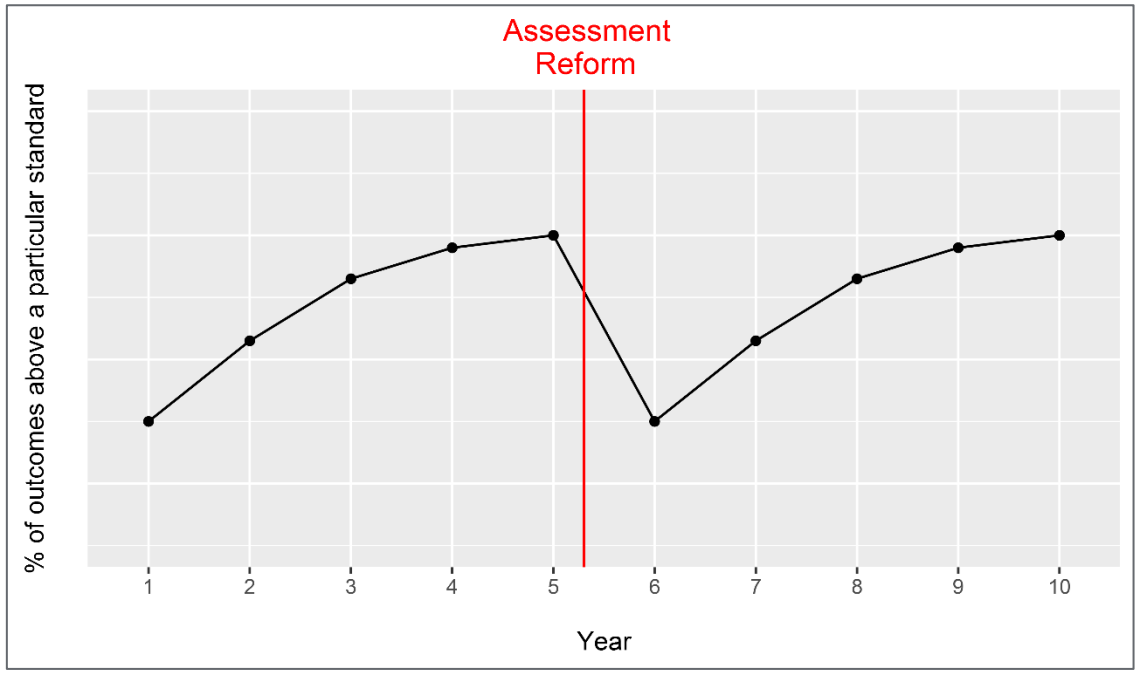
Using a comparative judgment methodology to investigate the ‘Sawtooth Effect’ in UK secondary school assessments

Ben Cuff
31/03/2017

The Sawtooth Effect

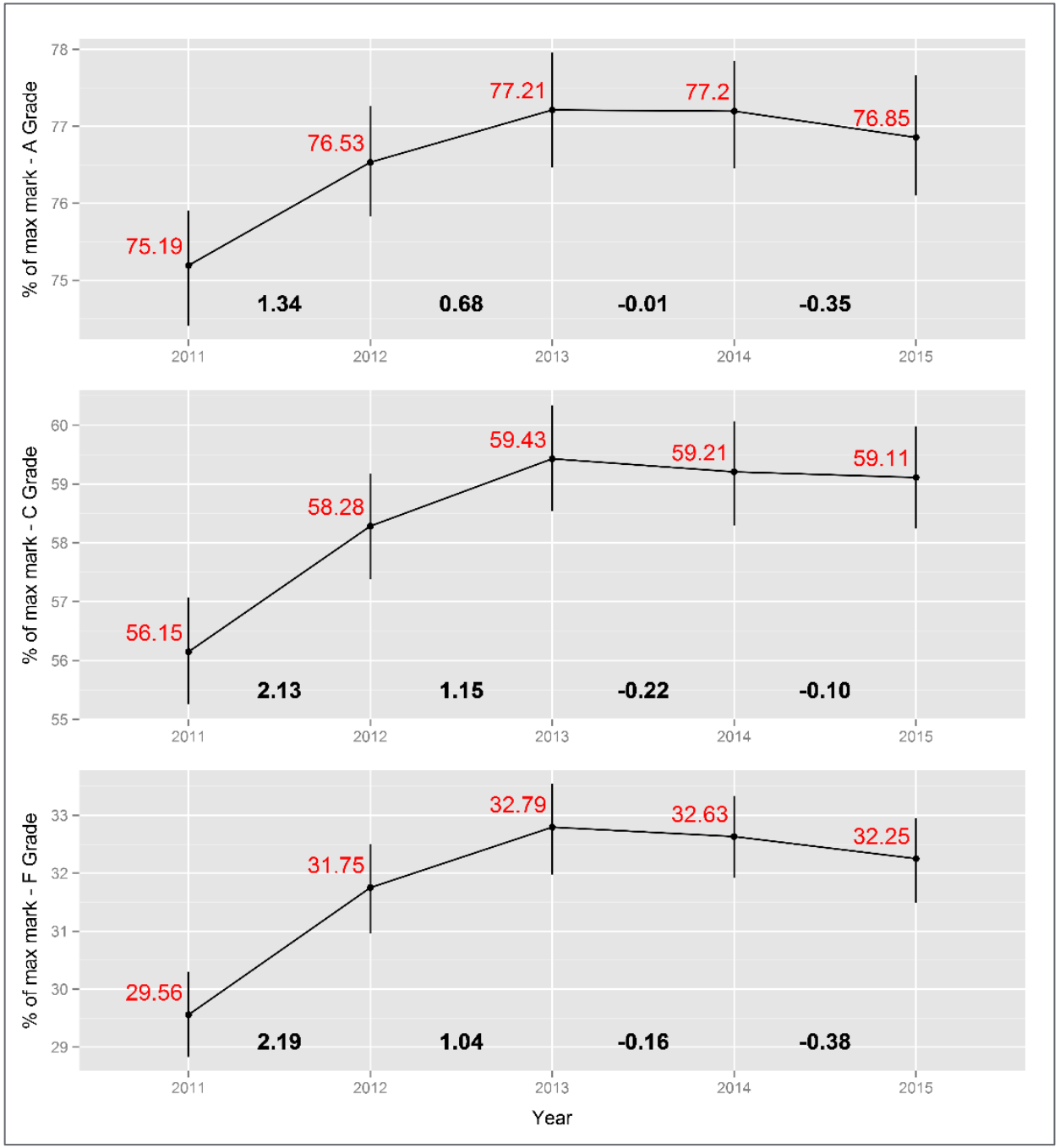
- Pattern of performance change following reform:
  - Relatively rapid changes in performance over the first few years of a new (unfamiliar) test
  - Less rapid changes once students / teachers become familiar with the new test



