

# Creating an analysis tool to inform assessment development

John Little  
Sarah Gott  
Gideon Copestake  
Robert Coe

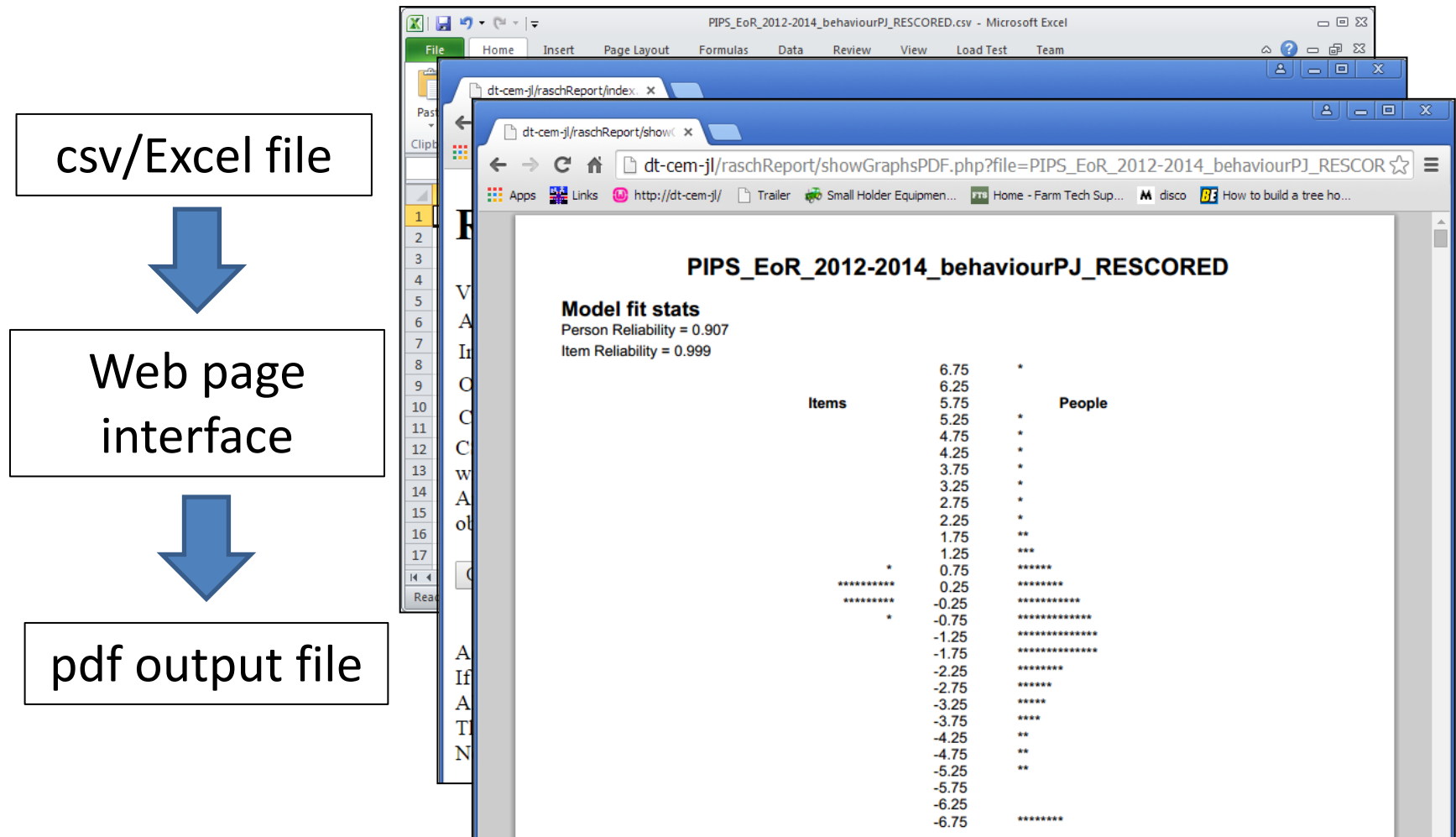
# Aims

- Build an easy to use analysis tool for those creating assessments.
- No knowledge of 'standard' Rasch tools or item response theory.
- Provide a detailed review of item level and test level performance.

