Using Rasch Modelling to Investigate Inter-board Comparability of Examination Standards in GCSE

Qingping He

March 23 2018

Inter-board comparability (IBC)

- Exams testing the same subject areas are provided by different exam boards (EBs)
- There is no pre-testing or equating between the exams due to their high-stakes nature
- Inter-board comparability (IBC) is defined as the extent to which examinees awarded the same grade by different exam boards in a subject have similar level of attainment. It can be viewed as one aspect of validity
- Inter-board comparability has been a concern for a wide range of stakeholders

Monitoring and maintaining inter-board comparability

- Methods used to study inter-board comparability:

  Judgemental methods (absolute and comparative) and Statistical methods

- Post-award inter-board statistical screening (IBSS) using mean GCSE score

Candidates from all boards (a specific subject):

Mean GCSE band	No of C. in band	Number of Can. in grade				
		A*	A	B	C	…
Band1	N_1	$N_{1,1}$	$N_{1,2}$	$N_{1,3}$	$N_{1,4}$	…
…	…	…	…	…	…	…
Band5	N_5	$N_{5,1}$	$N_{5,2}$	$N_{5,3}$	$N_{5,4}$	…
…	…	…	…	…	…	…
Band10	N_{10}	$N_{10,1}$	$N_{10,2}$	$N_{10,3}$	$N_{10,4}$	…

Expectation matrix: $\alpha_{i,j} = \dfrac{N_{i,j}}{\sum_{i=1}^{10} N_i} = \dfrac{N_{i,j}}{N}, \quad \alpha = (\alpha_{i,j})$

Candidates from a particular board (board x):

Mean GCSE band	No of Can. in band	Expected Number of Can. in grade				
		A*	A	B	C	…
Band1	N_{x1}	$N_{x1,1}$	$N_{x1,2}$	$N_{x1,3}$	$N_{x1,4}$	…
…	…	…	…	…	…	…
Band5	N_{x5}	$N_{x5,1}$	$N_{x5,2}$	$N_{x5,3}$	$N_{x5,4}$	…
…	…	…	…	…	…	…
Band10	N_{x10}	$N_{x10,1}$	$N_{x10,2}$	$N_{x10,3}$	$N_{x10,4}$	…
Expected total		N_{px1}	N_{px2}	N_{px3}	N_{px4}	…
Observed		N_{ox1}	N_{ox2}	N_{ox3}	N_{ox4}	…

Expected: $N_{xi,j} = N_{xi}\alpha_{i,j}, \quad N_{pxj} = \sum_{i=1}^{10} N_{xi,j}$

vs Observed

Potential issues with existing IBSS using mean GCSE score

- The appropriateness of the use of mean GCSE score for screening
- Not all candidates took the same subjects that are used to calculate the mean GCSE score. Different candidates may have taken a different set of subjects
- Variability in difficulty between subjects
- Variability in entry pattern between the exam boards

Aims of study

- To gain further understanding of the issues with inter-board comparability
- To explore the potential of using differential category functioning (DCF) analysis with Rasch modelling to investigate the comparability of examination standards in GCSEs between exam boards as partial validation of the current post-award inter-board statistical screening (IBSS) approach
- To explore the potential of using Rasch modelling to enhance the existing inter-board statistical screening approach

Differential category functioning analysis using partial credit Rasch model (PCM)

- The Partial Credit Model for polytomous items: $\ln \frac{p_k}{p_{k-1}} = \theta - \delta_k$

- Differential category functioning (DCF): Test takers with the same ability from different subgroups perform differently at specific score (category) levels of an item
- Methods used to investigate DCF:
  - Calibrate items for different subgroups separately
  - Calibrate items using persons from all subgroups and compare average abilities from different subgroups
  - Re-estimate item parameters for individual subgroups using their ability distributions estimated with the population

Method

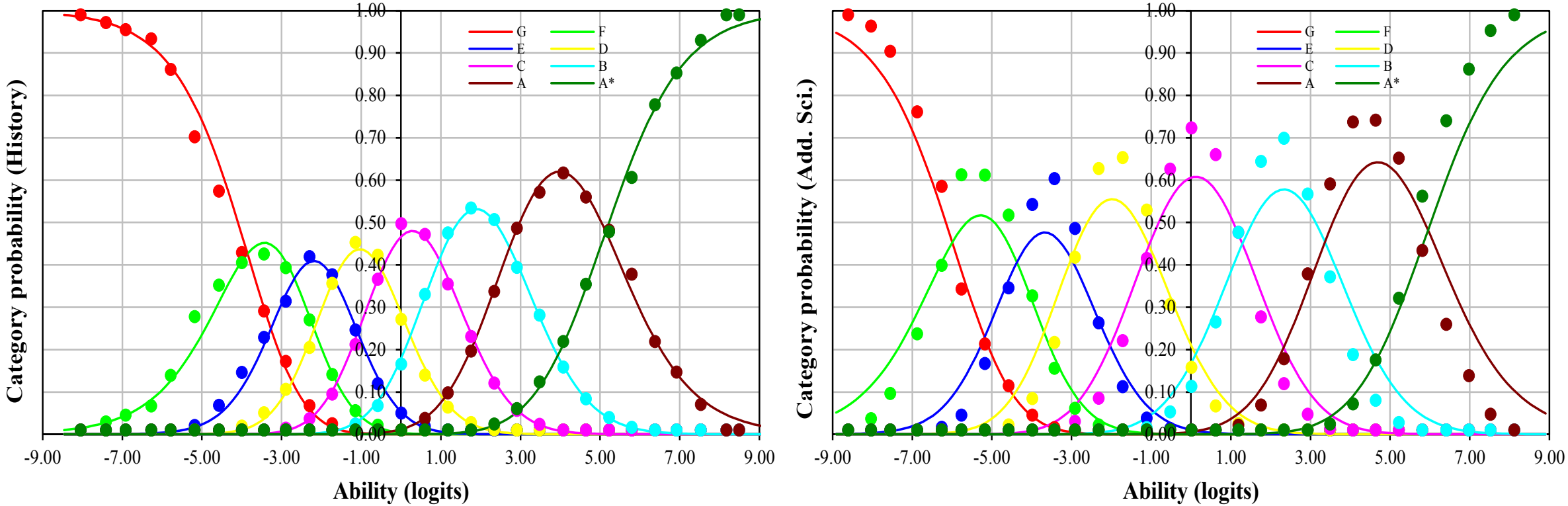
- The GCSE letter grades were converted into ordered numerical grades (U to 0, G to 1, …, A* to 8)
- Each subject was treated as a polytomous item and the grade a candidate received on the subject as a score (or performance category) on the item
- The subjects included in the analysis were assumed to define a shared construct which is closely related to the constructs measured by the individual examinations
- Candidates taking the exams that test the same subject areas but provided by different exam boards were treated as different subgroups
- WINSTEPS was used for Rasch analysis and R code was developed to re-estimate the item category parameters for individual exam boards and standard error of estimation
- The existence of significant DCF effect at specific grades in a subject between the exam boards (subgroups) was assumed to indicate inconsistency in standards at these grades between the exam boards
- DCF measures were converted into changes in grade boundary scores required for aligning standards between the exam boards to investigate potential impact on grade outcomes
- Results from Rasch analyses were compared with those from the existing inter-board statistical screening using mean GCSE score

The data (from the 2015 exam series)

Subject name	Sample size			
	Board A	Board B	Board C	Board D
Additional Science	180638	75720	51340	7987
Applications of Mathematics	5156	2284	3812	1179
Biology	74074	39027	11551	5080
Chemistry	73196	38995	11516	5016
English	236791	25687	34854	130959
English Literature	231965	26675	36889	109420
French	80568	8868	44369	14484
Further Additional Science	14462	3072	5383	
Geography	107895	33780	42921	29231
German	4987	3252	28025	15233
History	58424	85077	69070	22133
Mathematics	74752	52592	430695	30177
Methods in Mathematics	4397	2126	4033	872
Physics	74030	39397	11483	4919
Science	141775	52715	37670	7305

