WILDERN SCHOOL PROFESSIONAL DEVELOPMENT BULLETIN

THE ROUND

MAY 2019

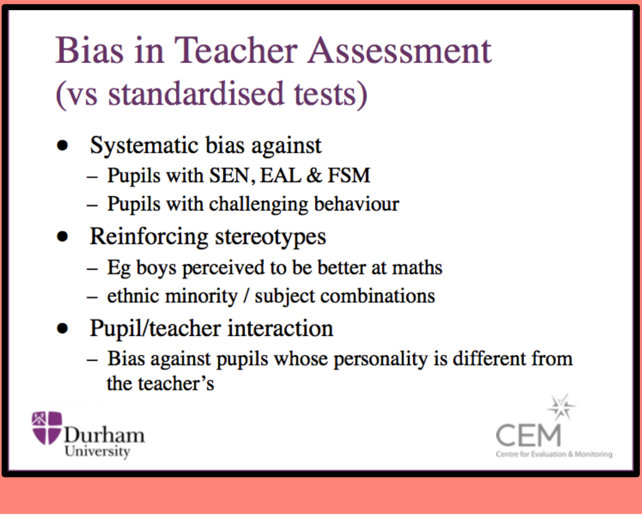
ASSESSMENT SPECIAL

BIAS IN TEACHER ASSESSMENT VS. TESTS

One of the things that has challenged me the most in recent years is the evidence that teachers are susceptible to unconscious bias when assessing.

This isn't a challenge to the professionalism of teachers. Rather it is merely a result of the very fact that we are humans. Bias is just human nature.

As such, researchers such as Professor Robert Coe and Daisy Christodoulou draw our attention to the importance of standardised tests as a fairer form of assessment. Christodoulou points out that the popular understanding is that disadvantaged students will do better within a system of teacher assessment, such as a GCSE which includes a coursework element, but there is a body of research evidence that suggests such pupils actually do better on standardised tests, due to a greater absence of human bias. The slide below from Professor Coe shows the key findings on teacher assessment bias. Click on the links in this article for more.



Bias in Teacher Assessment
(vs standardised tests)

- Systematic bias against
  - Pupils with SEN, EAL & FSM
  - Pupils with challenging behaviour
- Reinforcing stereotypes
  - Eg boys perceived to be better at maths
  - ethnic minority / subject combinations
- Pupil/teacher interaction
  - Bias against pupils whose personality is different from the teacher's

Durham University

CEM Centre for Evaluation & Monitoring

THE PRINCIPLES

DAISY CHRISTODOULOU

In her 2017 book, *Making Good Progress*, Daisy Christodoulou begins with a discussion of how Assessment for Learning has failed to have the impact that its creators had hoped for.

Christodoulou refers to Professor Robert Coe's damning conclusion that "During the fifteen years of the intensive intervention to promote AfL, despite its near universal adoption and strong research evidence of substantial impact on attainment, there has been no (or at best limited) effect on learning outcomes nationally."

One of the authors of the original paper on AfL, Dylan Wiliam, suggests that its failure to have the promised impact is down to implementation. Christodoulou argues that government support for the policy was counter-productive because it meant AfL became about high-stakes monitoring and tracking and so what was meant to be formative became summative.

Christodoulou argues that it is rare that an assessment can fulfil both roles: formative assessment is about methods, whereas summative assessment is about aims. The aim might be that a pupil can write an essay about the causes of WWI by the end of their studies, but the methods that will lead them to be able to write that essay will differ. An exam question should be able to summatively assess those aims, but is unlikely to be able to formatively assess the methods effectively.

In the book, Christodoulou identifies the key concepts needed to design assessment: **validity** and **reliability**.

Validity is perhaps the most important component. It refers to the ability to make inferences from an assessment. We aren't really interested in how well a student did on a particular assessment; rather, we are more interested in what we can say about what the student can do in other situations or contexts, at other times based on that assessment. This is what validity in assessment refers to: is the assessment designed well enough for us to make valid inferences from it.

The second concept is **reliability**: whether the marks generated are reliable. The table opposite shows the main sources of reliability and unreliability in assessment.

JNT

OF ASSESSMENT

Sources of unreliability	
Sampling	• Most tests do not directly measure the domain; they only sample from it • Students do better or worse depending on the particular sample ***For a test to have sampling reliability****: if a pupil were to take different versions of the same test, they should get approximately the same mark.*
Marker	• Different markers may disagree on quality • Applying standards consistently is difficult, even for one marker ***For a test to have marker reliability****: if a pupil's answer paper were submitted to ten different markers, it should return each time with the same mark.*
Student	• Student performance on the day varies • Students can perform differently depending on illness, time of day, whether they have eaten beforehand, etc. ***For a test to have reliability for student performance****: if a pupil were to take the test at different times of the day, they should get approximately the same mark.*

ASSESSMENT SPECIAL

MAY 2019

BLOGROLL

ARTICLES AND BLOGS YOU MIGHT FIND INTERESTING

'What to do after a mock? Assessment, sampling, inferences and more' is a great blog from Science teacher Adam Boxer, using the principles of assessment to challenge the more common approaches to mock exam reflection.

'How do we know pupils are making progress? Part 3: Assessment' is one of a series of blogs on measuring progress by writer David Didau; it's worth reading the whole series but this one on assessment is thorough, thought-provoking, and definitely worth a read if you are designing assessment to measure pupil progress.

'Five Things I've Learning about the Importance of Good Assessment' is an excellent blog by Alex Quigley for the Centre for Evaluation & Monitoring at Durham University. Alex touches on 'the testing effect', teacher bias and mock exams.

If you want to know a little bit more about any of the ideas in this edition, please don't hesitate to email me – j.theobald@wildern.org – or come and find me in Block 9! James