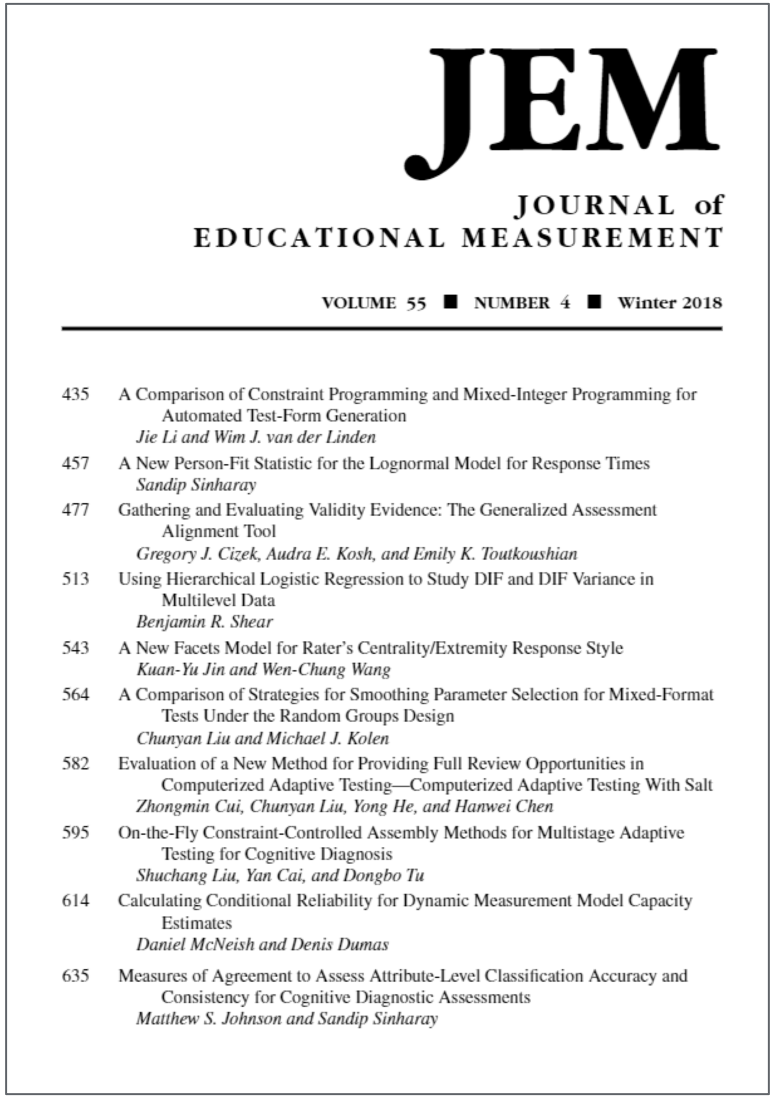
ofqual

VTQ Research: Getting Technical

Paul E. Newton

AO Forum, London, 14 May 2019

Technical research into GQ-like tests and exams be like…



JEM

JOURNAL of
EDUCATIONAL MEASUREMENT

VOLUME 55 ■ NUMBER 4 ■ Winter 2018

- Using Hierarchical Logistic Regression to Study DIF and DIF Variance in Multilevel Data
- A Comparison of Strategies for Smoothing Parameter Selection for Mixed-Format Tests Under the Random Groups Design
- Calculating Conditional Reliability for Dynamic Measurement Model Capacity Estimates

Technical research into VTQ-like assessments be like…

Credit: Bill Applin

But isn't VTQ assessment inherently non-technical?

The assumption [underpinning NVQs] has always been that **assessment will be unproblematic because it simply involves comparing behaviour with the transparent 'benchmark' of the performance criteria**.

(Wolf, 1995, p.24)

What I am proposing is that **we should just forget reliability altogether, and concentrate on validity**, which is ultimately all that matters.