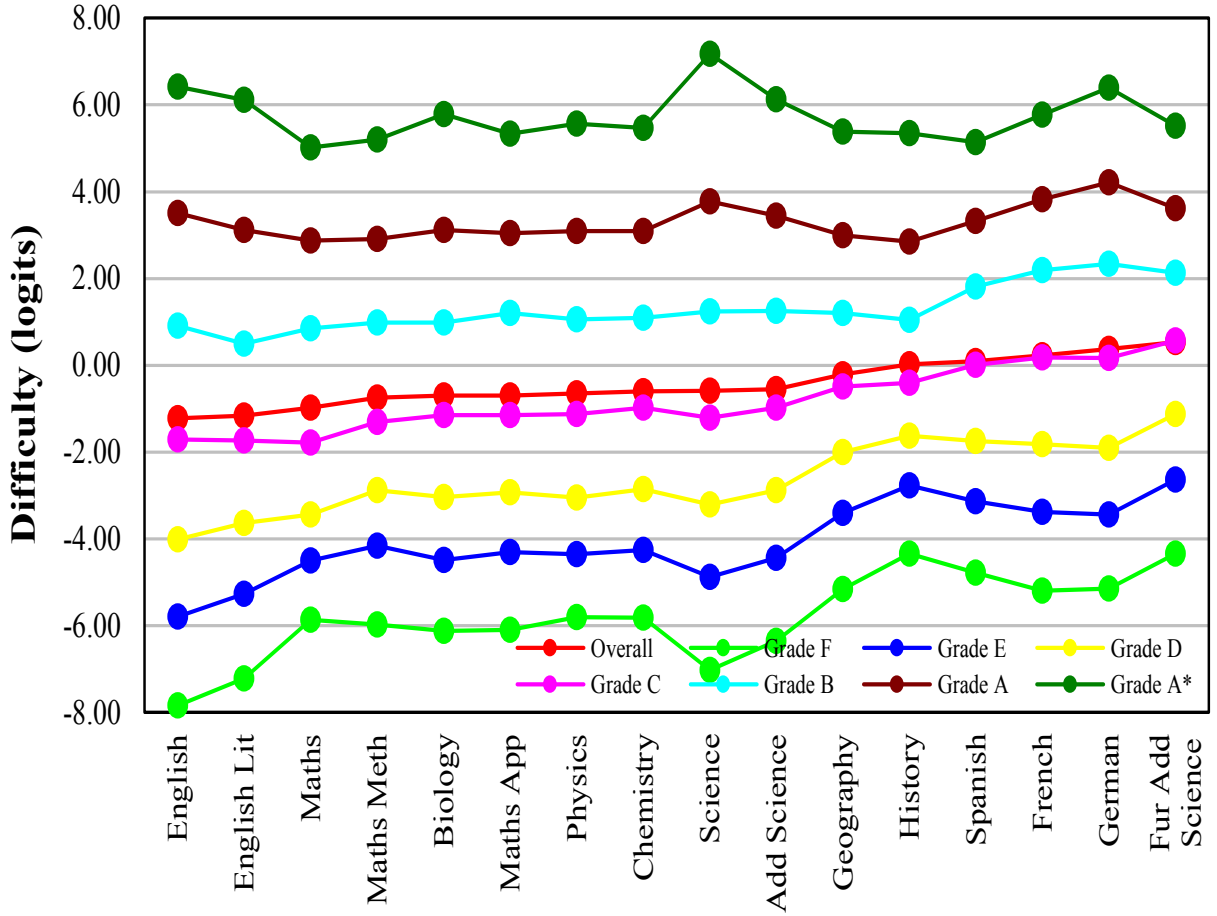
Model assumptions and model fit

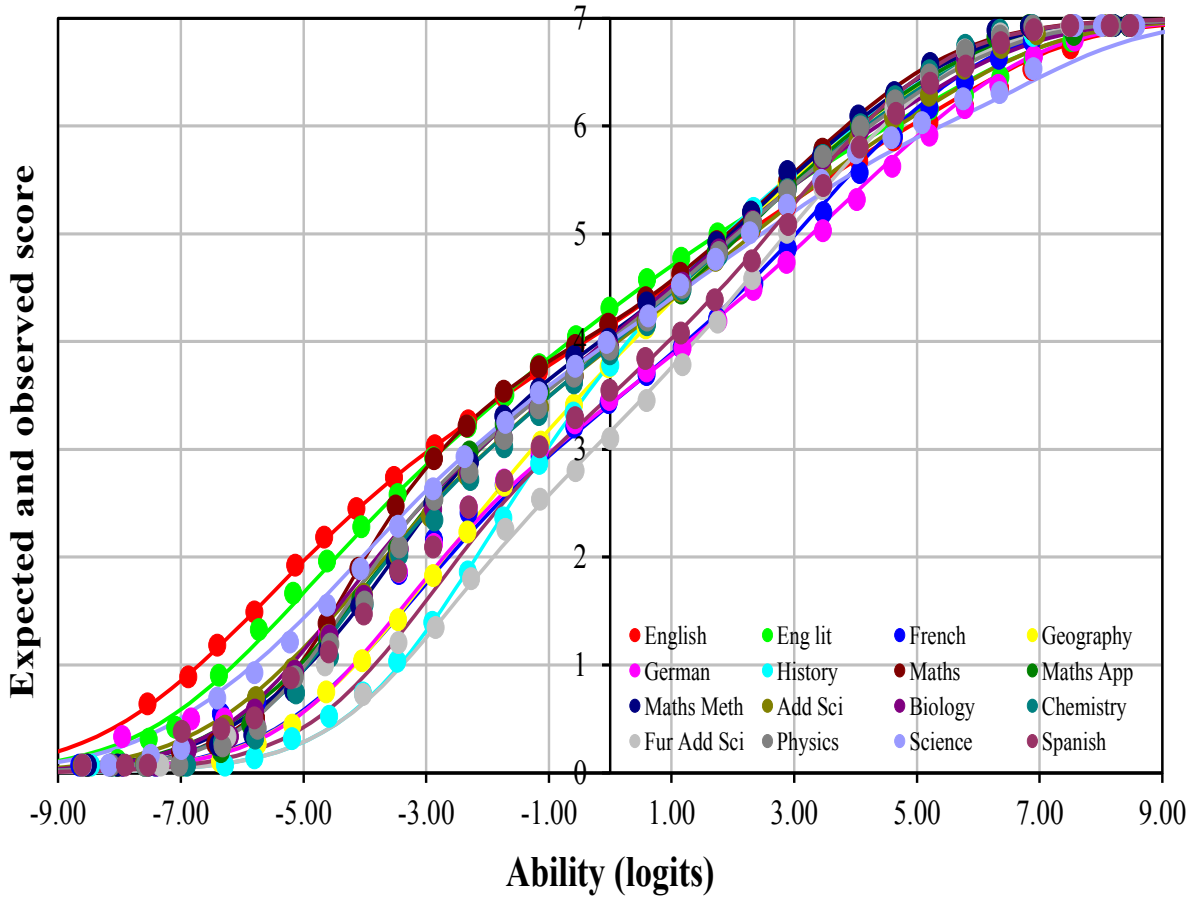
- Grade U did not fit the PCM well and was removed from analysis. Candidates taking fewer than two subjects were excluded. The final dataset fitted the model reasonably well
- Unidimensionality assumption of the PCM held reasonably well



