



CEM

Centre for Evaluation & Monitoring

 @ProfCoe

Why assessment may tell you less than you think

Robert Coe, Durham University

researchED Durham, November 2018

THE **BIG** EVIDENCE DEBATE

4th JUNE 2019, DURHAM

193
days

00
hours

12
minutes

28
seconds

Register for the **ballot**

How should
we be using
meta-analysis
and effect
sizes?

<http://thebigevidencedebate.com/>

4 June 2019, Durham



CEM

Centre for Evaluation & Monitoring

Quiz

How precise is a test score?

1. You have created a 20-item right-wrong test and given it to your class of 30 pupils. How far apart do two scores have to be before you can be confident one is really better than the other?
 - a) 3 marks
 - b) 6 marks
 - c) 8 marks

(Assume scores cover the full range 0-20, $SD=5$, Cronbach alpha = 0.7)

How much information is there in a ‘hinge question’?

2. You have taught and seen a student working on this topic and estimate their probability of mastery at 80%, but they get the hinge question wrong.

What is your new estimate of their probability of mastery of the concept?

- a) 20%
- b) 50%
- c) 70%

How should you respond to an exit ticket?

3. You ask a question as an exit ticket at the end of a lesson. 25 of a class of 30 get it right. Next lesson, would you
 - a) Move on (maybe review/revisit later)
 - b) Re-teach to all
 - c) Re-teach to those who got it wrong
 - d) Do none of these (ie something else)
4. If the number getting it right had been lower than 25, would you have done something different?
5. How much lower would it have to be to change your response?



