SQE assessment methodology: assessing day one competence

Dr Eileen Fry, Director of Assessment, Kaplan
Dr Lee Coombes, Director of Psychometrics and Assessment Development, Kaplan

Assessing day one competence

- Current Good Assessment Practice: Fundamental Psychometric Principles
- Marking and competency-based exams
- Setting standards: ‘cut scores’ and pass marks

Fundamental psychometric principles

- High stakes exams which lead to admission into a profession have to fulfil fundamentally different criteria from eg university exams

- To provide a technically competent licensing exam, which tests people appropriately and fairly, you need:

  - Reliability
  - Precision
  - Validity

Fundamental psychometric principles

- **Reliability**
  Numerical – to do with consistency and predictive utility. *Does the test rank order candidates in a way that would be replicated in another exam?* Measured most simply by the alpha co-efficient

- **Precision**
  Numerical– statistically related to reliability. *How accurate is any candidate's score?* Estimated by the Standard Error of Measurement (SEm)

- **Validity**
  A unifying framework for a "good test": testing the right things in appropriate ways; includes reliability and precision

Why?

- "*I was lucky what I had revised came up*" should have no place in a high stakes professional exam

- In a high stakes professional exam whether or not candidates qualify must be based on an assessment of appropriate competencies and knowledge which would be replicated if they took a similar exam and which reaches a very high standard of accuracy

- Responsibility to the consumer
- Responsibility to the profession
- Responsibility to the candidates

- The exam must be defensible in the courts

Reliability: how it is measured

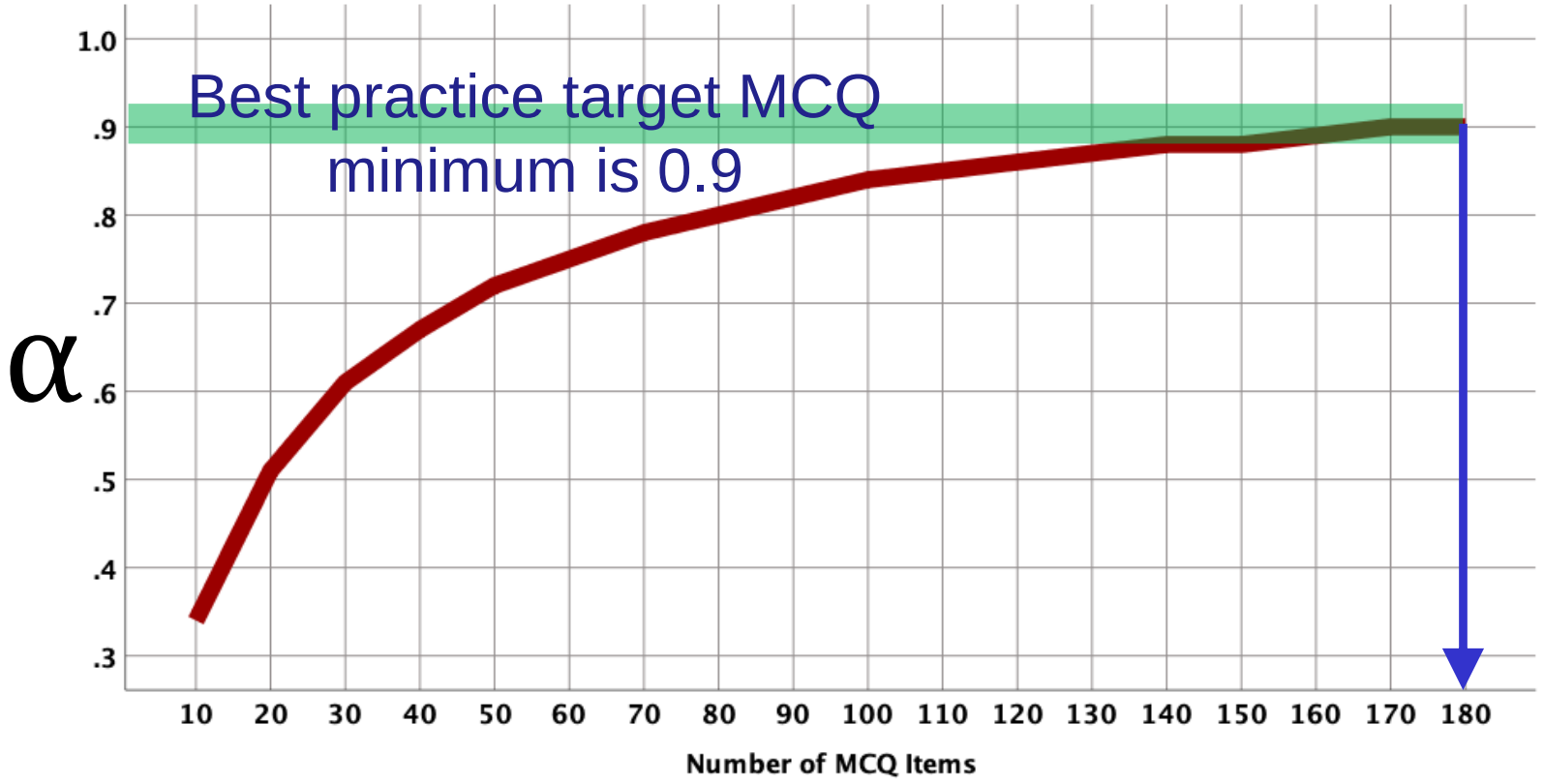
- Reliability: *Does the test rank order candidates in a way that would be replicated in another exam?*
- Ideally, measured by candidates taking two exams
- Alternative: divide the test into two parts, as many ways as possible; compare performance on the halves
- *Coefficient alpha* averages the result of dividing the test into all possible halves and comparing performance (the correlation) on the halves
- α ranges from zero to one
- Best practice targets are at least 0.8 for an OSCE; at least 0.9 for an MCQ. This is because high stakes professional exams have to get it right

What does this mean for exam design?