Category probability curves (CPCs) for History and Additional Science

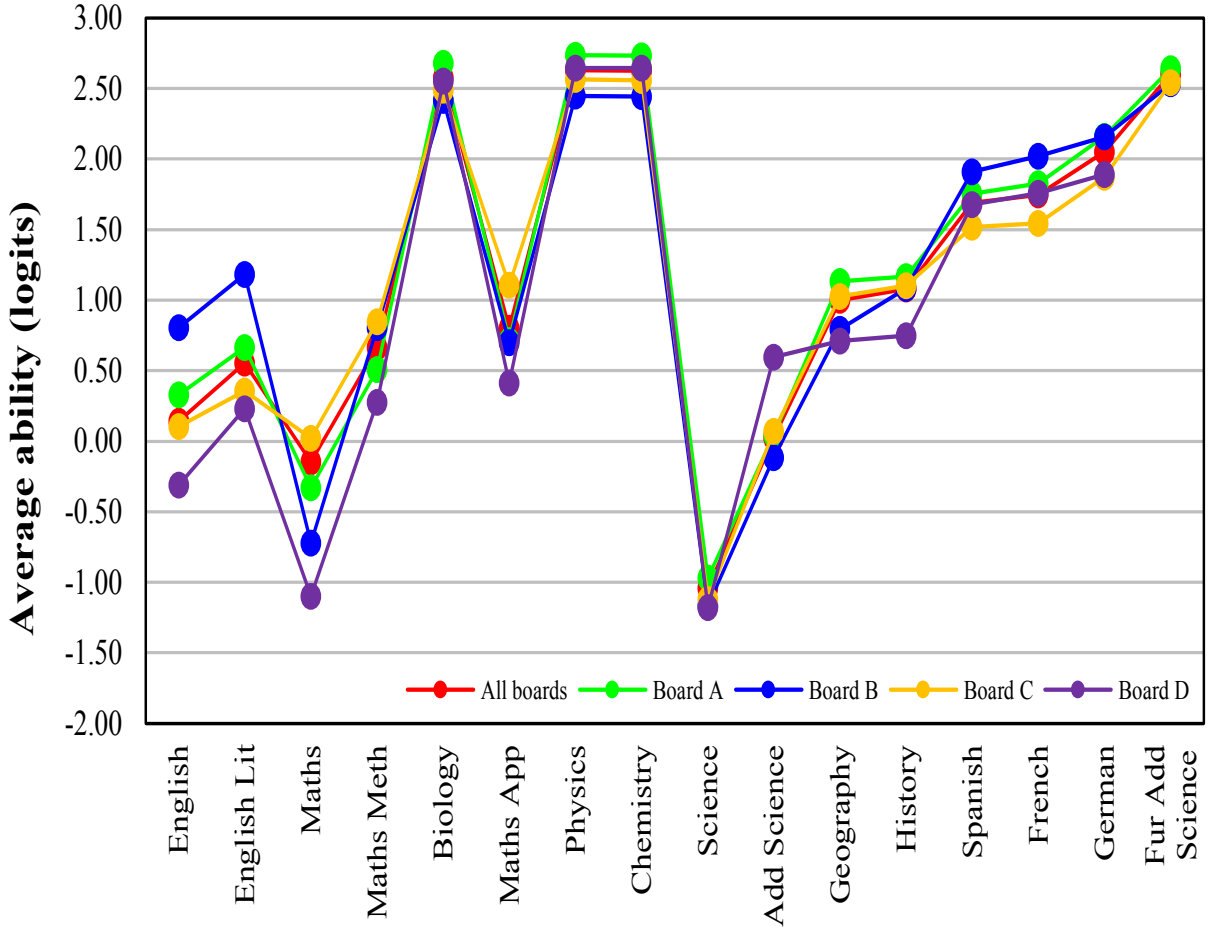
Subject relative Rasch grade difficulty (all boards) and item characteristic curves (ICCs)

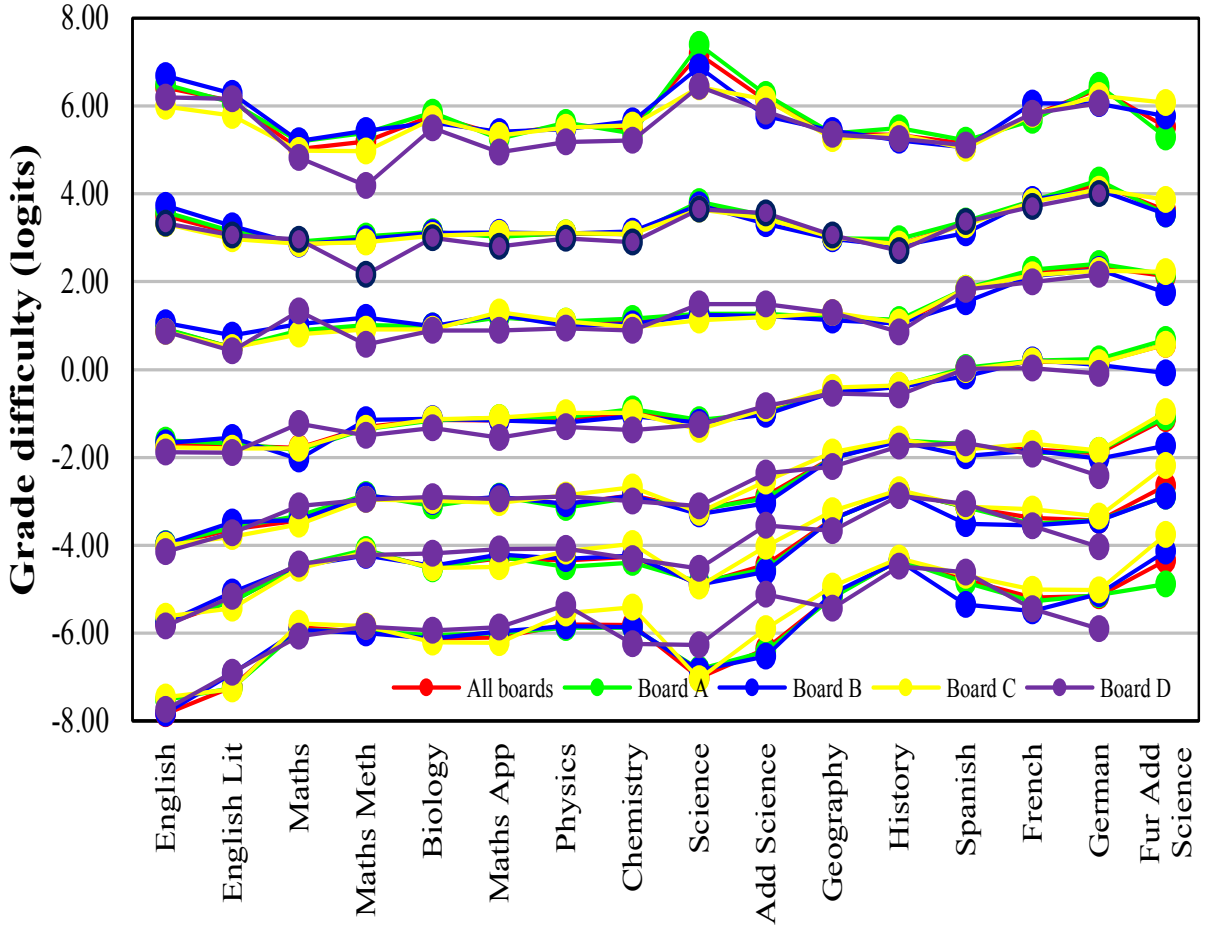




Grade gap in unit of logits: $\Delta = \frac{1}{N_G N_S} \sum_{i=1}^{N_S} (d_{i,A} - d_{i,E})$

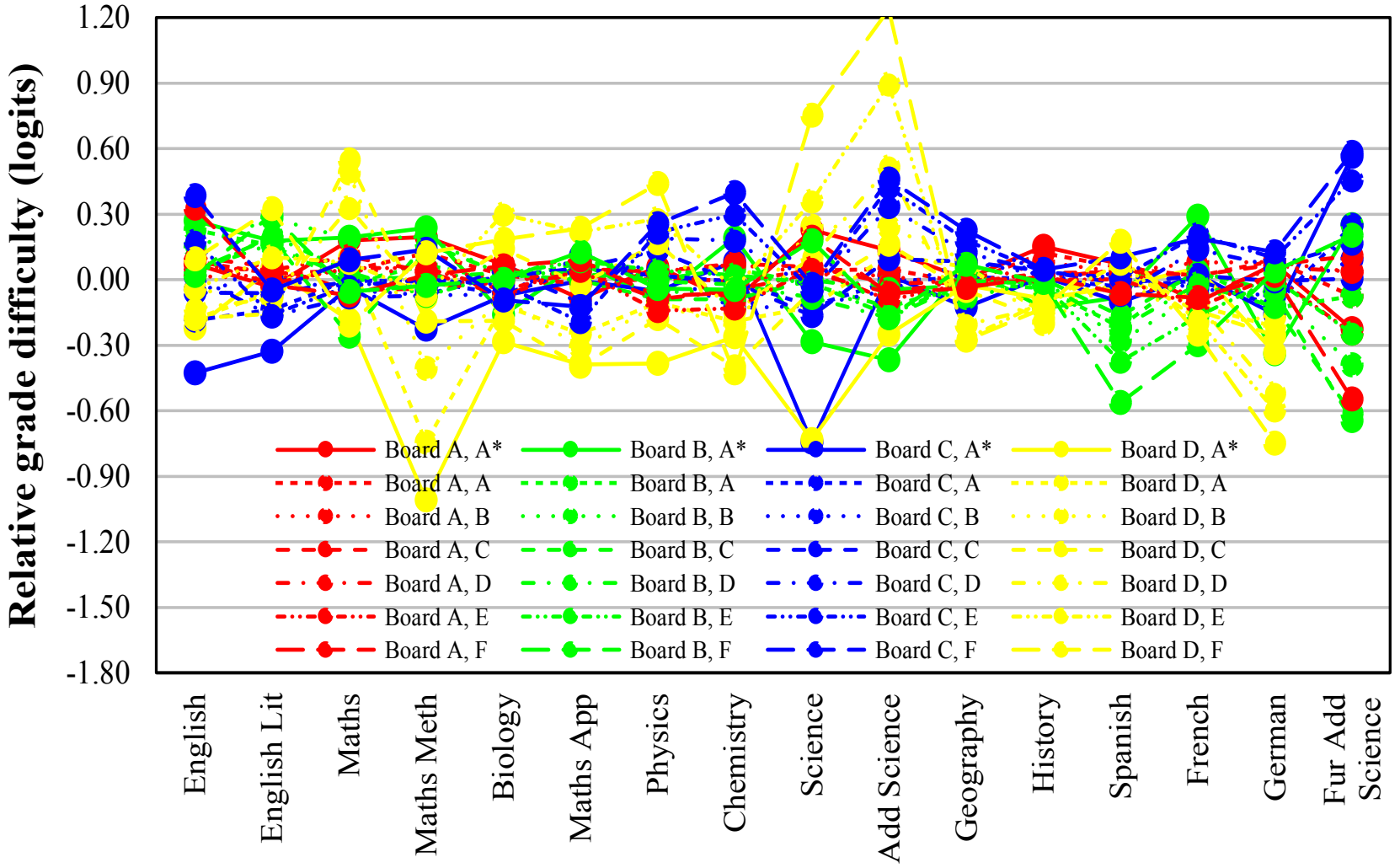
Average ability of candidates taking different subjects and grade Rasch difficulty (all and individual boards)