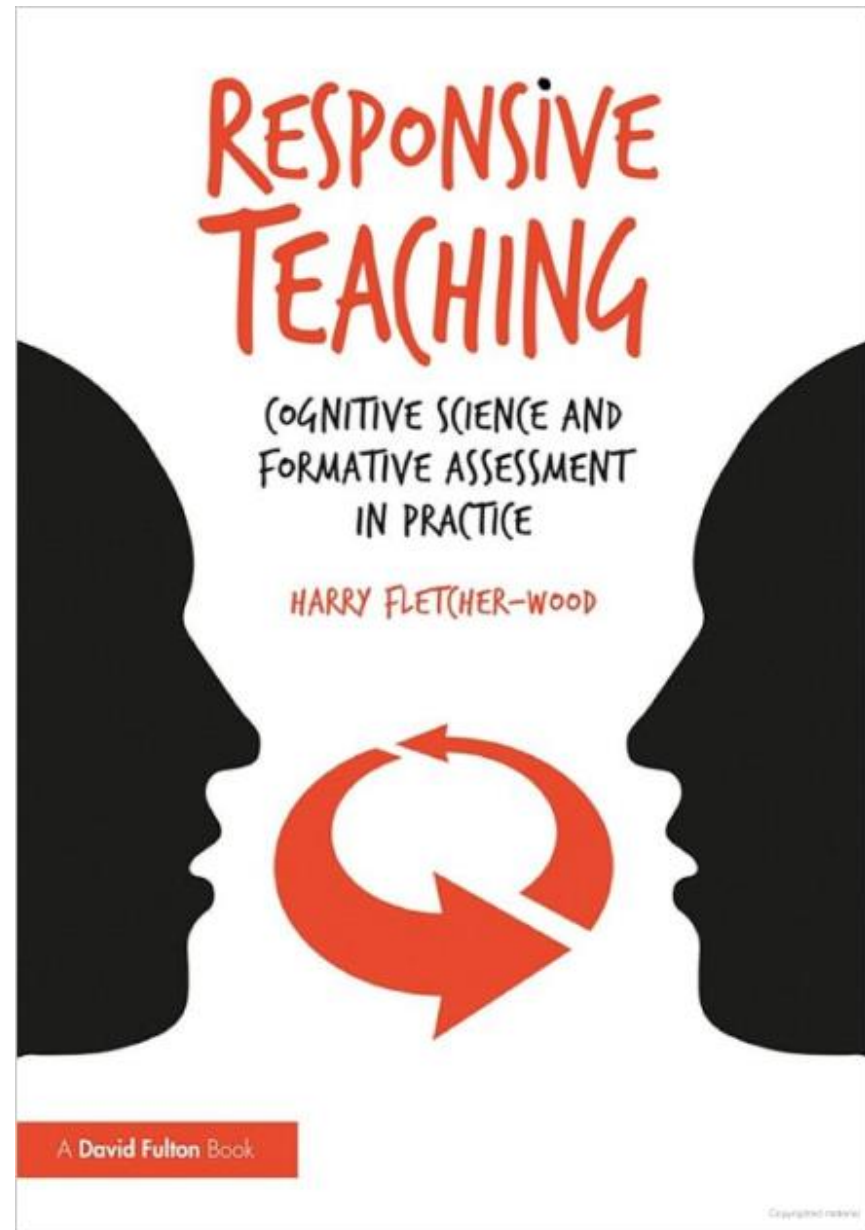
CEM

Centre for Evaluation & Monitoring

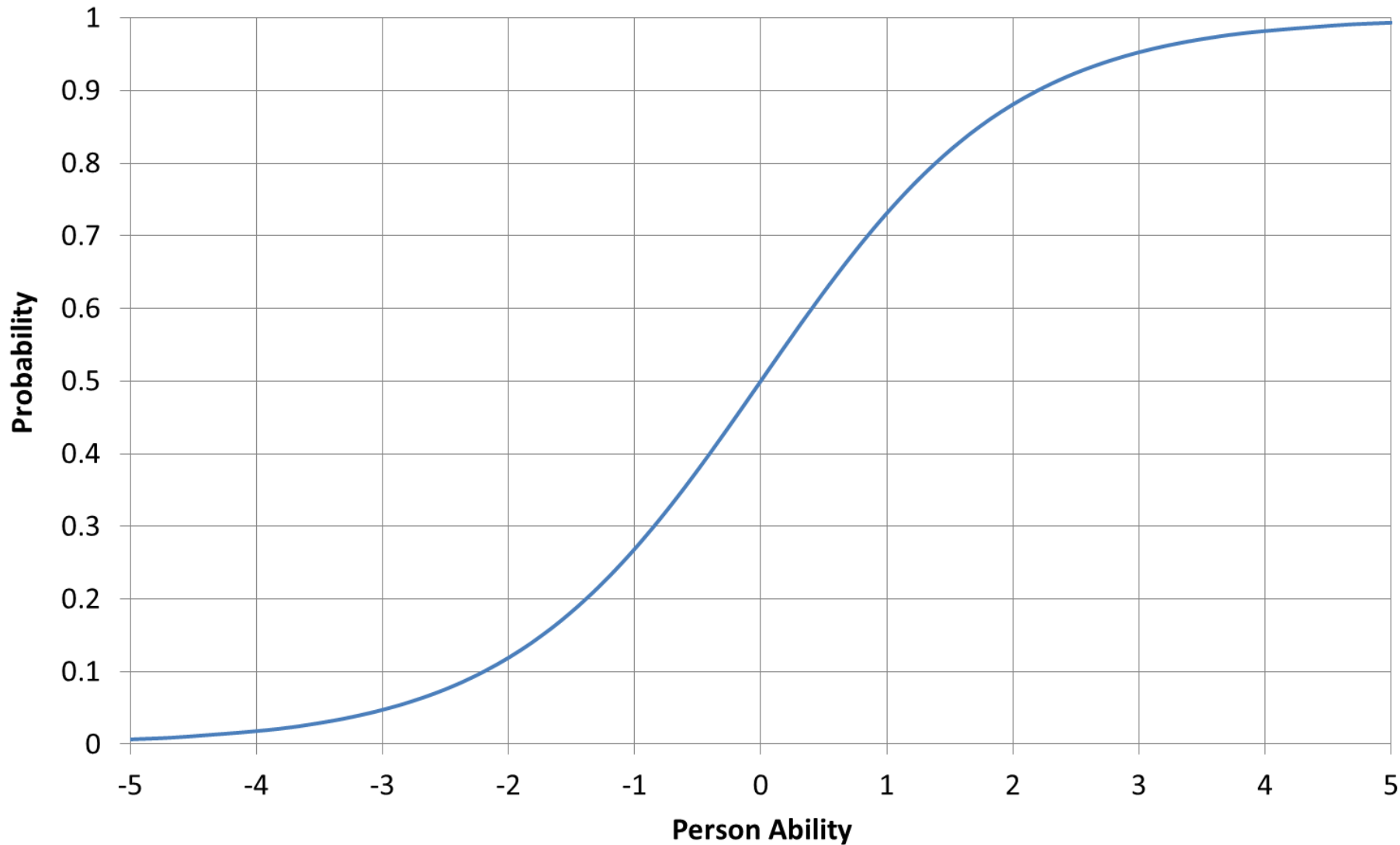
Why does this matter?