Grade Boundaries as a Proxy Measure for Performance Change

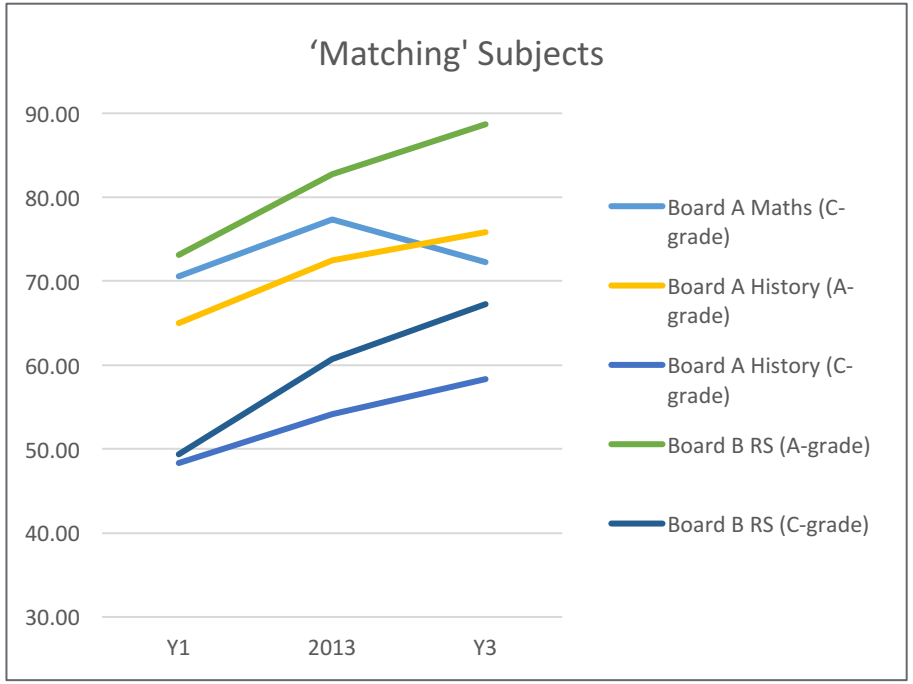
- We sought to investigate the nature of this effect in the UK (limited evidence)
- We couldn't observe changes in outcomes directly, because outcomes are made comparable over time, against prior attainment
- Changes in *test-specific* performance = changes in grade boundaries
  - T-S Performance goes up, boundaries go up
  - T-S Performance goes down, boundaries go down