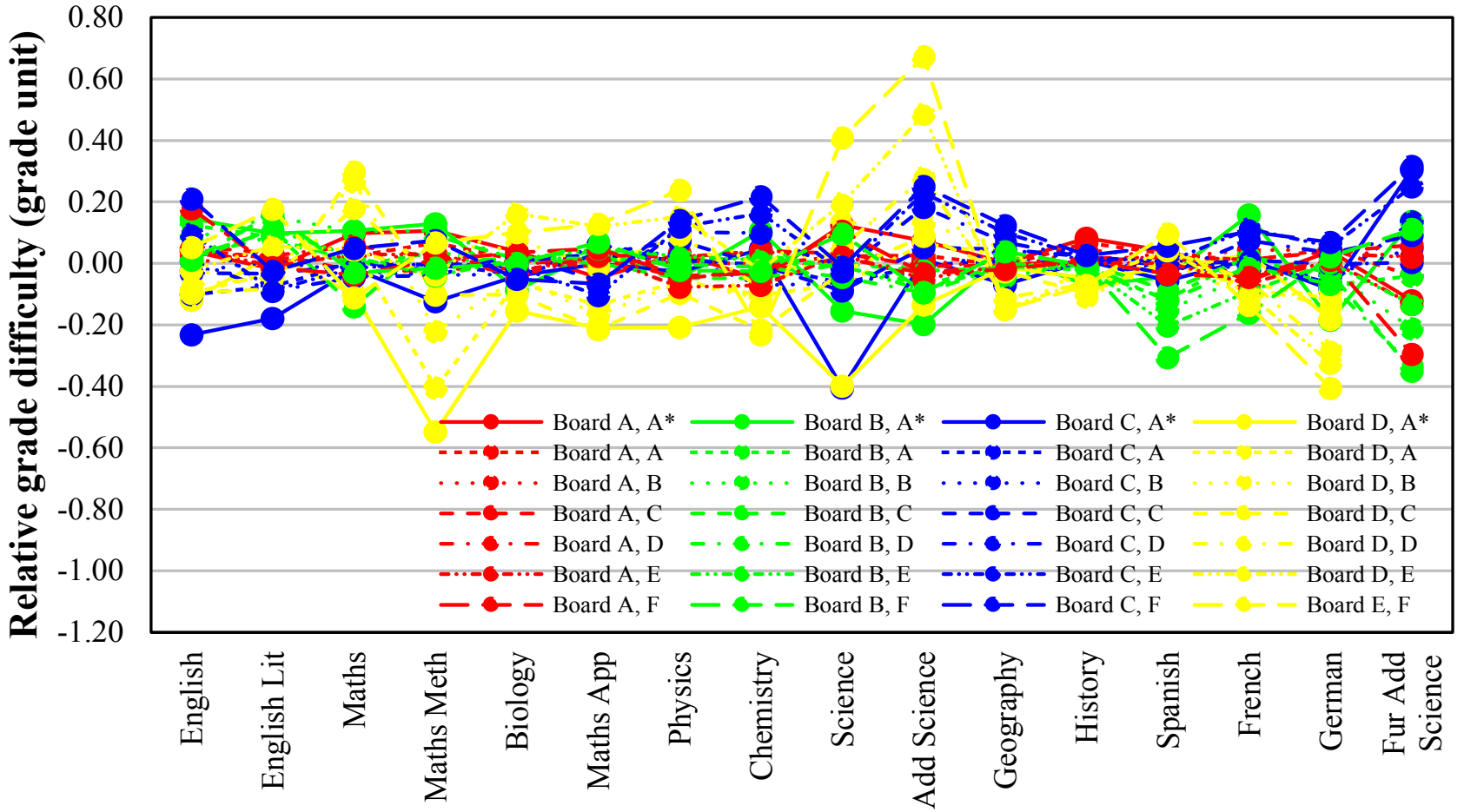
Relative grade difficulty (difference in standards – DCF effect, in unit of logits)

Relative grade difficulty in logits and unit of grade: $d_{k,R} = d_k - d_{k,ALL}$



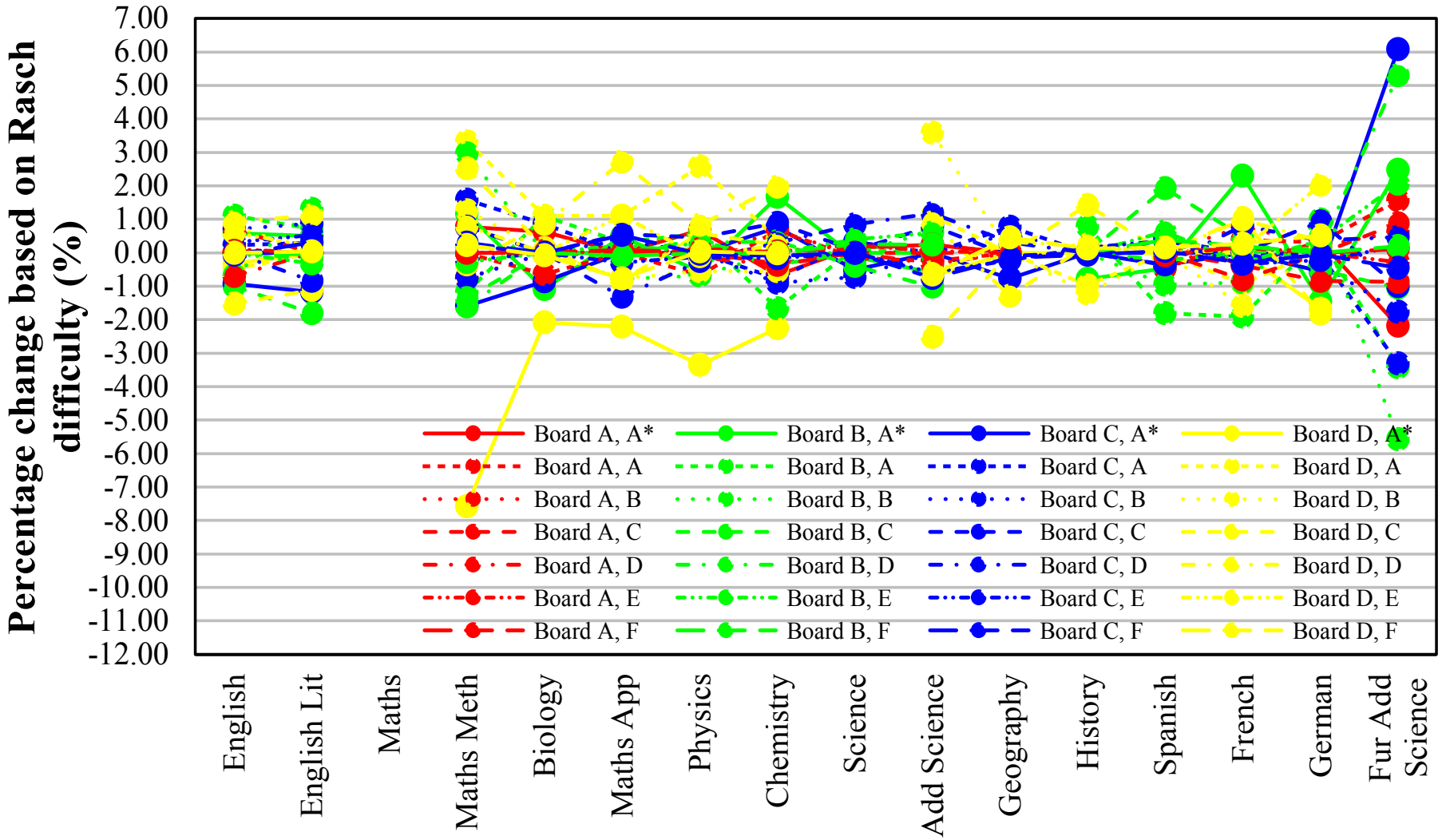
Relative grade difficulty (difference in standards – DCF effect, in unit of grade)