A chapter each on

- Hinge questions
- Exit tickets

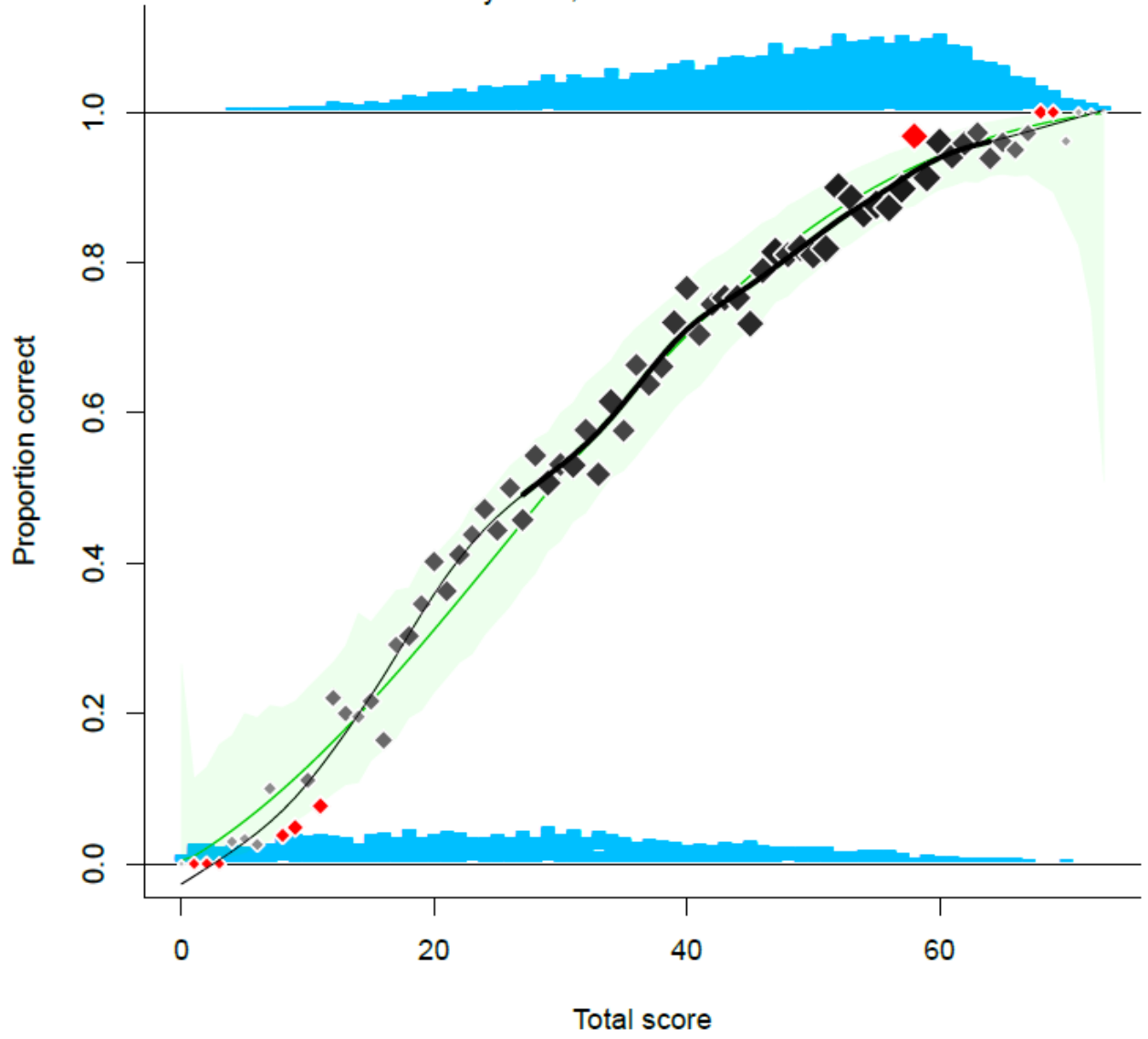


Item Characteristic Curve



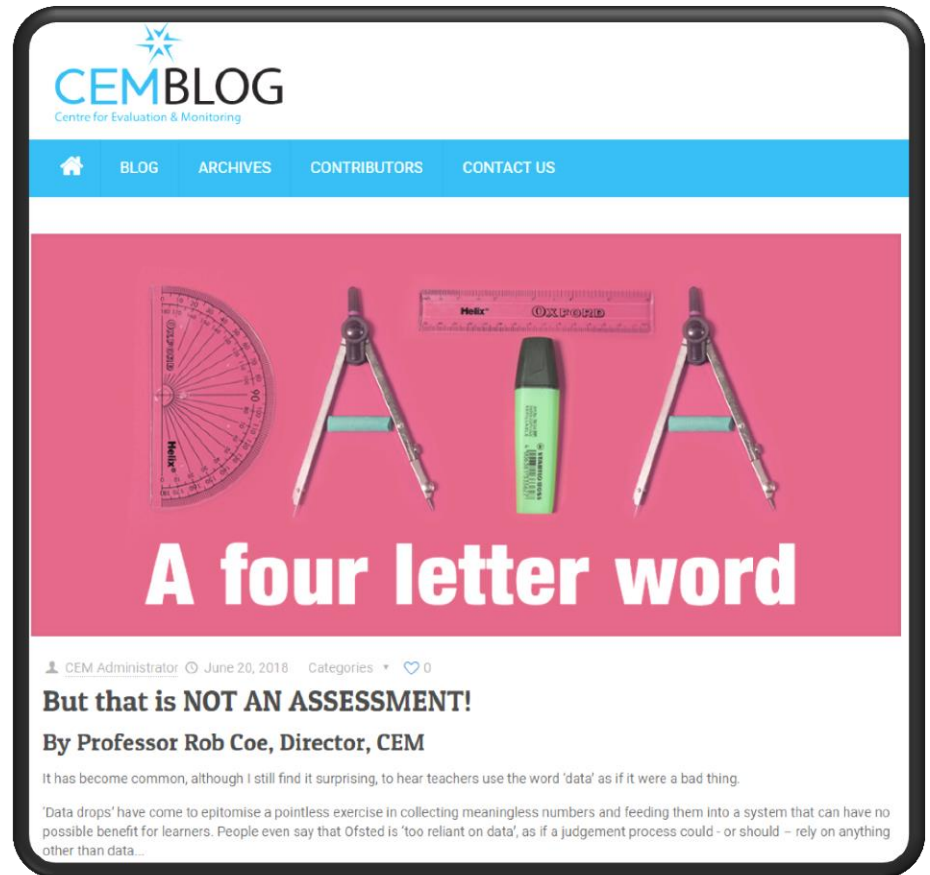
ICCAMS Number: num05dd_r

Facility=0.67; Item-rest correln=0.53



<http://www.cem.org/blog/but-that-is-not-an-assessment/>

- Assessment must be
 - Informative
 - Accurate
 - Independent
 - Generalisable
 - Replicable



The image shows a screenshot of a blog post from the Centre for Evaluation & Monitoring (CEM). The header includes the CEMBLOG logo and navigation links for Home, Blog, Archives, Contributors, and Contact Us. The main content area features a pink background with a photograph of school supplies: a protractor, two compasses, a ruler, and a highlighter. Below the image, the title 'A four letter word' is written in large white letters. The post is dated June 20, 2018, and is written by Professor Rob Coe, Director of CEM. The text of the post discusses the use of the word 'data' in education and criticizes the practice of 'data drops'.

CEMBLOG
Centre for Evaluation & Monitoring

Home BLOG ARCHIVES CONTRIBUTORS CONTACT US

A four letter word

CEM Administrator June 20, 2018 Categories 0

But that is NOT AN ASSESSMENT!
By Professor Rob Coe, Director, CEM

It has become common, although I still find it surprising, to hear teachers use the word 'data' as if it were a bad thing.

'Data drops' have come to epitomise a pointless exercise in collecting meaningless numbers and feeding them into a system that can have no possible benefit for learners. People even say that Ofsted is 'too reliant on data', as if a judgement process could - or should - rely on anything other than data...

A bit of assessment theory ...

- Assessment is the process of capturing and scoring aspects of task performance in order to support inferences
- Inferences are usually about a person and are time-bound
 - ‘She doesn’t currently understand this’
 - ‘He will make a good employee’
- Decisions may be informed by those inferences
 - ‘I need to re-teach it’
 - ‘Offer him the job’
- A binary distinction between ‘formative’ and ‘summative’ is pretty much nonsense

Myths about assessment

Common sense view	Reality
Learners either understand something or they don't, in a binary sort of way	Understanding is best seen as a continuum, imperfectly observed
Once you 'get it' you never go back (threshold concepts)	There is no observable behaviour that corresponds to having 'got it', but what is observable is very erratic
The subjective feeling of 'mastery' is a good guide to learning	It just isn't
Whether someone gets a question right or completes a task satisfactorily mostly depends on their knowledge, understanding, competence, etc	So many other factors affect responses and scores. Signal:noise ratio is often woeful

Not all questions are assessment

Questions may not be intended to provide information as a basis for inference or decision-making

- Rhetorical device
- Pedagogical device
 - Retrieval practice
 - Maintain attention
- To elicit thinking
- To provoke dialogue



