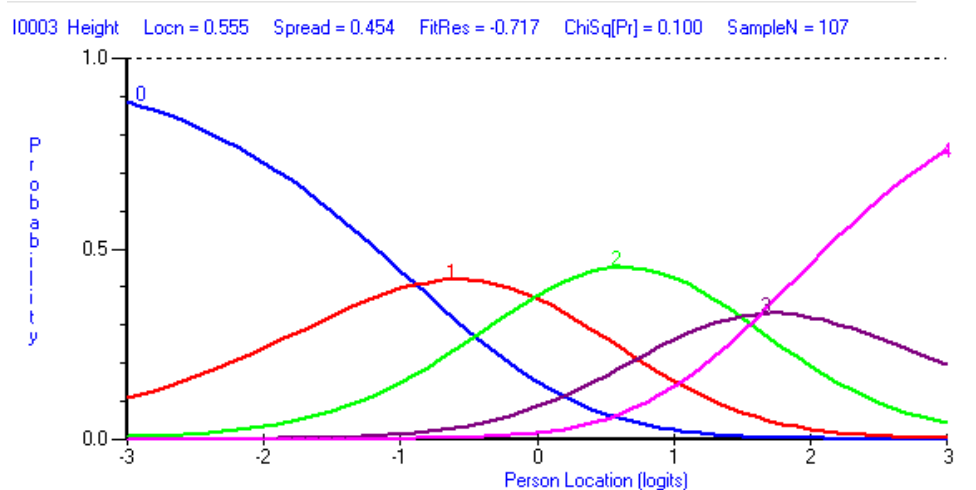
#### Item 1



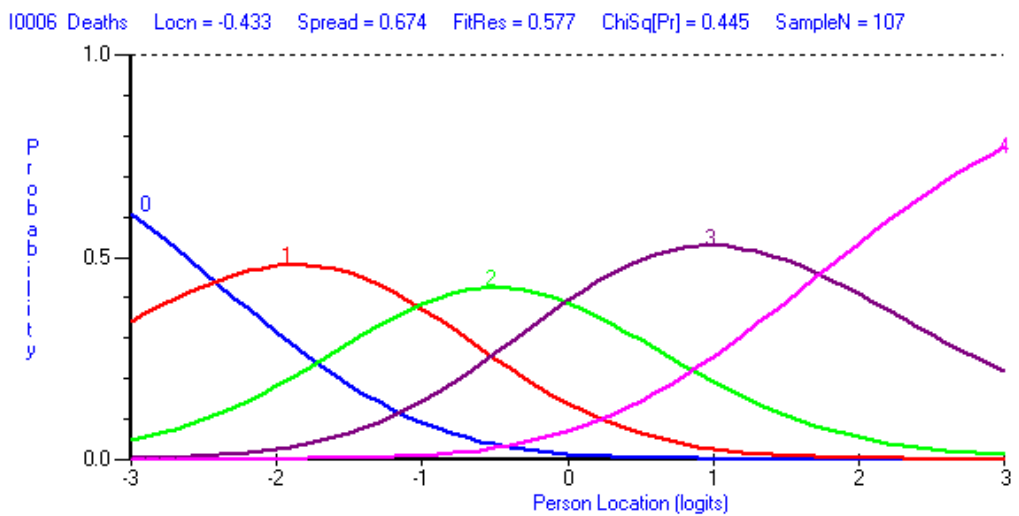
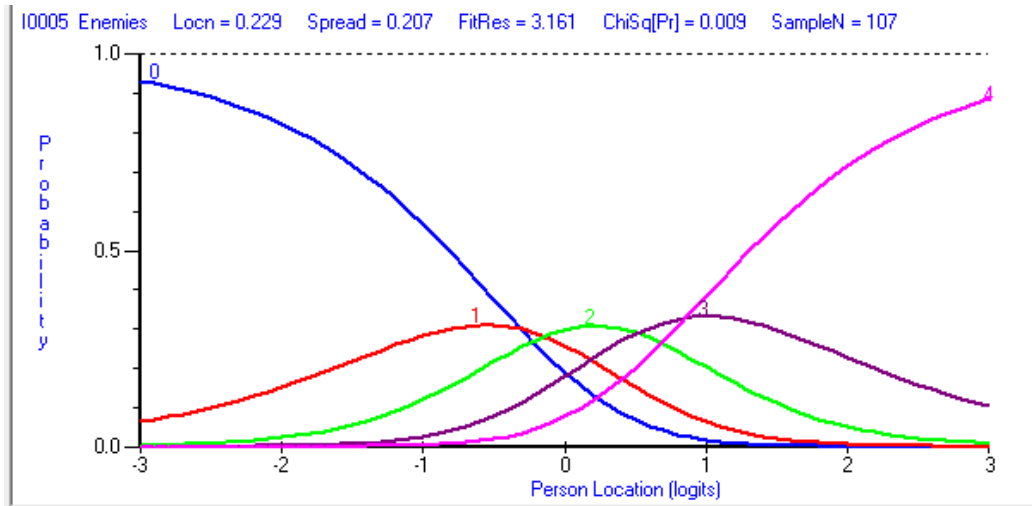
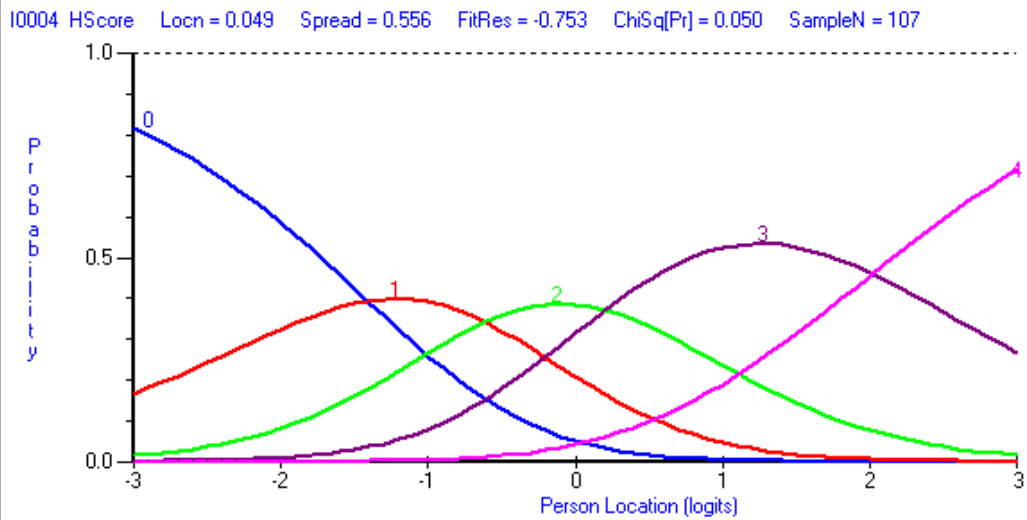
Item 2