Relative grade difficulty in unit of grade: $d_{k,RG} = \frac{d_{k,R}}{\Delta}$

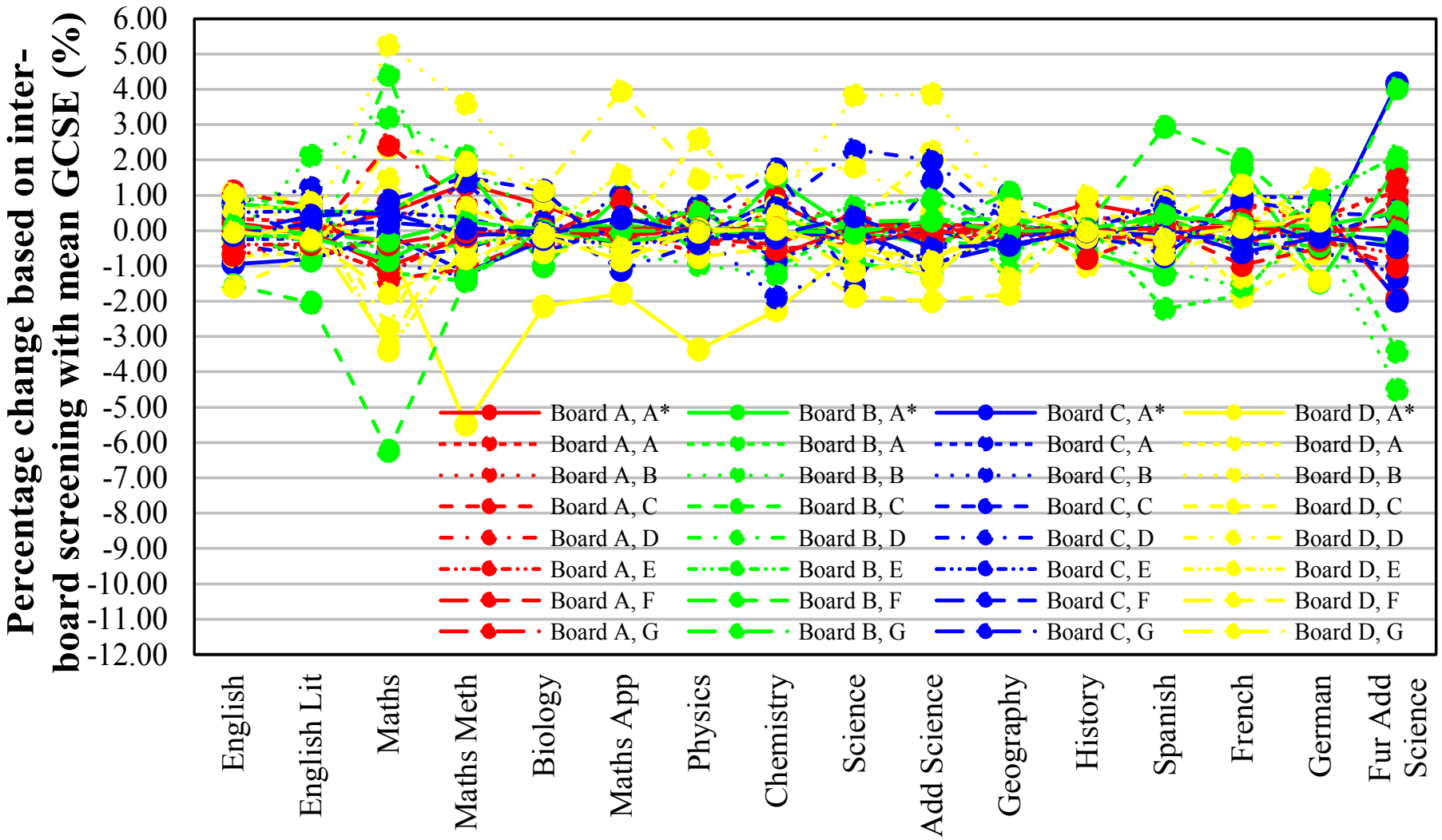


Changes in boundaries required for aligning standards: $b'_k = b_k - wd_{k,RG}$

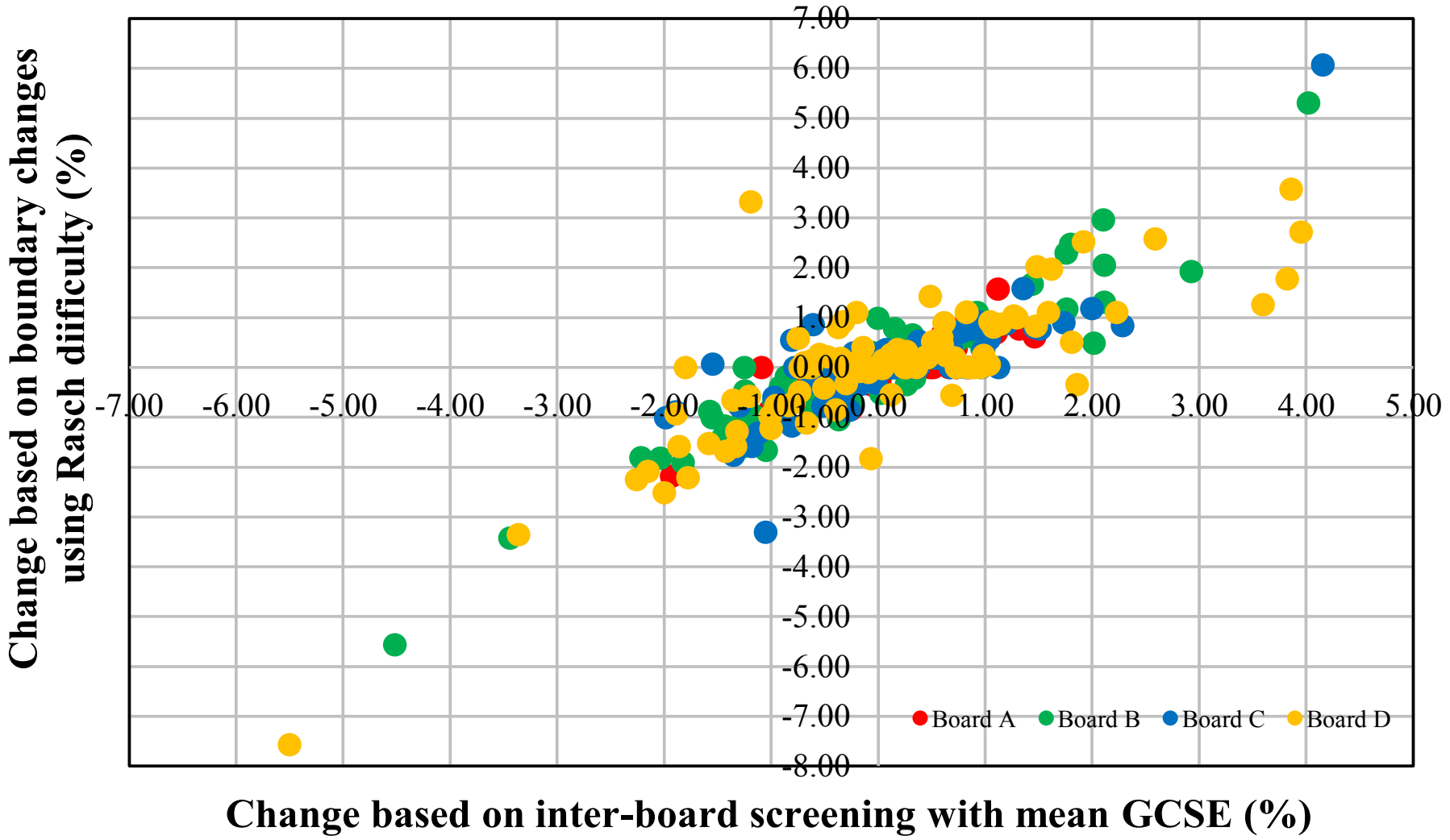
Potential impact of aligning standards between exam boards: Changes in grade outcomes based on Rasch analysis (changing grade boundaries)