CEM

Centre for Evaluation & Monitoring

Test information, accuracy, reliability and generalisability

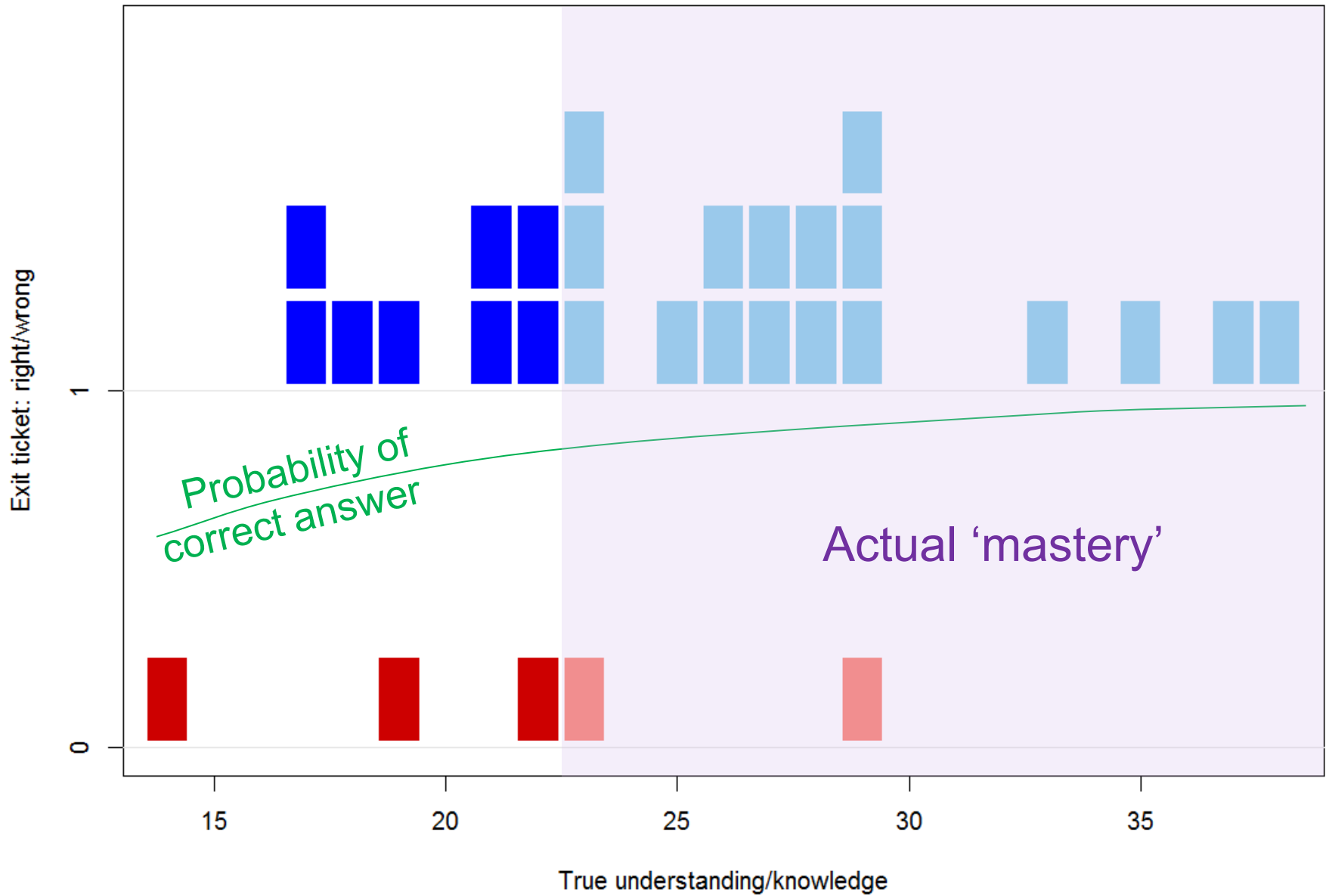


Durham
University

Problems with exit tickets

1. If 25 out of 30 get a question right, what does this tell us about their true knowledge/understanding?

Exit ticket responses



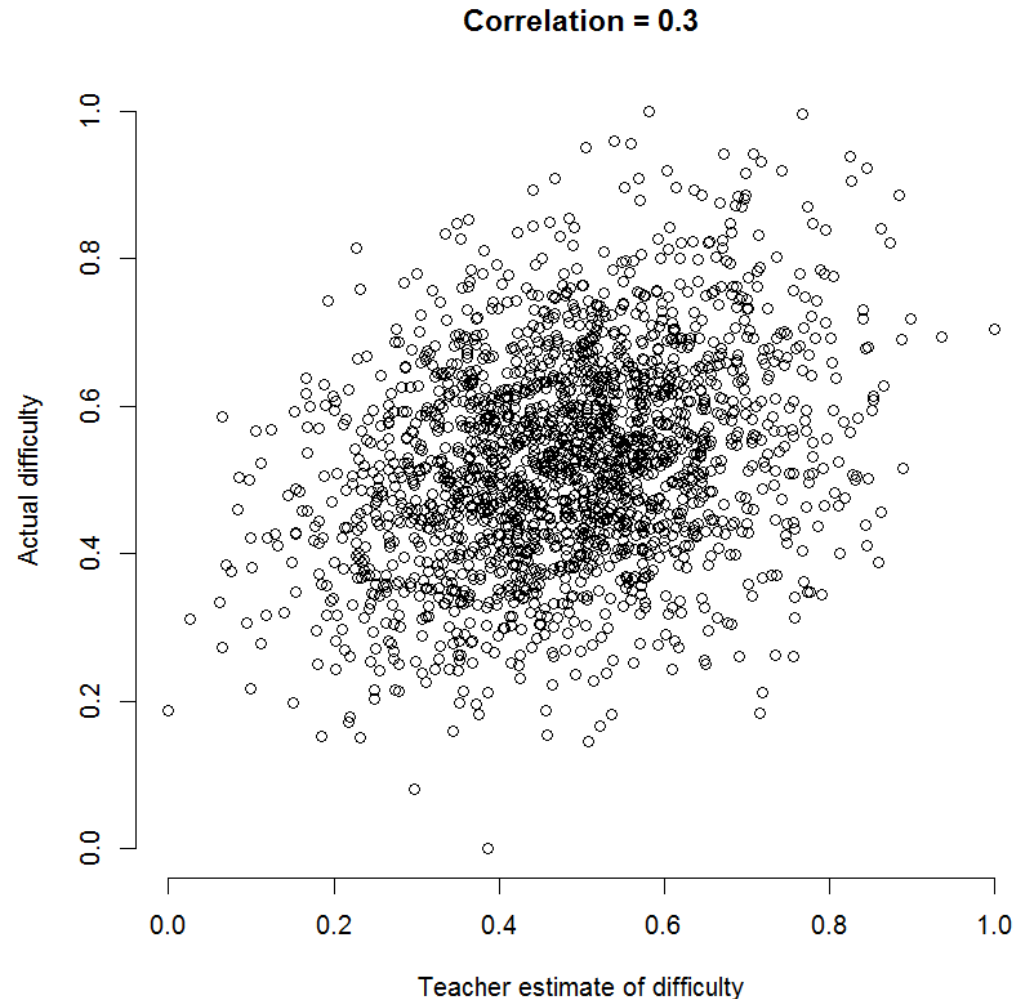
Problems with exit tickets

2. If 25 out of 30 get a question right, it might tell you that
 - a) 25 of 30 have reached the required level
 - b) The question was easier than you thought (so the number who have actually reached the required level is lower than 25)
 - c) The question was harder than you thought (so the number who have actually reached the required level is higher than 25)

How well can teachers estimate the difficulty of a question?