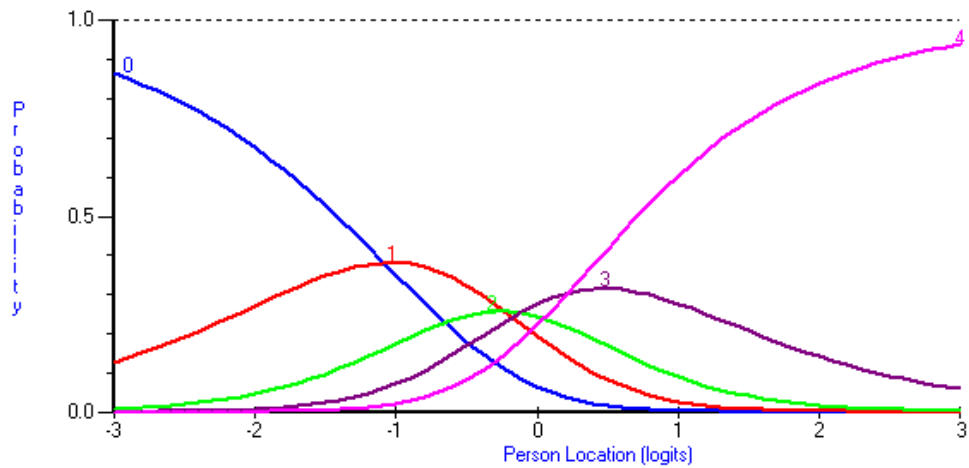
Item 3



Item 4



I0007 Jumps Lochn = -0.302 Spread = 0.201 FitRes = 3.195 ChiSq[Pr] = 0.001 SampleN = 107



t-test for unidimensionality

Summary Table of t-test analyses for this Subset pair

Test	Subset Pair	No. < 5%	No. < 1%	PerC < 5%	PerC < 1%	Total
1	pos; neg	16	3	14.95%	2.80%	107

Sample statistics

Mean of pos	-0.1019418
Std dev of pos	1.1416310
Sample size of pos	107
Mean of neg	-0.1141675
Std dev of neg	0.6312470
Sample size of neg	107