- If you ask one question on contract, you will have little idea of whether a candidate will get the next question right
- If you ask more questions, the predictive power of the result will increase.
- Longer tests (unless there is something very wrong with them) are more reliable than shorter tests.
- But test design is a compromise– a much longer exam will also be more expensive

Longer tests are more reliable (= alpha is bigger)

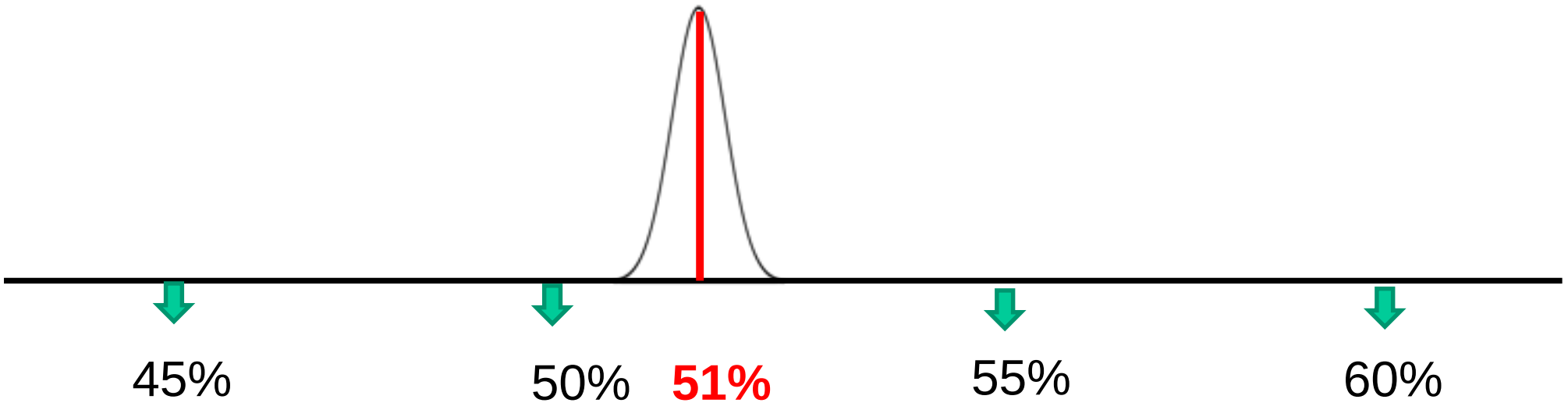
Example is SQE Pilot MCQ



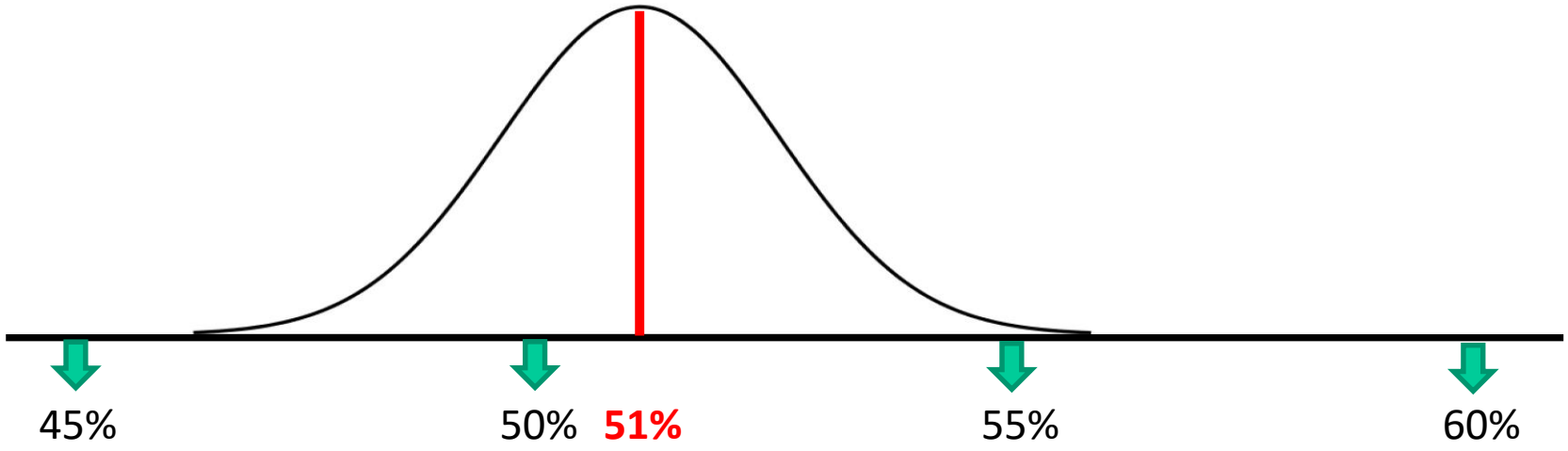
What is precision?