(Jessup, 1991, p.191)

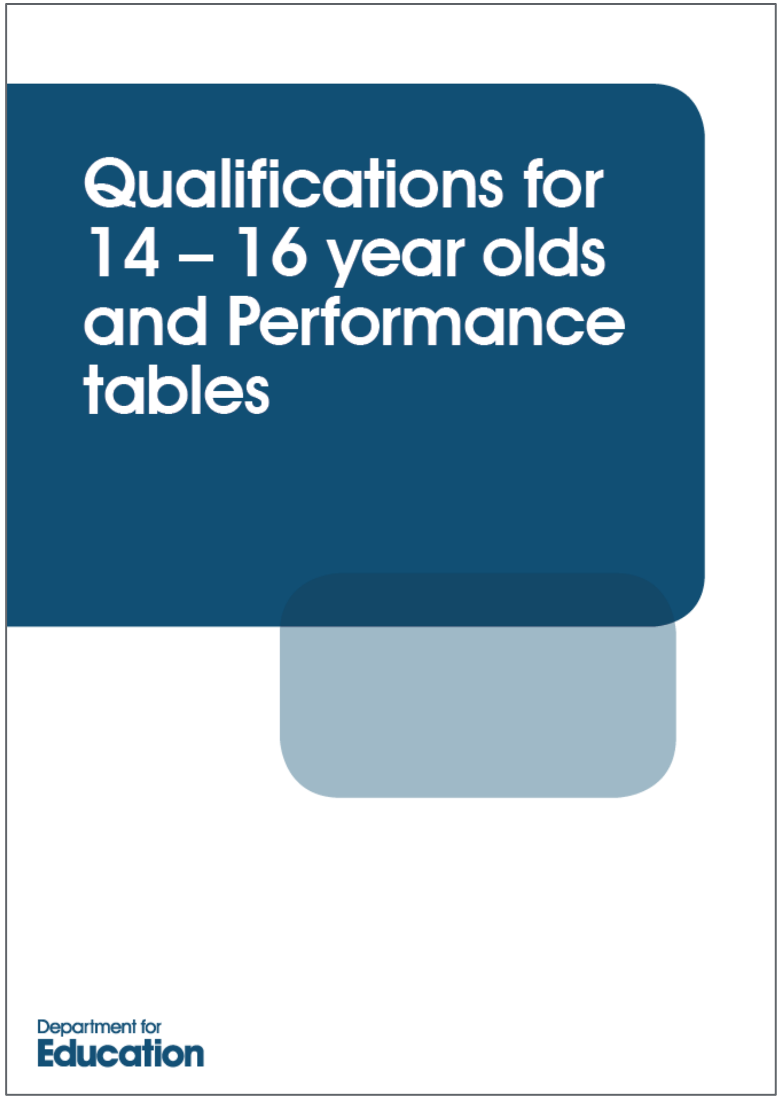
But isn’t VTQ-like assessment inherently non-technical?

The assumption [underpinning NVQs] has always been that **assessment will be unproblematic because it simply involves comparing behaviour with the transparent ‘benchmark’ of the performance criteria**. **Unfortunately, in practice this turns out not to be the case.**

(Wolf, 1995, p.24)

- … even NVQ-like Competence-Based Qualifications are far from unproblematic, and require a technical eye;
- … and the more ‘GQ-like’ a Competence-Based Qualification becomes, the more of a technical eye it’s likely to require.

Becoming ‘GQ-like’?