Total Number of Extreme Scores

	pos	neg
Minimum score	1	0
Maximum score	2	0

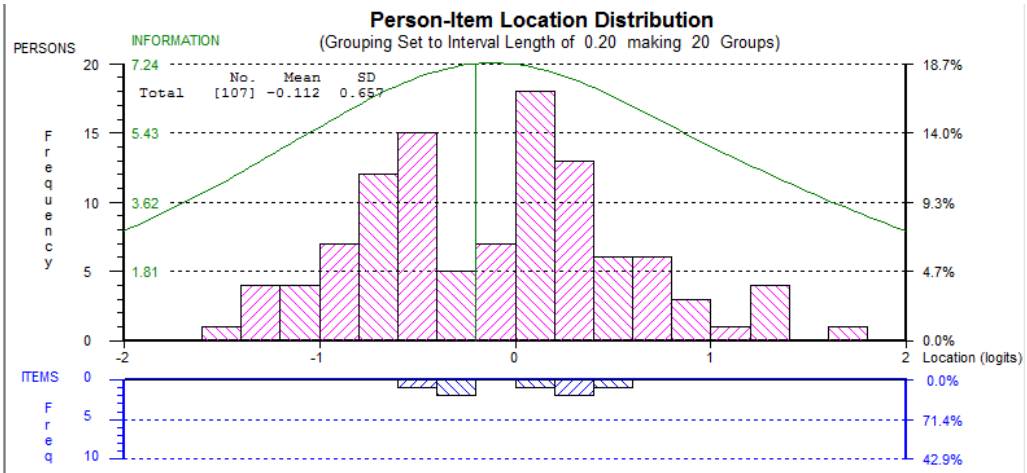
Dependent Sample t-test

Mean of differences	0.0122256
Std Dev of differences	1.2159930
Std Error of differences	0.1175545
Sample size	107
t-value	0.1039998

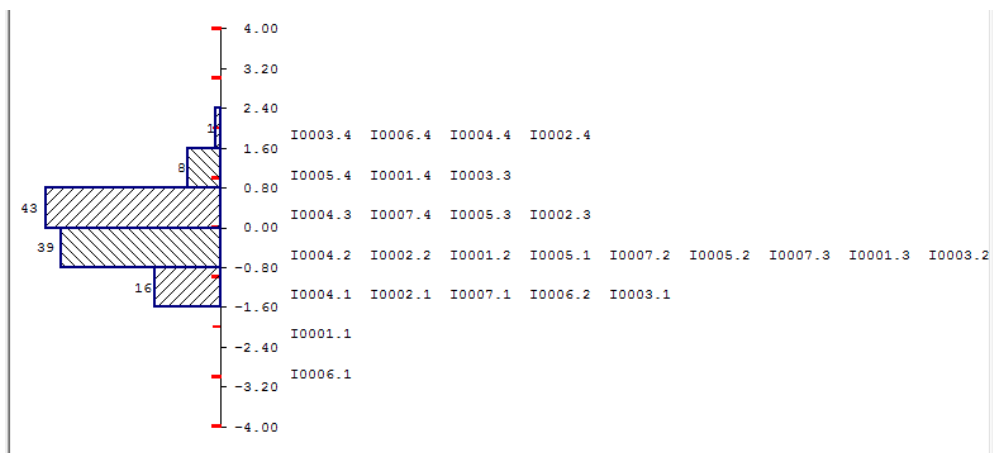
Correlation between pos and neg

0.154829

Person-item distribution



Item map



Game 2

Class interval

Item	CI1	CI2	CI3
I0001	32	36	38
I0002	32	36	38
I0003	32	36	38
I0004	32	36	38
I0005	32	36	38
I0006	32	36	38
I0007	32	36	38

Summary statistics

SUMMARY STATISTICS for Analysis Name GALAX

ITEM - PERSON INTERACTION

ITEMS				PERSONS			
	Location		Fit Residual		Location		Fit Residual
Mean	0.0000		0.3798	Mean	-0.3765		-0.2298
Std Dev	0.3283	Std Dev	0.8615	Std Dev	1.0913	Std Dev	1.0495
Skewness	-0.4483	Skewness	0.5581	Skewness	-0.5434	Skewness	-0.1062
Kurtosis	0.2276	Kurtosis	-1.4444	Kurtosis	0.8458	Kurtosis	0.0284
		Correlation (location/stdResidual)	-0.1180	<input checked="" type="checkbox"/> Include Extremes	N = 107	Correlation (location/stdResidual)	-0.1390

ITEM - TRAIT INTERACTION

Total - Item Chi Square: 36.2973  
 Degrees of Freedom: 14  
 Chi Square Probability: 0.000942

RELIABILITY INDICES

PerSepIdx: Galax  
 \* with extms: 0.80442  
 \* NO extms: 0.79402  
 CoeficAlpha \* with extms: 0.81096  
 \* NO extms: 0.80152

LIKELIHOOD RATIO TEST

Analysis Likelihood ChiSq  
 anaName1  
 anaName2

POWER OF ANALYSIS OF FIT

Excellent  
 Good  
 Reasonable  
 Low  
 Too Low

**GOOD**

This display is intended as a guide ONLY and should be used in conjunction with other analysis indicators