- All measurement has some error
- There is difference between a candidate's *actual score* on the test and their *true score* which would result from their taking *an infinite number of similar tests*
- We know their actual score but can clearly never *know* their true score. We can estimate its value statistically, within a range using the Standard Error of Measurement (SEm)

An illustration of SEm: A mark of 51% might be very precise, say very probably between 50% and 52%



Another illustration: A mark of 51% might be quite inaccurate, say very probably between 46% and 56%



The Standard Error of Measurement SEm

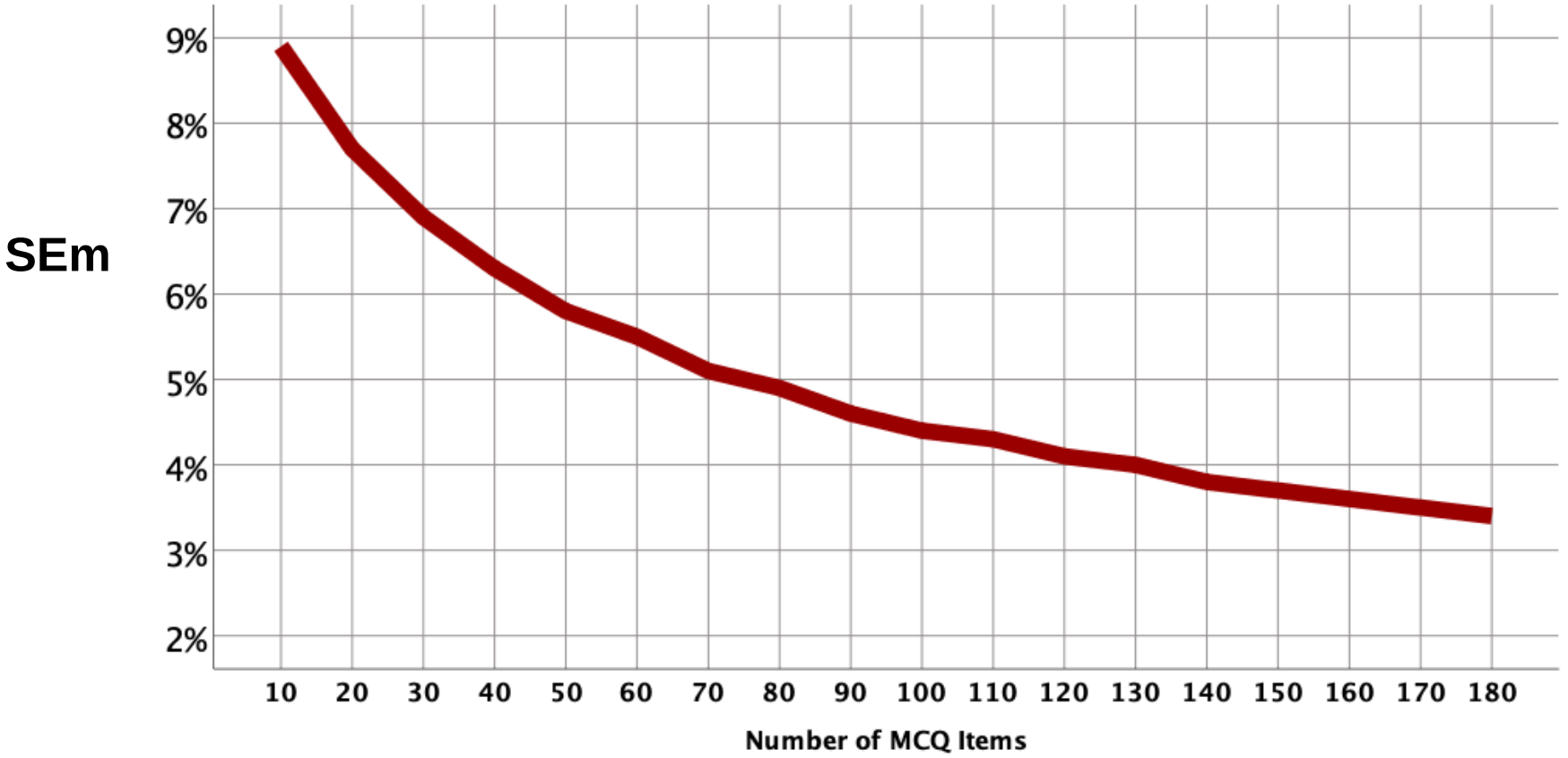
SEm (Precision) depends upon:

- Test length
- Test reliability (α)
- The spread of candidates' scores

The lower the SEm the higher the quality of the exam

Test precision – standard error of measurement (the SEm) – decreases as test length increases

Example is SQE Pilot MCQ



What does this mean for exam design?

- Unless there is something very wrong with the test candidates' scores are more precise from longer tests

- But test design is a compromise and this must be balanced with competing factors such as cost

Validity

- Kane Validity Framework: 5 key domains of validity evidence:
  - Assessment content (appropriate content, sufficient sampling etc)
  - Assessment response process (marking, quality control etc)
  - Internal structure of the assessment (reliability, precision, standard setting etc)
  - Relationship to other variables (performance on other tests – statistical concurrent validity etc)
  - Consequences of the assessment outcome – (for all stakeholders - reasonableness and reliability of pass/fail determination, appeal procedures etc)

For further information, see the Kane Validity Framework. Eg Kane, MT (2013) Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1), 1-73

Reflections on marking

Exercise: working individually as examiners, please give this student a mark out of 10

The student's task:
"Multiply 269 by 63"

Answer:

```
   269
x   63
-----
   807
 16040
-----
 16847
-----
```

Marking legal skills stations

Solicitors Regulation Authority

Some issues:

- The results and the approach
- A small mistake that could result in a negligence claim given the fact pattern and vice versa
- In looking at the detail the examiners may miss the whole:
  - The scattergun approach
  - The answer which makes the right points but lacks understanding
- Mark schemes that require examiners to exercise professional judgment have greater validity for entry to a profession *provided* there is a level descriptor and examiners have adequate training, preparation and monitoring

Competency based marking of legal skills assessments in the SQE

- Marking is based on the professional judgement of examiners informed by competency as defined by the level descriptor (the Threshold standard) and the Functioning Legal Knowledge