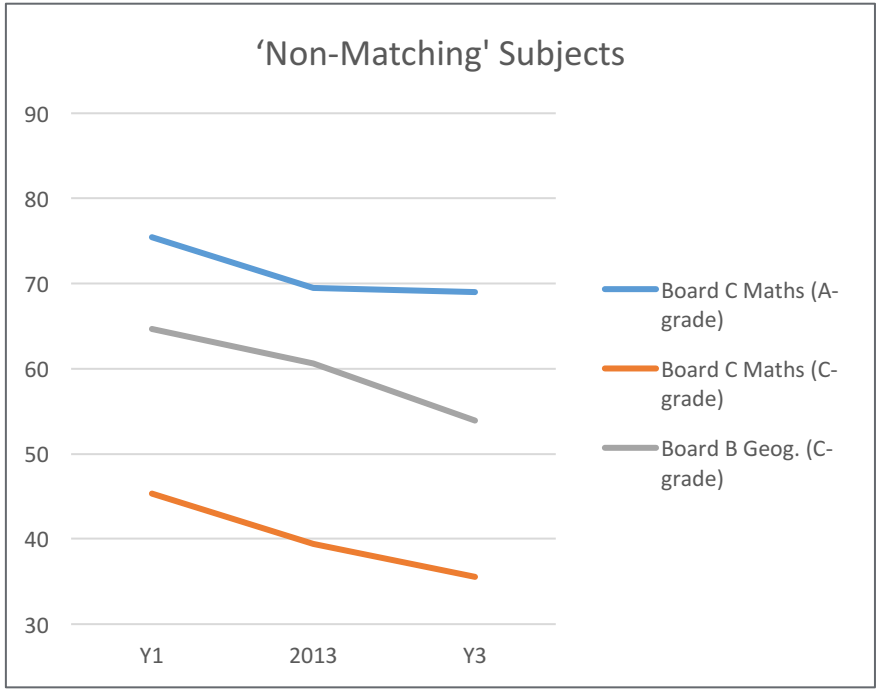
Study 1 –
GCSE Grade Boundaries



Study 2
Design

- 5 GCSE specifications (2 x maths, geography, history, RS)
  - Patterns of mean grade boundaries (Examined units only):





- 'Thurstone Triples' comparative judgement design
  - Rank order 3 scripts – "best", "middle", "worst"

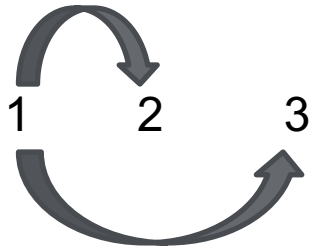
Materials and Judges

- 5 scripts per unit per year (that fall on each grade boundary)
  - Board A Maths (3 unit) – 45 [C]
  - Board C Maths (3 unit) – 45 [C] + 45 [A]
  - History (2 unit) – 30 [C] + 30 [A]
  - RS (2 unit) – 30 [C] + 30 [A]
  - Geography (3 unit) – 45 [C]

- Each script fully ‘cleaned’ of…
  - All marker’s annotations
  - Anything that identified the board or year

- 6 judges per subject area

Pack Design

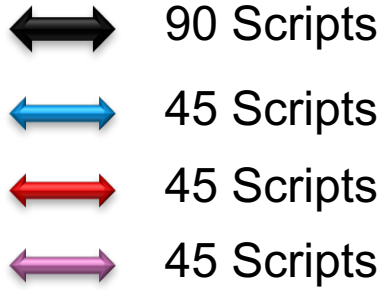


- ‘Thurstone Triples’ design (packs of 3)
  - Each script appears in 15 packs (= 30 paired comparisons per script)
- We wanted to conduct analyses at a subject level, rather than a unit level
  - The Sawtooth Effect doesn’t matter if patterns ‘cancel out’ when unit marks are aggregated to arrive at subject level outcomes
- Estimates of script quality (theta scores) produced by Rasch analyses are on an arbitrary scale (dependent upon the items included)
  - If we treated units separately, the scales would be different
  - Needed some ‘linking packs’ across the different units

Pack Design

- Cross-unit comparisons would be difficult (eg different content areas), so we needed to ease the judges in gently

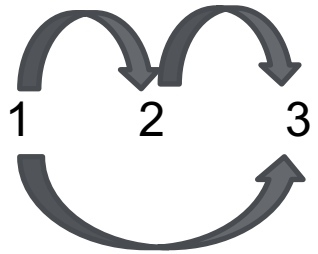
	2011	2013	2015
Unit 1	5 Scripts	5 Scripts	5 Scripts
Unit 2	5 Scripts	5 Scripts	5 Scripts
Unit 3	5 Scripts	5 Scripts	5 Scripts



