Experience of Setting the Standard

Robert Wright MD FESC

UEMS Cardiac Section

Cardiologist, James Cook University Hospital, UK

Chair of Standard Setting for the EEGC

Chair of MRCP Part 2 Written Examination

Exam Validity

- Are we testing the right thing?
- Are we testing it the right way?
- Are our processes robust?

Exam Validity

- Are we testing the right thing?
- Are we testing it the right way?
- Are our processes robust?

- Is a candidate who passes the exam able to apply knowledge in such a way as to indicate that they are a competent practitioner?

Validity and Reliability

- A valid test must be reliable
- An unreliable test cannot be valid
- A reliable test is not necessarily valid

- Reliability is necessary but not sufficient
- Reliability measures consistency or the likelihood of test-retest agreement

Content Validity

- Examination Blueprint
- Sampling from across the syllabus and curriculum
- Establishing the broad domains to be tested and the categories within each domain to be tested
- Within each domain to balance for difficulty
- Inclusion of previously used ‘anchor’ items representative of domains, categories, difficulty and question type

MRCP Pt2 Content by Specialty

Cardiology	10%
Dermatology	5%
Endo/Diabetes	10%
G-I	10%
Haematology	5%
Infectious Diseases	10%
Neurology/Opth/Psych	10%
Oncology	5%
Renal	10%
Respiratory	10%
Rheumatology	5%
Therapeutics	10%

MRCP Pt2 content by category

- Diagnosis – including symptoms & signs, associated features etc
- Investigation – includes interpretation of results
- Management – acute and chronic, prognosis and prevention
- Others – rehab, occupation, DVLA, adolescent medicine, pregnancy, ethics

Validity of Process

- Question Writing
- Question Bank
- Selection of questions for exam
- Review of exam selection
- Standard setting of pass mark
- Analysis of results

Analysis of Results

- Item difficulty – p – percentage of candidates answering correct (20-100%)
- Item discrimination – performance tables for each question
- Item performance and internal consistency – point biserial (item:total score correlation)
  - Does performance on this question correlate with performance on examination overall
  - How to handle negative point biserial

Performance Tables

- 5 groups

- Clear Pass – top 10%
- Pass – 40%
- Just Pass – 20%
- Fail - 20%
- Clear Fail – 10%

Thinking about performance

- In a Pass-Fail exam it is the performance of the exam and the candidate around the cut score that is paramount
- Think of the ‘just passing’ candidate

Poor performance and –ve PBS

- Check the answer key!

Poor performance and –ve PBS

- Check the answer key!
- Decide on any items to withdraw

Setting the Pass Mark

- Norm – referenced
- Criterion – referenced
- Test equated with Item Response Theory

Norm referenced

- A fixed pass rate (common historically)
- Problems
  - Does not take into account variation in the difficulty of the exam or the ability of candidates

Norm referenced

- A fixed pass rate (common historically)
- Problems
  - Does not take into account variation in the difficulty of the exam or the ability of candidates
  - The candidates should not set the pass mark
  - The pass mark should vary with the test difficulty

Criterion referenced

- Pass mark set by a Standard Setting Group based upon the expected performance of a ‘just-passing’ candidate
- Adjusts for variation in the difficulty of the exam assessed by an expert panel
- Problems
  - Significant workload
  - Reliability of the expert panel judgement
  - How to define the ‘just-passing’candidate

Modified Angoff and Hofstee

- Used in the EEGC and other high stakes MCQ assessments of specialist trainees in the UK
- Standard Setting Group is composed of trainees, generalists and specialists. None have been involved in Question Selection.
- N= 6-12
- Receive the questions, separate answer key and instructions 2 weeks before the meeting.