- This will involving grading performance against the assessment criteria in terms of the required level of competency. Typically this might involve grading as follows:
  - A. Superior performance: well above the competency requirements of the assessment (5 marks)
  - B. Clearly satisfactory: clearly meets the competency requirements of the assessment (4 marks)
  - C. Marginal pass: on balance, just meets the competency requirements of the assessment (3 marks)
  - D. Marginal fail: on balance, just fails to meet the competency requirements of the assessment (2 marks)
  - E. Clearly unsatisfactory: clearly does not meet the competency requirements of the assessment (1 mark)
  - F. Poor performance: well below the competency requirements of the assessment (0 marks)

- This approach ensures the connection is maintained between marking and the purpose of the assessment and so helps maintain its Validity

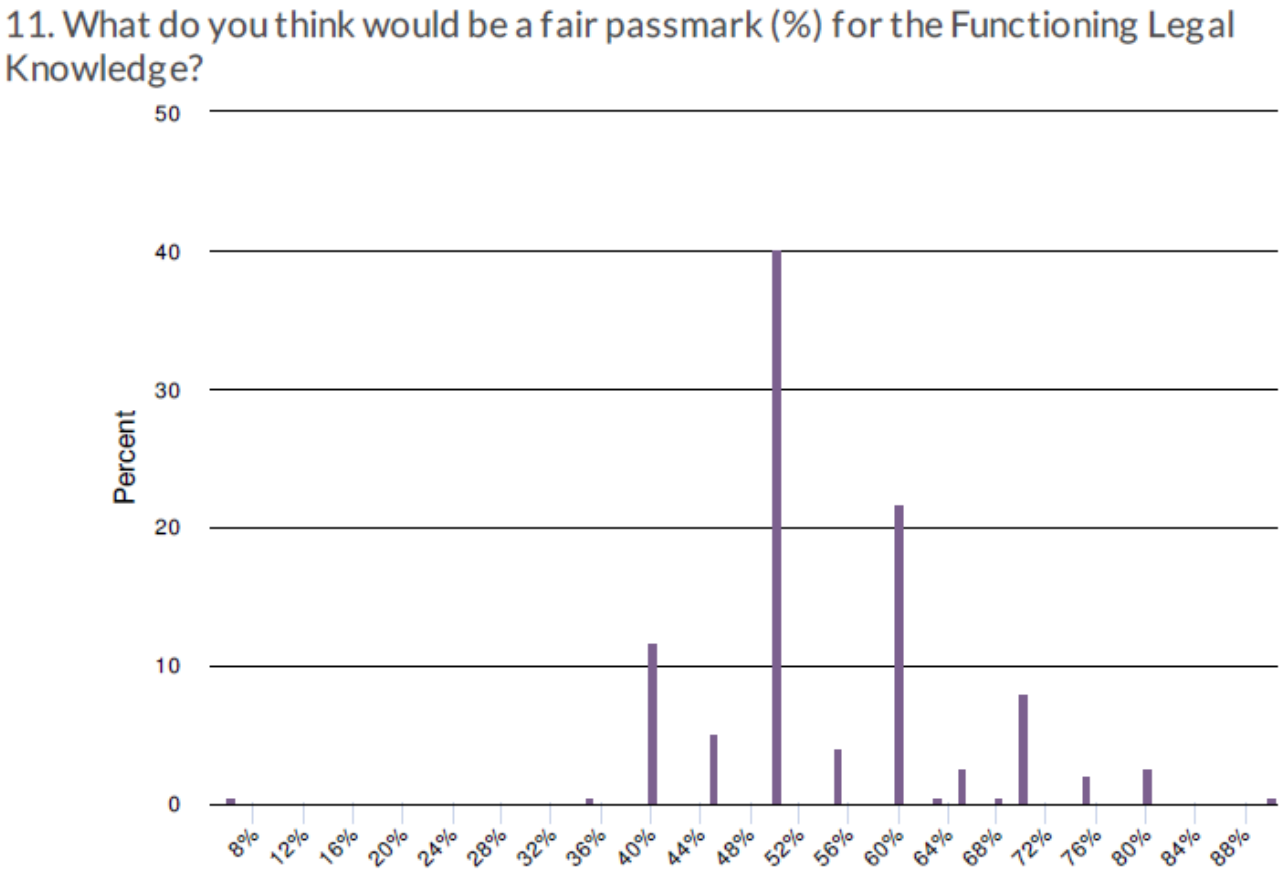
Solicitors Regulation Authority

Setting the pass mark for the SQE: cut scores and pass marks

What should the pass mark for the SQE be? The traditional approach

Choose an appropriate mark

SQE1 Pilot:



Why can’t we use the traditional approach

- What pass mark should we choose and how can we justify our approach?
- 40%? 50%? 60%?
- The exam might be more difficult or easier than usual

- Fairness to candidates
- Protecting the consumer
- Maintaining the standards of the profession

- Anticipated legal challenges

What is the alternative?

- Don’t start from a number and apply it to the script
- Start from a description of the just passing candidate – a level descriptor – and apply it to the exam
- The pass mark is based on the professional judgement of the examiners about the application of the level descriptor (the threshold standard) to that particular assessment
- Results in a justifiable pass mark which adjusts to the difficulty of the exam and maintains standards between cohorts even if the ability of cohorts is different

How is the level descriptor applied to the exam?