Method

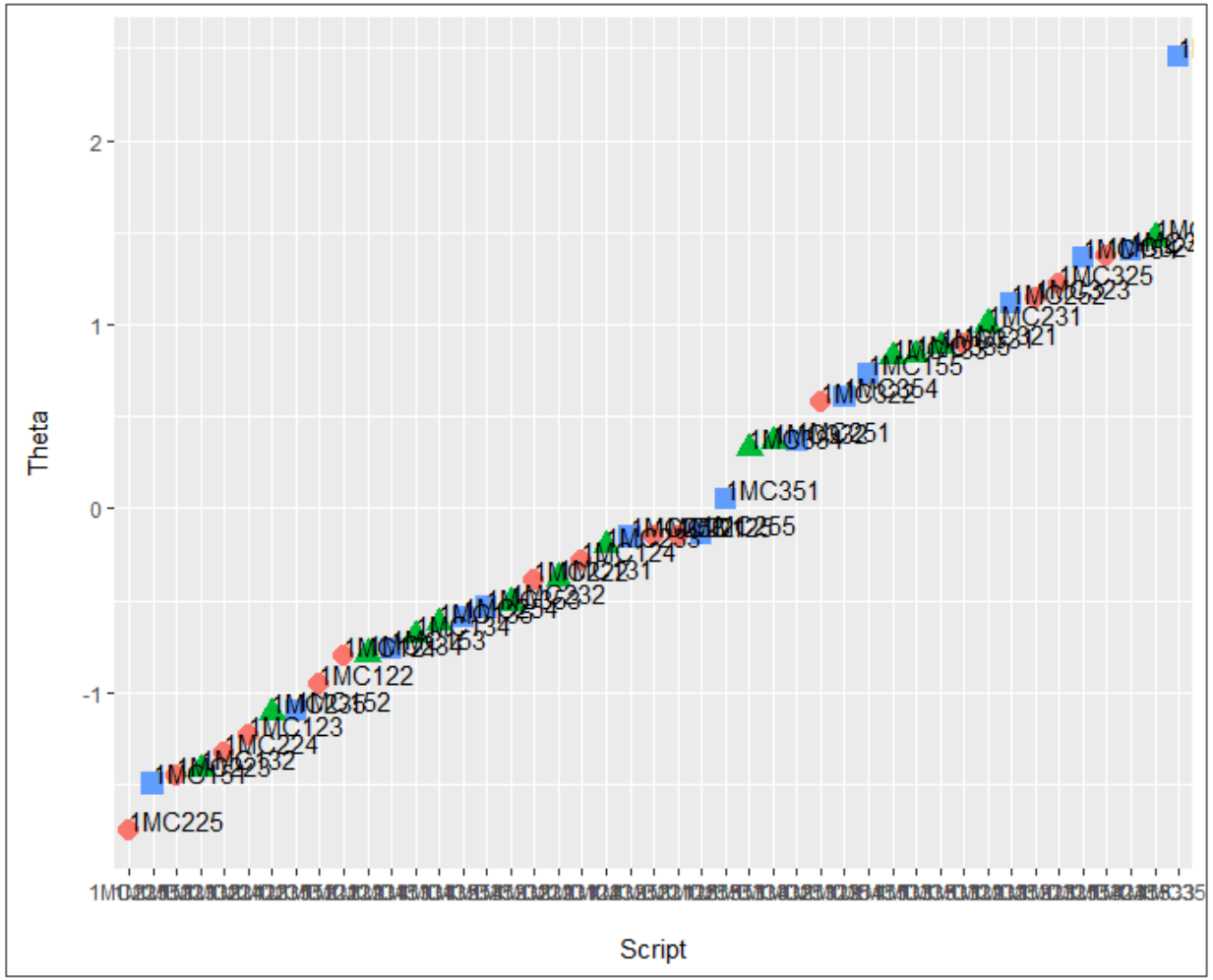
- Started with a short familiarisation task
  - Rating blank question papers on difficulty

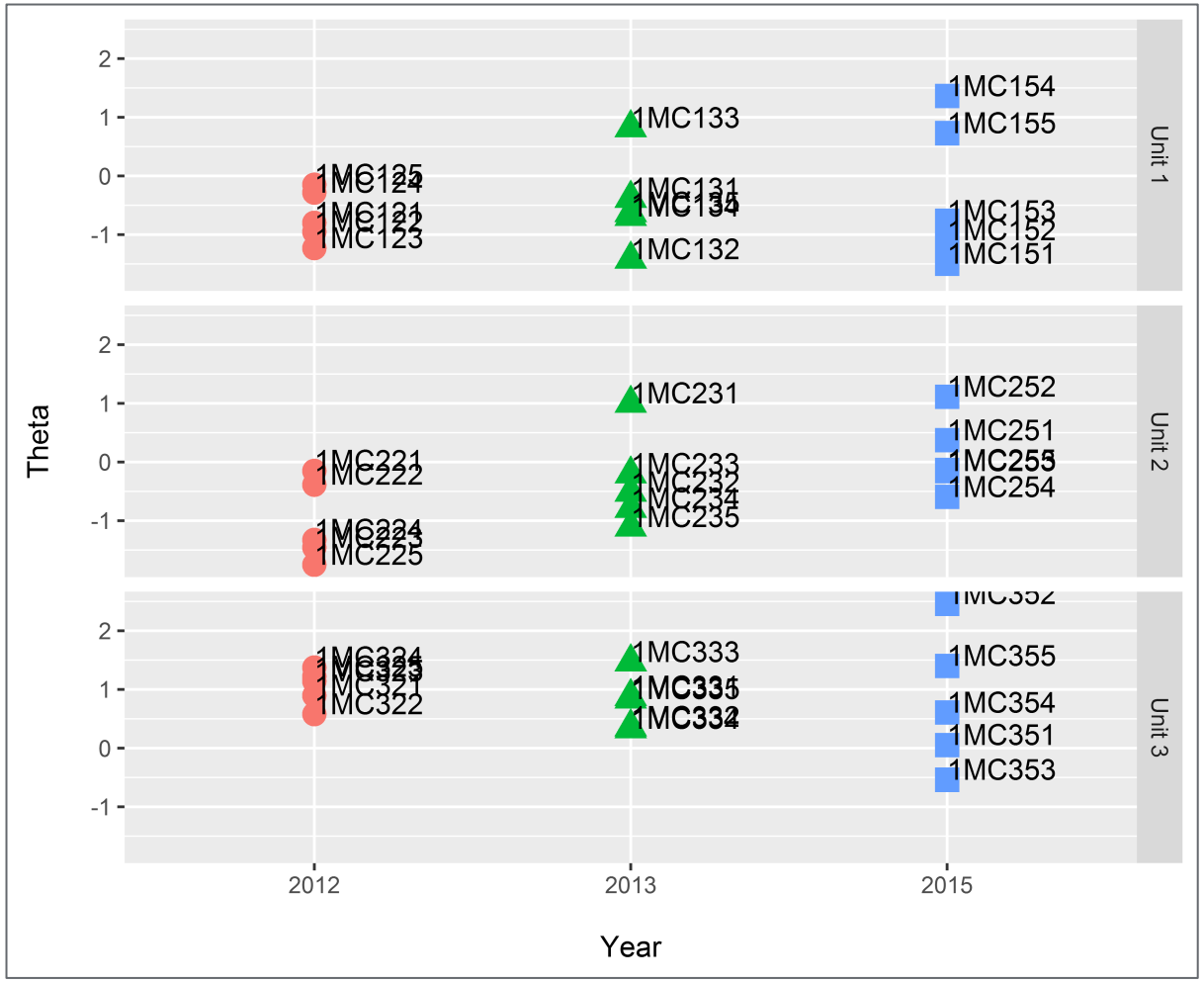
- Script judgements
  - “While taking the difficulty of each paper into account, please rank these three scripts in order of best to worst, in terms of the quality of each student’s work… In essence, which is the better mathematician?”