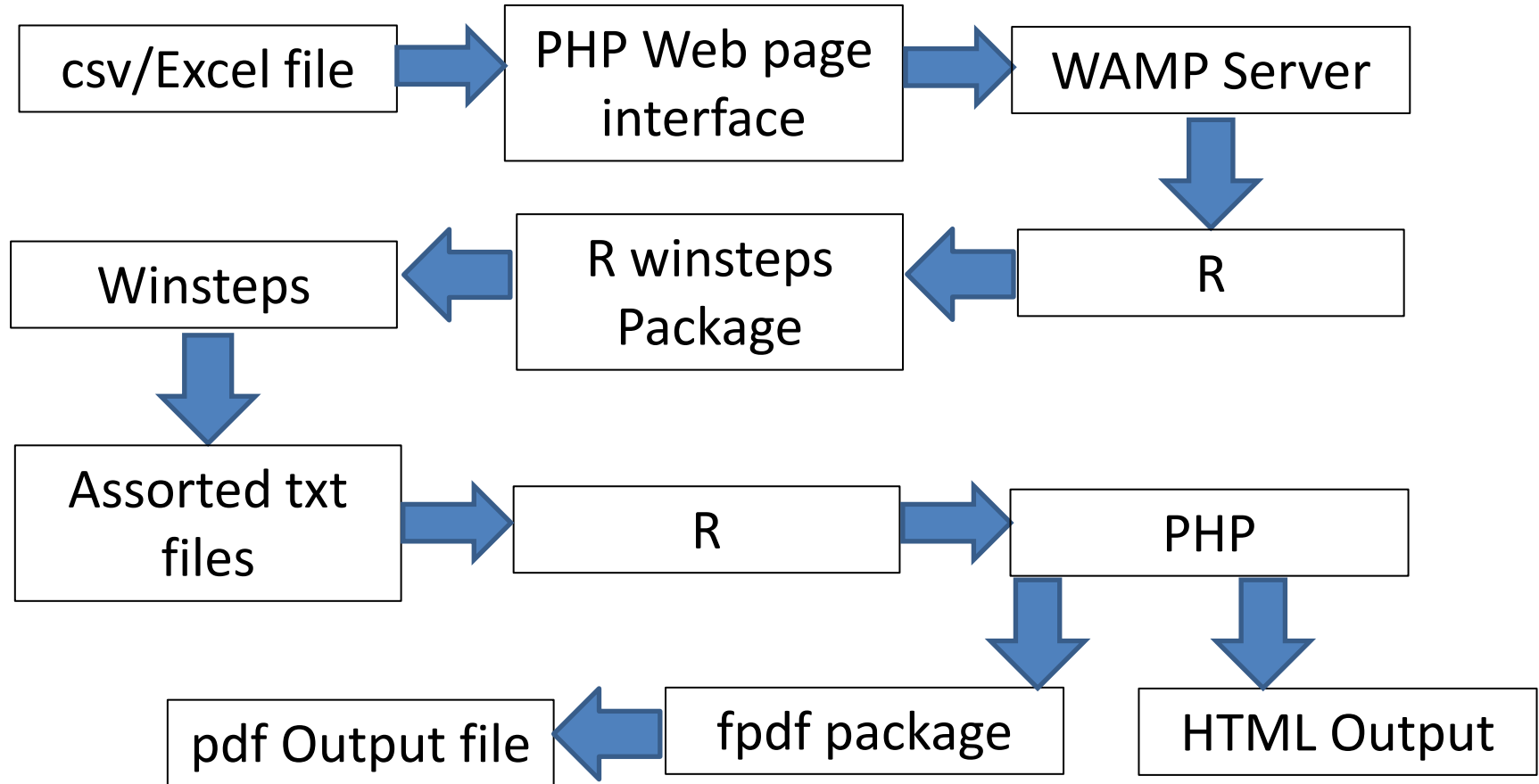
# Requirements

- Provide a user friendly report.
- Suitable for dichotomous and polytomous items.
- Report item difficulties and fit statistics for each marking category.
- Produce plots that allow for visual inspection of results to complement item statistics.