< Display Control

File Text Format  
 Fixed  Tab Delimit Save

### Individual person fit

recID	Tot/Exp Sc	MaxSc	Items	Extm	Location	SE	FitResid	DegFree	Data Pts	PersID
11	13	27	7		-0.108	0.427	2.418	5.8	7	11
59	13	27	7		-0.108	0.427	2.155	5.8	7	59
19	6	27	7		-1.466	0.487	1.838	5.8	7	19
97	13	27	7		-0.108	0.427	1.803	5.8	7	97
67	7	27	7		-1.248	0.468	1.791	5.8	7	67
107	14	27	7		0.071	0.426	1.719	5.8	7	107
99	8	27	7		-1.043	0.454	1.304	5.8	7	99
96	9	27	7		-0.848	0.445	1.162	5.8	7	96
38	3	27	7		-2.295	0.607	1.144	5.8	7	38
94	15	27	7		0.249	0.425	1.002	5.8	7	94
95	10	27	7		-0.658	0.438	0.999	5.8	7	95
26	15	27	7		0.249	0.425	0.955	5.8	7	26
66	12	27	7		-0.289	0.429	0.929	5.8	7	66
63	11	27	7		-0.472	0.433	0.925	5.8	7	63
87	16	27	7		0.425	0.427	0.837	5.8	7	87
25	9	27	7		-0.848	0.445	0.750	5.8	7	25
23	16	27	7		0.425	0.427	0.712	5.8	7	23
89	9	27	7		-0.848	0.445	0.687	5.8	7	89
106	13	27	7		-0.108	0.427	0.684	5.8	7	106
85	8	27	7		-1.043	0.454	0.601	5.8	7	85

39	17	27	7	0.603	0.431	0.499	5.8	7	39
73	9	27	7	-0.848	0.445	0.499	5.8	7	73
86	5	27	7	-1.705	0.513	0.487	5.8	7	86
93	10	27	7	-0.658	0.438	0.436	5.8	7	93
79	21	27	7	1.384	0.490	0.435	5.8	7	79
68	20	27	7	1.168	0.466	0.430	5.8	7	68
74	15	27	7	0.249	0.425	0.402	5.8	7	74
46	15	27	7	0.249	0.425	0.356	5.8	7	46
61	12	27	7	-0.289	0.429	0.356	5.8	7	61
102	21	27	7	1.384	0.490	0.354	5.8	7	102
90	3	27	7	-2.295	0.607	0.330	5.8	7	90
4	2	27	7	-2.701	0.701	0.304	5.8	7	4
62	10	27	7	-0.658	0.438	0.301	5.8	7	62
21	23	27	7	1.921	0.572	0.292	5.8	7	21
50	12	27	7	-0.289	0.429	0.270	5.8	7	50
47	8	27	7	-1.043	0.454	0.233	5.8	7	47
71	17	27	7	0.603	0.431	0.223	5.8	7	71
10	7	27	7	-1.248	0.468	0.217	5.8	7	10
104	15	27	7	0.249	0.425	0.196	5.8	7	104

51	1	27	7	-3.286	0.883	0.188	5.8	7	51
92	10	27	7	-0.658	0.438	0.182	5.8	7	92
42	3	27	7	-2.295	0.607	0.149	5.8	7	42
80	12	27	7	-0.289	0.429	0.144	5.8	7	80
58	8	27	7	-1.043	0.454	0.092	5.8	7	58
82	17	27	7	0.603	0.431	0.080	5.8	7	82
78	19	27	7	0.970	0.449	0.073	5.8	7	78
77	19	27	7	0.970	0.449	0.046	5.8	7	77
84	8	27	7	-1.043	0.454	0.044	5.8	7	84
81	12	27	7	-0.289	0.429	-0.021	5.8	7	81
54	20	27	7	1.168	0.466	-0.121	5.8	7	54
52	2	27	7	-2.701	0.701	-0.124	5.8	7	52
100	5	27	7	-1.705	0.513	-0.137	5.8	7	100
91	16	27	7	0.425	0.427	-0.144	5.8	7	91
55	6	27	7	-1.466	0.487	-0.207	5.8	7	55
6	18	27	7	0.783	0.438	-0.228	5.8	7	6
41	10	27	7	-0.658	0.438	-0.245	5.8	7	41
32	13	27	7	-0.108	0.427	-0.275	5.8	7	32
105	14	27	7	0.071	0.426	-0.294	5.8	7	105