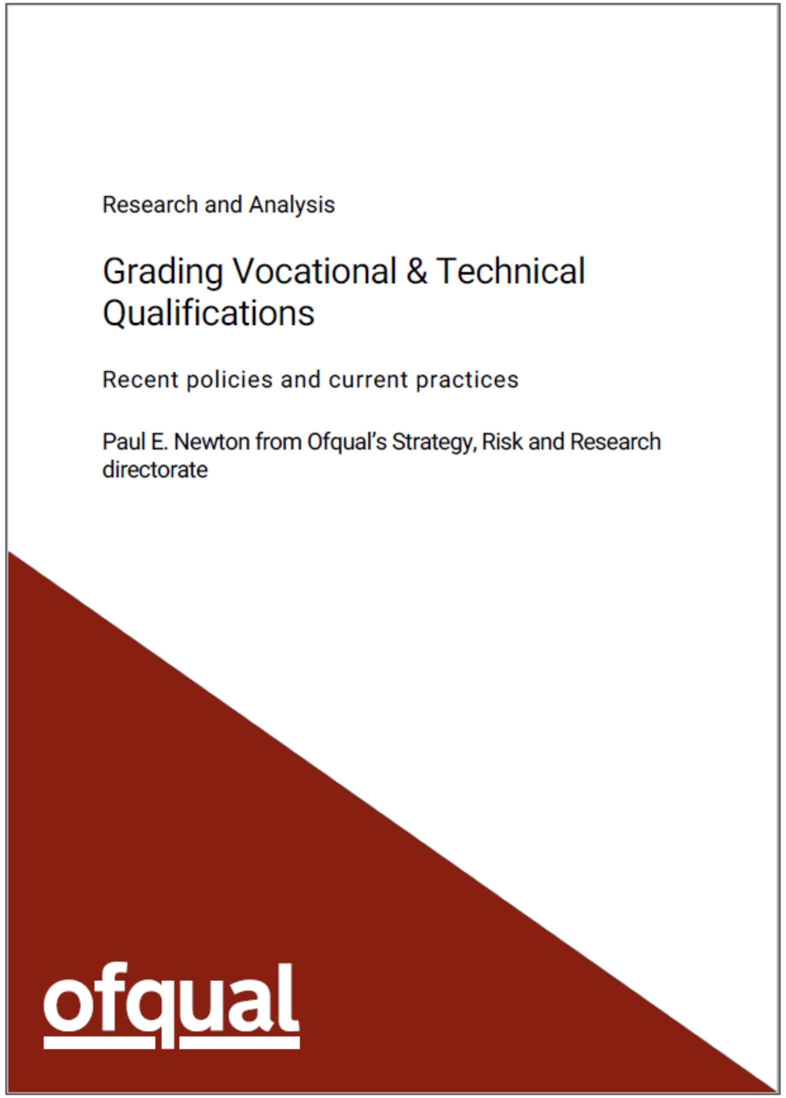


Qualifications for 14 – 16 year olds and Performance tables

Department for Education

- “Professor Wolf’s report is very clear that assessment methods for many vocational qualifications need to be strengthened […] Therefore, **only those qualifications that provide evidence of substantial amount of external assessment**, together with synoptic assessment […] will be counted in the tables.”
- “So **we will only include those qualifications that are graded** – as opposed to a pass/fail – in the tables in the future. Qualifications may have a pass, merit, distinction structure or a more detailed scale.”

https://www.gov.uk/government/publications/grading-vocational-and-technical-qualifications