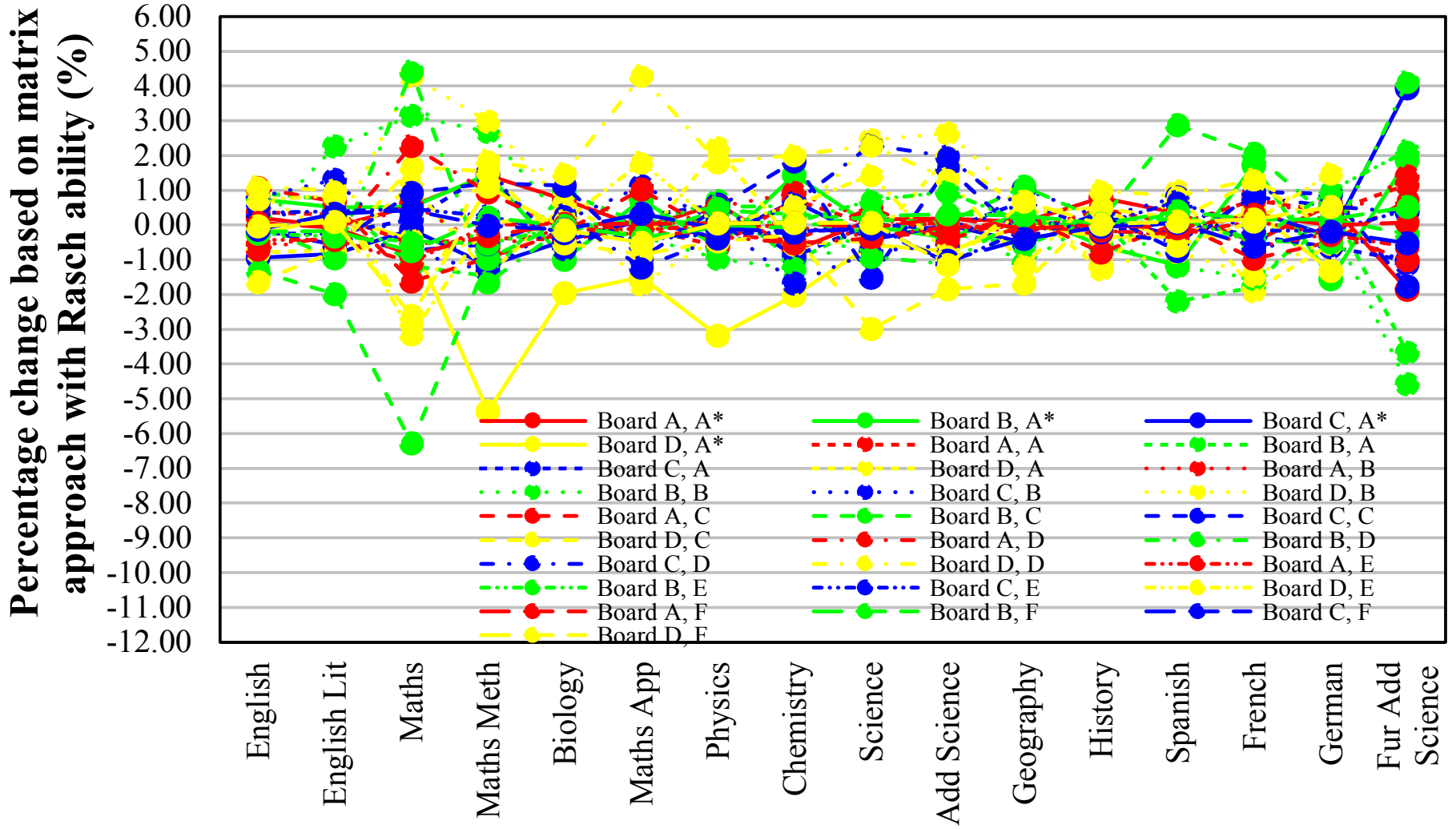
Changes in grade outcomes after aligning standards between exam boards based on existing IBSS with mean GCSE score (no change in boundaries)



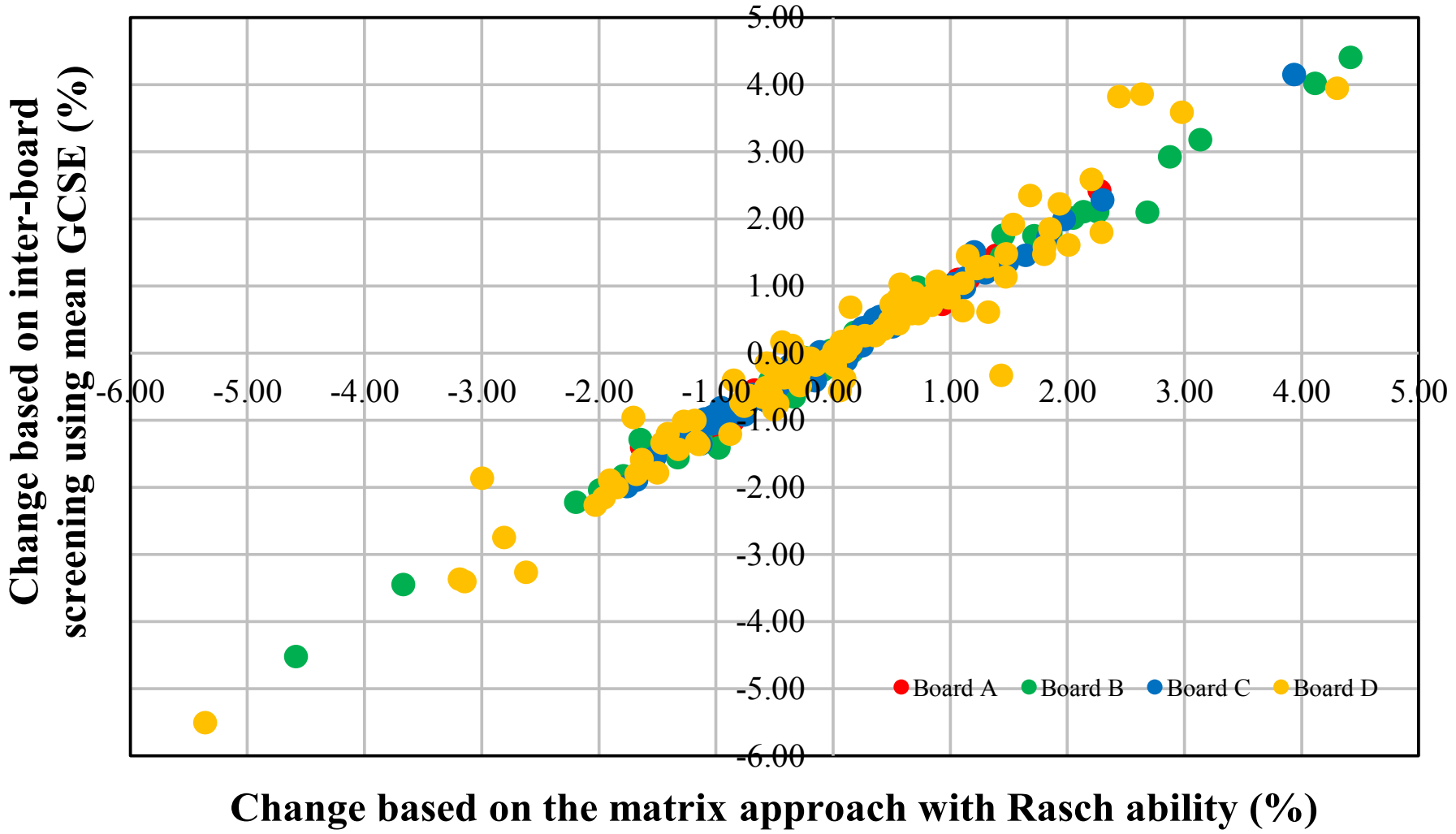
Comparison of changes in grade outcomes based on Rasch grade difficulty and those based on IBSS with mean GCSE score



