

Investigating the Comparability of Examination Difficulty Using Comparative Judgement and Rasch Modelling

Steve Holmes, Qingping He
and Michelle Meadows



The context

- GCSEs taken in England by students aged 16 are currently being reformed with respect to content and the associated assessment.
- Examinations in the same subject areas are provided by several exam boards.
- There is no pre-testing or equating between exams due to security considerations and their high-stakes nature.
- Small differences in difficulty can be dealt with by adjusting grade boundaries. Large differences may wash-back on teaching and learning.
- A need arose for Ofqual to evaluate reformed GCSE Maths sample assessment difficulties.

Aim of study

To explore the potential of using comparative judgement and Rasch modelling to investigate the relative difficulty between examinations

Method

- Items in six mathematics question papers designed by three exam boards (Boards A, B and C) for 16-year olds in England were judged in paired comparison by experts. Three of the papers were for the more able students (Higher Tier) and the other three for the less able students (Foundation Tier)
- The Rasch model for dichotomous items was applied to the paired comparison data to establish the scale of expected difficulty.
- The six papers were also taken by 2933 students using an equivalent-groups design for each tier, allowing the difficulties of the items to be compared and placed on the same measurement scale using the Partial Credit Model (PCM).
- The actual item difficulties derived from the test data using the Partial Credit Model were compared with the expected item difficulties derived from the comparative judgement data and the Rasch model to validate the comparative judgement approach.

The online comparative judgement process

◀ Left

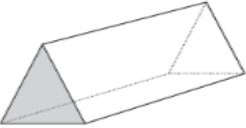
Which question is the more mathematically difficult to answer

Right ▶

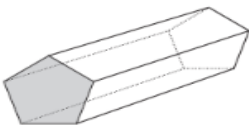
🏠

Page: 1 of 1 Automatic Zoom

These prisms are named after the shape of their end face.



Triangular



Pentagonal

(a) Complete this table.

Shape of end face	Number of faces	Number of edges	Number of vertices
Triangle (3 sides)	5	9	6
Rectangle (4 sides)	8
Pentagon (5 sides)	15	10
Hexagon (6 sides)	8	18

(b) How many edges and vertices does a prism with a 100-sided end face have?

Edges.....
Vertices.....

(c) Write down a formula connecting the number of faces F of a prism and the number of sides of its end face n .

(c).....

Page: 1 of 1 Automatic Zoom

Calculator allowed.

Person A is in a class of 28 students, 3 of whom are left-handed. There are 1250 students in the school.

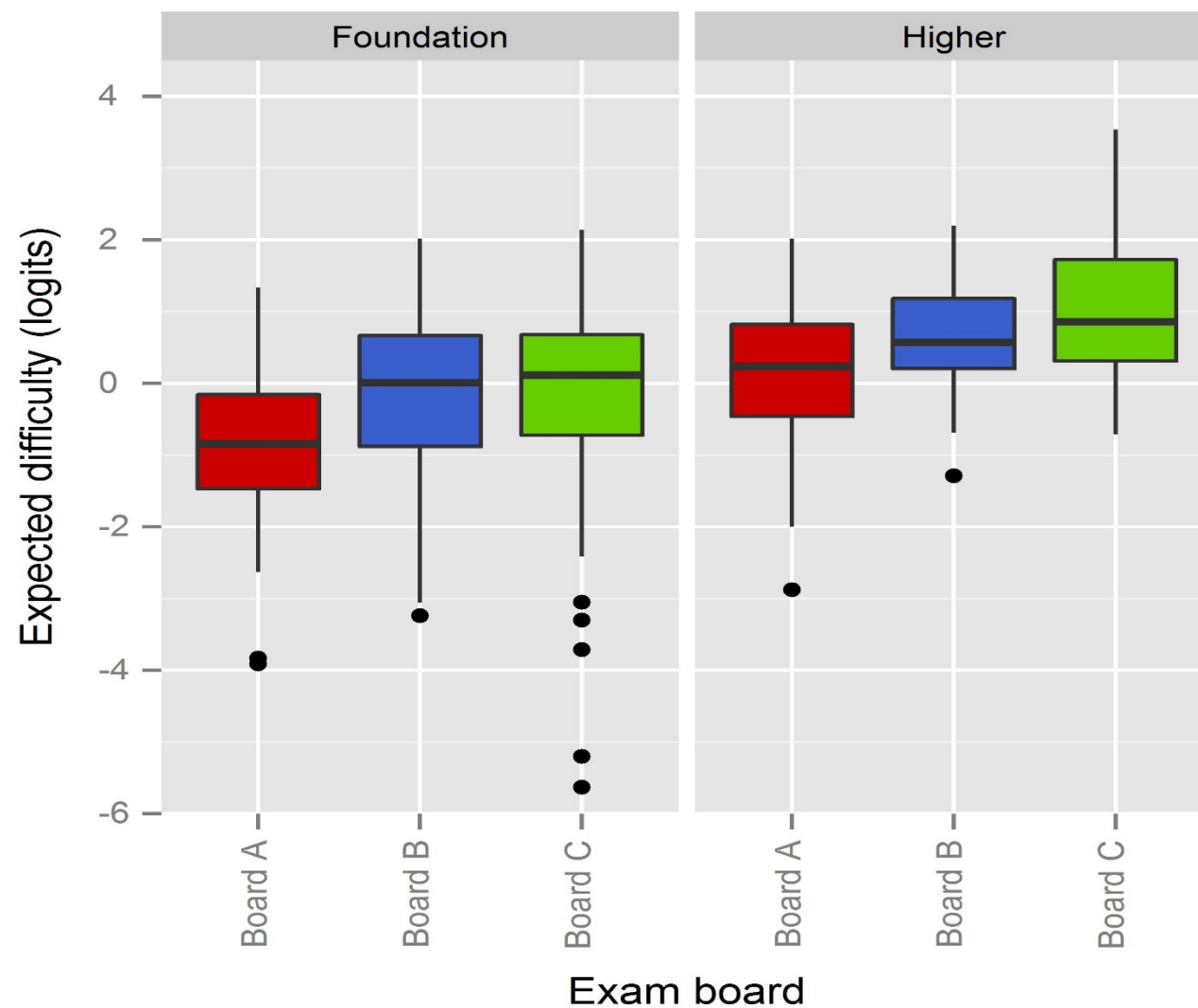
(a) Use this information to estimate how many students in the school are left-handed.