Correlations between estimated and actual facility often 0.2 – 0.3

(Bramley & Wilson, 2016; Attali et al, 2014)



How should you respond to an exit ticket?

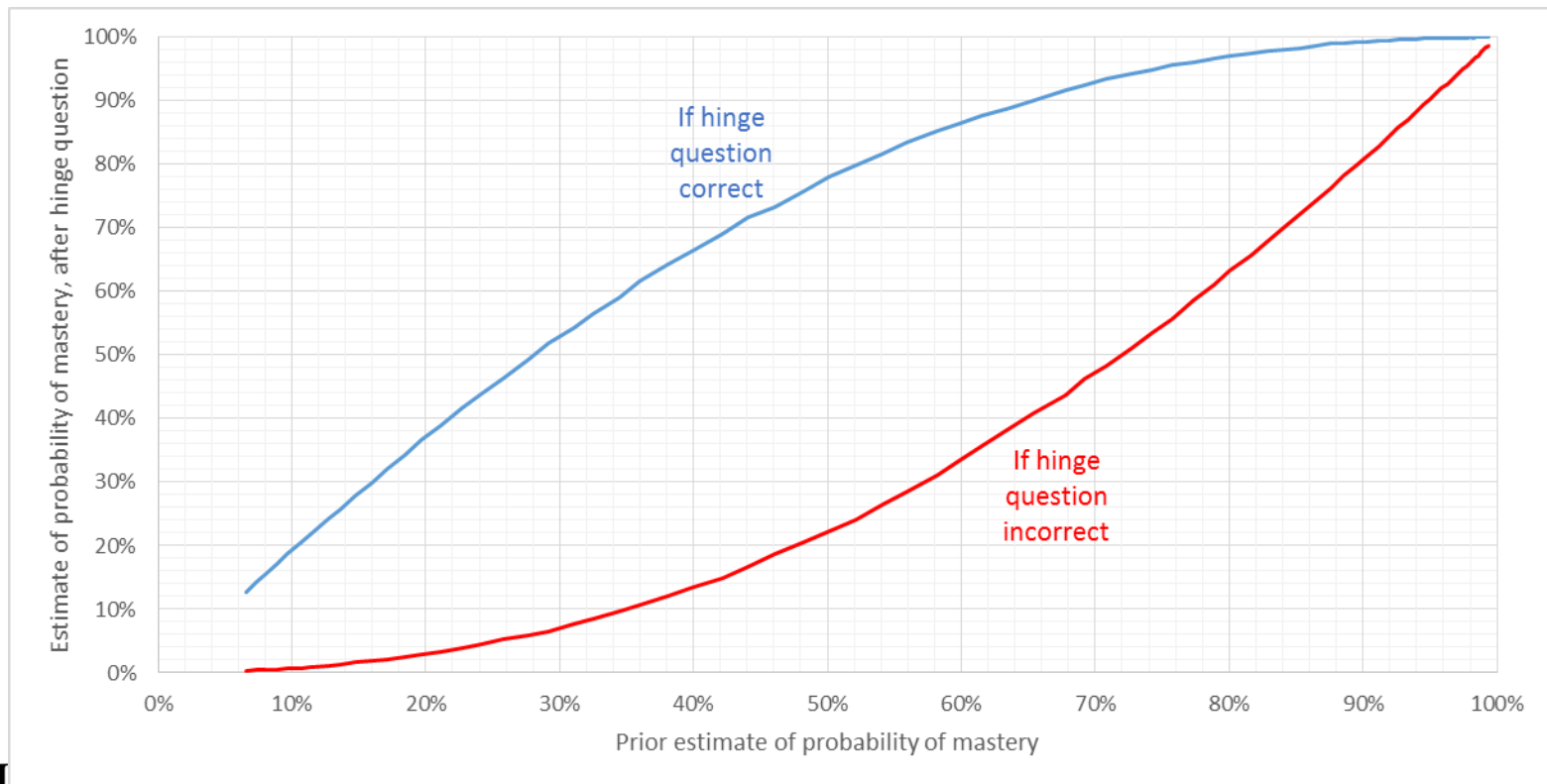
3. You ask a question as an exit ticket at the end of a lesson. 25 of a class of 30 get it right. Next lesson, would you
 - a) Move on (maybe review/revisit later)
 - b) Re-teach to all
 - c) Re-teach to those who got it wrong
 - d) Do none of these (ie something else)
4. If the number getting it right had been lower than 25, would you have done something different?
5. How much lower would it have to be to change your response?

Hinge questions

- Bayes' Theorem
 - If they understand it, what is the chance they will get a question right?
 - If they got the question right, what is the probability they understand it?
- We have to start with a 'prior':
 - Before we find out whether they got the question right, what is our estimate of the probability they understand it?
 - How confident/precise are we about this?