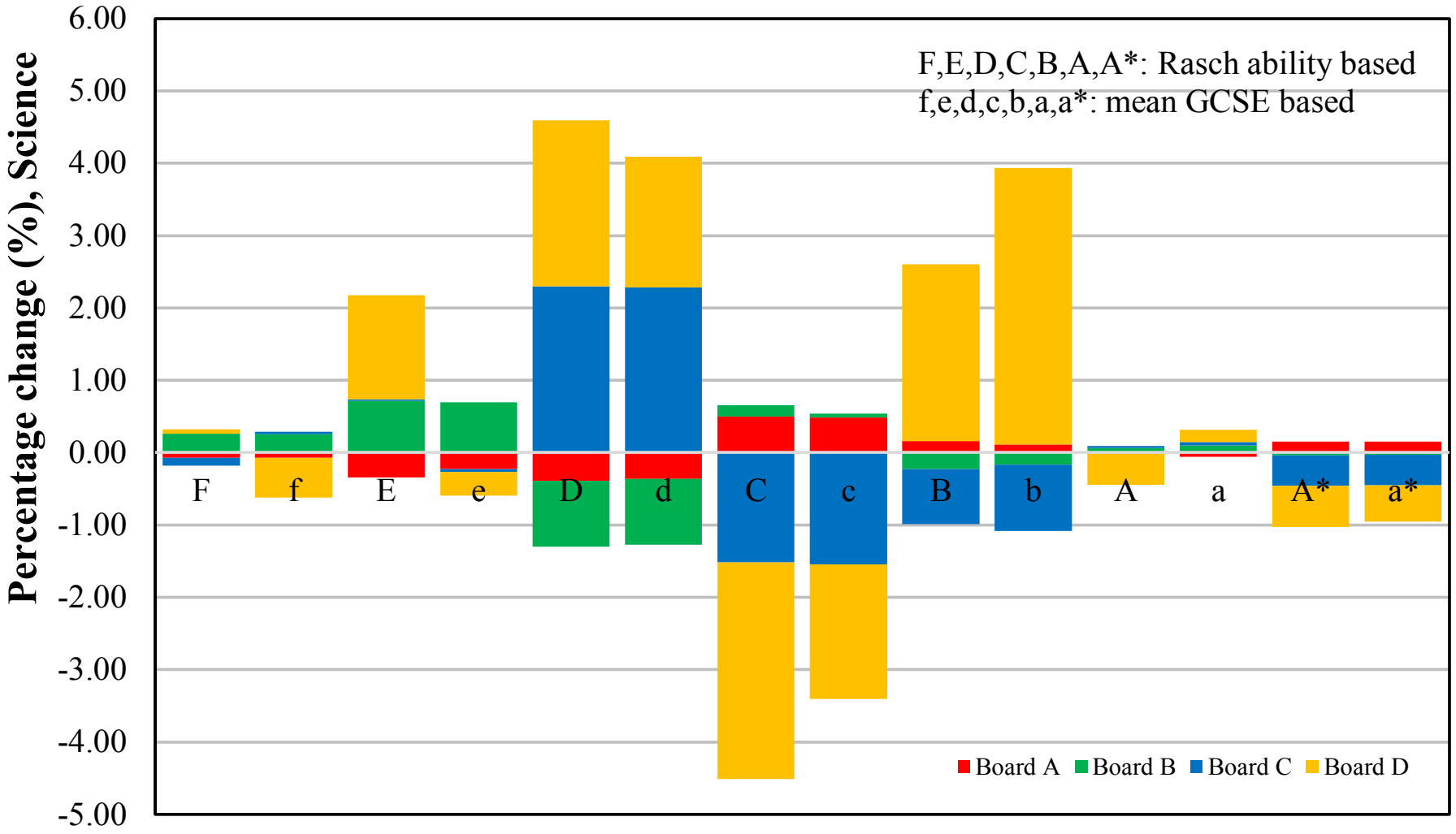
Changes in grade outcomes after aligning standards between exam boards based on matrix approach with Rasch ability instead of mean GCSE



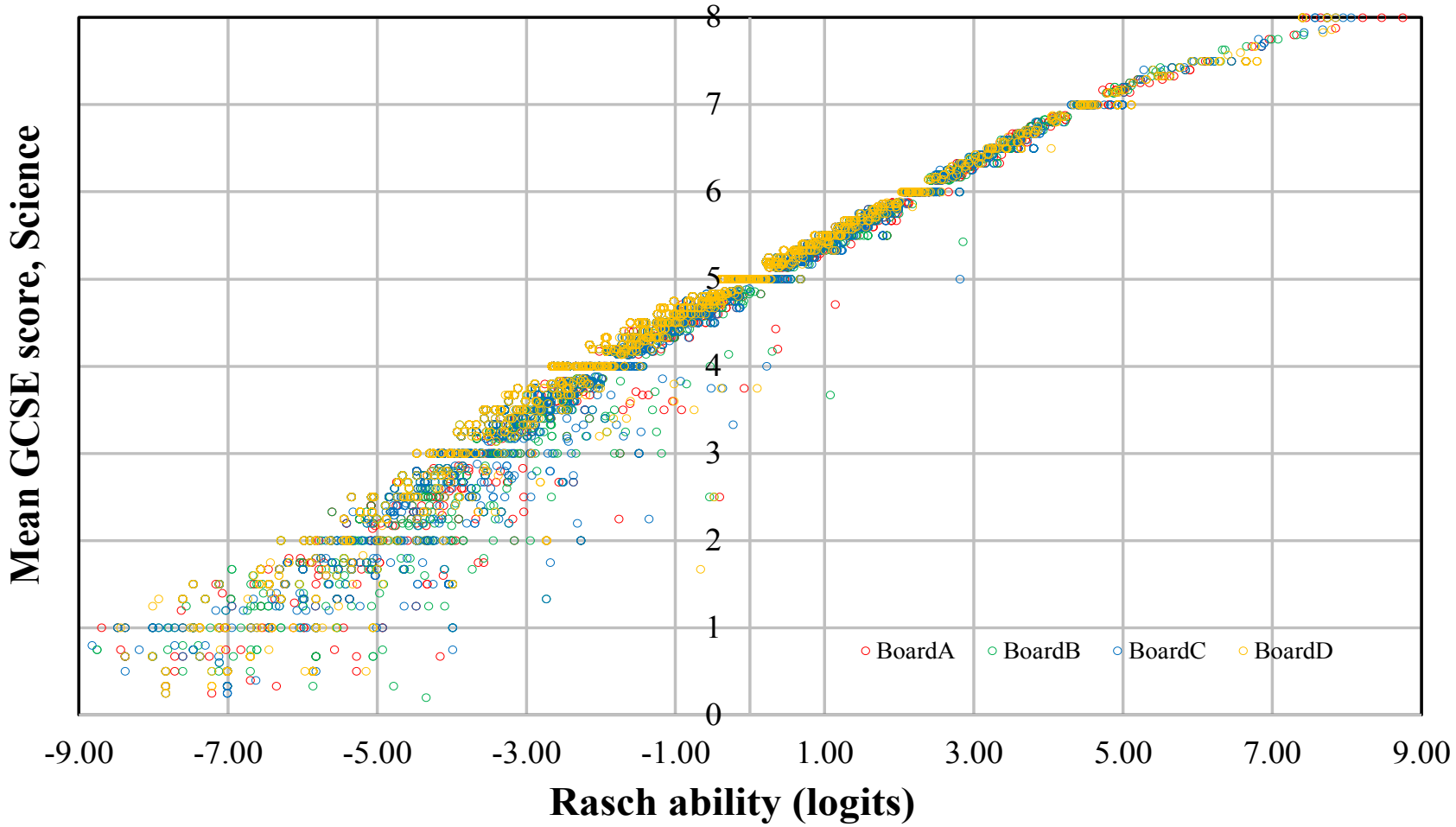
Comparison of changes in grade outcomes based on IBSS with mean GCSE score and the matrix approach with Rasch ability



GCSE Science: aligning standards based on IBSS with mean GCSE score and Rasch ability; effect of variability in subject difficulty and entry pattern (1)