# The user process



# Background Process



# Raw Data Requirements

- csv/xls/xlsx file
- Data columns start with a “q”
- Label columns DO NOT start with a “q”
- Missing data should be missing...
- Other parameters are calculated directly from data

# An example from ADHD measure

- Sample consists of over 12,000 primary school pupils from 550 schools in England in the academic year 2007/2008.
- Class teachers completed a survey based on a 10 point scale relating to the frequency with which students met each of 21 criterion relating to a combination of inattention, hyperactivity and impulsivity (based on DSM-IV scale).

# Some items from the behaviour rating scale

- Makes careless mistakes in school work or other activities.
- Has difficulty sustaining attention in tasks or play activities.
- Does not seem to listen when spoken to directly.
- Does not follow through instructions, fails to finish work.
- Has difficulty organising tasks and activities.
- Is reluctant to engage in tasks which require sustained mental activity.



# Response format

Pupil: Test1 Test1 - Date: 10/03/2015

**Back an Item**


**Back to Start**

To complete this assessment, consider whether or not each statement applies to the child.

Move the cursor to the position on the slider that shows the extent to which the statement applies.

Try this now on the example below and click on the slider to continue.

Never ..... Always



Does this apply? Click on the slider to continue.

# Item example

Pupil: Test1 Test1 - Date: 10/03/2015

Back an Item

Back to Start

Does not follow through instructions, fails to finish work.

Never ..... Always

To what extent does this apply?

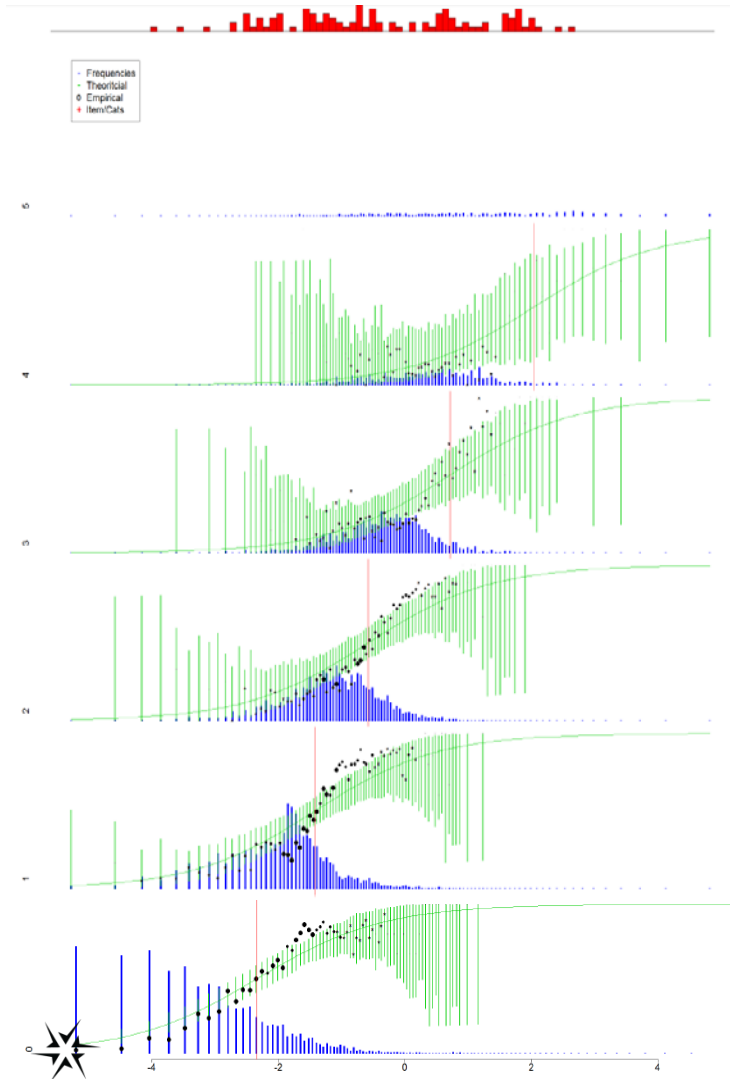
# Collapsing the scale

Between 25% and 50% of responses were in the 0, 1 or 2 mark category.

The original 10 point scale was therefore collapsed to 6 categories.