- Angoff method for objective testing (e.g.MCQs)
  - A panel of solicitor ‘judges’ *assesses individual test items* and estimates the performance of a ‘just passing candidate’ on each
- Borderline regression (for legal skills stations)
  - In addition to *providing detailed marks on the candidate’s OSCE station performance*, the examiner gives *a global estimate of outcome on the station* (eg Clear Pass). The latter is used to set a cut score on the former
  -
- Accommodating test unreliability
  - Modifying the cut score to a pass mark
- Triangulating’ – when more than one method is used

Angoff method (for objective testing including MCQs)

- Standard setting group of 10 – 16 ‘judges’ = solicitors
- All qualified solicitors, should include some newly qualified
- NOT people with unrealistic expectations

1. Use of level descriptor of the just passing candidate to discuss characteristics of a candidate who will just pass
2. Each panel member assesses the likely performance of 10 just passing candidates on each of the questions: how many would get it right?
3. The average rating of the panel members becomes the difficulty rating of each question
4. The item difficulty ratings are then averaged over all questions to calculate the borderline cut score

Completed spreadsheet from 15 judges

Solicitors Regulation Authority

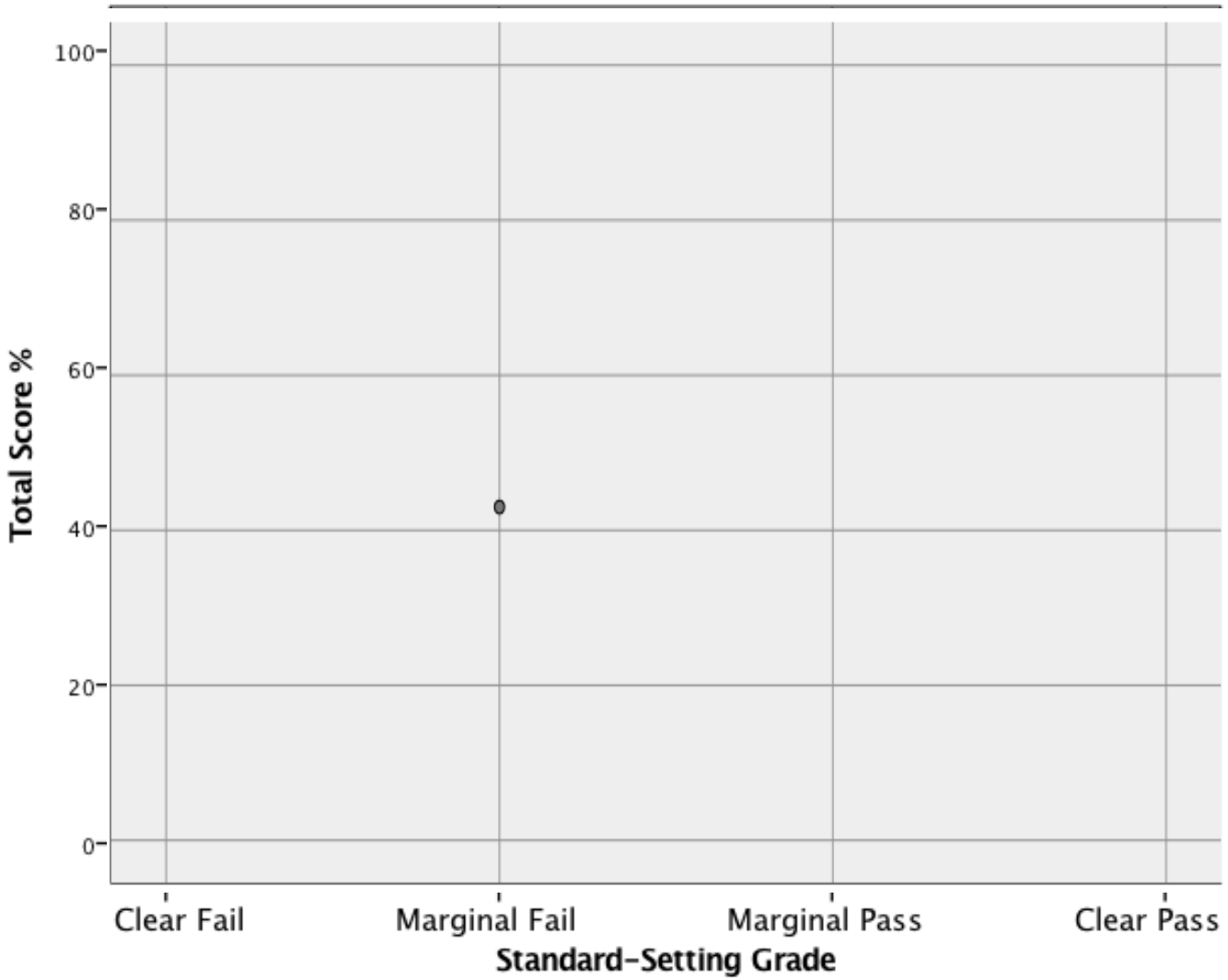
Q No	Examiners															OUTCOMES				
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	std dev	mean	range	min	max
1	1	2	3	2	3	3	3	2	1	4	3	3	2	1	1	0.96	2.3	3	1	4
2	5	5	6	6	7	7	6	6	5	5	7	9	8	8	7	1.25	6.5	4	5	9
3	9	9	8	6	7	8	7	7	6	6	7	7	6	7	8	1.01	7.2	3	6	9
4	1	2	3	2	3	3	3	2	1	4	3	3	2	1	1	0.96	2.3	3	1	4
5	5	5	6	6	7	7	6	6	5	5	7	9	8	8	7	1.25	6.5	4	5	9
6	7	7	7	6	7	8	7	7	6	6	7	7	6	7	8	0.64	6.9	2	6	8
7	5	5	6	2	3	3	3	2	1	4	3	3	2	1	1	1.53	2.9	5	1	6
8	5	5	6	6	5	7	6	6	5	5	7	9	8	8	7	1.29	6.3	4	5	9
9	9	9	8	3	7	8	7	7	6	6	7	7	6	7	8	1.46	7.0	6	3	9
10	2	3	2	3	4	3	2	3	7	3	2	4	2	3	3	1.28	3.1	5	2	7
Total	49	52	55	42	53	57	50	48	43	48	53	61	50	51	51	11.64	50.9			