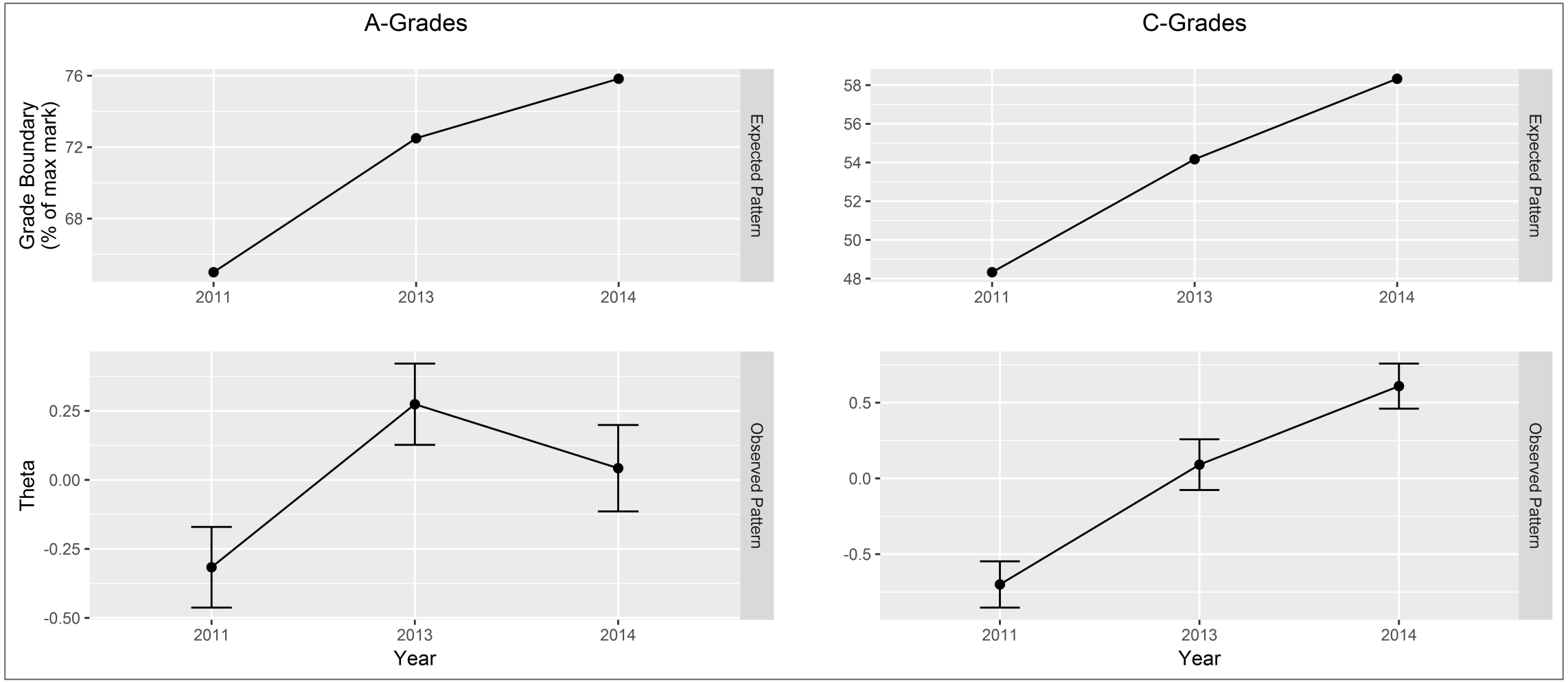
- After we got the judgments back…
  - Each triples comparison was converted into 3 paired comparisons (win/loss)
  - Analysed with Rasch (Bradley-Terry)

EB1 Math (Foundation Tier)