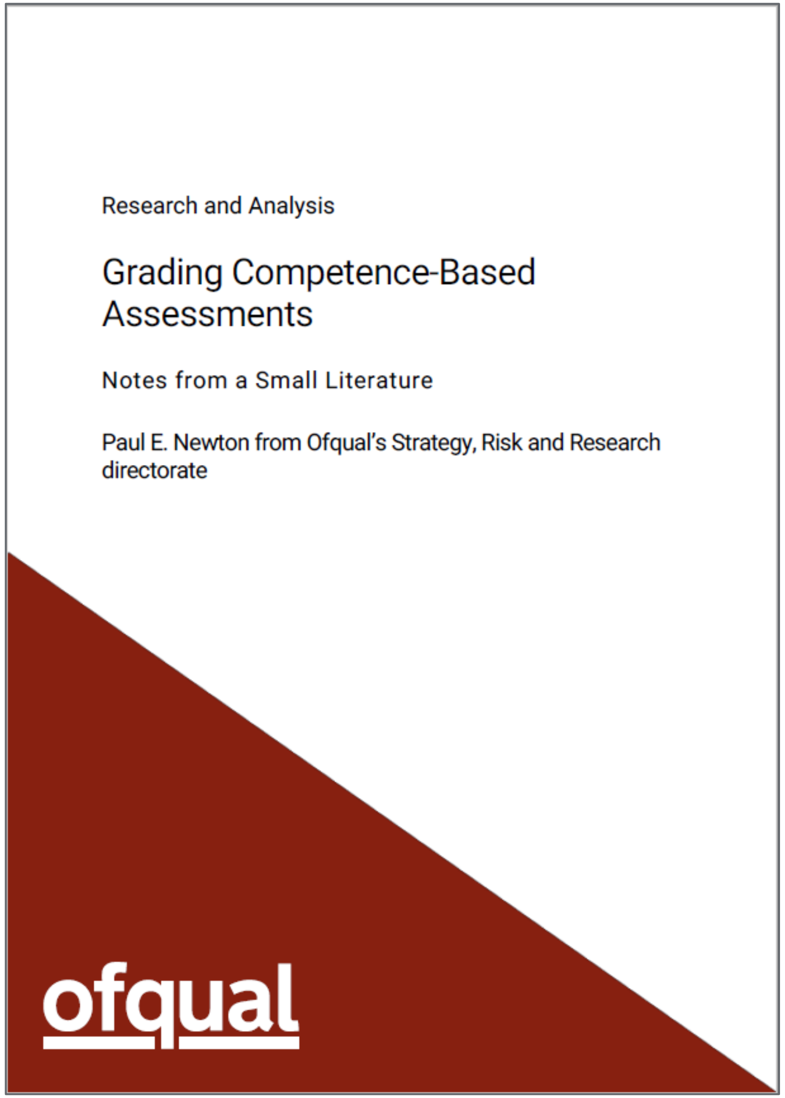


Research and Analysis

Grading Vocational & Technical Qualifications

Recent policies and current practices

Paul E. Newton from Ofqual's Strategy, Risk and Research directorate



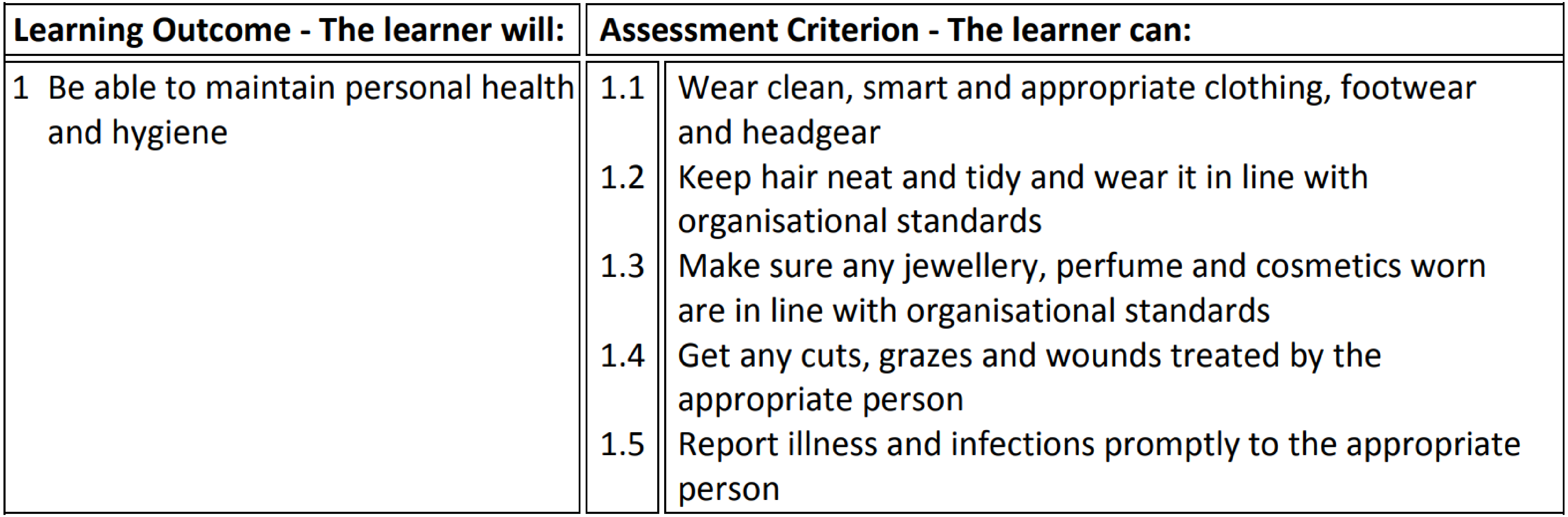
Research and Analysis

Grading Competence-Based Assessments

Notes from a Small Literature

Paul E. Newton from Ofqual's Strategy, Risk and Research directorate

Level 1 NVQ Certificate in Hospitality Services



Learning Outcome - The learner will:	Assessment Criterion - The learner can:	
1 Be able to maintain personal health and hygiene	1.1	Wear clean, smart and appropriate clothing, footwear and headgear
	1.2	Keep hair neat and tidy and wear it in line with organisational standards
	1.3	Make sure any jewellery, perfume and cosmetics worn are in line with organisational standards
	1.4	Get any cuts, grazes and wounds treated by the appropriate person
	1.5	Report illness and infections promptly to the appropriate person

… comparing behaviour with the transparent ‘benchmark’ of the performance criteria?