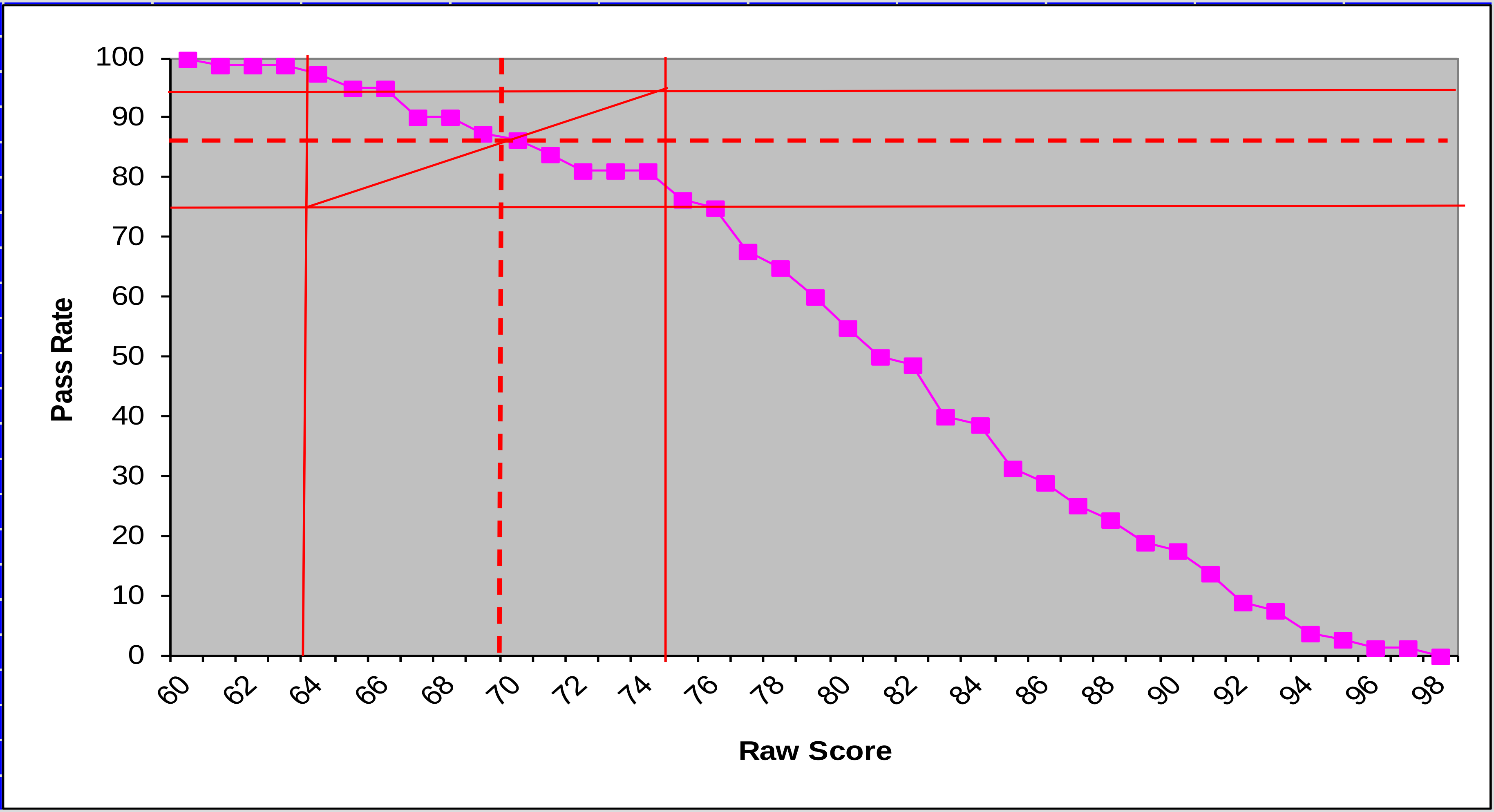
Standard Setting Meeting

- Pre-Meeting scores displayed on a spreadsheet
- The question +/- image is reviewed
- Discussion is lead by the highest and lowest scorer
- Revised scores are entered on to a spreadsheet
- Rejected questions are replaced from a selection of spare questions

Standard Setting Meeting

ID	RW	CL	EMcF	RvdB	MJA		RW	CL	EMcF	RvdB	MJA
115	70	40	50	70	75		70	50	55	70	80
116	75	70	70	70	70		75	70	70	70	70
117	60	40	40	50	55		55	50	50	50	55
118	55	60	65	60	65		55	60	65	60	65
119	45	50	50	70	60		50	50	50	60	60
120	35	40	50	35	60		35	40	45	35	60
Mean	58.5	64.9	58.7	63.1	59.7		56.3	61.9	57.0	60.8	58.4
SD	15.1	12.9	13.9	13.1	11.9		13.8	12.7	12.9	13.7	12.6

EEGC Pilot 2012
Setting the Hofstee Limits



Setting the Pass Mark

- The range of acceptable pass marks is defined by the Trimmed Mean +/- 1.96 SD of the scores of the whole group
- The Trimmed Mean excludes the highest and lowest scorers
- The range of acceptable pass rates is set by the Examination Board