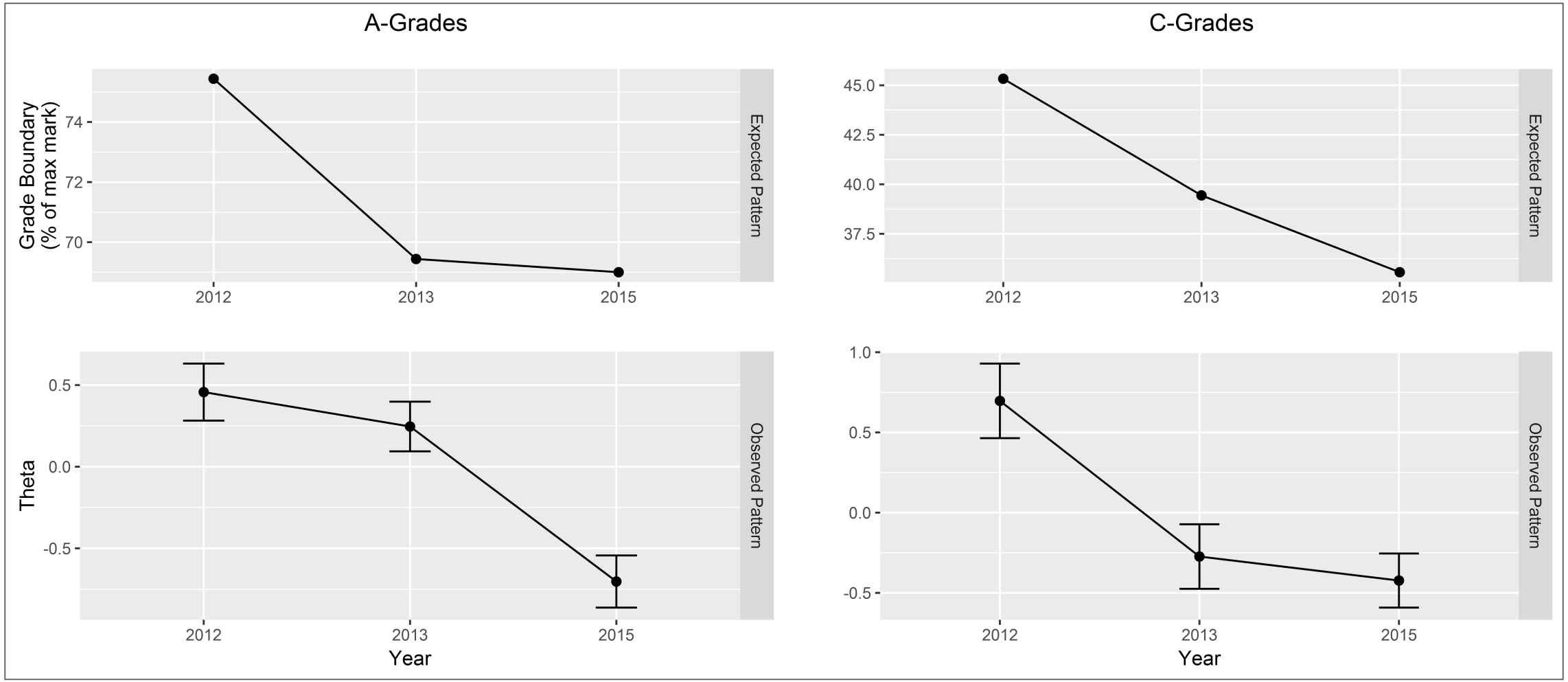




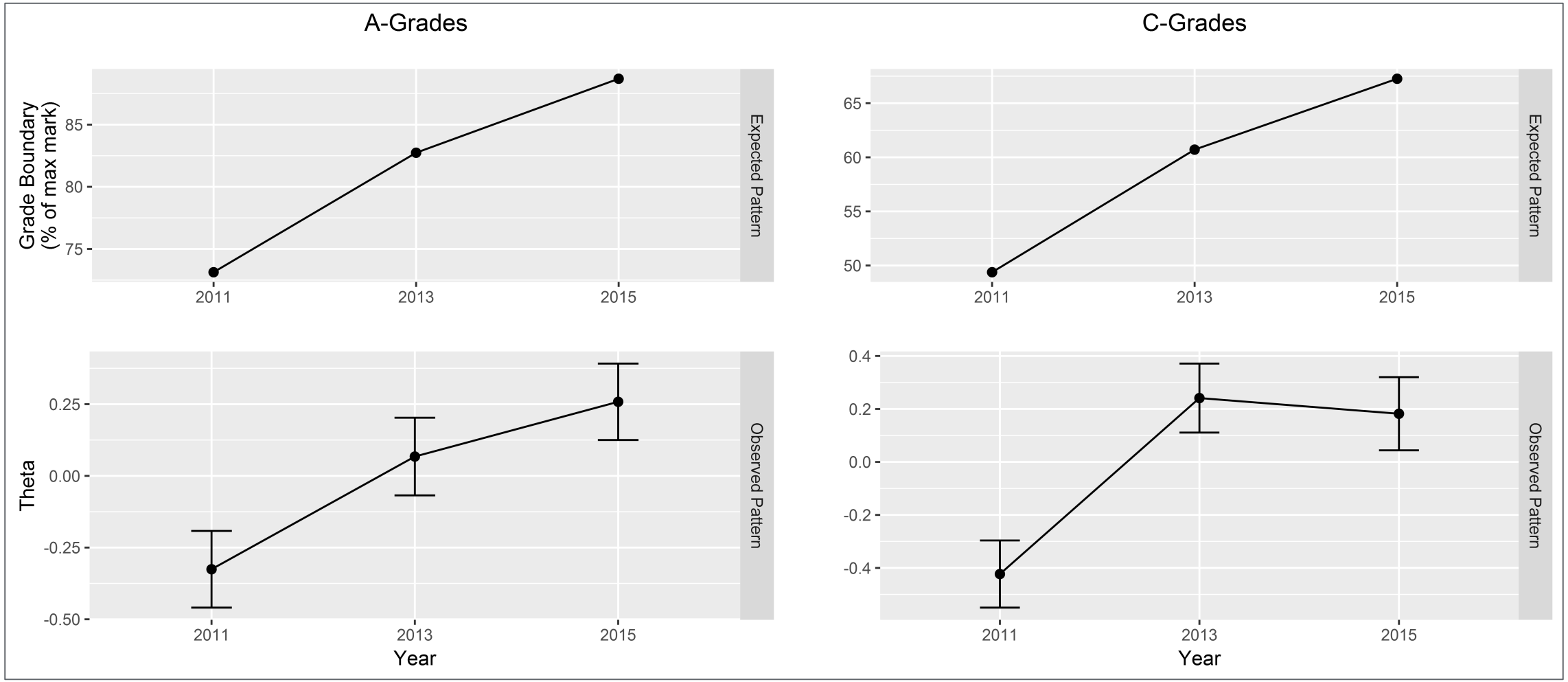
History


A-Grades
C-Grades
Grade Boundary (% of max mark)
76
72
68
58
56
54
52
50
48
2011
2013
2014
Expected Pattern
Theta
0.25
0.00
-0.25
-0.50
0.5
0.0