Hinge questions

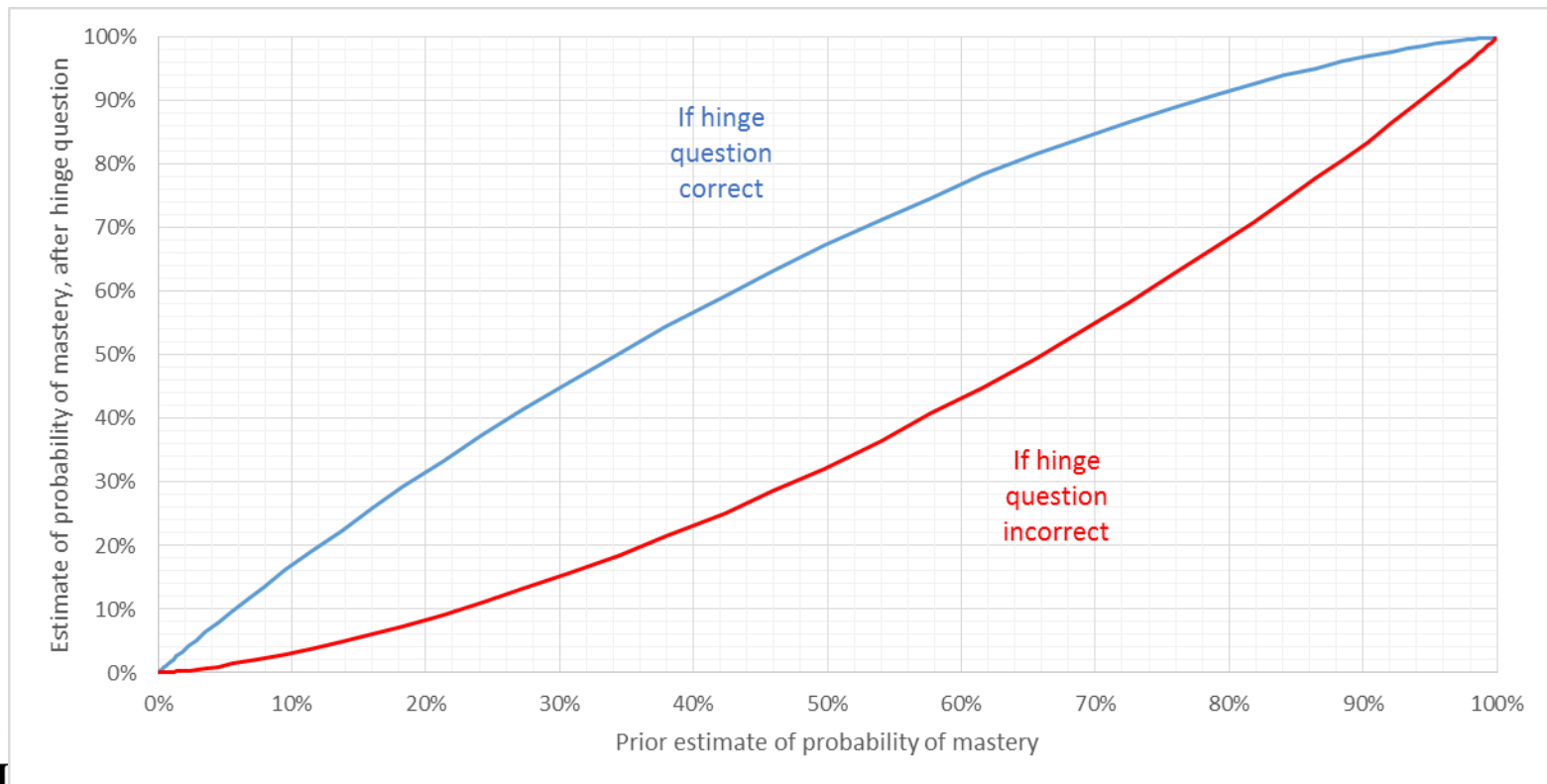
How much does a correct/incorrect answer affect your estimate of whether a student has 'mastered' the concept?
Case 1: You know nothing about the student before asking



Hinge questions

How much does a correct/incorrect answer affect your estimate of whether a student has 'mastered' the concept?

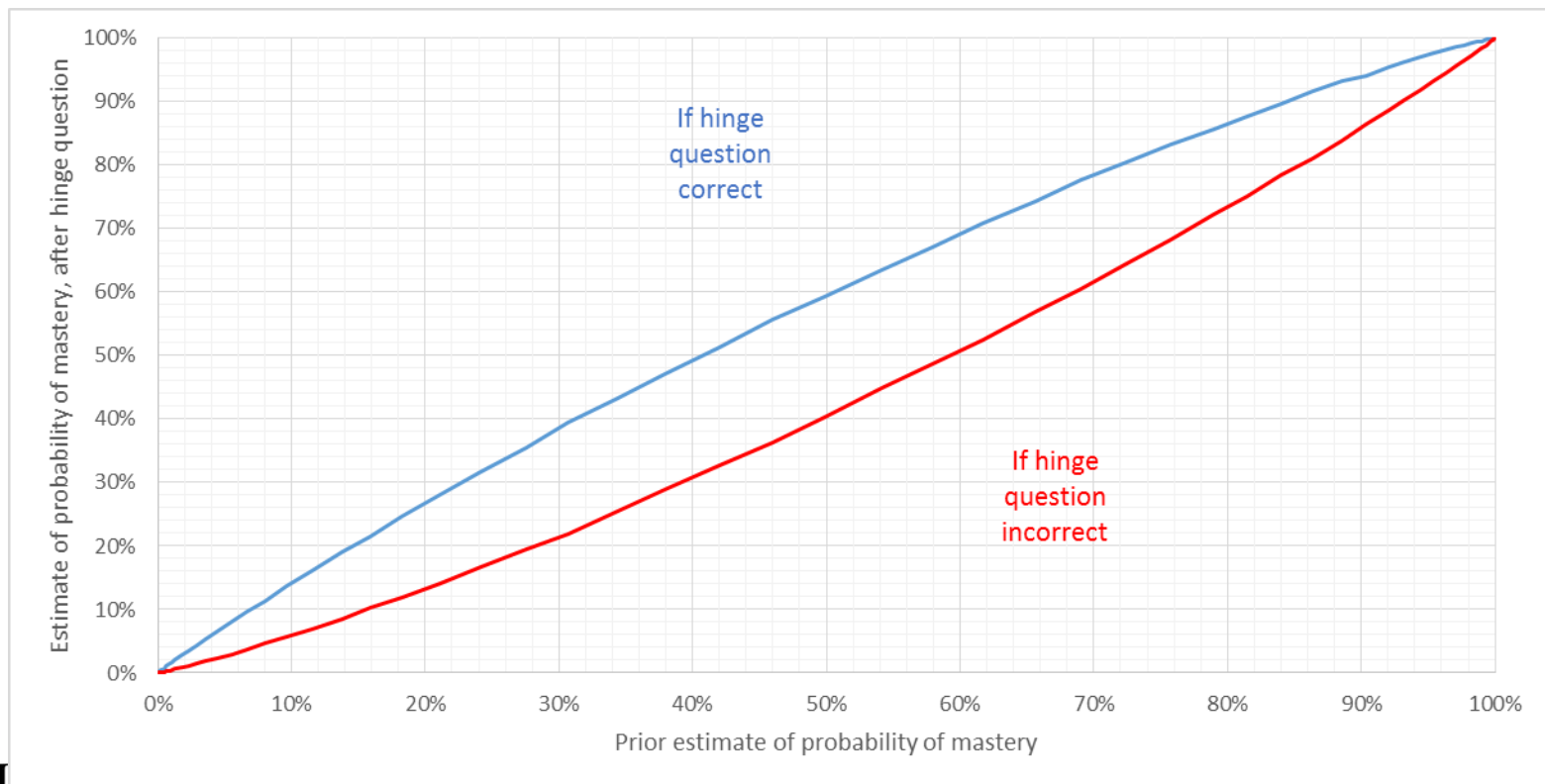
Case 2: You have good prior attainment data