(a).....

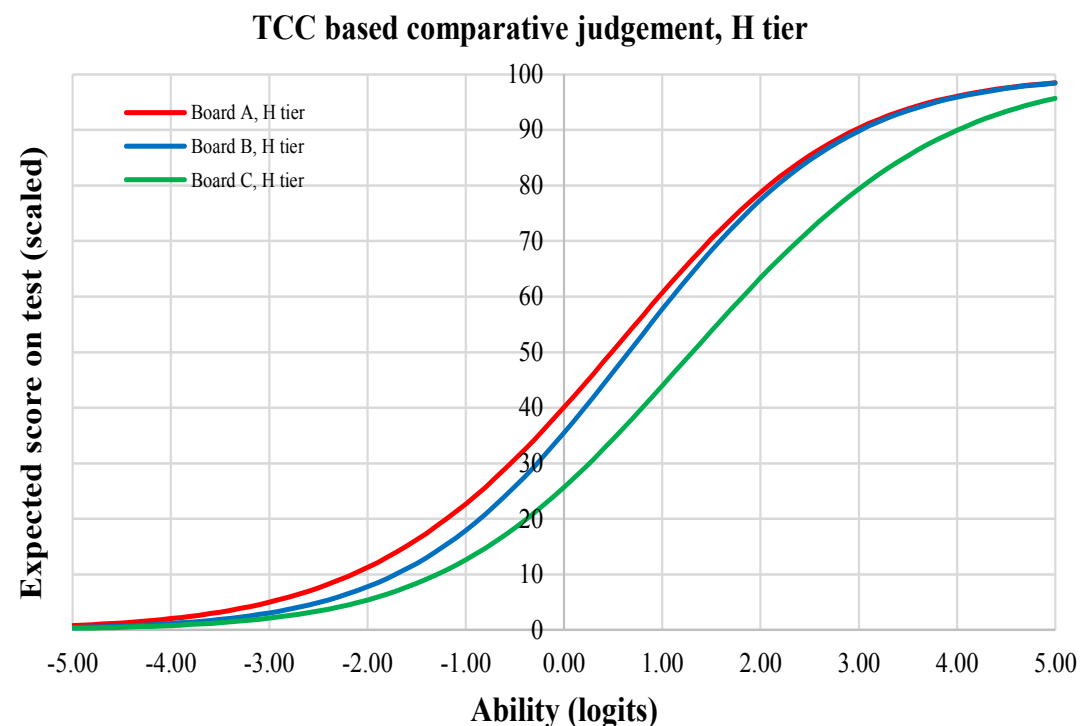
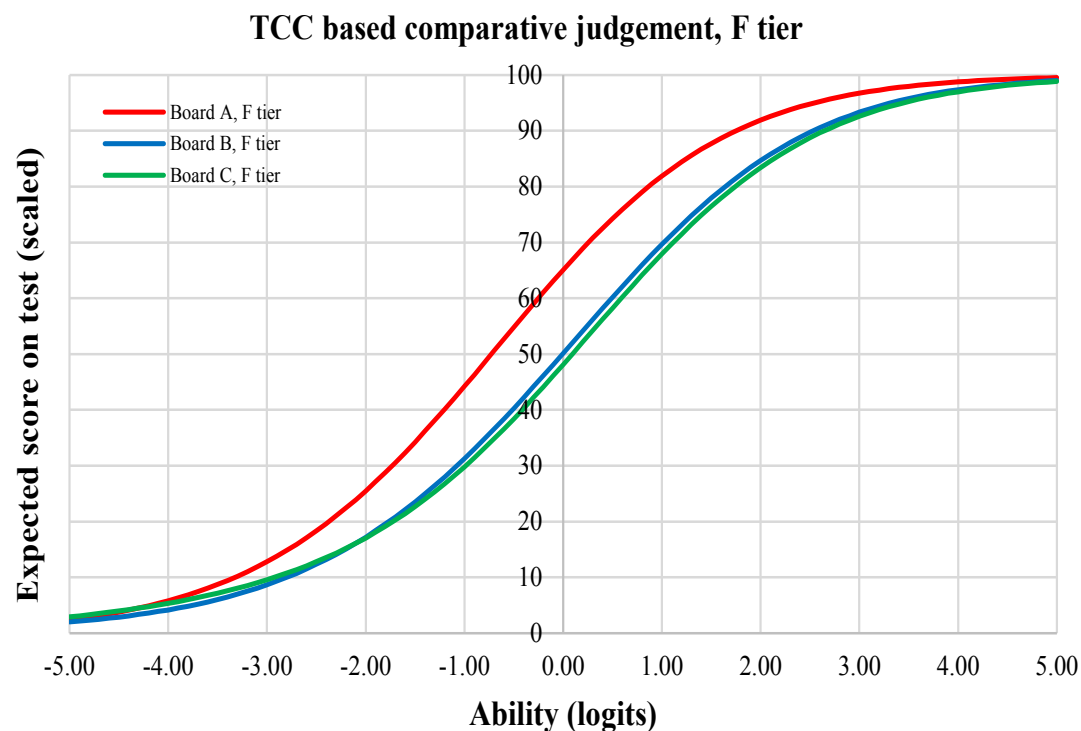
(b) Is your solution to (a) likely to be an overestimate or an underestimate? Explain your reasoning.