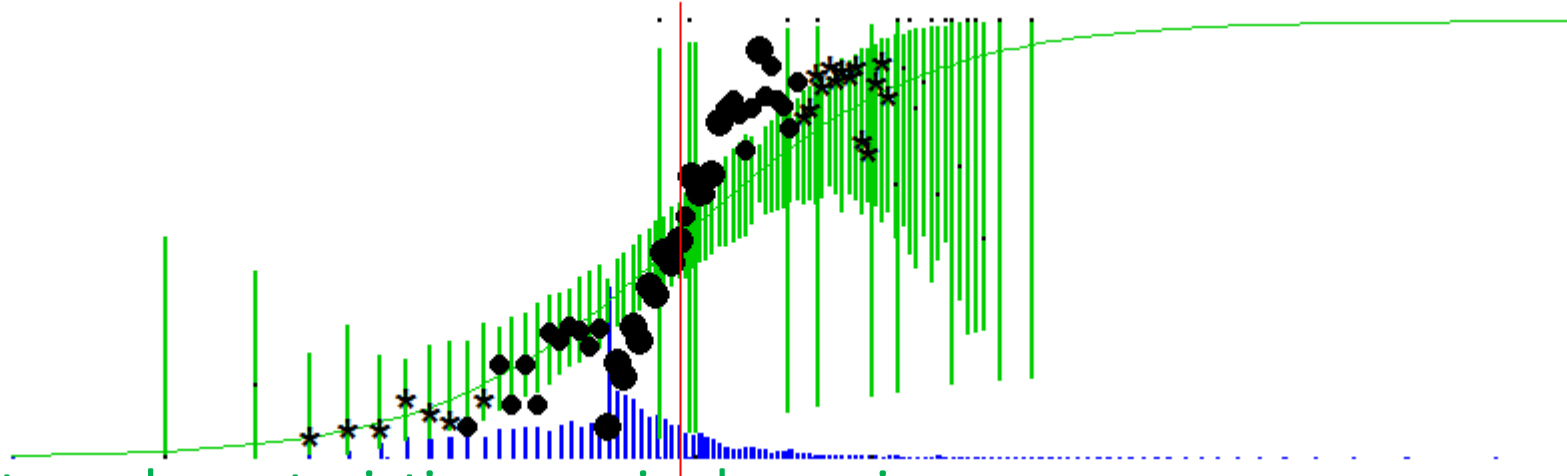
10 point scale	5 point scale
9	5
7, 8	4
4, 5, 6	3
2, 3	2
1	1
0	0

# Item level output



Mark category	Freq	Infit	Outfit
5	161		
4	693	.345	.175
3	1701	.579	.372
2	2118	.690	.552
1	2217	.717	.615
0	1964	.655	.523

# Understanding the output for each mark category



The item characteristic curve is shown in green.