- If the range for any item is large or if there are outliers, discuss reasons and possibly re-score
- Mean estimate for each item is summed to get borderline cut score for exam (red)

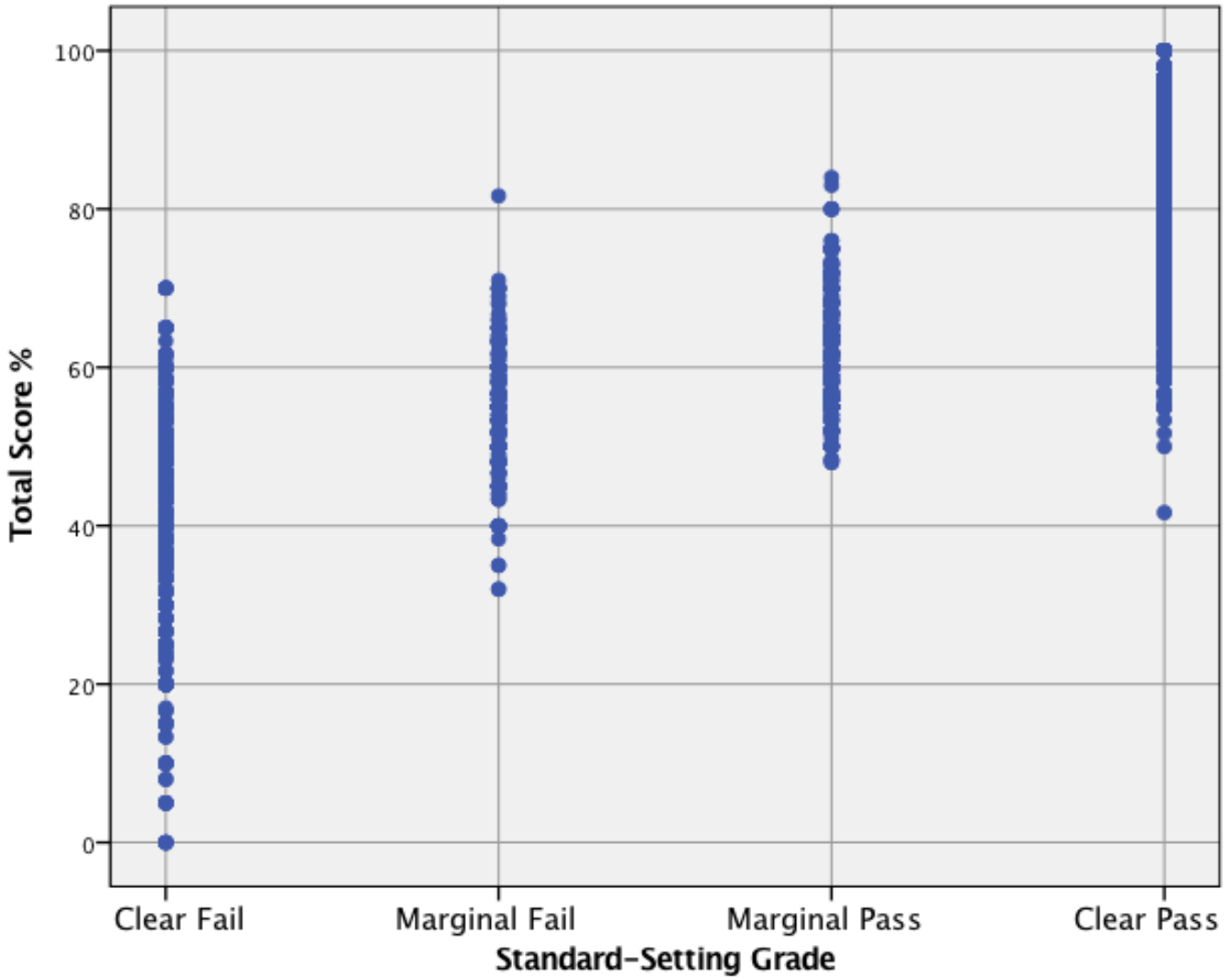
Borderline Regression (for legal skills stations)

- Stations (individual assessments) are the building block for calculating test reliability and setting the passing standard
- In addition to providing detailed marks on the candidate's performance on each assessment (station), the examiner gives a global estimate of whether or not the candidate reaches the standard of a newly qualified solicitor of England and Wales (Pass, Marginal Pass, Marginal Fail, Fail)
- A statistical calculation reviewing all candidate marks in the light of all global judgements will set the cut score on each station