(c) Person B is at a different school. Person B is in a class of 26 students, 6 of whom are left-handed. Person B says to Person A, "In our two classes, there are 54 students, 9 of whom are left-handed. We can use this bigger sample to improve the estimate". What assumption has Person B made? Explain whether you think that Person B's argument is correct.

Distribution of expected item difficulties derived using comparative judgement and the Rasch model



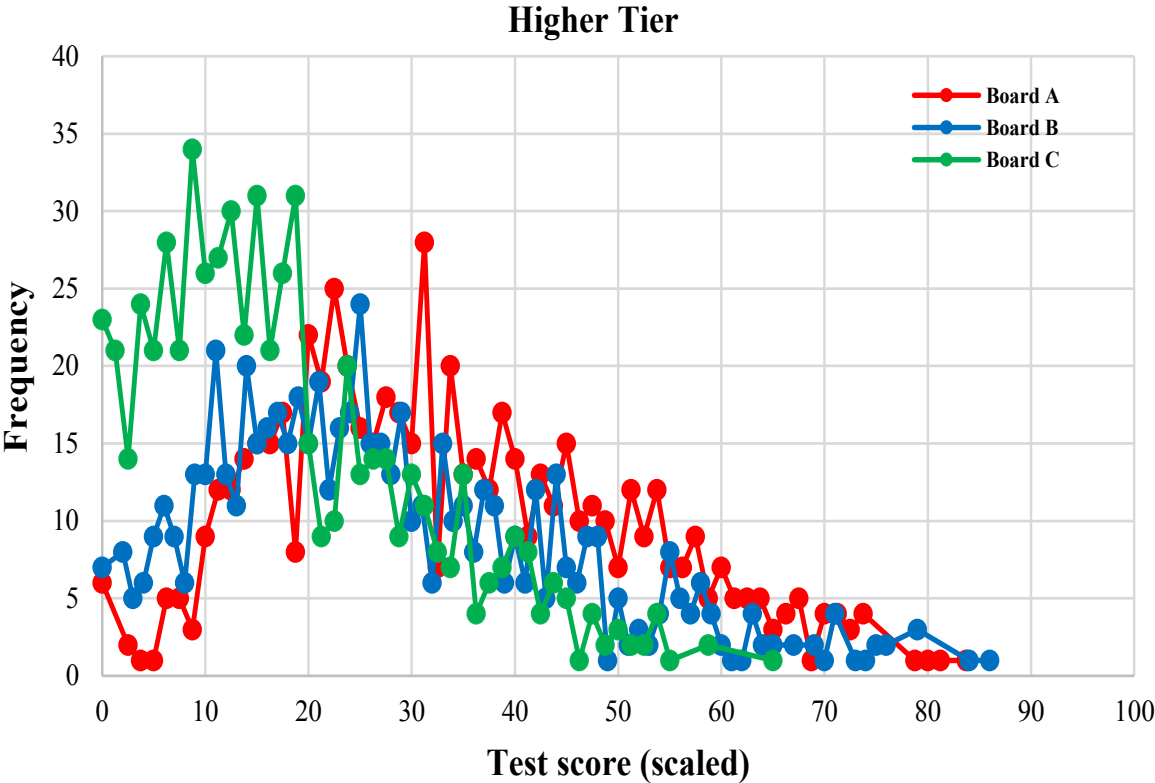
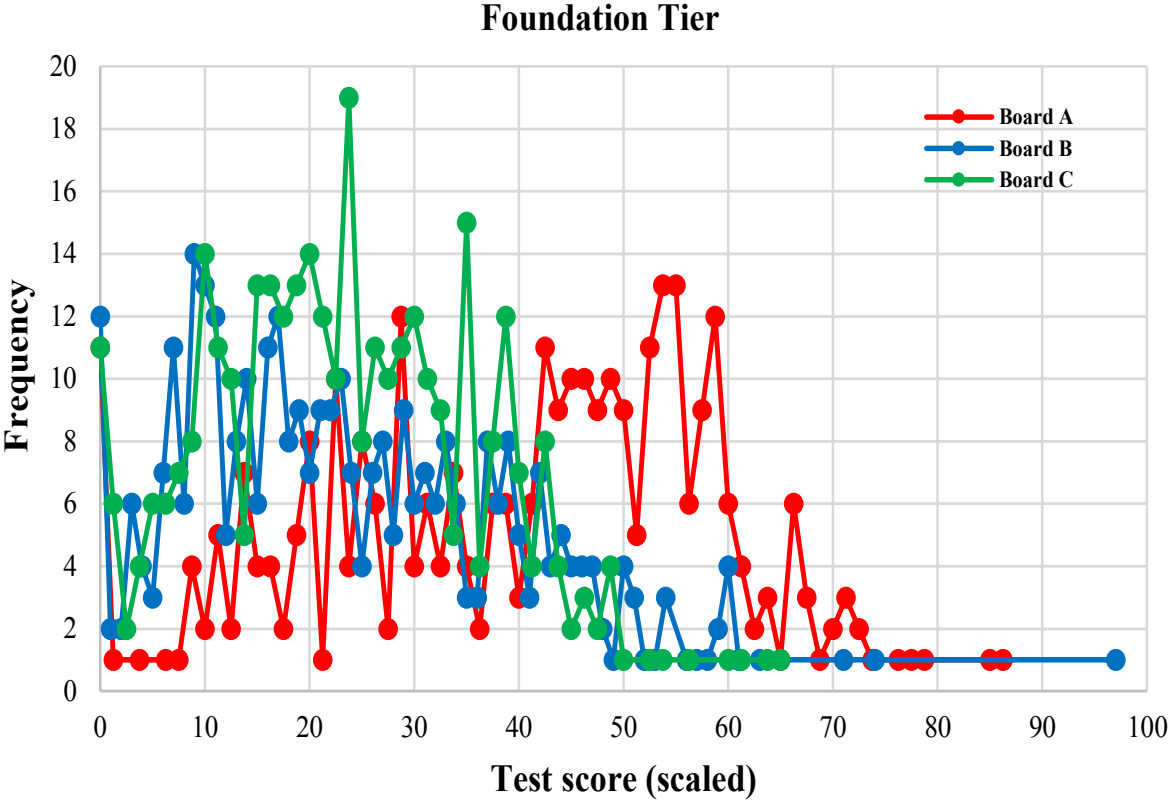
Test characteristic curves (TCCs) derived using comparative judgement data and Rasch model for dichotomous items