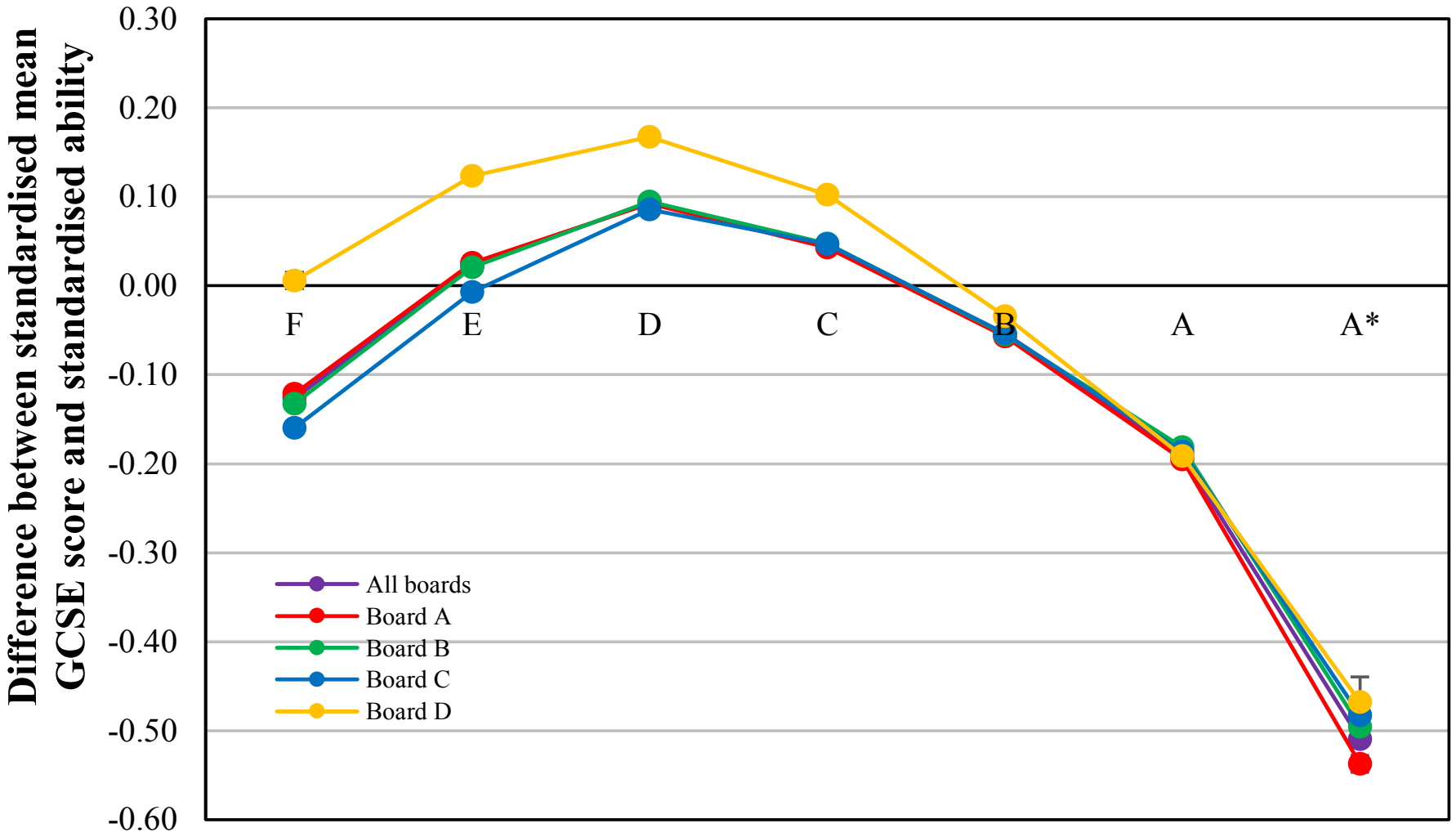


GCSE Science: aligning standards based on IBSS with mean GCSE score and Rasch ability; effect of variability in subject difficulty and entry pattern (2)

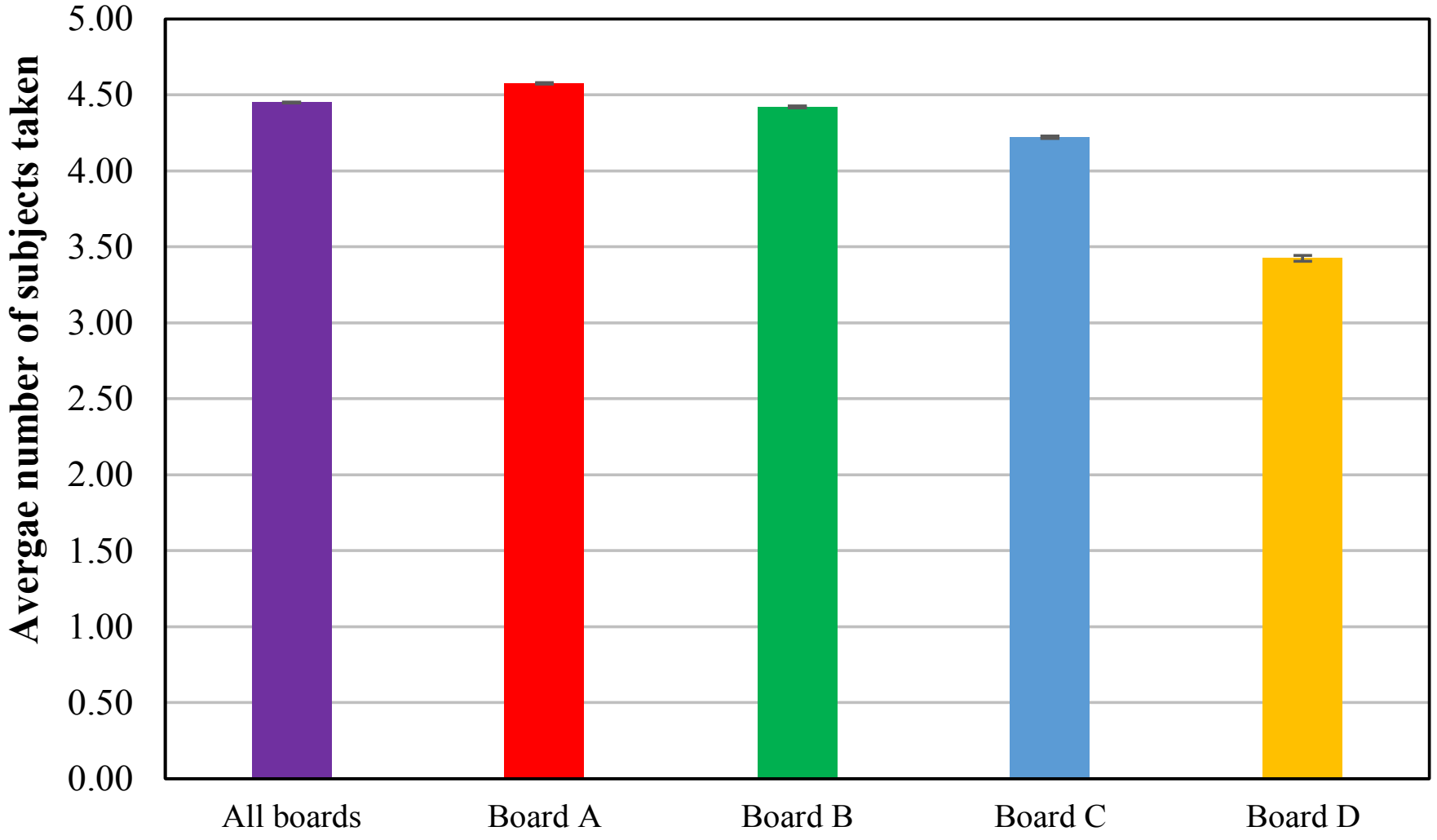


"Value added" measure: $v_i = x_{i,GCSE} - x_{i,Ability}$

GCSE Science: aligning standards based on IBSS with mean GCSE score and Rasch ability; effect of variability in subject difficulty and entry pattern (3)



GCSE Science: aligning standards based on IBSS with mean GCSE score and Rasch ability; effect of variability in subject difficulty and entry pattern (4)



GCSE Science: aligning standards based on IBSS with mean GCSE score and Rasch ability; effect of variability in subject difficulty and entry pattern (5)

- When the total number of subjects taken by students from a board was small, the subject itself would make a substantial contribution to the overall mean GCSE score. And science was a relatively easy subject. These factors may partly explain the higher value added estimated for students from Board D and the differences in estimated changes in grade outcomes between IBSS with mean GCSE score and IBSS with Rasch ability
- The use of mean GCSE score as a performance measure with the existing inter-board statistical screening procedure would not be able to take into account the different difficulties of the subjects and the variability in entries between the exam boards

Conclusions

- For most of the grades across the 16 subjects studied, the effect of DCF was small or moderate. Comparability of standards in the exams between the exam boards for the higher grades was found to be higher than that for the lower grades
- Changes in grade outcomes after aligning standards between the exam boards based on Rasch grade difficulty and those estimated using existing inter-board statistical screening procedure with mean GCSE score were broadly consistent
- The use of Rasch ability as a performance measure with the existing IBSS procedure produced results which were closely similar to those produced using the mean GCSE score. However, since the ability measure takes into account the difference in difficulty between subjects, its use with the existing inter-board screening procedure is likely to produce fairer results than the use of mean GCSE score