The problem solving problem: Can comparative judgement help?

Ian Jones & Matthew Inglis

Mathematics Education Centre
Loughborough University

I.Jones@lboro.ac.uk

Loughborough University
Mathematics Education Centre

Problem solving in mathematics

HOW MUCH CAN
WE **TRUST**
OPINION POLLS!!??

15 A new bicycle is priced at £240

In a sale it is reduced by 35%.

Calculate the sale price.

Answer £ ___________ [3]

---

16 **(a)** Write down the *n*th term for the sequence

6, 12, 18, 24, ...

Answer ____________ [1]

(b) Write down the *n*th term for the sequence

4, 9, 14, 19, ...

Answer ____________ [2]

Plan

- Marking and Comparative Judgement;
- The study:
  - Designing the paper;
  - Evaluating the paper;
  - Assessing the paper;
  - Judge feedback.

Marking

- Assumes precise, predictable responses
- Validity grounded in detailed criteria
- Low inter-rater reliability for sustained problem solving

Murphy (1982)
Newton (1996)
Willmott & Nuttall (1975)

Comparative Judgement

- Assumes varied, unpredictable responses
- Validity grounded in collective expert opinion
- High inter-rater reliability for sustained problem solving?

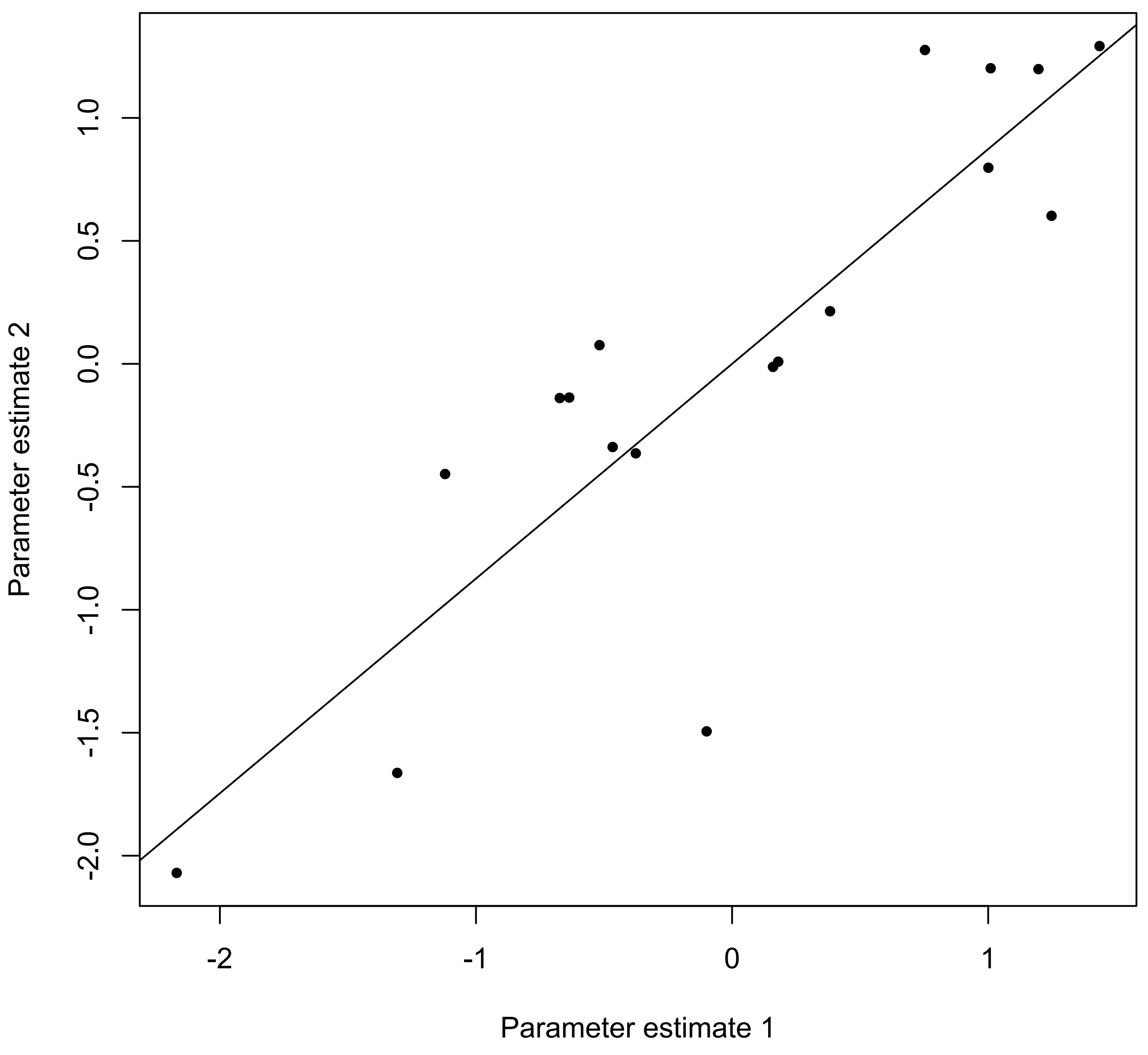
Bramley (2007)
Pollitt (2012)
Thurstone (1927)

Pilot study

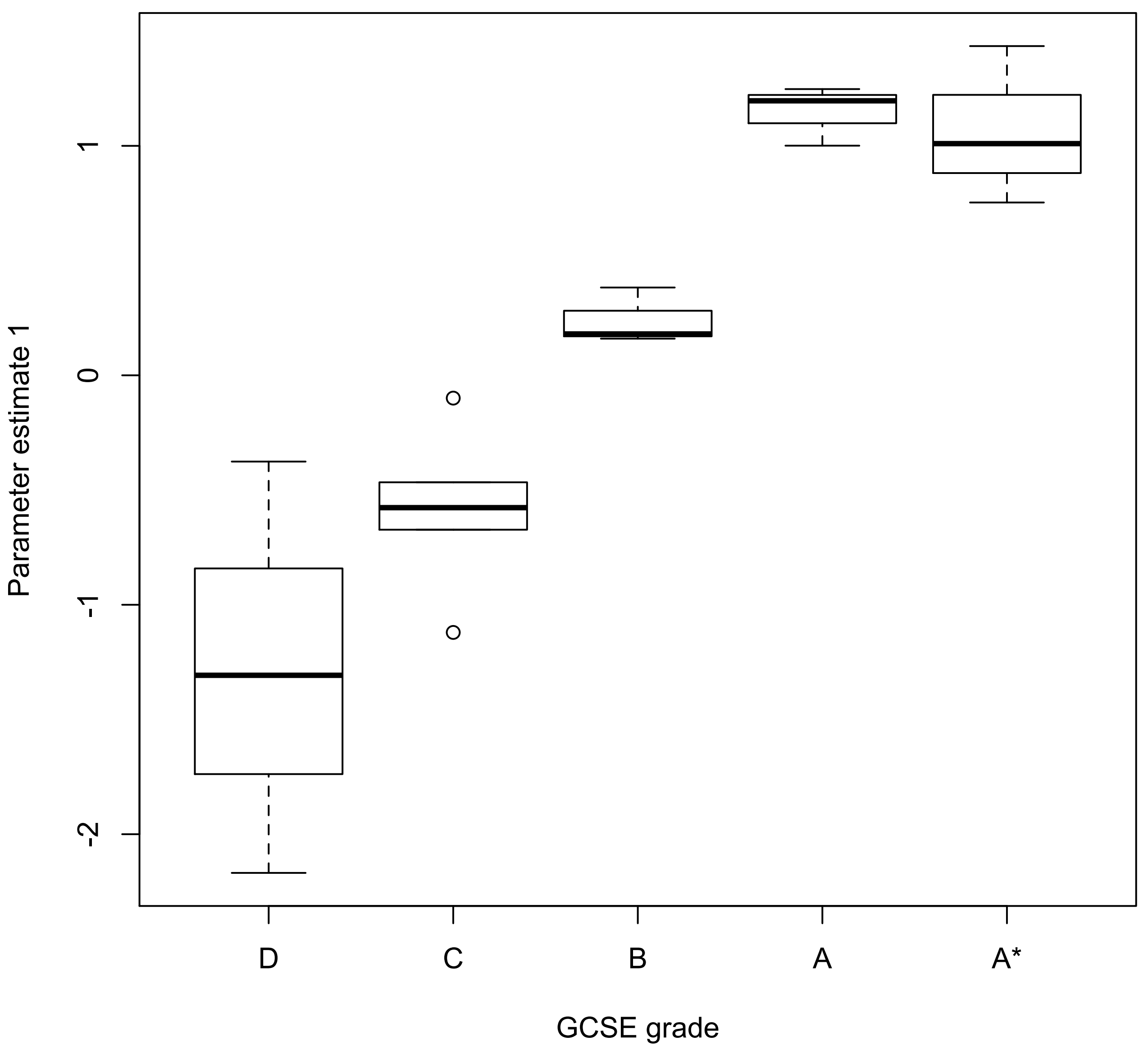
- 18 scripts, three awarding bodies
- Two tiers, grades A* to D
- Two groups of judges ($N_1 = 12$, $N_2 = 12$)