Problems with Angoff / Hofstee

- Time-consuming and costly
- Requires training
- Can be unstable (use Hofstee)
- Is it what candidates *would* or *should* know?
- Difficult for standard setters to derive the acceptable pass rates (use Exam Board)

Problems with Angoff / Hofstee

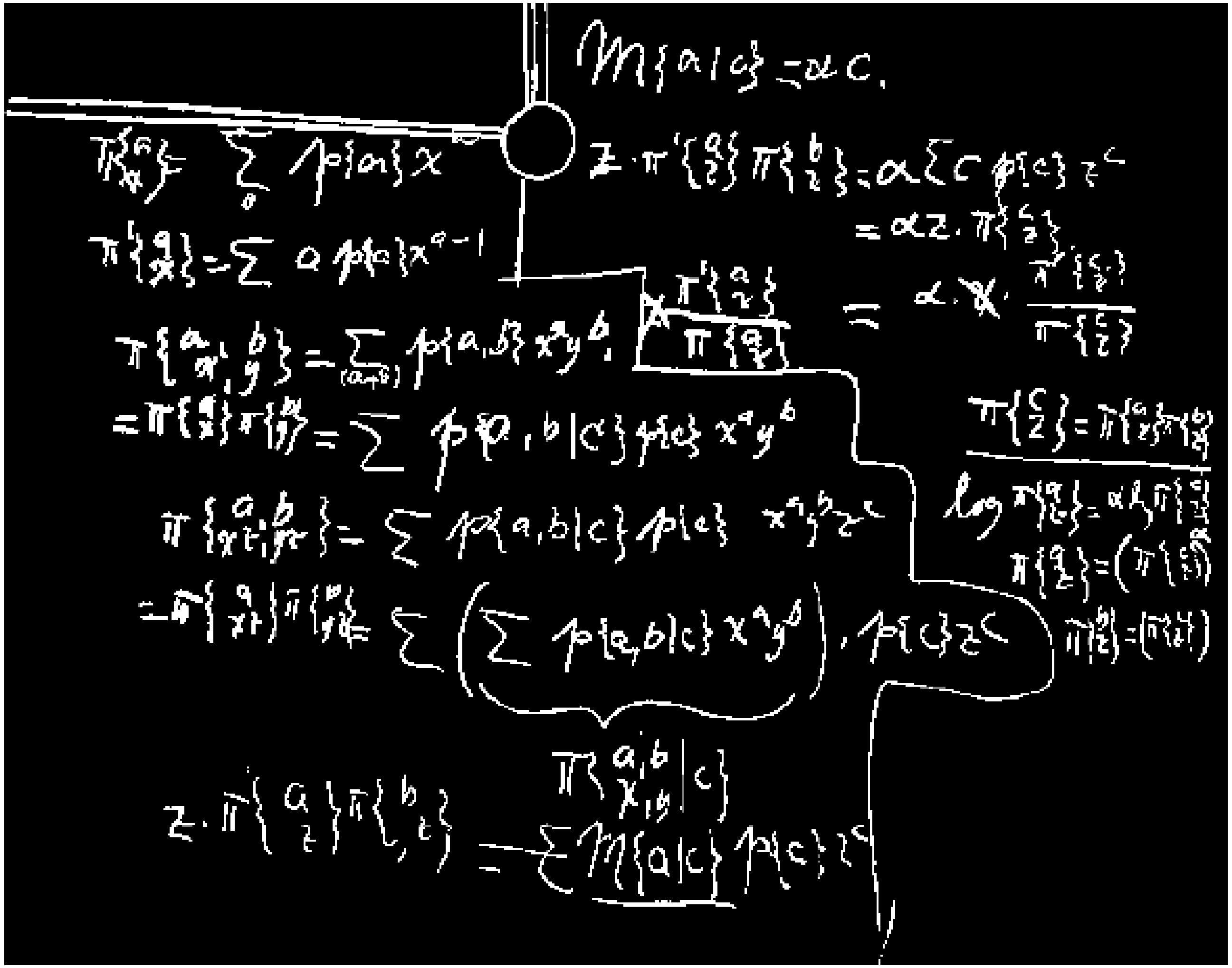
- Time-consuming and costly
- Requires training
- Can be unstable (use Hofstee)
- Is it what candidates *would* or *should* know?
- Difficult for standard setters to derive the acceptable pass rates (use Exam Board)

- It is excellent CPD and Quality Control

Test Equating – used in MRCP(UK)