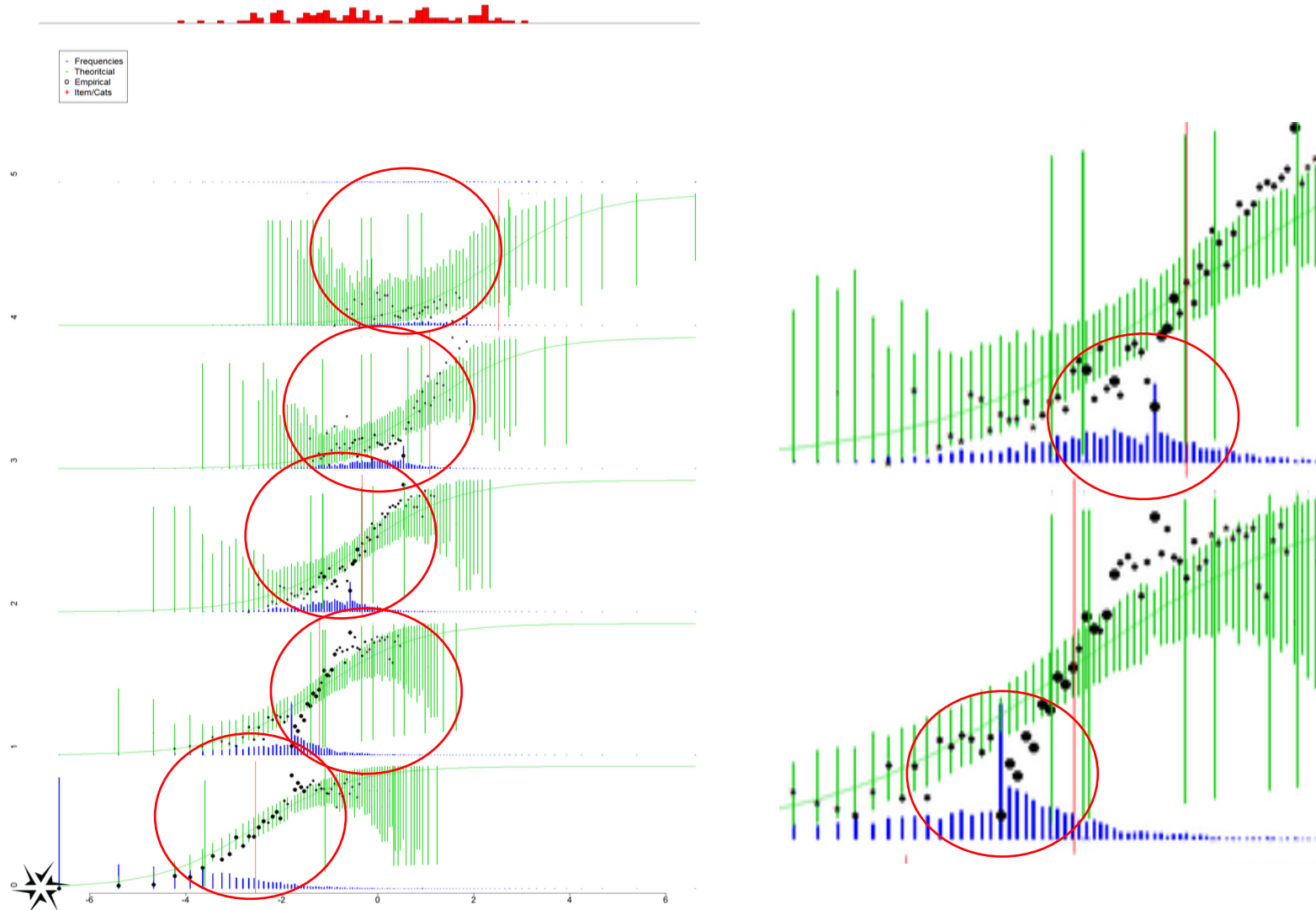
The blue 'histogram' shows the frequency of candidates of particular abilities who achieved that score.

The green vertical bars show