Results

Inter-rater reliability
r = .873



Validity
r = .900



Designing the paper

Evaluating the paper

Assessing the paper

Judge feedback

Design brief

- Four GCSE exam writers, two awarding bodies
- Familiar with Comparative Judgement
- Constraints:
    - “GCSE like” exam paper;
    - no mark scheme, no marks;
    - suitable for both tiers;
    - to be administered early in Year 10;
    - candidates allowed 50 minutes.

Outcome

- 11 pages
- Included a “Resource sheet”
- Pupils write on question paper
- No marks!
- Questions have names not numbers
- Most questions contextualised

Resource Sheet

You may find **some** of the information on this sheet useful for some questions.

You must decide which information to use – you will not need all of it!

Coins

Coin	Value	Diameter	Thickness	Weight
	50p	27.3mm	1.78mm	8.0g
	20p	21.4mm	1.7mm	5.0g
	10p	24.5mm	1.85mm	6.5g
	5p	18.0mm	1.7mm	3.25g

Some Imperial Measures

1 foot = 12 inches (This can be written: 1' = 12")

1 pound = 16 ounces (This can be written: 1lb = 16oz)

1 gallon = 8 pints (This can be written: 1gal = 8pts)

Some Imperial to Metric conversions (to 3 significant figures)

1 foot = 30.5 centimetres

1 pound = 454 grams

1 gallon = 4.55 litres

Nines

This calculation uses **three 9**s:

9 × 9 – 9

The answer is **72**

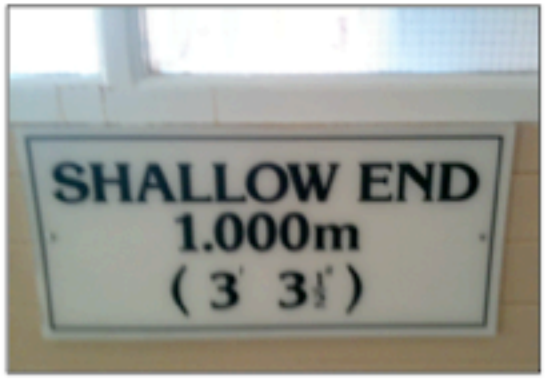
Use three 9s to write a calculation with the biggest possible answer.

You can use any mathematical symbols, but no other digits.

You must write the calculation, but you do not have to work out the answer unless you want to.

Pool

This notice was at one end of an indoor swimming pool.



Explain why the notice is silly.

Look at these cooking instructions from a packet of rice and a pack of lamb chops.

grill minutes

16-20

lamb chops

Heat up the grill.

Put the lamb chops under the grill for 16-20 minutes.

Turn them over half way through the cooking time.

hob minutes

12-15

easy cook rice

Wash the rice.

Put the rice into a saucepan with double the amount of water.

Bring it to the boil.

Turn down the heat, cover the pan and cook for 12-15 minutes.

Tom is going to cook the lamb chops, the rice and the sprouts.

***Cooking* continues on the next page.**

(b) Make a timetable to show when Tom should do each task.

You should allow:

10 minutes for the grill to heat up for the lamb chops,

2 minutes to wash the rice

5 minutes for the water to boil after he has put the washed rice in.

All the food must be just ready at 1 o'clock.

Designing the paper

Evaluating the paper

Assessing the paper

Judge feedback

Teacher survey

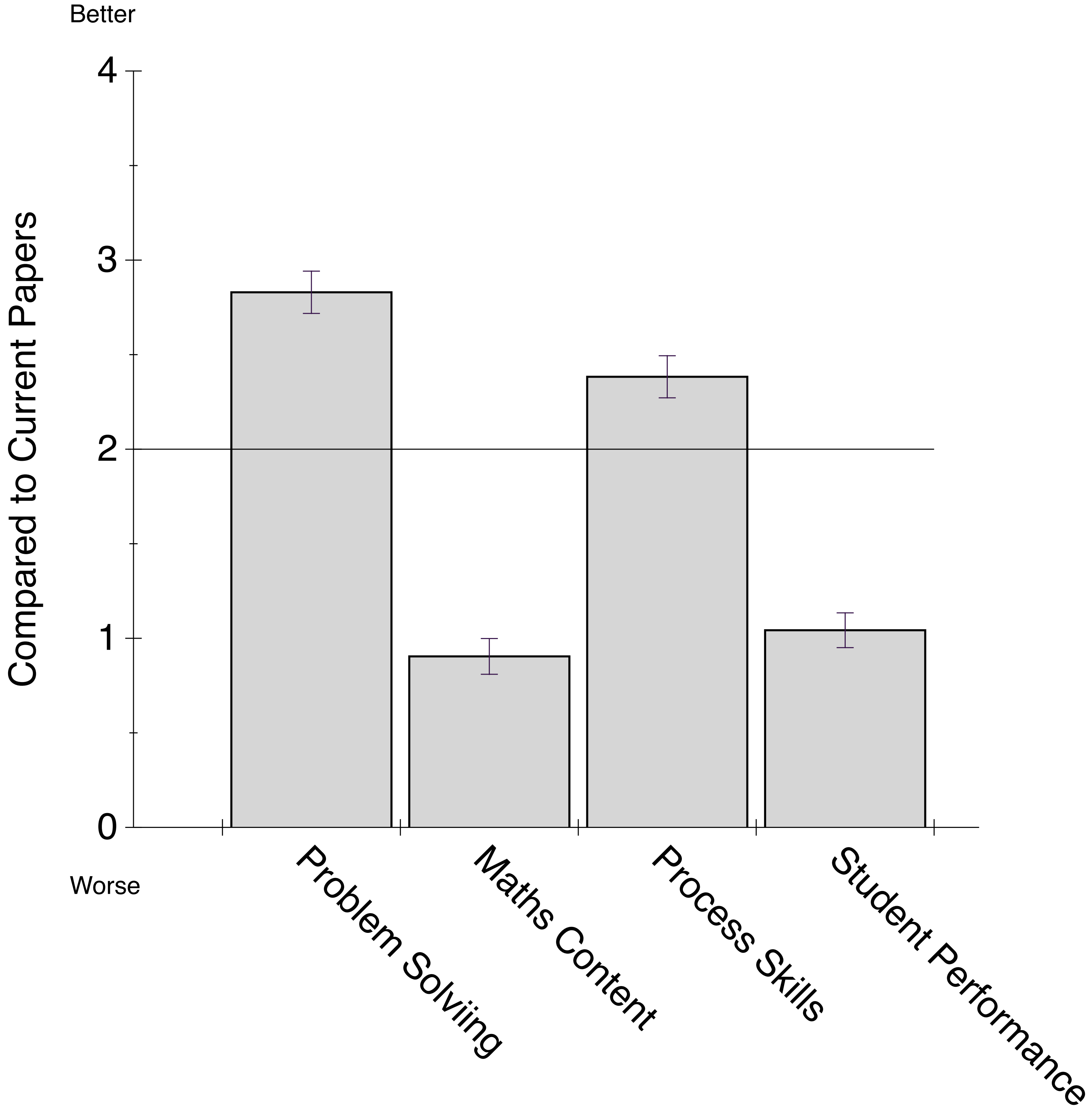
1. How well do you think the paper assesses mathematical problem solving?
2. How well do you think the paper assesses mathematical content?
3. How well do you think the paper assesses the Key Stage 4 Process Skills in mathematics?
4. How well do you think your students would perform on this paper?