-0.5
Observed Pattern
Year

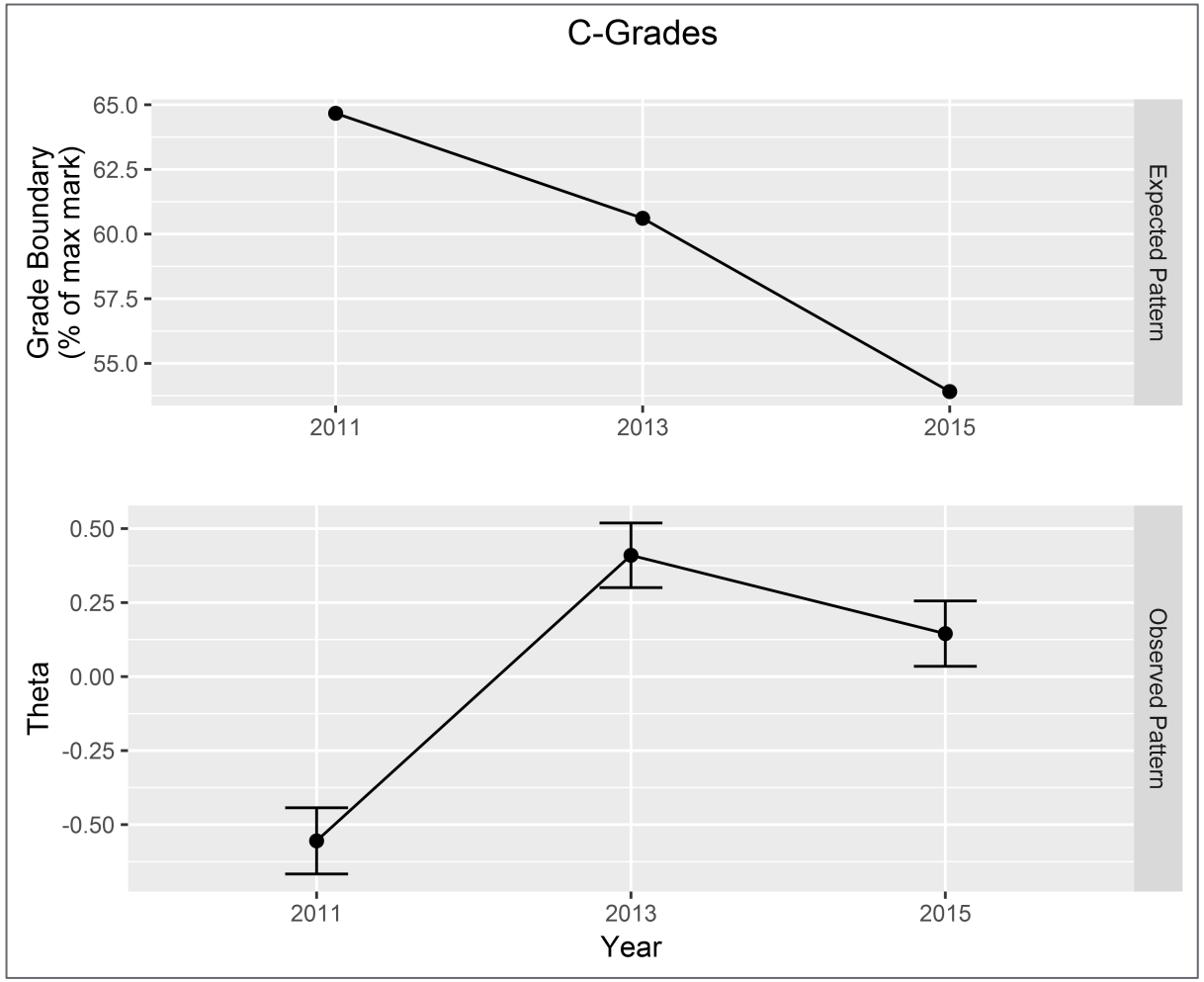
EB2 Maths


A-Grades
C-Grades
Grade Boundary (% of max mark)
74
72
70
45.0
42.5
40.0
37.5
2012
2013
2015
Expected Pattern
Theta
0.5
0.0
-0.5
1.0
Observed Pattern
Year