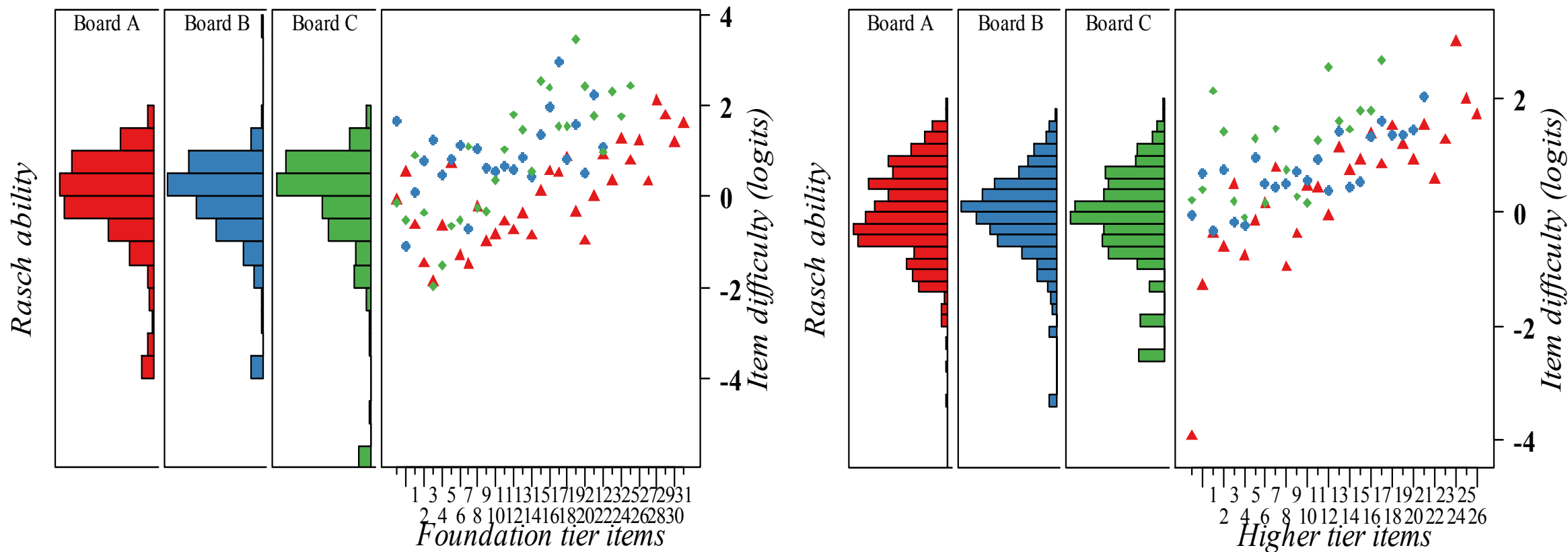
Question paper statistics based on test data using classical test theory

	Board A		Board B		Board C	
	F	H	F	H	F	H
Number of students	325	618	326	648	353	627
Number of items	44	37	48	36	37	28
Scaled maximum available mark	100	100	100	100	100	100
Mean scaled score	40.16	33.91	24.44	27.98	23.62	18.04
Standard deviation of scaled score	18.53	16.89	15.82	16.96	13.14	13.01
Cronbach's alpha	0.90	0.88	0.87	0.88	0.84	0.83
McDonald's omega_t	0.91	0.90	0.90	0.90	0.88	0.85