A lot less than a typical current GCSE paper

↕

A lot more than a typical current GCSE paper

Teacher survey



N = 94
All significantly different to GCSE at *p* < .001

Open text feedback

Open text feedback

Please do not continue with the project which appears to be watering down the course even more than the current version does

Where is the assessment of mathematical rigour? This obsession with functionality ignores the need for study of algebraic manipulation as training for further study

Open text feedback

I don’t see much testing of algebra, it’s better for practical mathematics but not as good for the academic

Love the paper and the focus on functional mathematics ... This style would ‘force’ the adoption of developing what is the most neglected element of the mathematics curriculum

Open text feedback

The literacy needs are quite high. There is a lot of questions that require a strong level of literacy. The literacy level is above the mathematical level

[some questions] look difficult to assess - it might be difficult to compare alternative, valid solutions. Markers would need to exercise more professional judgement

Designing the paper

Evaluating the paper

Assessing the paper

Judge feedback

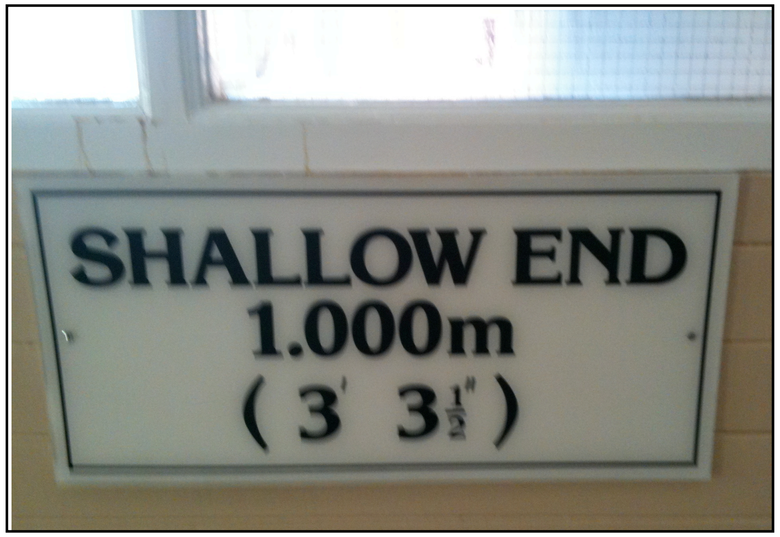
- Administered to 750 Y10 pupils of all abilities
- Retrospective mark scheme constructed
- 750 scripts marked, sample 250 remarked
- 750 scripts judged, sample 250 rejudged
- Predicted grades

Mark scheme

- Retrospective mark scheme (16 pages)
- One examiner commissioned
- Based on sample of student scripts ($N \approx 30$)
- Trialled with two experienced teachers

Pool

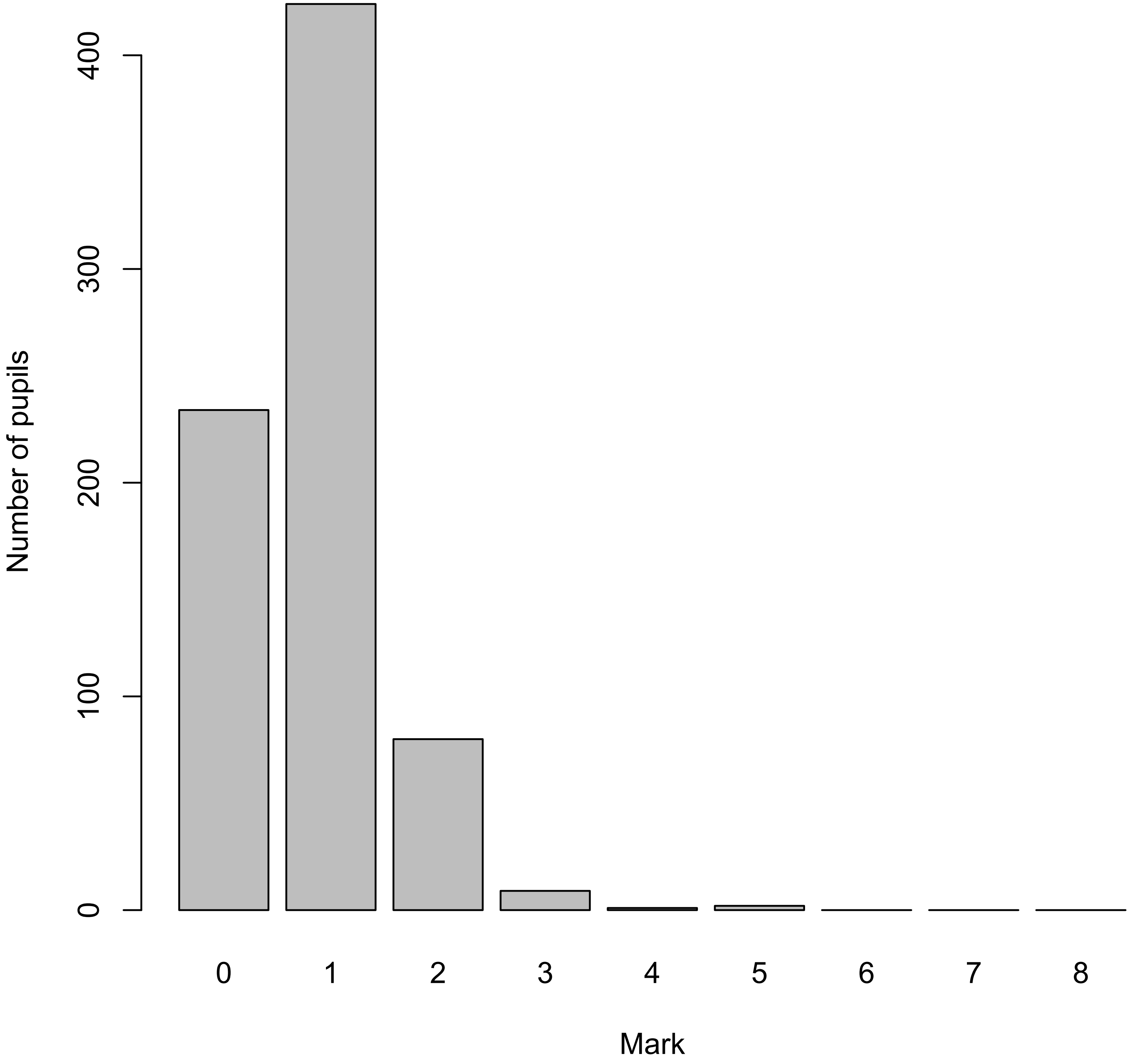
This notice was at one end of an indoor swimming pool.



Explain why the notice is silly.