the confidence intervals (1.96 SD) around the expected theoretical probabilities.

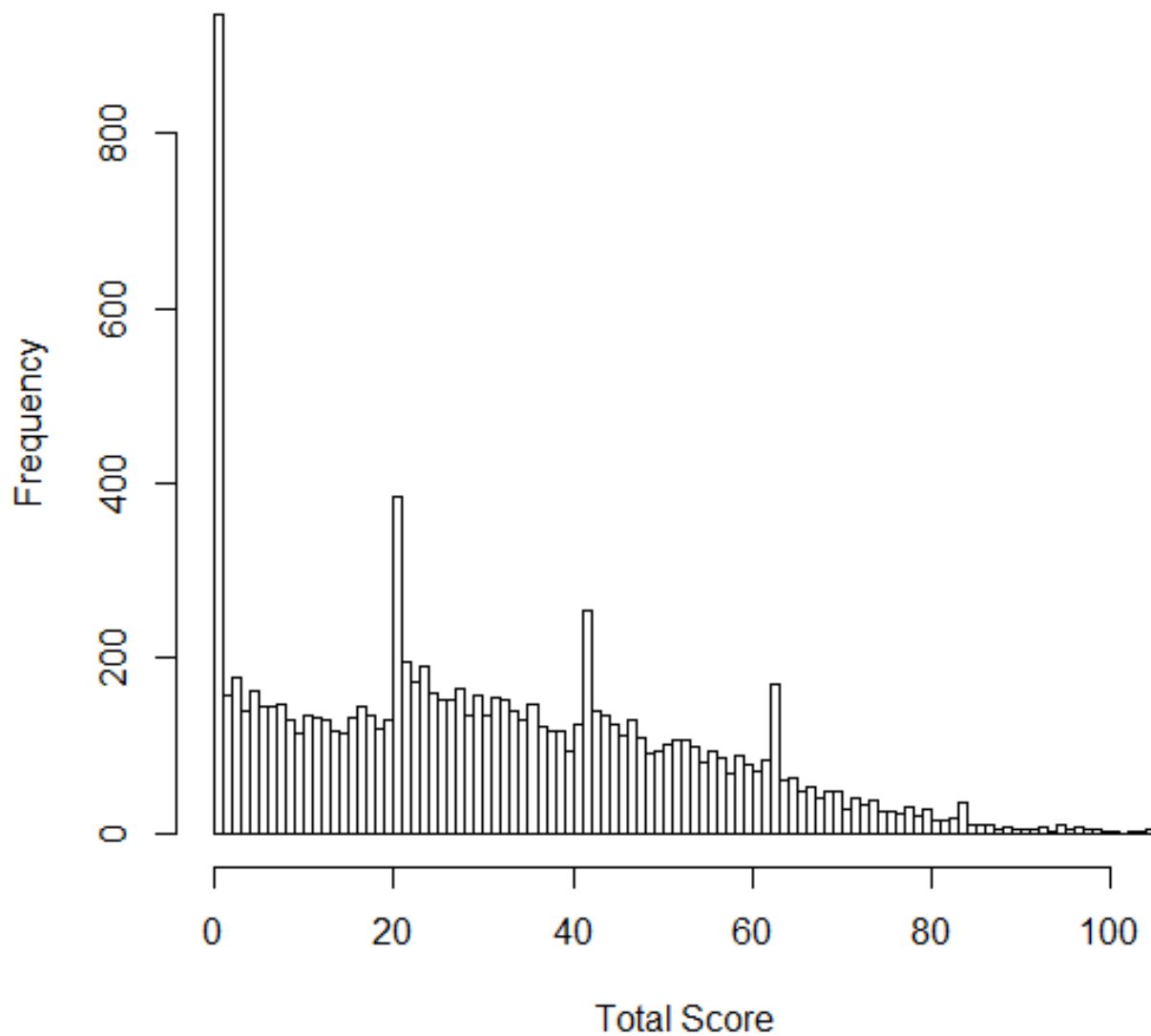
The black dots indicate empirical observations.