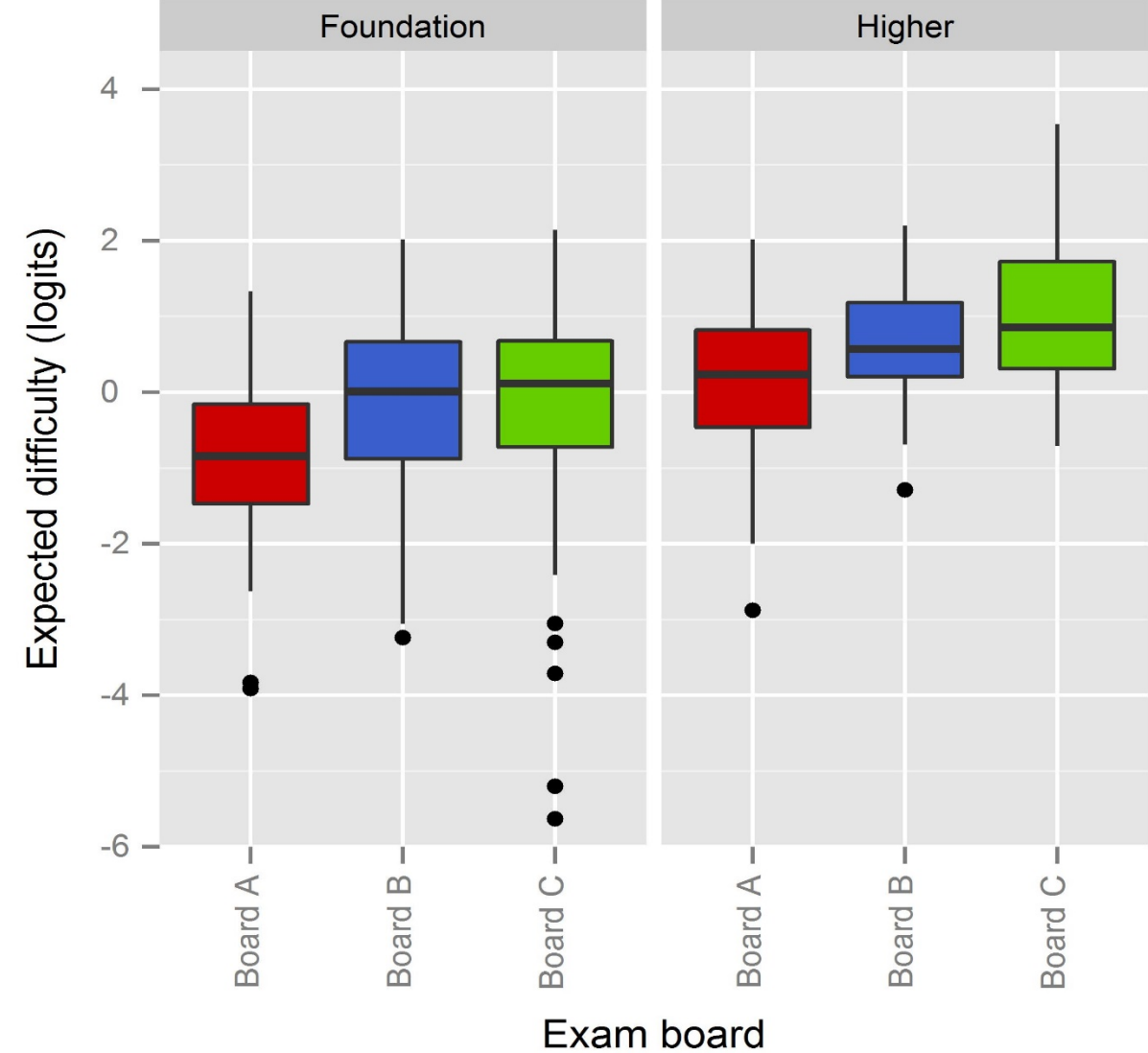
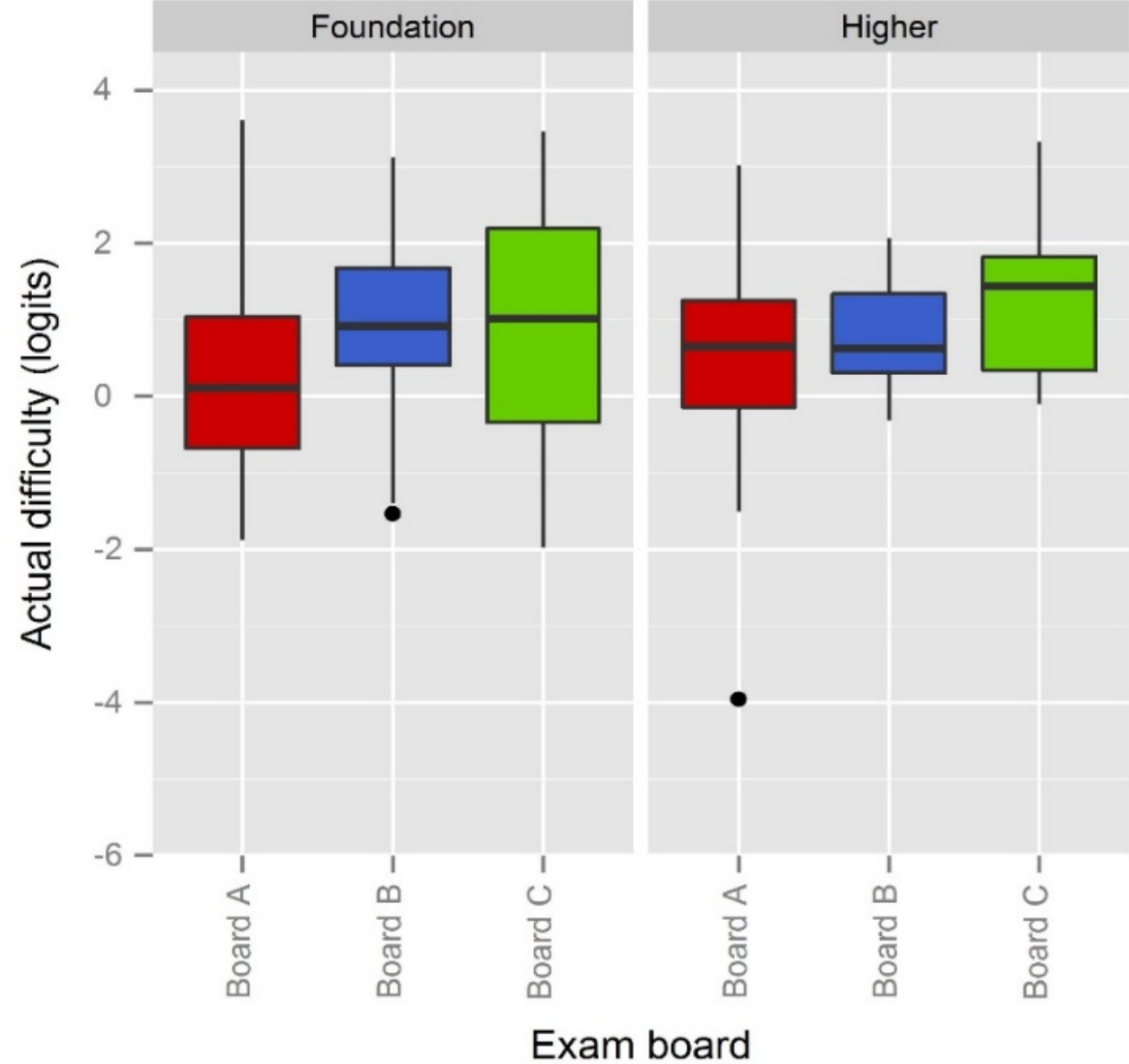
Distributions of scores (scaled to 100)