31	19	27	7	0.970	0.449	-0.304	5.8	7	31
65	11	27	7	-0.472	0.433	-0.317	5.8	7	65
17	11	27	7	-0.472	0.433	-0.317	5.8	7	17
15	11	27	7	-0.472	0.433	-0.320	5.8	7	15
29	18	27	7	0.783	0.438	-0.324	5.8	7	29
45	10	27	7	-0.658	0.438	-0.334	5.8	7	45
33	12	27	7	-0.289	0.429	-0.356	5.8	7	33
14	10	27	7	-0.658	0.438	-0.377	5.8	7	14
83	6	27	7	-1.466	0.487	-0.411	5.8	7	83
70	9	27	7	-0.848	0.445	-0.559	5.8	7	70
44	12	27	7	-0.289	0.429	-0.610	5.8	7	44
1	3	27	7	-2.295	0.607	-0.739	5.8	7	1
2	10	27	7	-0.658	0.438	-0.755	5.8	7	2
35	5	27	7	-1.705	0.513	-0.772	5.8	7	35
98	12	27	7	-0.289	0.429	-0.797	5.8	7	98
49	5	27	7	-1.705	0.513	-0.805	5.8	7	49
24	13	27	7	-0.108	0.427	-0.826	5.8	7	24
69	22	27	7	1.630	0.523	-0.833	5.8	7	69
18	10	27	7	-0.658	0.438	-0.836	5.8	7	18

56	17	27	7		0.603	0.431	-0.900	5.8	7	56
8	17	27	7		0.603	0.431	-0.900	5.8	7	8
37	8	27	7		-1.043	0.454	-0.936	5.8	7	37
88	13	27	7		-0.108	0.427	-0.961	5.8	7	88
20	18	27	7		0.783	0.438	-0.963	5.8	7	20
48	3	27	7		-2.295	0.607	-0.979	5.8	7	48
75	11	27	7		-0.472	0.433	-1.001	5.8	7	75
57	15	27	7		0.249	0.425	-1.008	5.8	7	57
13	12	27	7		-0.289	0.429	-1.121	5.8	7	13
7	6	27	7		-1.466	0.487	-1.131	5.8	7	7
30	16	27	7		0.425	0.427	-1.192	5.8	7	30
76	10	27	7		-0.658	0.438	-1.238	5.8	7	76
43	16	27	7		0.425	0.427	-1.302	5.8	7	43
103	8	27	7		-1.043	0.454	-1.310	5.8	7	103
53	17	27	7		0.603	0.431	-1.346	5.8	7	53
101	15	27	7		0.249	0.425	-1.403	5.8	7	101
60	11	27	7		-0.472	0.433	-1.407	5.8	7	60
12	11	27	7		-0.472	0.433	-1.407	5.8	7	12
40	15	27	7		0.249	0.425	-1.559	5.8	7	40

9	14	27	7		0.071	0.426	-1.806	5.8	7	9
64	22	27	7		1.630	0.523	-1.848	5.8	7	64
16	22	27	7		1.630	0.523	-1.848	5.8	7	16
28	11	27	7		-0.472	0.433	-1.903	5.8	7	28
22	8	27	7		-1.043	0.454	-2.091	5.8	7	22
27	8	27	7		-1.043	0.454	-2.091	5.8	7	27
34	21	27	7		1.384	0.490	-2.179	5.8	7	34
36	11	27	7		-0.472	0.433	-2.434	5.8	7	36
5	17	27	7		0.603	0.431	-2.628	5.8	7	5
72	13	27	7		-0.108	0.427	-3.109	5.8	7	72
3	0.190	27	7	extm	-4.127	1.275	...			3