Religious Studies



Geography (Foundation Tier)



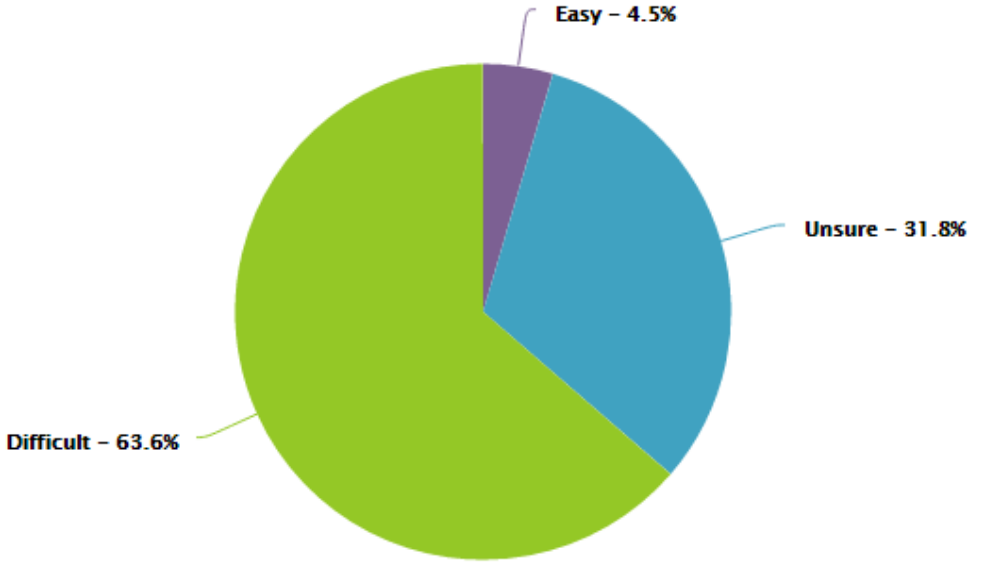
- Spelling and grammar marks in 2013
- Demands raised in 2015

- Maybe these changes affected the grade boundaries (causing them to drop), but underlying performance still followed the sawtooth pattern?

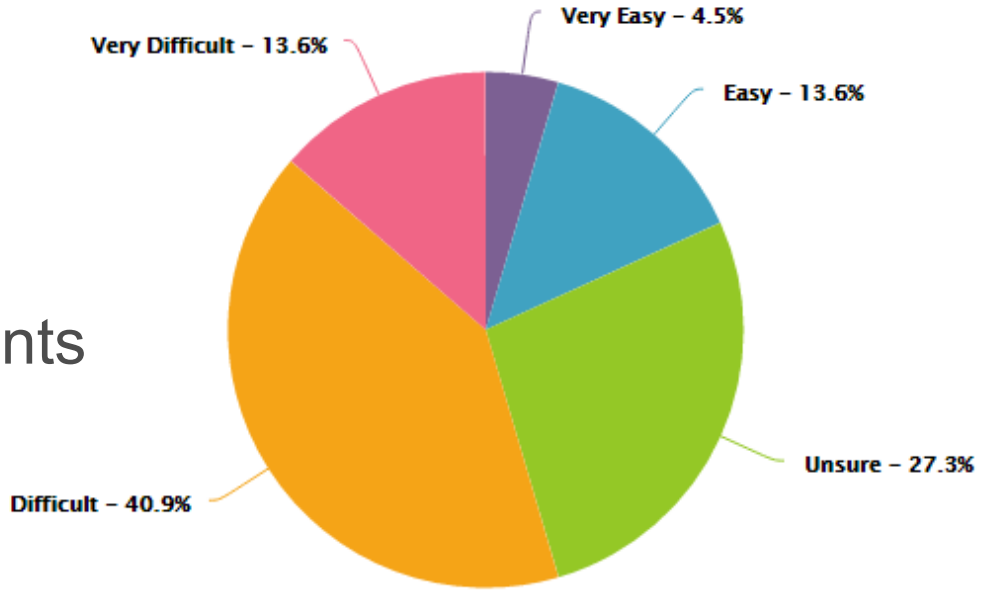
- Suggests there is a degree of disconnect between boundaries and performance change for this specification

Survey Findings

- How easy or difficult did you find the task overall?



- How easy or difficult was it to take question paper difficulty into account when making your judgements on candidate performance?



Conclusions

- Overall (with one major + a few minor inconsistencies), ratings of performance generally matched the patterns of grade boundary change
  - What we were asking judges to do was quite difficult!

- Helps to validates the results of Study 1, by broadly supporting our assumption that grade boundary change = performance change

- However – we didn't find perfect matches, which adds more caution to our interpretation of the results from Study 1

Thank you for listening!

Any questions?

More info:

ben.cuff@ofqual.gov.uk

Search “Sawtooth Effect” on www.gov.uk