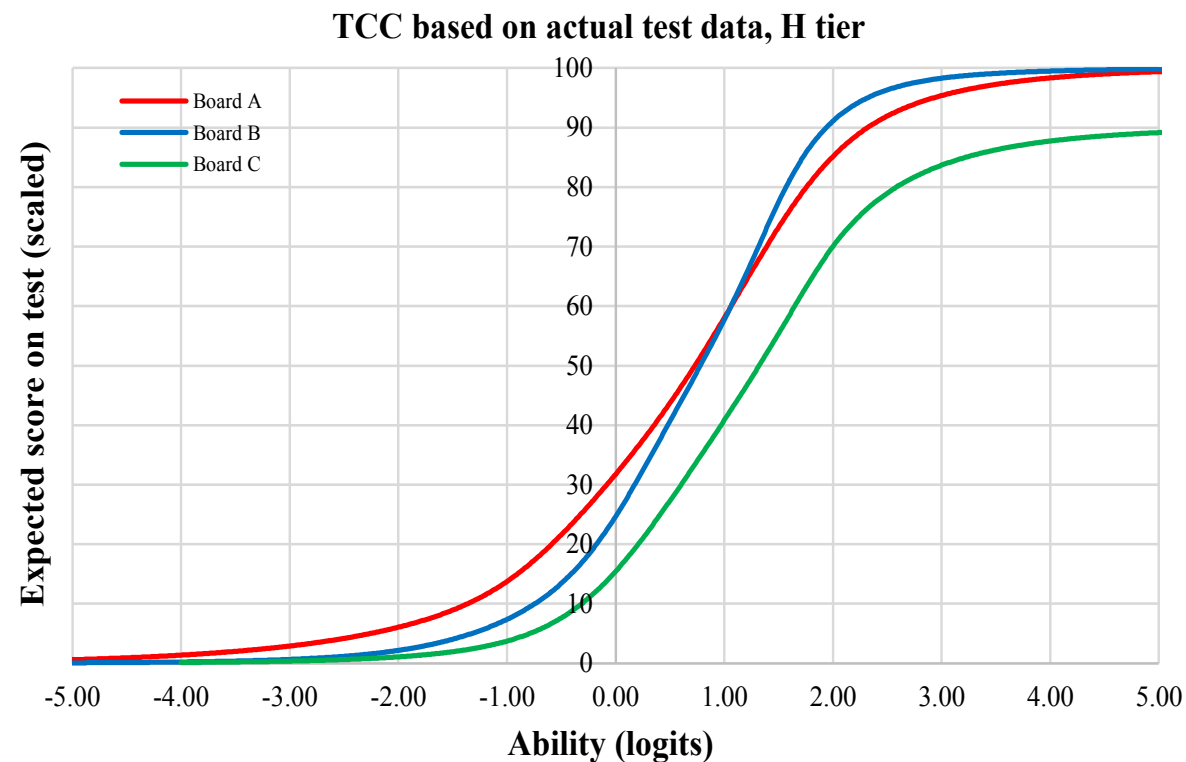
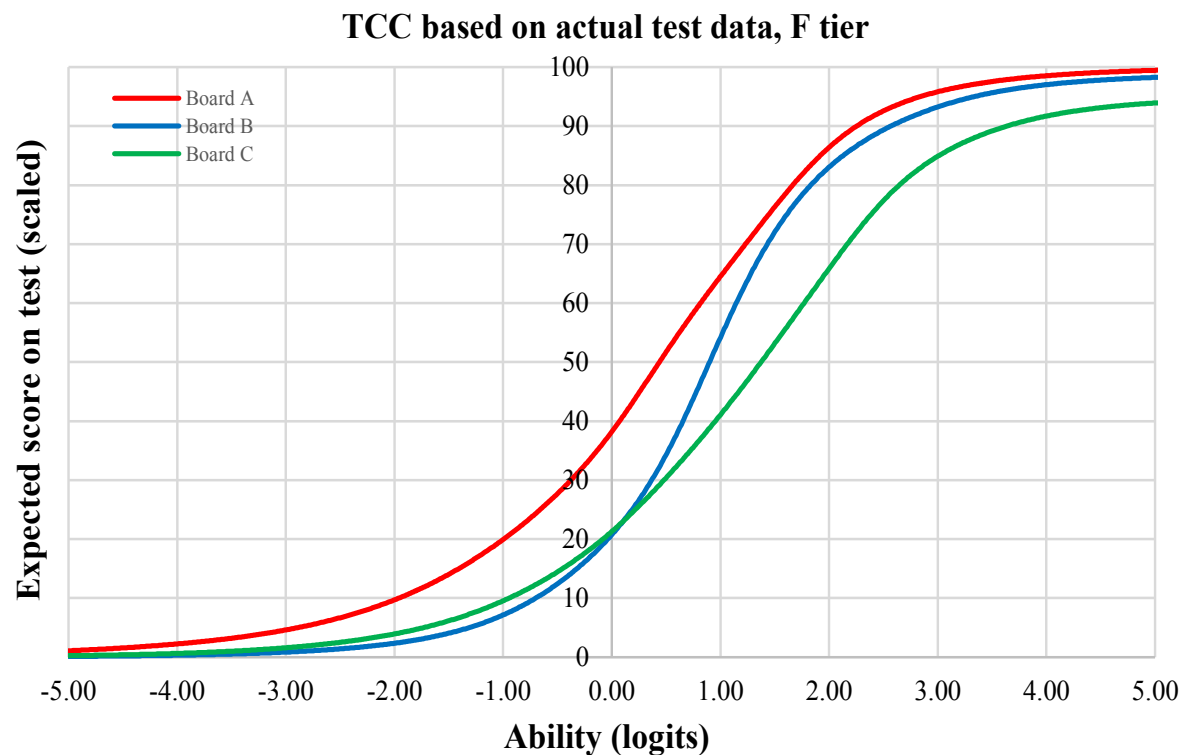
Ability and difficulty distributions based on PCM analysis of test data