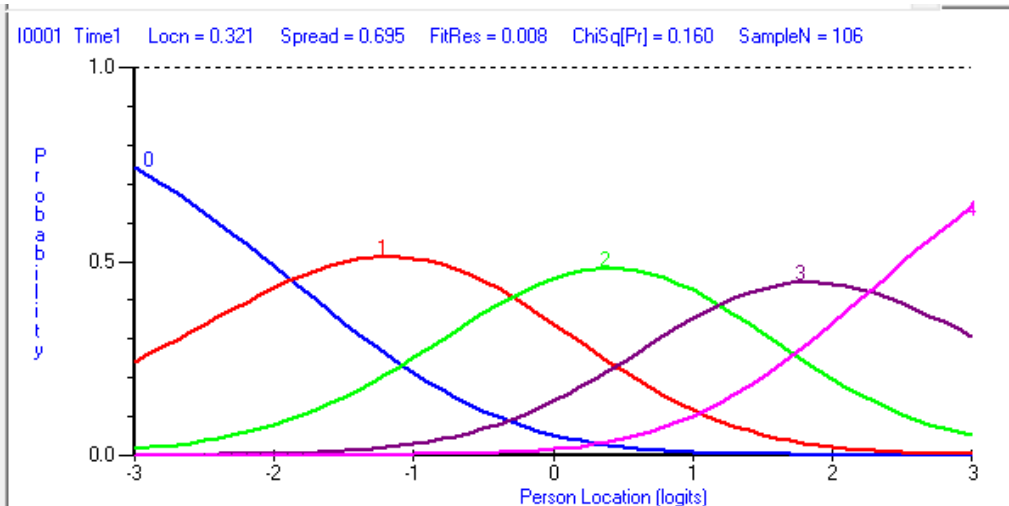
### Individual Item Fit

	Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob	F-stat	DF-1	DF-2	Prob
1	1	I0001	Poly	0.321	0.124	0.008	87.14	3.663	2	0.160202	...	...	...	..
2	2	I0002	Poly	0.056	0.124	0.041	87.14	3.171	2	0.204828	...	...	...	..
3	3	I0003	Poly	-0.165	0.107	-0.608	87.14	8.699	2	0.012913	...	...	...	..
4	4	I0004	Poly	0.426	0.108	-0.174	87.14	7.807	2	0.020171	...	...	...	..
5	5	I0005	Poly	0.057	0.132	1.809	87.14	2.491	2	0.287733	...	...	...	..
6	6	I0006	Poly	-0.556	0.110	0.273	87.14	3.951	2	0.138659	...	...	...	..
7	7	I0007	Poly	-0.140	0.119	1.309	87.14	6.514	2	0.038500	...	...	...	..