- Statistical Methods Based upon Item Response Theory
- Refers to previous performance of candidates on a ‘base form’ and on previously sat ‘anchor’ questions then assigns difficulty value to all questions and performance value to all candidates
- Independent of expert clinician panel
- Needs expert statistical input
- Favoured by NBME and ABIM
- Needs relatively large number of candidates

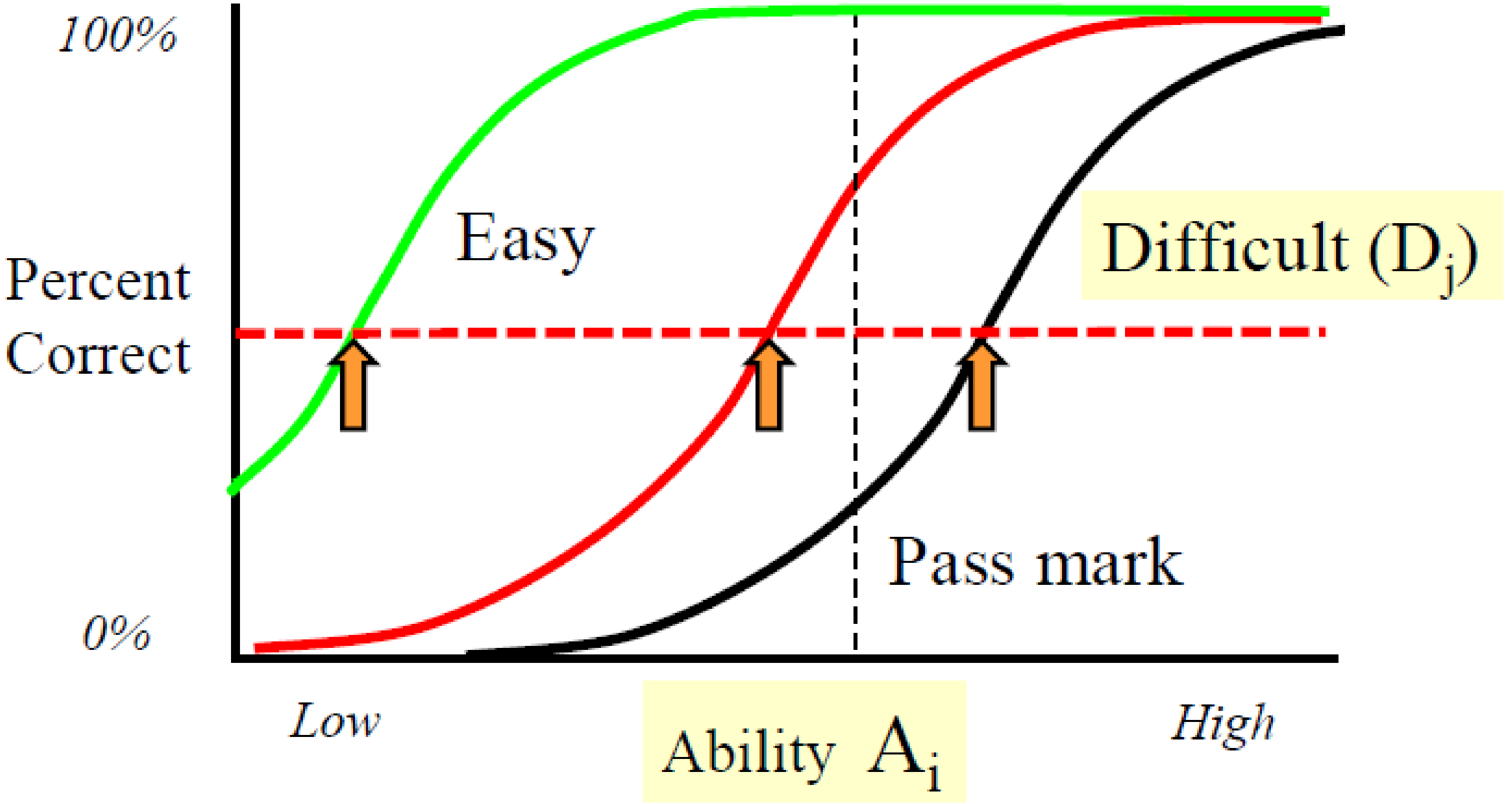
Georg Rasch



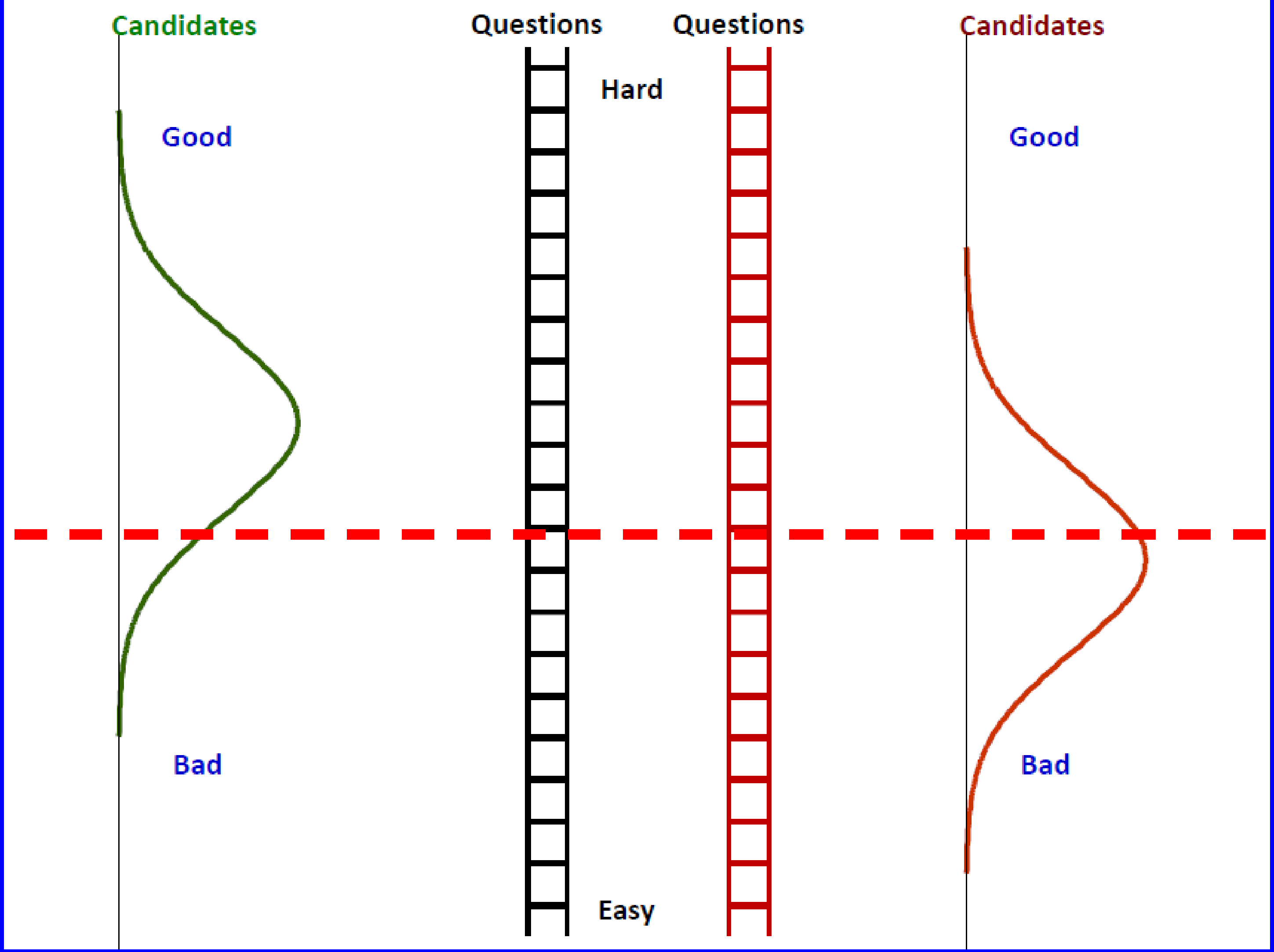
Statistical analysis

- Item Response Theory (IRT)
  - Difficulty