Distribution of observed item difficulties derived using test data and PCM

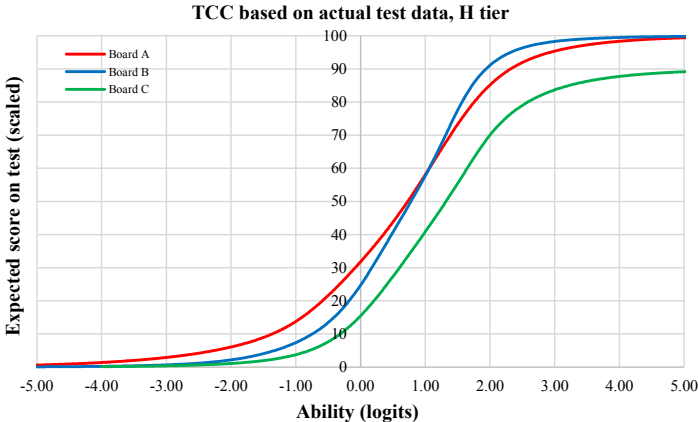
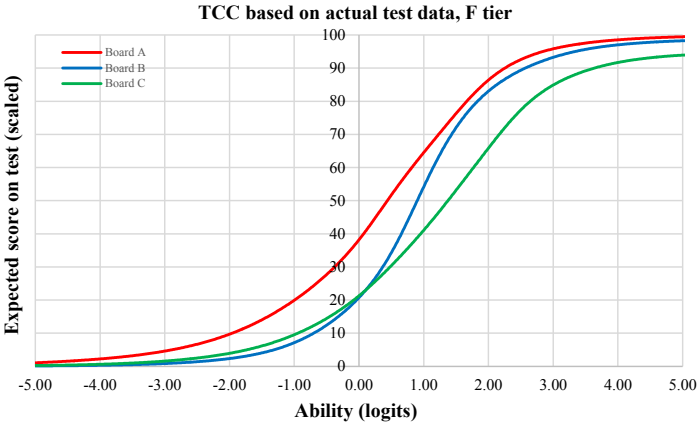


Test characteristic curves (TCCs) based on PCM analysis of test data

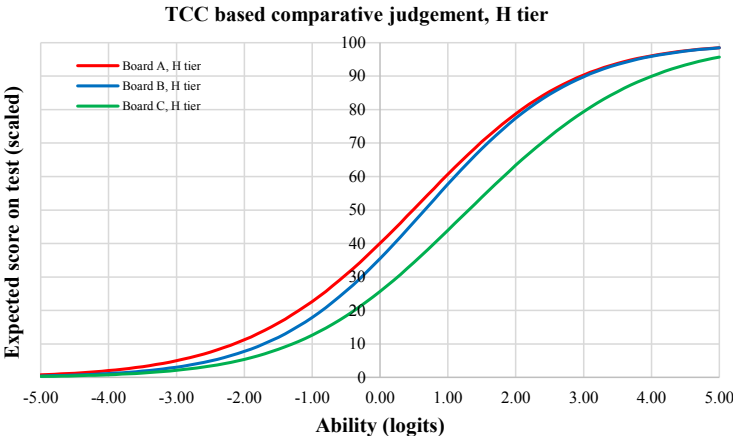
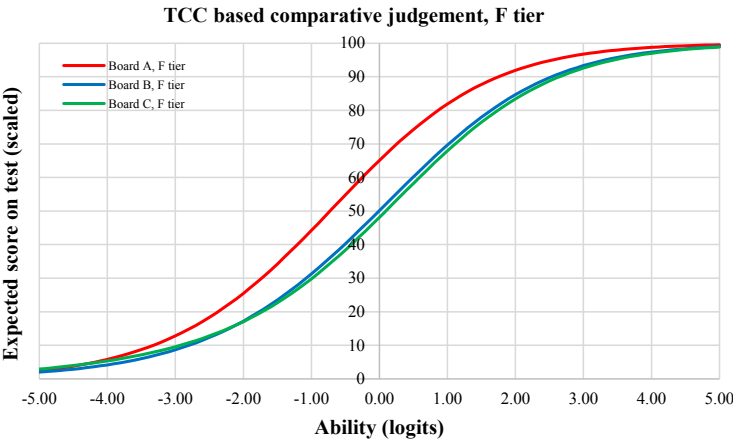