The red vertical line indicates the item difficulty.

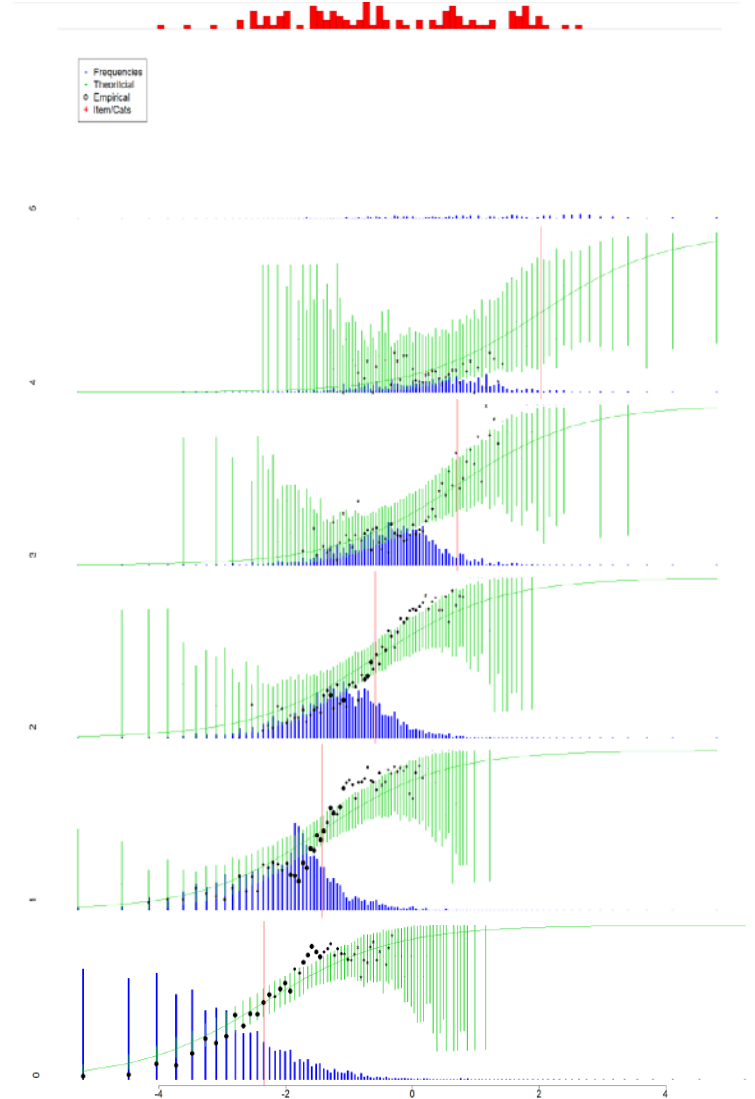
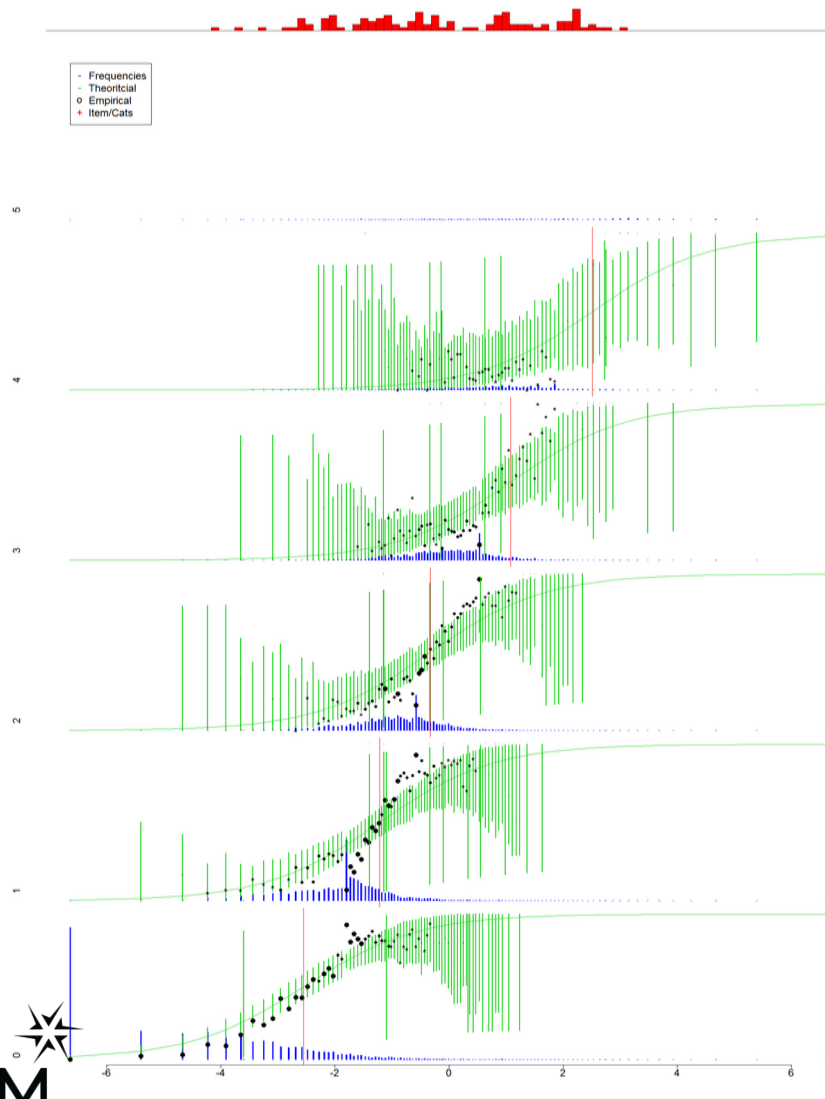
# Digging deeper into the data



# Histogram of scores



# Data before and after cleaning





# Feedback for assessment development

Results	Interpretation	Possible action
Some categories with very low frequencies	Too many categories	Collapse number of categories
Thresholds mis-ordered	$\uparrow$ score $\neq$ $\uparrow$ trait	Collapse number of categories
Overfit in mark category 1	Responses are too predictable	Use some items as a screener – don't use all items for all children
Systematic response patterns {1,1,1,1...1,1}	'Just clicking'	Change format of responses

# Future Developments

- Increased robustness
- Wider availability
- DIF
- Factor analysis

# Thank you

[john.little@cem.dur.ac.uk](mailto:john.little@cem.dur.ac.uk)

[sarah.gott@cem.dur.ac.uk](mailto:sarah.gott@cem.dur.ac.uk)