	Answer	Marks	*Examples* and Comments
	Pool		Marks may be awarded for each point relevant to the response.
	1st point: Accuracy Indicates that 1.000m is too accurate **or** Explains why 1.000m is too accurate a measurement	1 2	*There are too many zeros* *You don't need the decimal places* *That would be to the nearest millimetre* *Only 100 cm in one m*
	2nd point: The social context Indicates that feet and inches are too unfamiliar to be useful **and/or** Indicates that the extra zeros could be confusing	1 1	Note: Both these marks may be awarded if appropriate. *People don't understand old measurements* *People might think it meant 1000 metres*
	3rd point: The physical context Indicates that 1000m is too deep for the shallow end **or** Explains why 1.000m is too accurate in this context	1 2	This answer gets one mark because, although irrelevant, it is a true statement and indicates that the student has at least engaged with the context *The water will be choppy so the exact depth will vary*
	4th point: Measurement Indicates that the two measurements are not exactly equal **or** Shows working comparing the measurements **or** Observes that the figures given are accurate to only 3 significant figures	1 2 3	*3ft 3½ inches is not exactly 1.000m* *3ft 3½ inches is a bit less than 1.000m* (with supporting working) Note: Using the figures given, 3ft 3 ½ inches = 1.004m; 1.000m = 3ft 3.34 inches *You can't really change the 1.000m to inches because it says 'to 3 significant figures'*
			Maximum marks available for *Pool*: 8

“Pool” marks



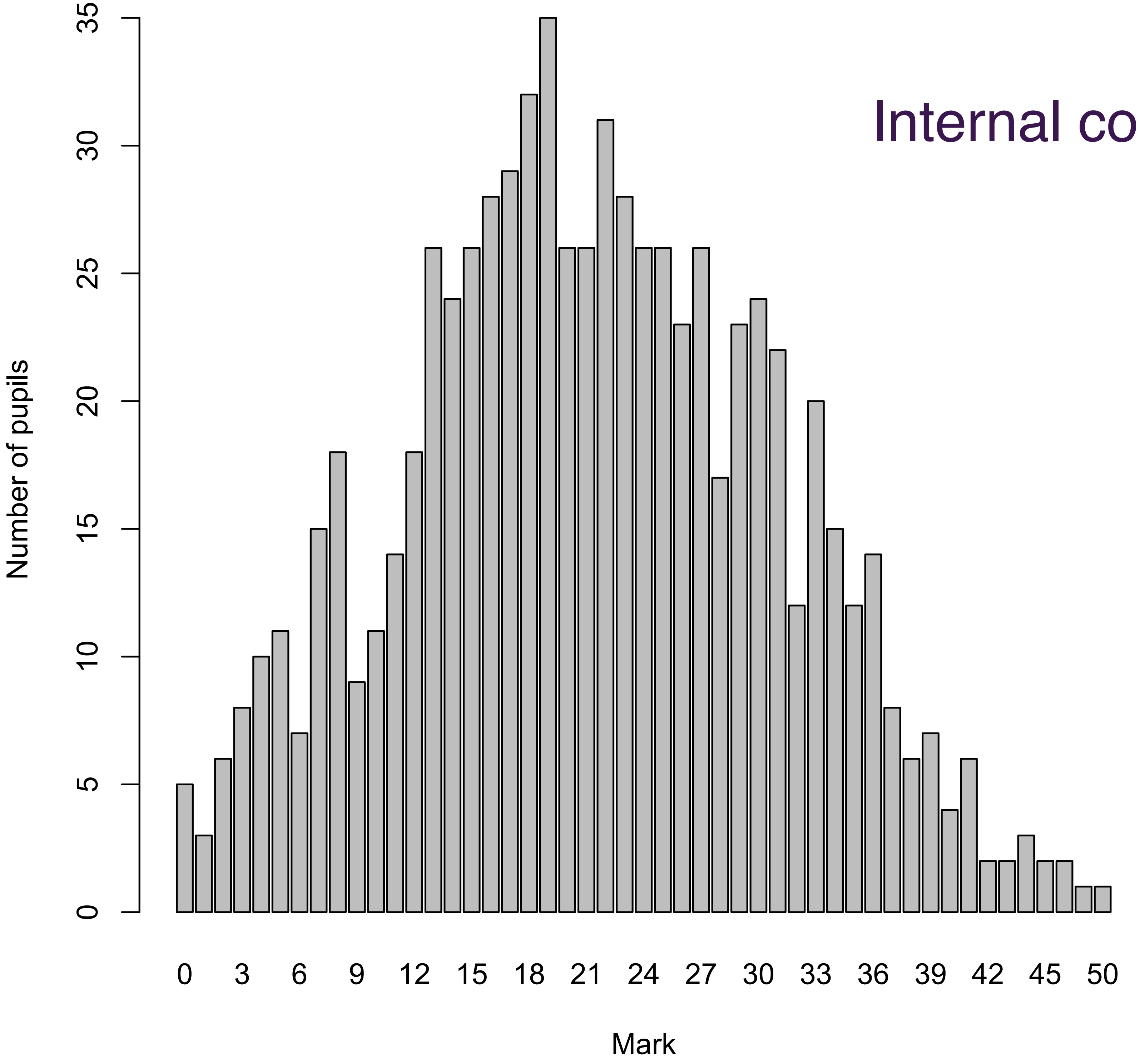
MARKING (750 scripts)

- Two highly experienced and one experienced teacher
- Two hours familiarisation and preparation
- Paid per script, assuming 6 minutes per script

REMARKING (249 scripts)