Comparison of Test characteristic curves (TCCs)

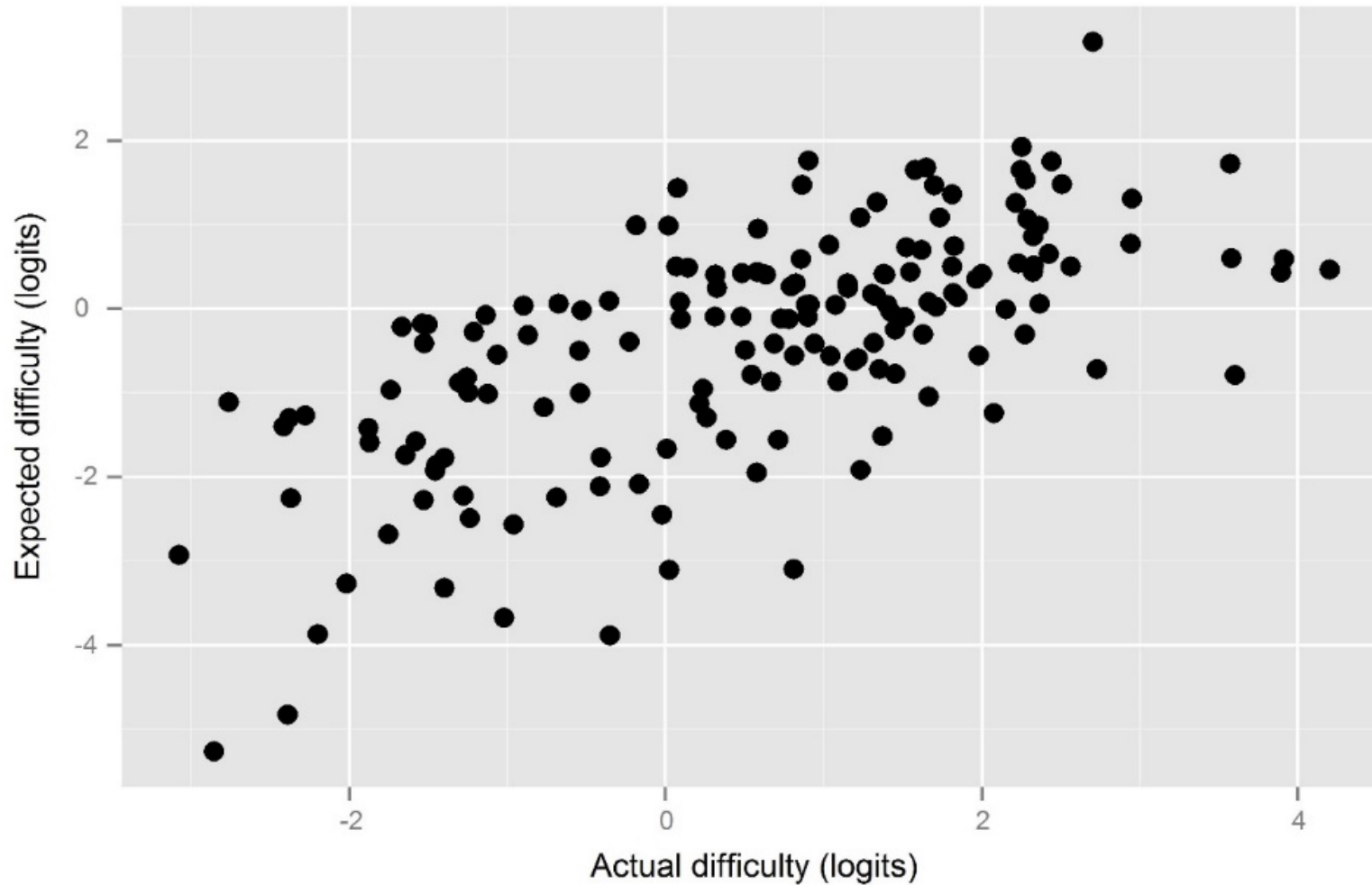
Test Data



CJ Data



Comparison of expected item difficulties derived using comparative judgement data and observed difficulties derived using test data



Correlation of 0.66 and disattenuated correlation of 0.76

Concluding remarks

- The expected item difficulties derived using the comparative judgement data and Rasch model and the actual item difficulties derived using the test data and PCM were reasonably strongly correlated.
- Comparative judgement appears to be an effective way to investigate the comparability of difficulty between examinations
- It could be used as a proxy for pretesting high-stakes tests in situations where pretesting is not feasible.
- There may be scope for refining the judging criteria.
- We have indeed used the comparative judgement approach alone to assess difficulty in a few assessments.

Thank you