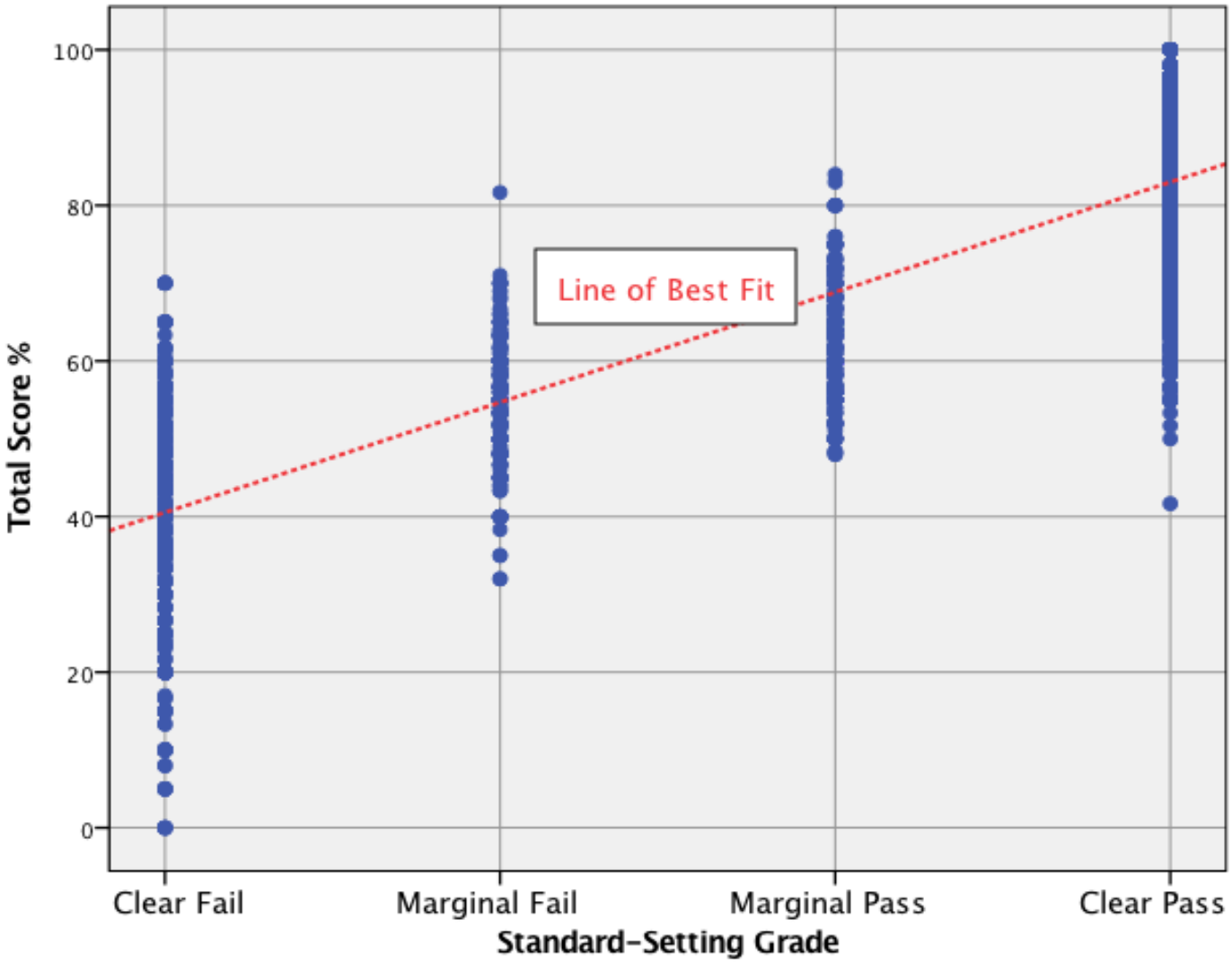
… for one candidate on one station



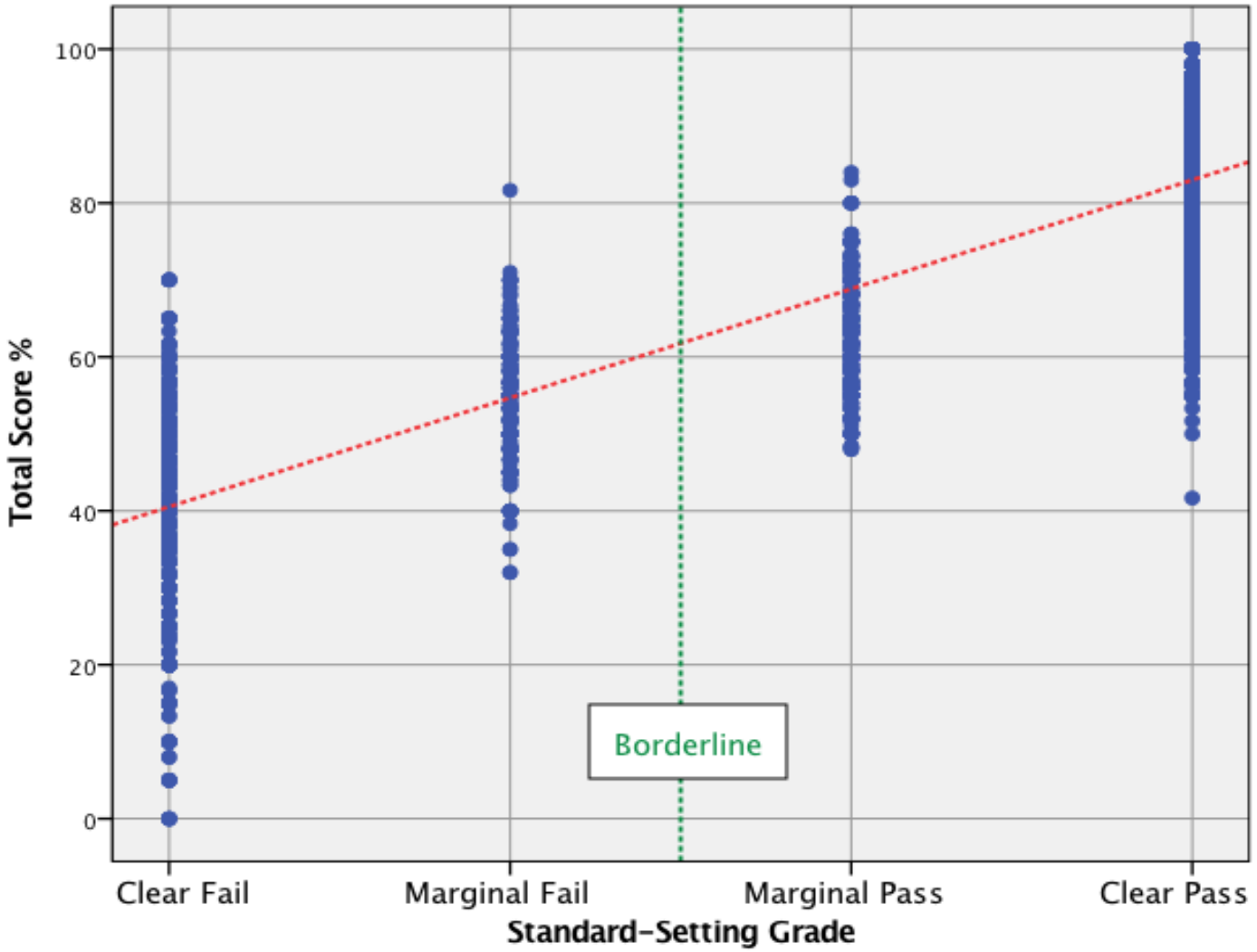
… now, for all candidates on one station