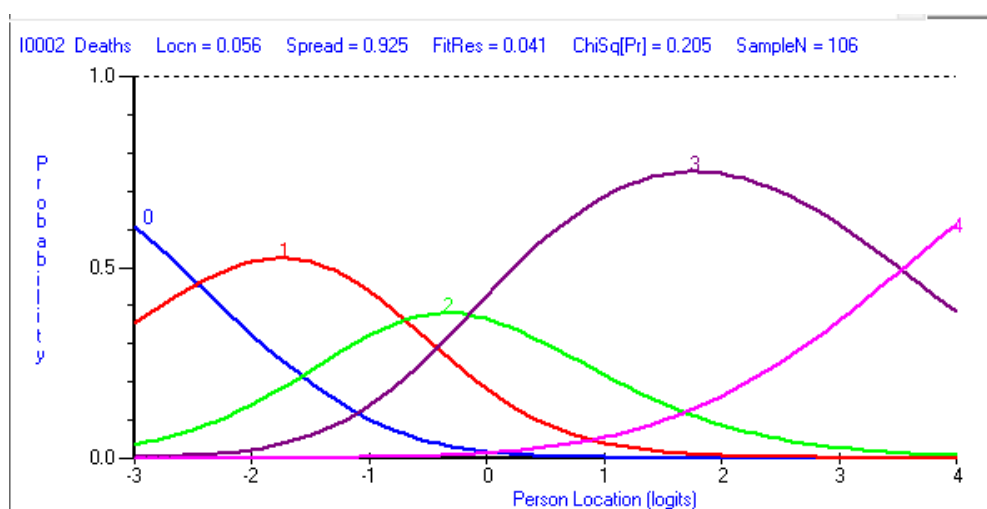
### Item category curves

#### Item 1

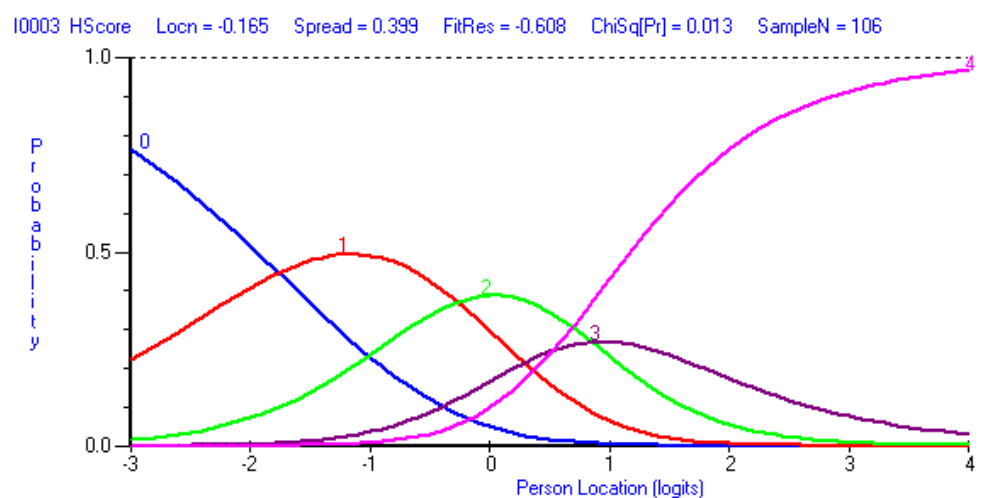




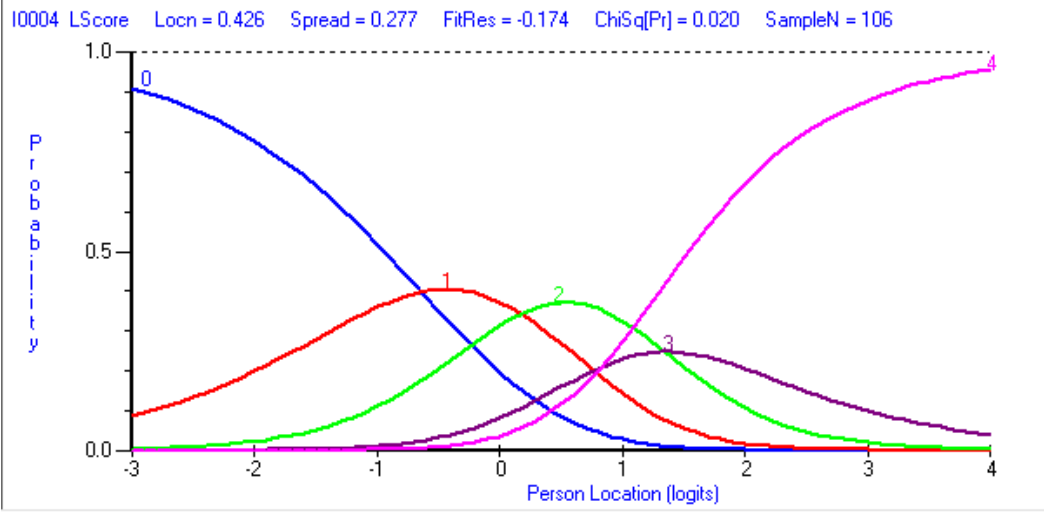
**Item 2**



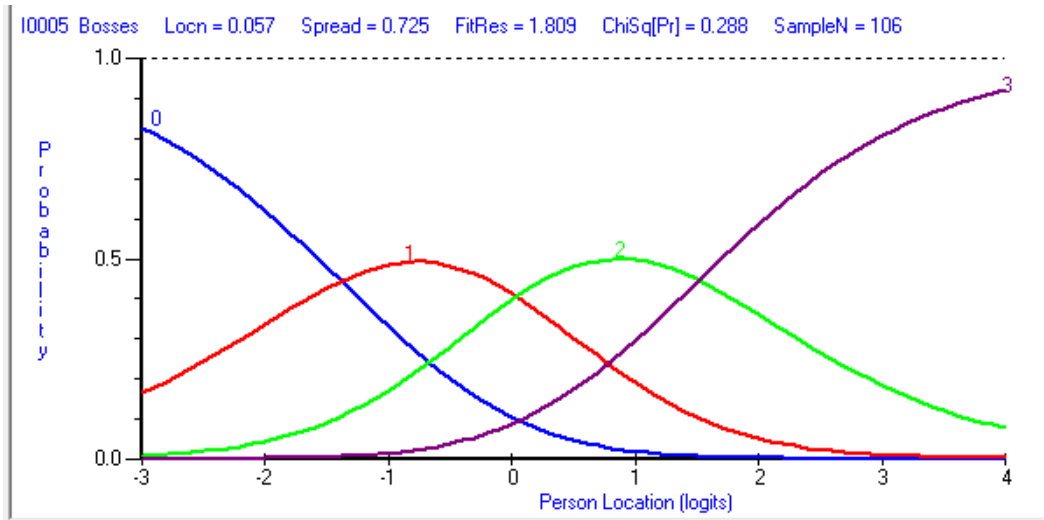
**Item 3**



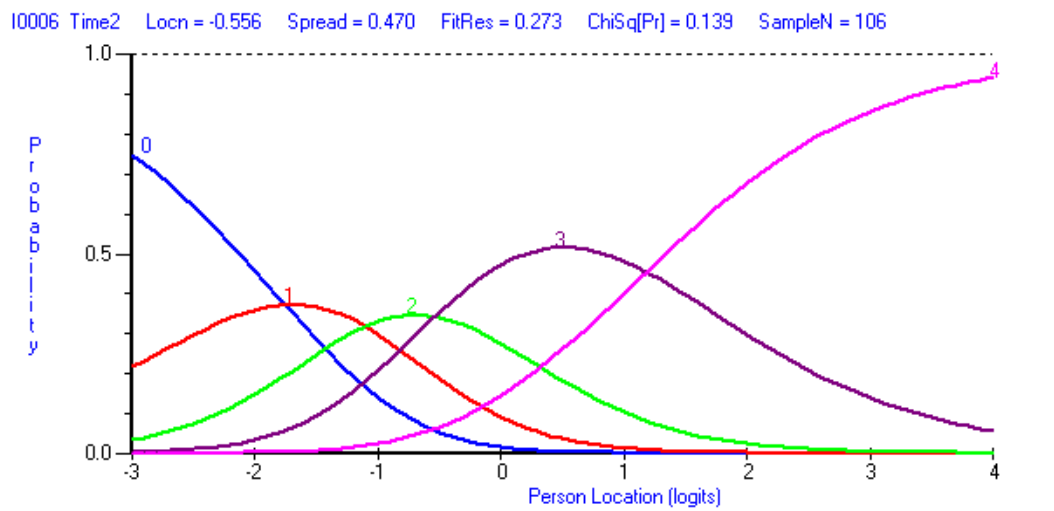
**Item 4**



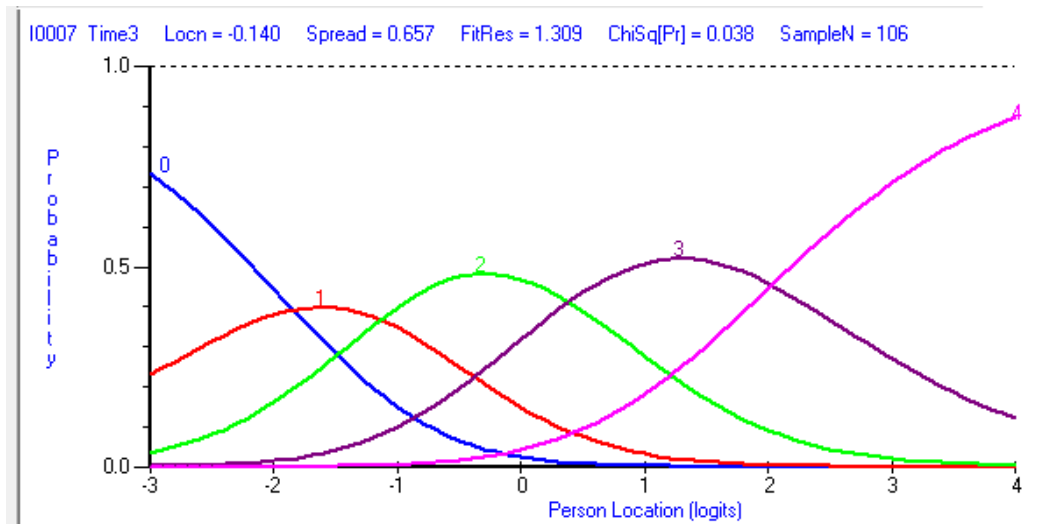
Item 5



Item 6



Item 7



t-test for unidimensionality

Summary Table of t-test analyses for this Subtest pair

Test	Subset Pair	No. < 5%	No. < 1%	PerC < 5%	PerC < 1%	Total
1	neg: pos	8	5	7.48%	4.67%	107