Level 2 & 3 Diploma in Skills for Business

Level 2 (unit 10)

LO1 Understand the reasons for change in business

AC1.1 **State** why it is important for a business to change

AC1.2 **State** the risks associated with a business changing too quickly

AC1.3 **State** the risks associated with a business changing too slowly

Level 3 (unit 10)

LO1 Understand change in business

AC1.1 **Explain** why it is important for a business to change

AC1.2 **Analyse** the positive and negative effects of change on a selected

AC1.3 **Compare** the risks of slow against rapid change within a business

AC1.4 **Compare** the benefits of slow against rapid change within a business

Unit 10	Pass	Merit	Distinction
L2 **AC1.1**	The candidate will **state** why it is important for a business to change	The candidate will **state** why it is important for a business to change, demonstrating **critical understanding**	[No D for this AC]
L3 **AC1.1**	The candidate will **explain** why it is important for a business to change	The candidate will **explain** **in detail** why it is important for a business to change	The candidate will give a **sophisticated** **explanation** of why it is important for a business to change

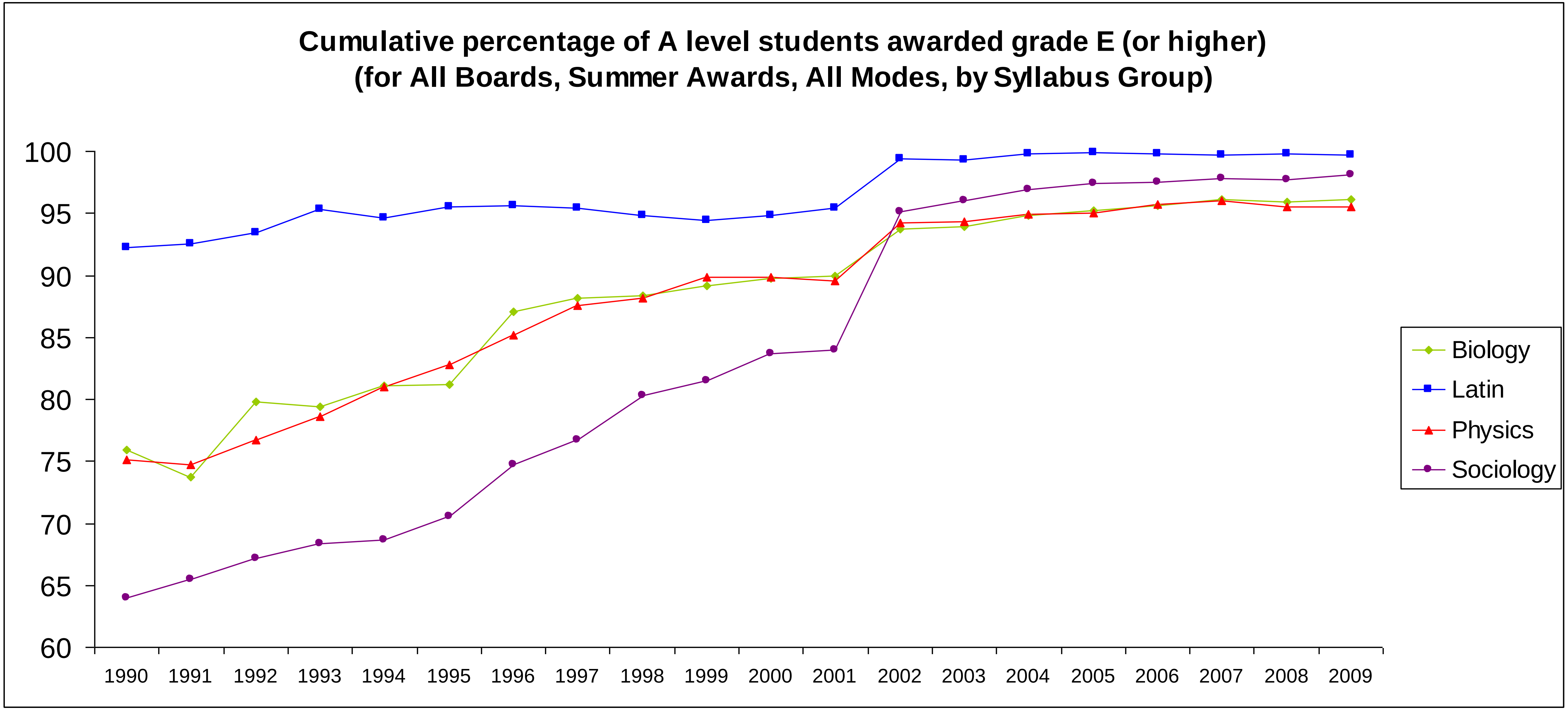
All sorts of technical questions related to grading