Hinge questions

How much does a correct/incorrect answer affect your estimate of whether a student has ‘mastered’ the concept?

Case 3: You have good feedback about the student’s learning of this concept



How much information is there in a ‘hinge question’?

2. You have taught and seen a student working on this topic and estimate their probability of mastery at 80%, but they get the hinge question wrong.

What is your new estimate of their probability of mastery of the concept?

- a) 20%
- b) 50%
- c) 70%

Hinge questions and exit tickets

- May be valuable for
 - Provoking thinking and discussion
 - Giving insight into students' thinking
 - Maintaining students' attention
- Is there ever a time when you should base a decision about what to do on the outcomes of a single question?
 - No

How precise is a test score?

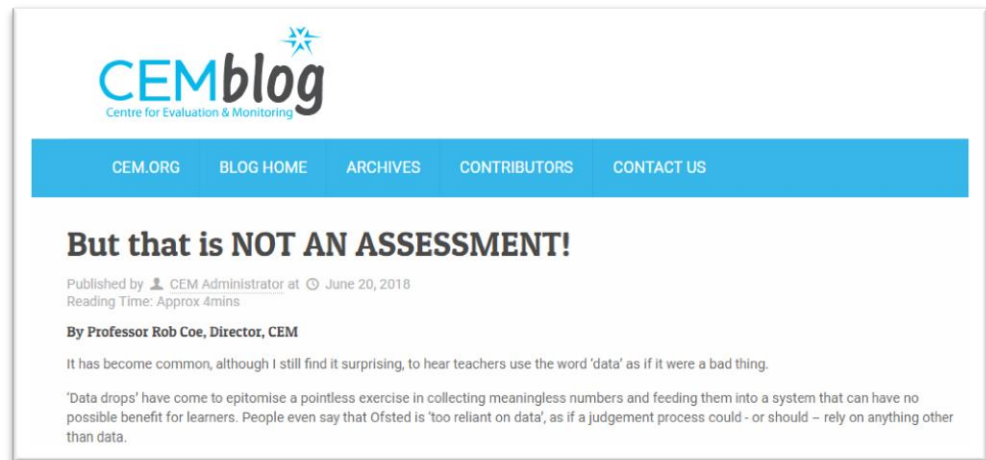
1. You have created a 20-item right-wrong test and given it to your class of 30 pupils. How far apart do two scores have to be before you can be confident one is really better than the other?
 - a) 3 marks
 - b) 6 marks
 - c) 8 marks

(Assume scores cover the full range 0-20, $SD=5$, Cronbach alpha = 0.7)

Reliability

Reliability can be interpreted as

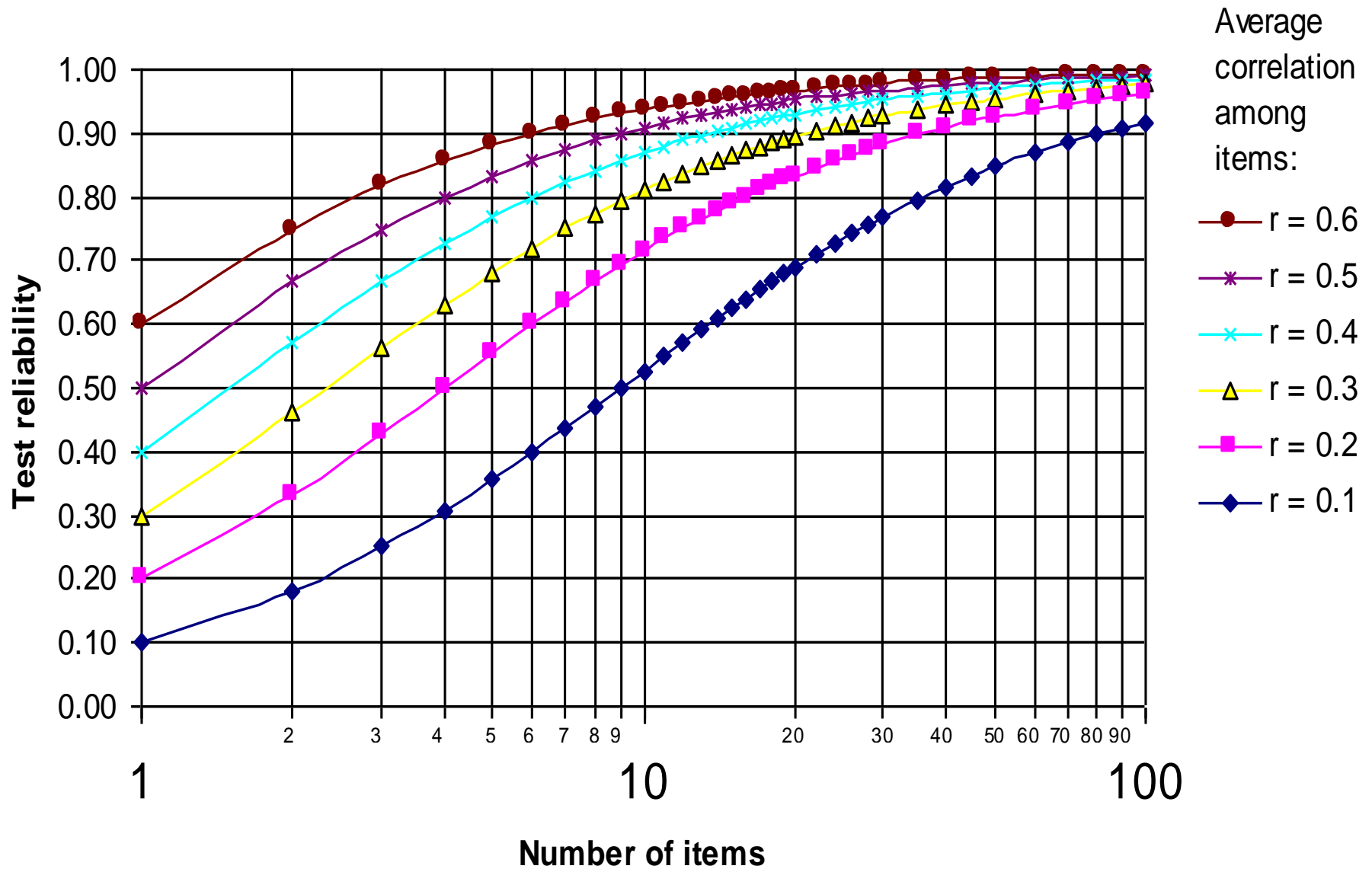
- **Replicability:** how consistently would an observed score be reproduced if elements of the measurement process that we wish to treat as arbitrary, unimportant and interchangeable were changed? (eg when taken, which paper sat, which questions came up, who marked it, etc)
- **Accuracy:** with what precision can an observed score be taken as an estimate of a true score?