- One highly experienced teacher

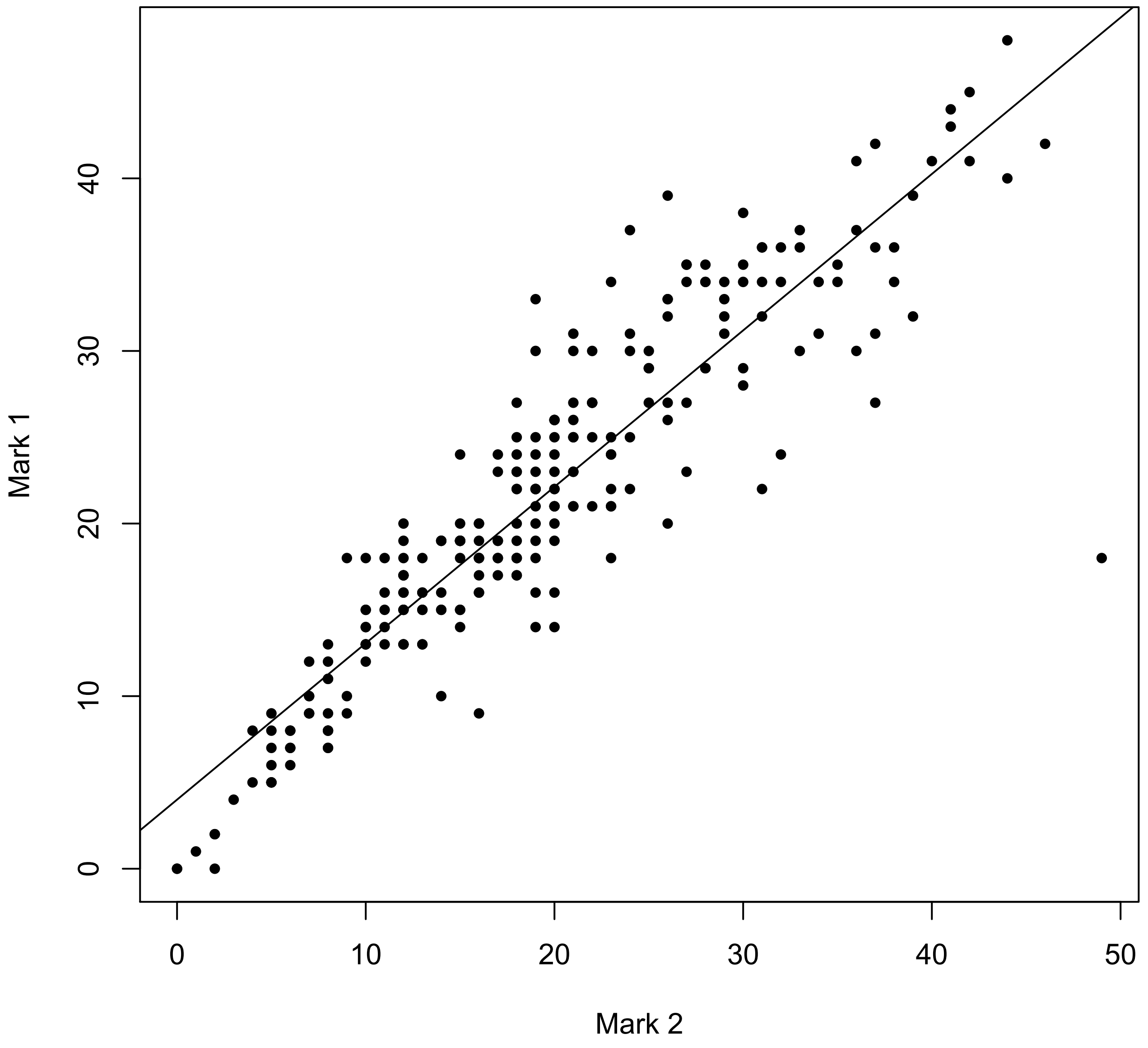
Marking outcome



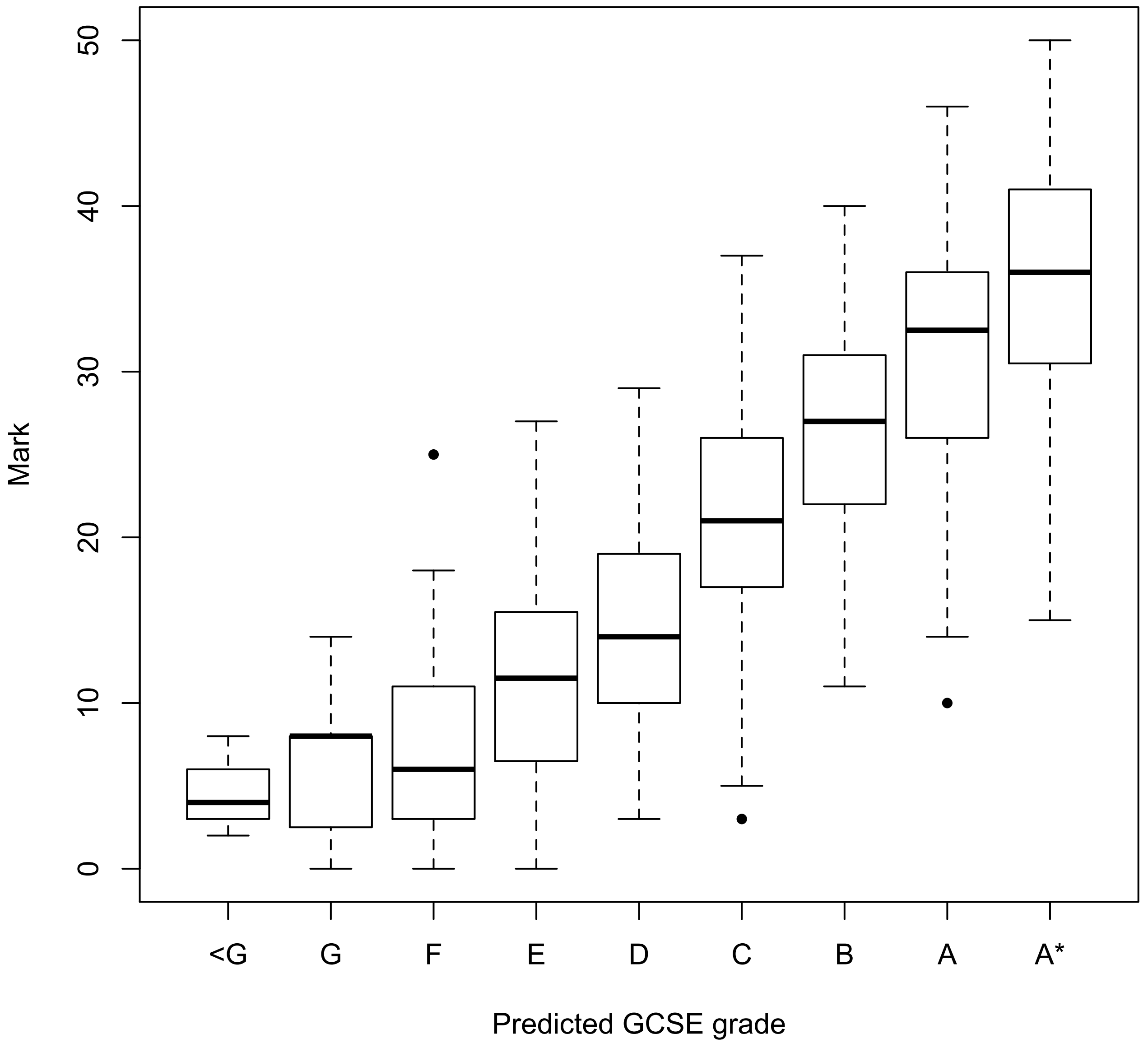
Internal consistency = .720
(Cronbach’s α)

Marking outcome

Inter-rater reliability (*N* = 249)
r = .907



Validity (*N* = 750)
r = .718



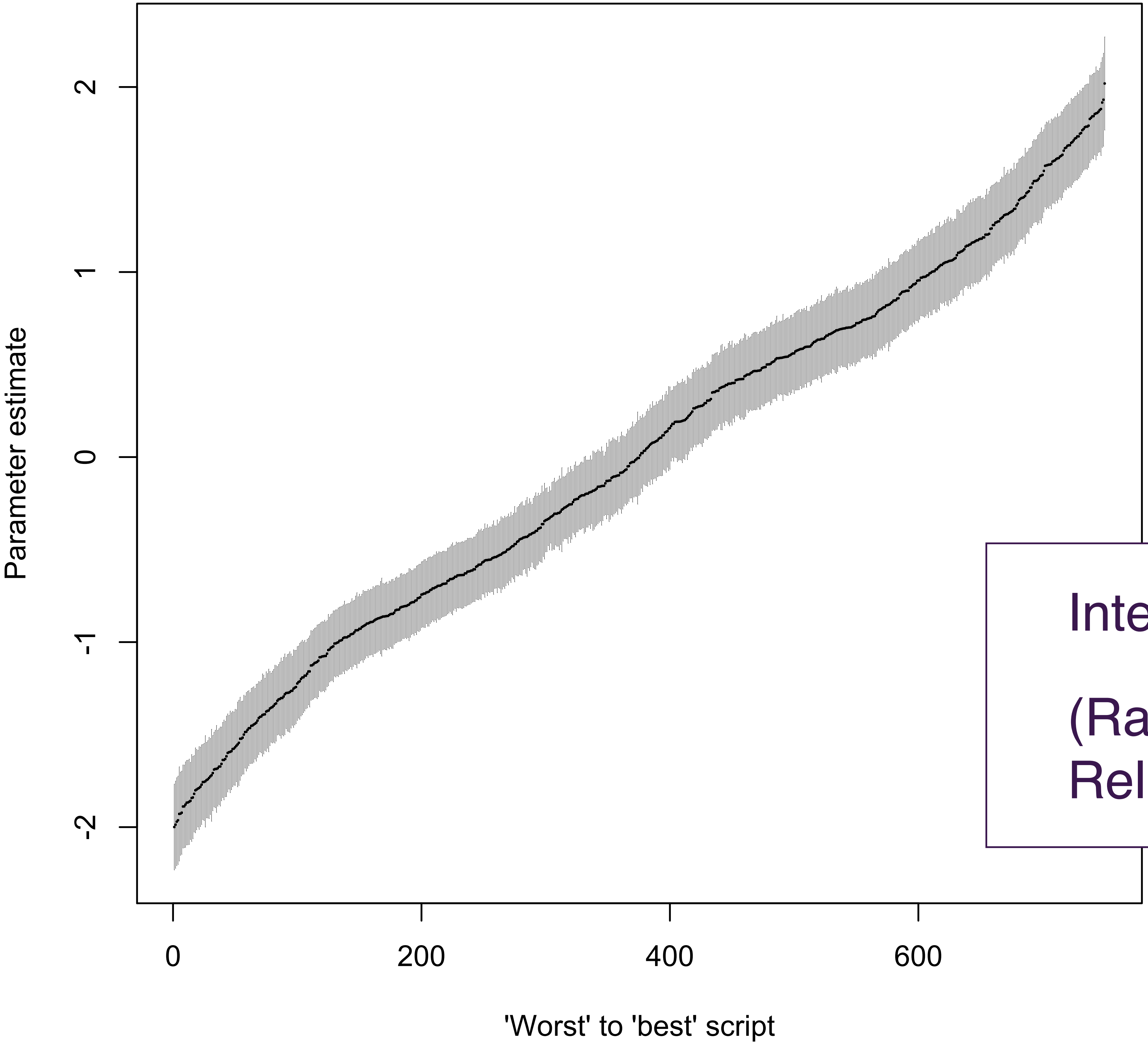
JUDGING (750 scripts)