- Standardisation
- Comparability
- Grading and levelling
- Weighting
- Burden and backwash
- Transparency

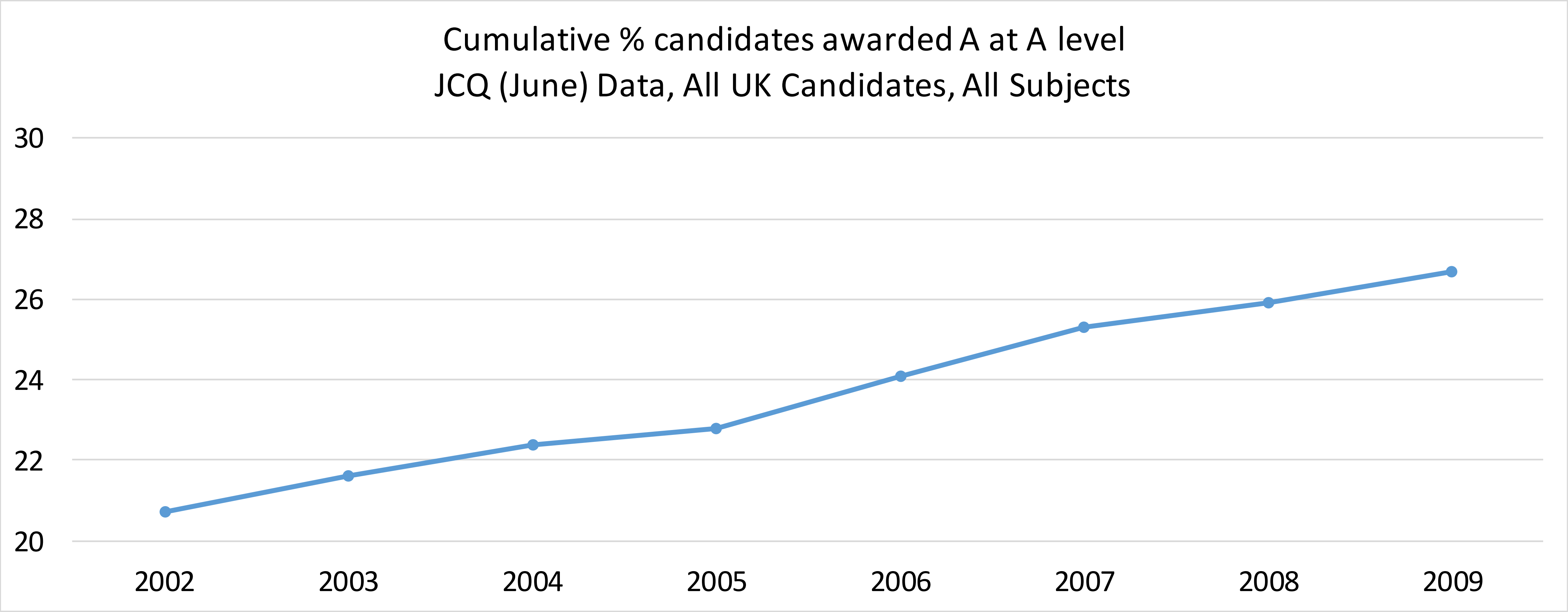
We know how hard challenges like standardisation/comparability can be