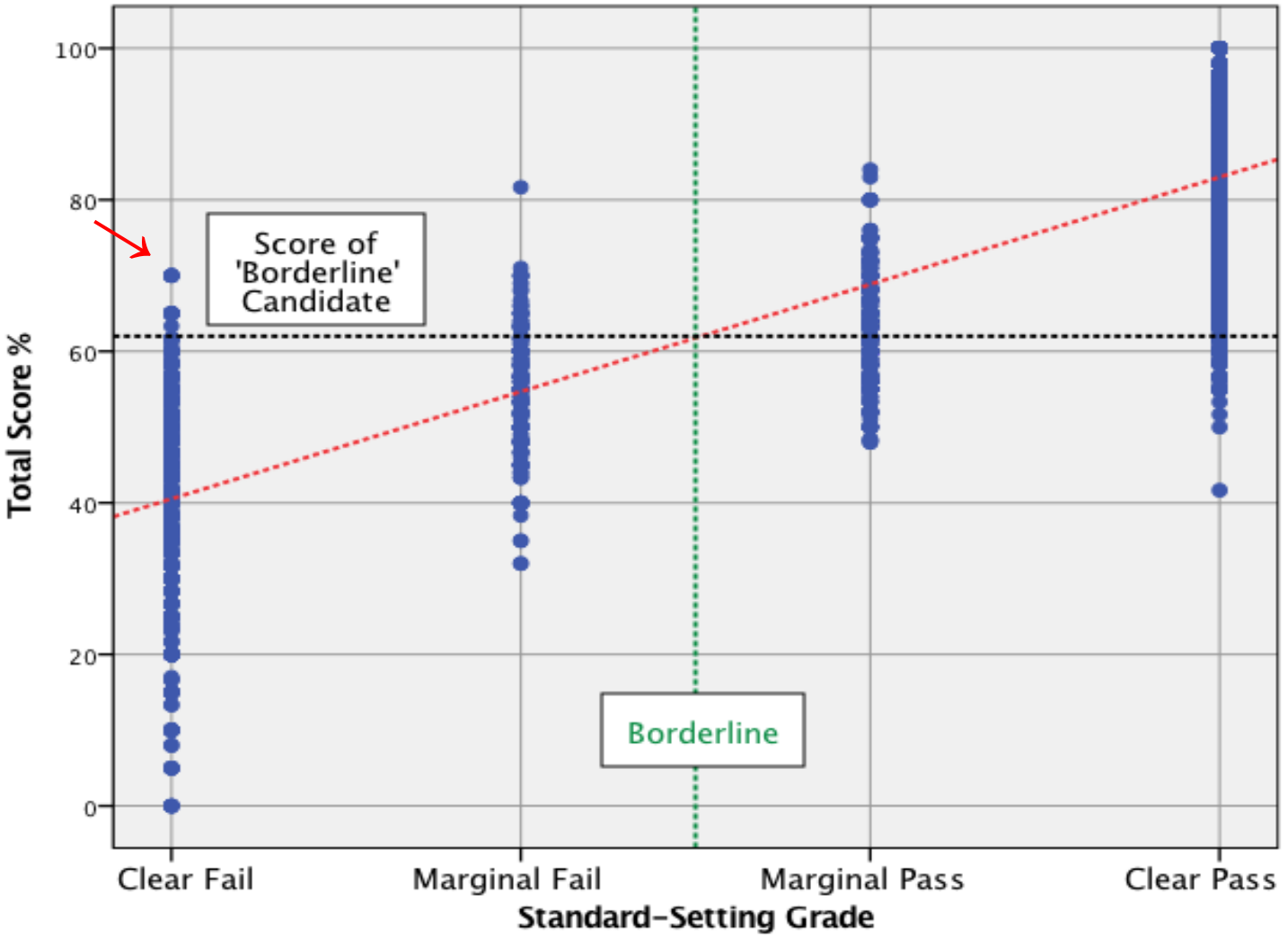
Insert a ‘line of best fit’



Set ‘borderline’ as midway between MF & MP



‘Borderline station score’ is at the intersection



Calculating the cut score for the exam

Solicitors Regulation Authority

- Averaging all the station borderline scores will give the cut score for the whole exam

From cut score to pass mark

- The Angoff method can be used to set the cut score for objective testing (eg MCQs) and borderline regression for legal skills stations
- But this takes no account of precision – the standard error of measurement
- Candidates' true score (on an infinite number of tests) may be higher or lower than their actual score on this test
- In high stakes professional exams the interests of the consumer are normally considered paramount and so an allowance for measurement error is added on to the cut score to arrive at the pass mark

‘Triangulating’

- Often, more than one method of standard-setting is used for a test
- In this case, the strengths and level of each can be reviewed by the Exam Board and some compromise agreed

End

Thank you for your attention

eileen.fry@kaplan.co.uk