- 15 teachers and researchers of varied experience
- One hour familiarisation
- 30 minute training session
- 250 - 300 judgements each, assuming 72 seconds per judgement

REJUDGING (250 scripts)

- 5 teachers of varied experience

Judging outcome

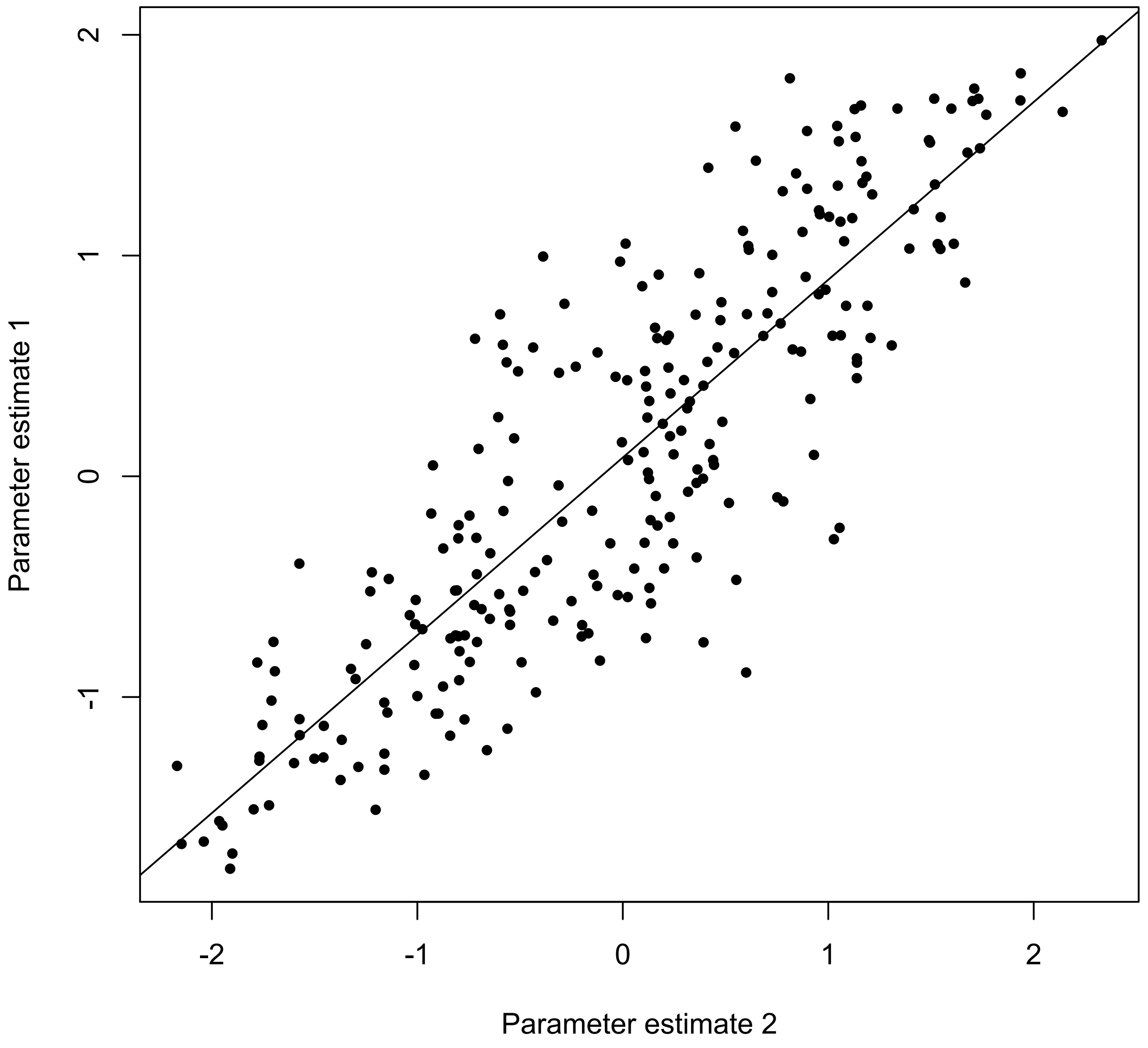


Internal consistency = .958

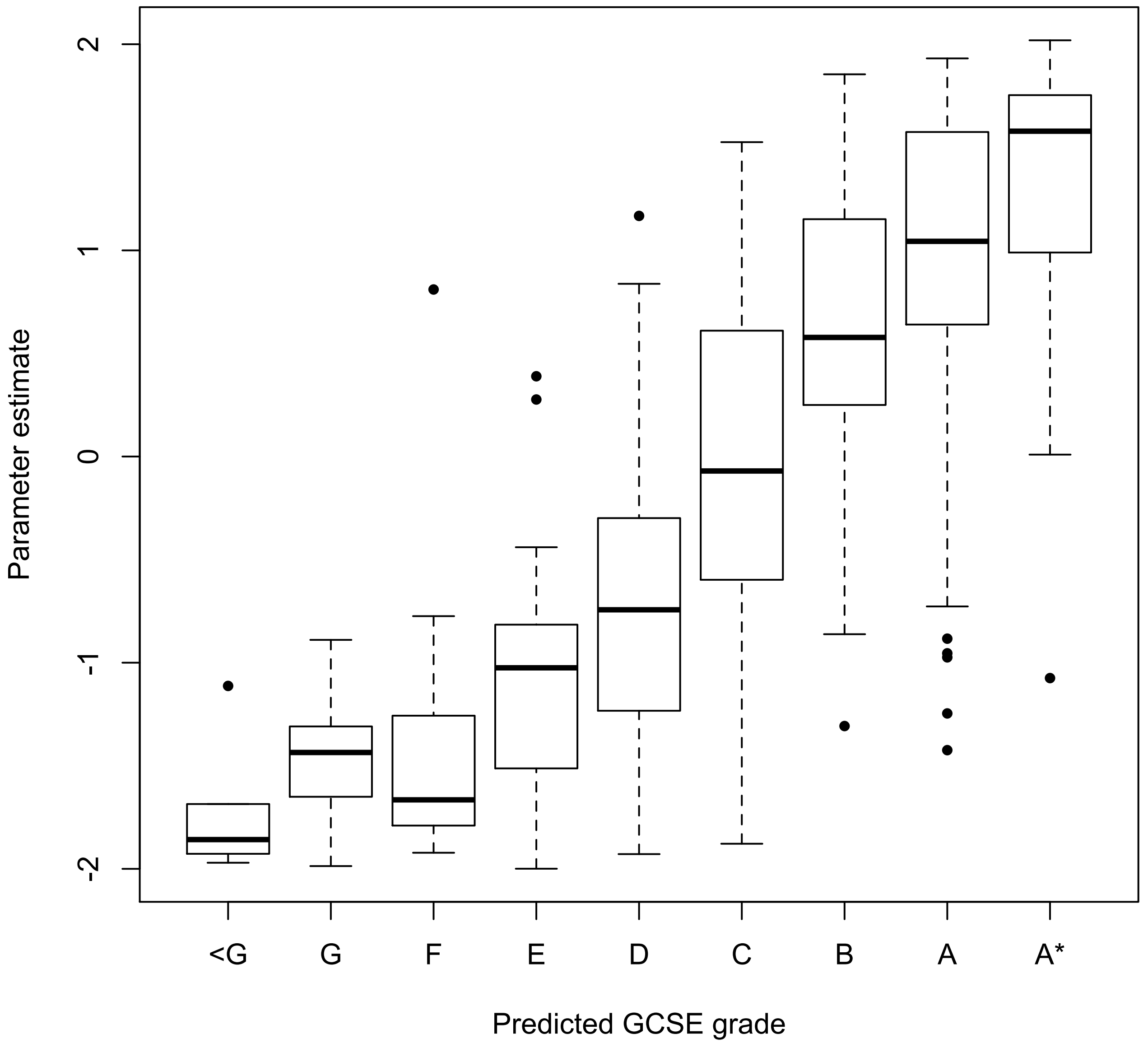
(Rasch Separation Reliability Coefficient)