**Sample statistics**

Mean of neg	-0.4065171
Std dev of neg	1.3658480
Sample size of neg	107
Mean of pos	-0.3531015
Std dev of pos	1.1066120
Sample size of pos	107

**Total Number of Extreme Scores**

	neg	pos
Minimum score	5	2
Maximum score	0	0

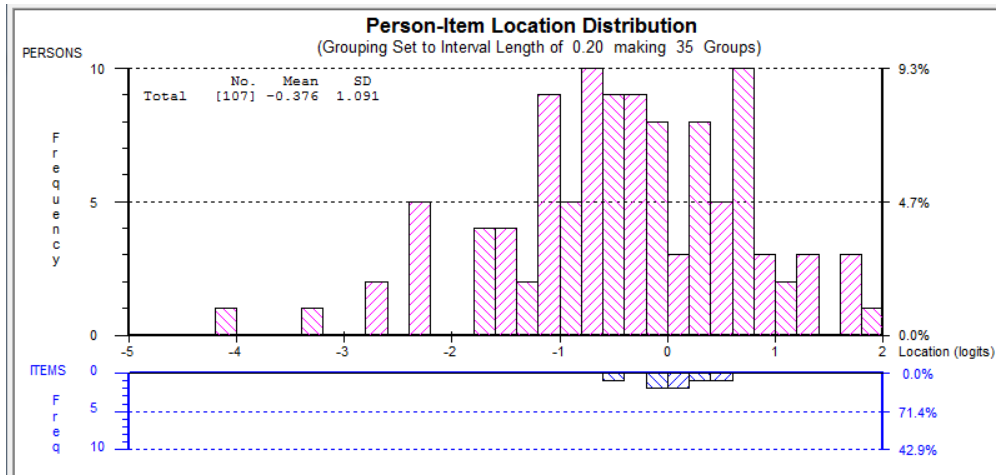
**Dependent Sample t-test**

Mean of differences	-0.0534157
Std Dev of differences	1.2263290
Std Error of differences	0.1185537
Sample size	107
t-value	-0.4505615

**Correlation between neg and pos**

0.524740

Person-item distribution



### Item map