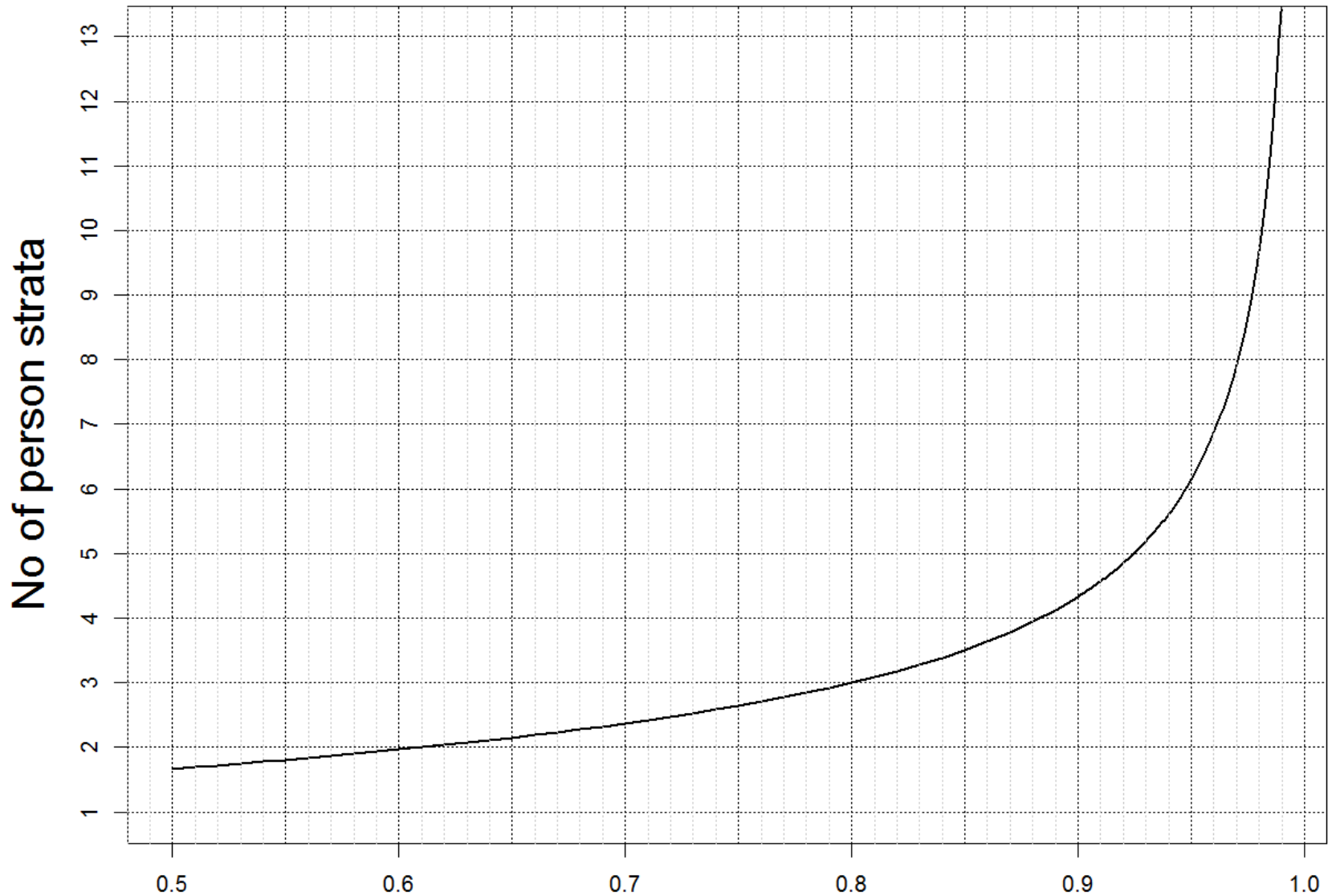
The screenshot shows the CEMblog website header with the logo and navigation links: CEM.ORG, BLOG HOME, ARCHIVES, CONTRIBUTORS, CONTACT US. The article title is "But that is NOT AN ASSESSMENT!". It is published by CEM Administrator on June 20, 2018, with a reading time of approximately 4 minutes. The author is Professor Rob Coe, Director of CEM. The article discusses the common use of the word 'data' by teachers and criticizes 'data drops' as a pointless exercise in collecting meaningless numbers.

Test length and reliability

(Spearman-Brown formula)



How many strata can persons be confidently separated into?



Reliability

Wright and Masters (1982)

Test reliability and precision

- Cronbach's alpha (and other indices of 'internal consistency') estimate the variation due to arbitrary selection of questions. Many other factors can arbitrarily affect scores (eg occasion, marker). These should be added to the error variance
- Even good classroom tests can only separate respondents confidently into 2-3 categories

Summary

People who use assessments (ie teachers) and those who help them to design and use assessment better (ie school assessment leads and those who train and support them) need to understand some of the limitations of what assessment can and cannot tell you. Often, it tells you less than you think.