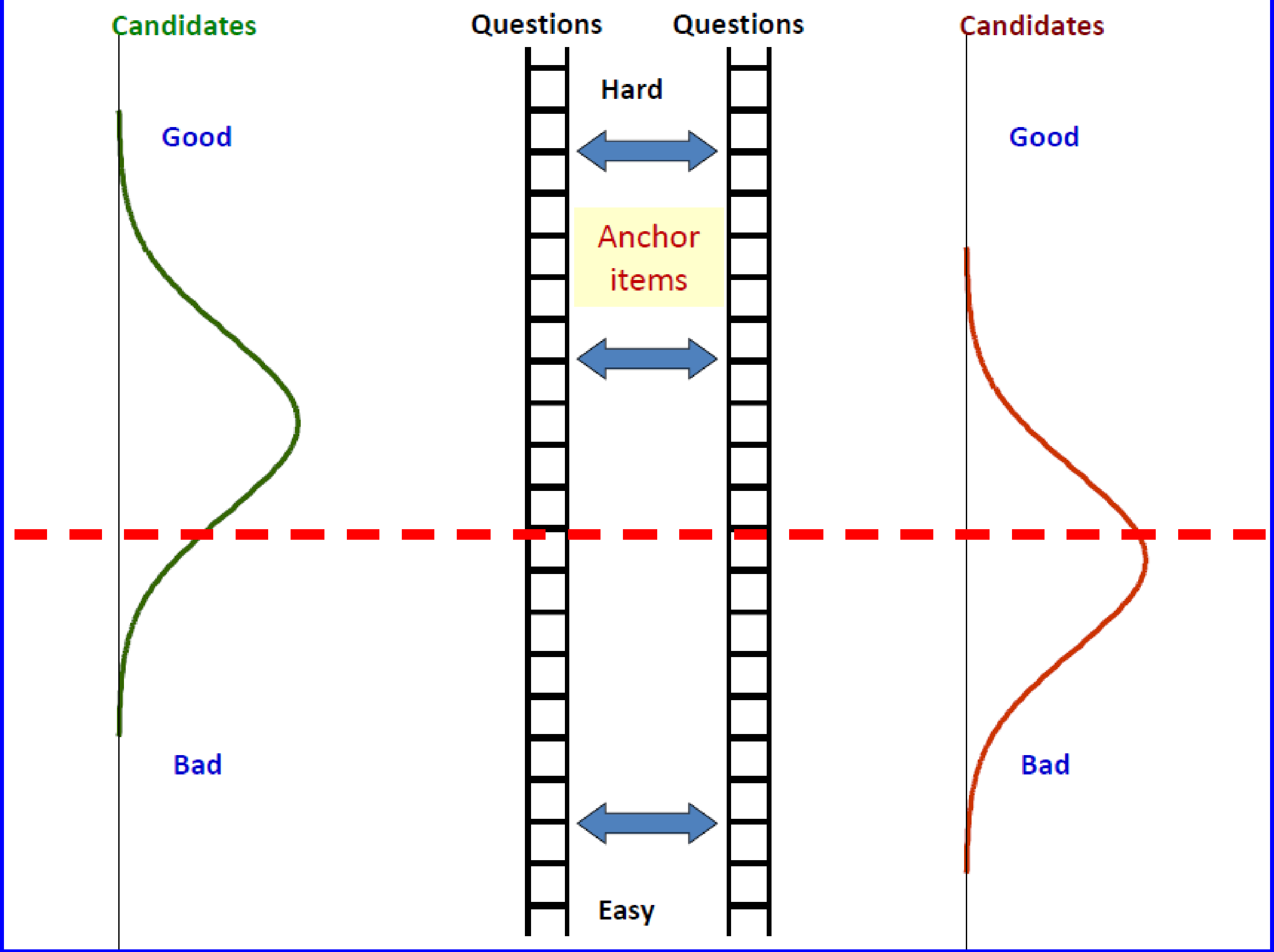
$$p(Correct) = \frac{e^{(A_i - D_j)}}{1 + e^{(A_i - D_j)}}$$



Are the candidates worse?
Are the questions more difficult?



Using Anchors to Equate



What is the output from Equating?

- Measure of overall candidate ability
- Measure of overall exam difficulty
- A pass mark related to a standard scale

Problems with Equating

- Requires large numbers of candidates
- Assumes that MCQ difficulty is fixed
- May need recalibration – a parallel standard setting meeting using Angoff / Hofstee should take place every 3 years

Conclusion

- There is no perfect system
- Psychometricians prefer Item-Response Theory
- Clinicians prefer Angoff / Hofstee

References

- Norcini JJ. Medical Education 2003;37:464
- Livingston SA, Zieky MJ. Passing Scores 1982
- Cohen-Schotanus J. Medical Teacher 2010; 32:154
- Burr SA. BMC Medical Education 2016; 16:34
- Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. Philadelphia, PA: National Board of Medical Examiners. 1996.
- Postgraduate Medical Education and Training Board (UK) PMETB 2007 Developing and maintaining an assessment system - a PMETB guide to good practice