Judging outcome

Inter-rater reliability (*N* = 249)
r = .861

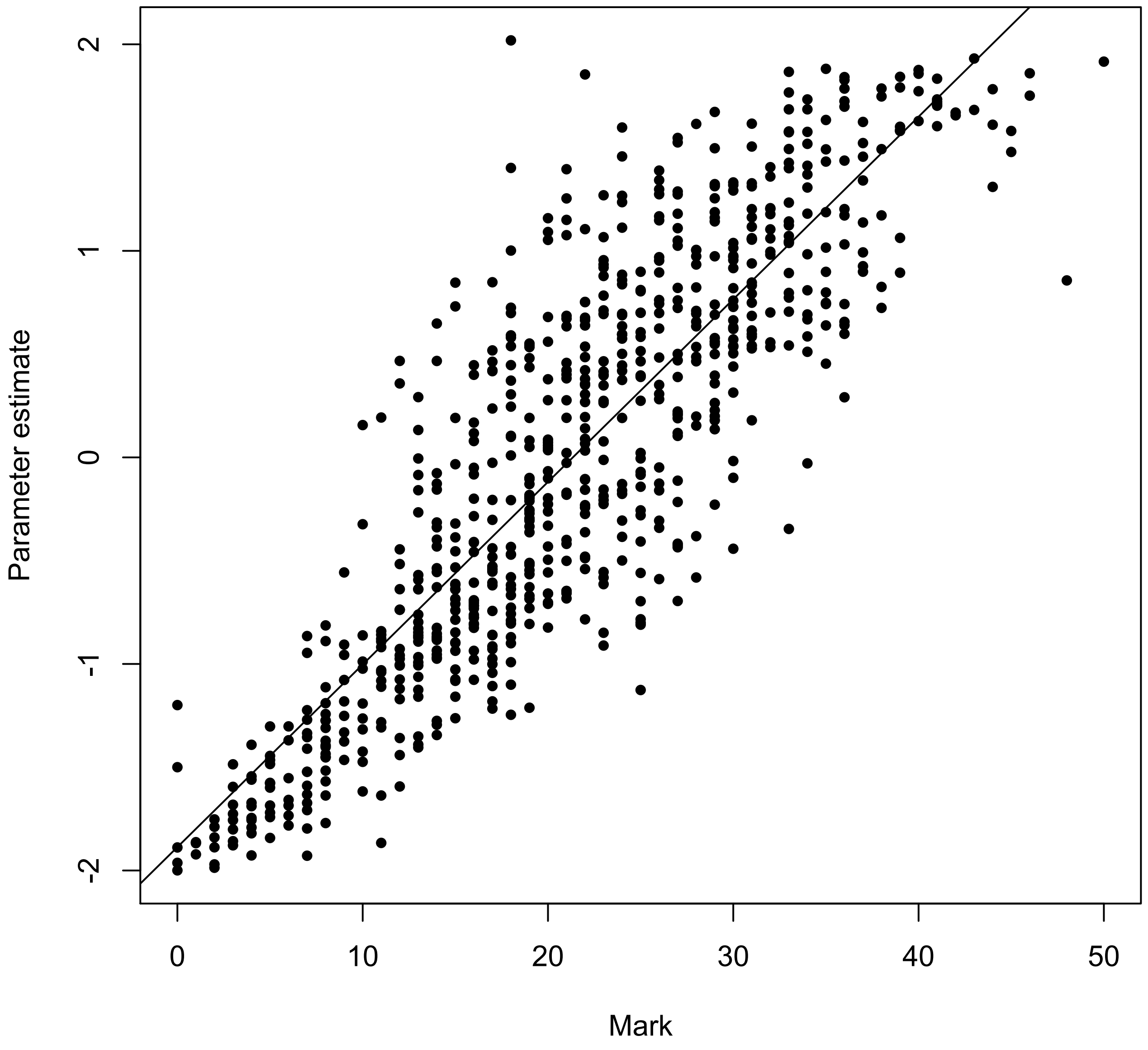


Validity (*N* = 750)
r = .708

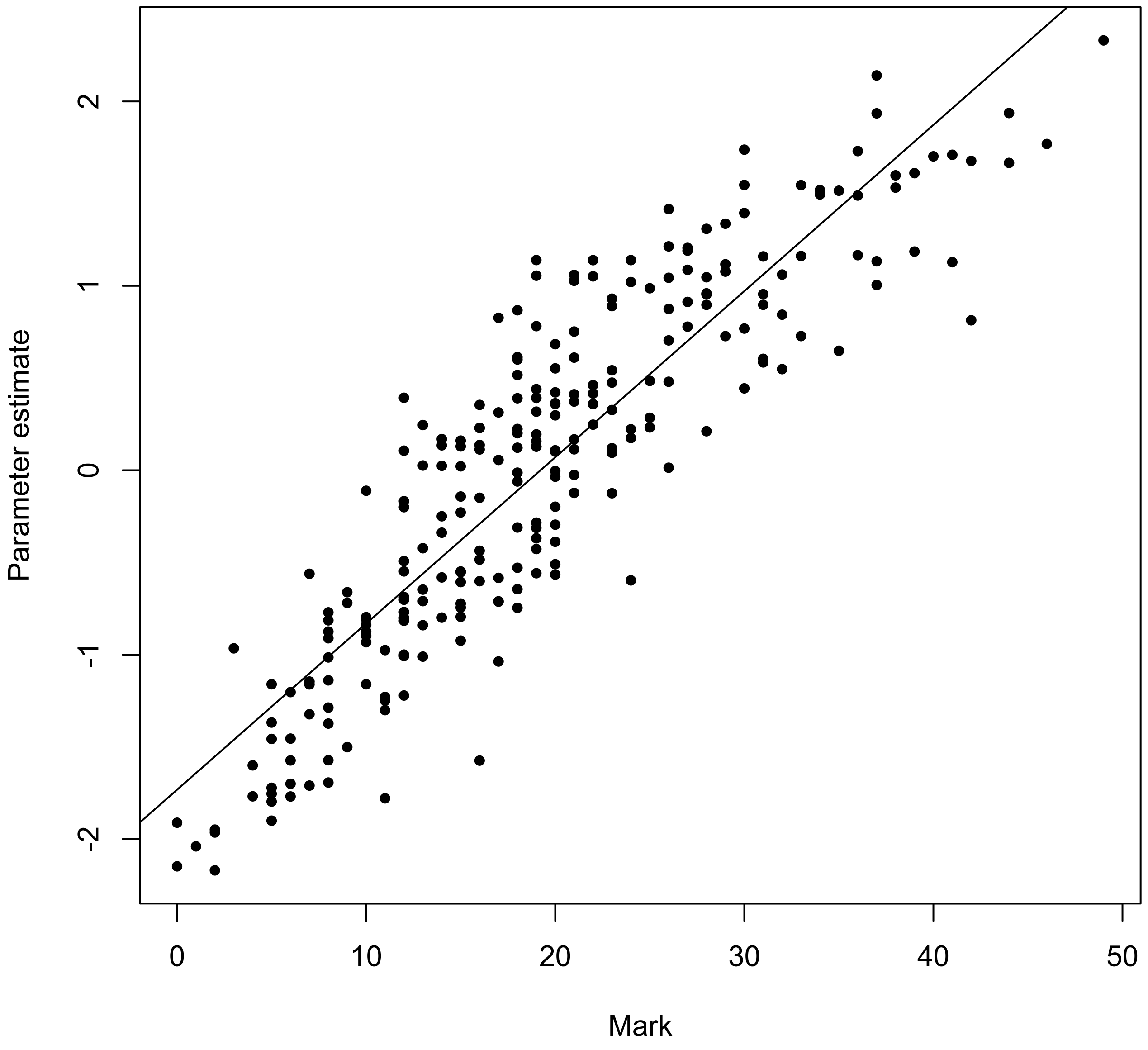


Judging and marking

750 scripts
r = .860



250 scripts
r = .891



Assessment summary

	marking	judging
‘internal consistency’	0.720	0.958
inter-rater reliability	0.907	0.861
validity (c.f. grade)	0.718	0.708
validity (judging vs. marking)	0.860	

Designing the paper

Evaluating the paper

Assessing the paper

Judge feedback

Please indicate the influence of the listed features when judging your allocated pairs of students' work.

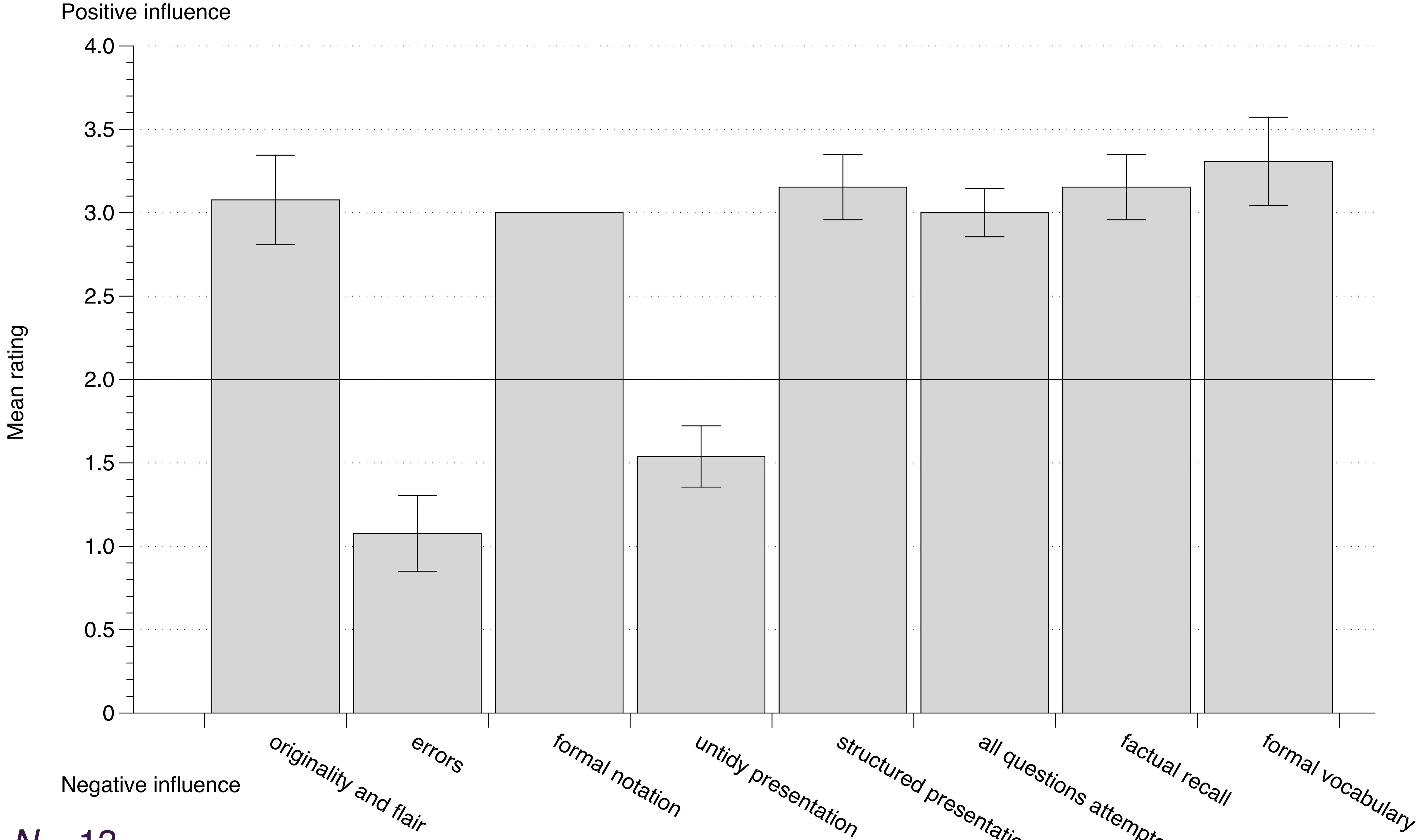
1. student displays originality and flair
2. presence of errors
3. use of formal notation
4. untidy presentation
5. structuredness of presentation
6. all questions attempted
7. student displays good factual recall
8. use of formal mathematical vocabulary

strong positive influence

↕

strong negative influence

Judge feedback



N = 13

Marginal difference between negative influences (*p* = .055)

No significant difference between positive influences (*p* = .165 to .771)

Open text feedback

Open text feedback

I really enjoyed it, it has created much discussion within my family and friends. I love the style of questions and thoroughly enjoyed the judging. I thought I may get bored but I didn’t! Does this mean I am a geek?

It has been very interesting! It was mind numbingly boring too, and I found that 50 was the most I could do in one sitting.

Open text feedback

If they made a rude comment about the question ("this is such a silly question") or drew a silly picture then I found it hard not to be negative towards them!

We can't do anything about students who choose to be silly/throw away marks, but it is in everyone's interests to have the student also believing in the paper, and I sensed that often this wasn't happening

Open text feedback

> The software was cumbersome, the downloading of the papers and the scroll through taking an age at times, there is no way you could judge 50 in one hour. Other than that fine

Conclusions

- Examiners produced a paper with less content and more problem solving when freed from marking constraints
- Comparative judgement performed reliably and validly as an assessment approach

Further work

- Improvements to the web interface
- Refinement of tasks appropriate for assessing by comparative judgement
- The potential for peer assessment
- Further work into judging processes

Acknowledgements

Funders
The Royal Society
The Nuffield Foundation

Thank you to all our judges, pupils and teachers

Ian Jones
I.Jones@lboro.ac.uk