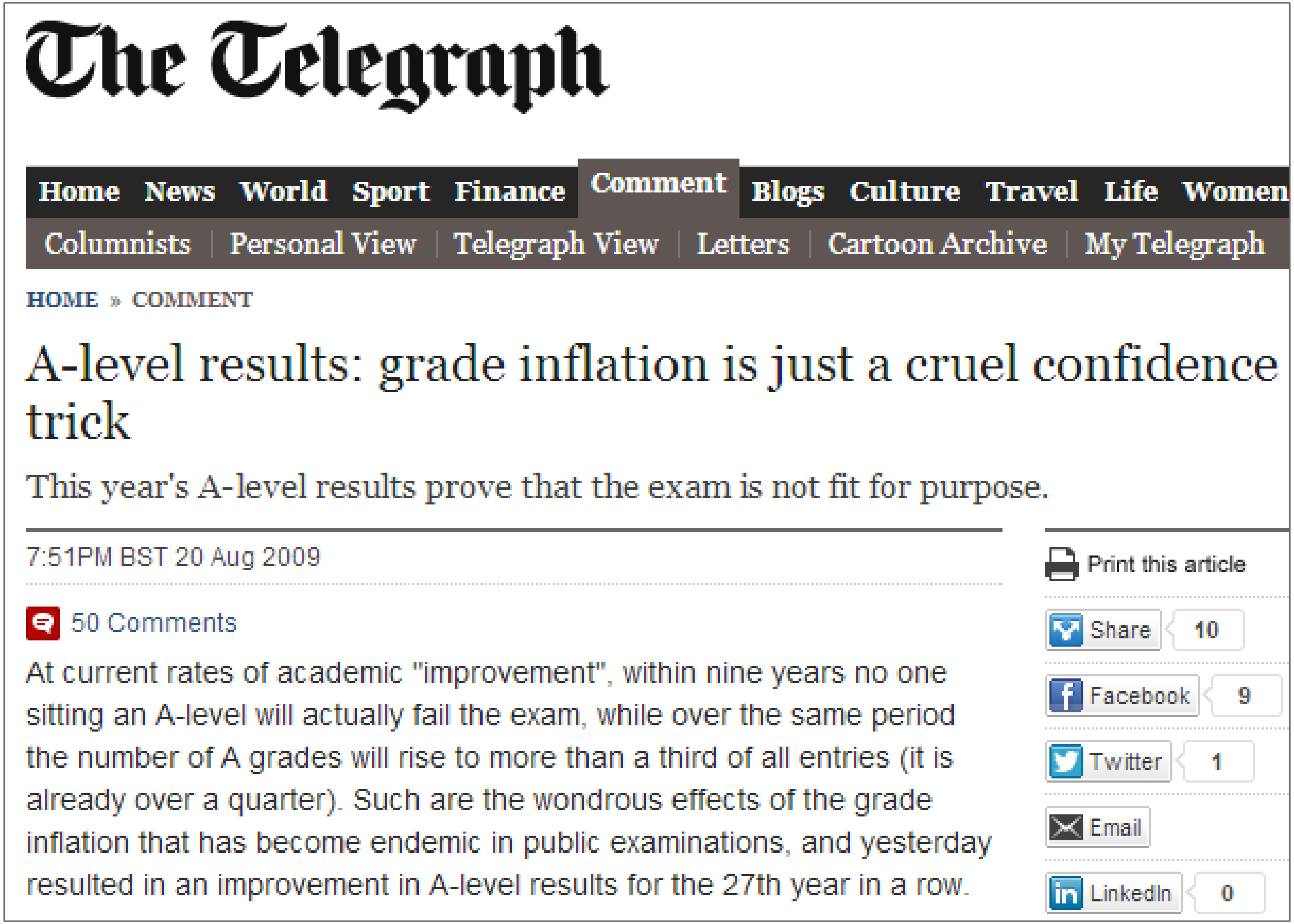
- Because we’ve done a lot of technical research into issues like these, for GQ-like qualifications.

Grade inflation at A level?



Grade inflation at A level?





A-level results: grade inflation is just a cruel confidence trick

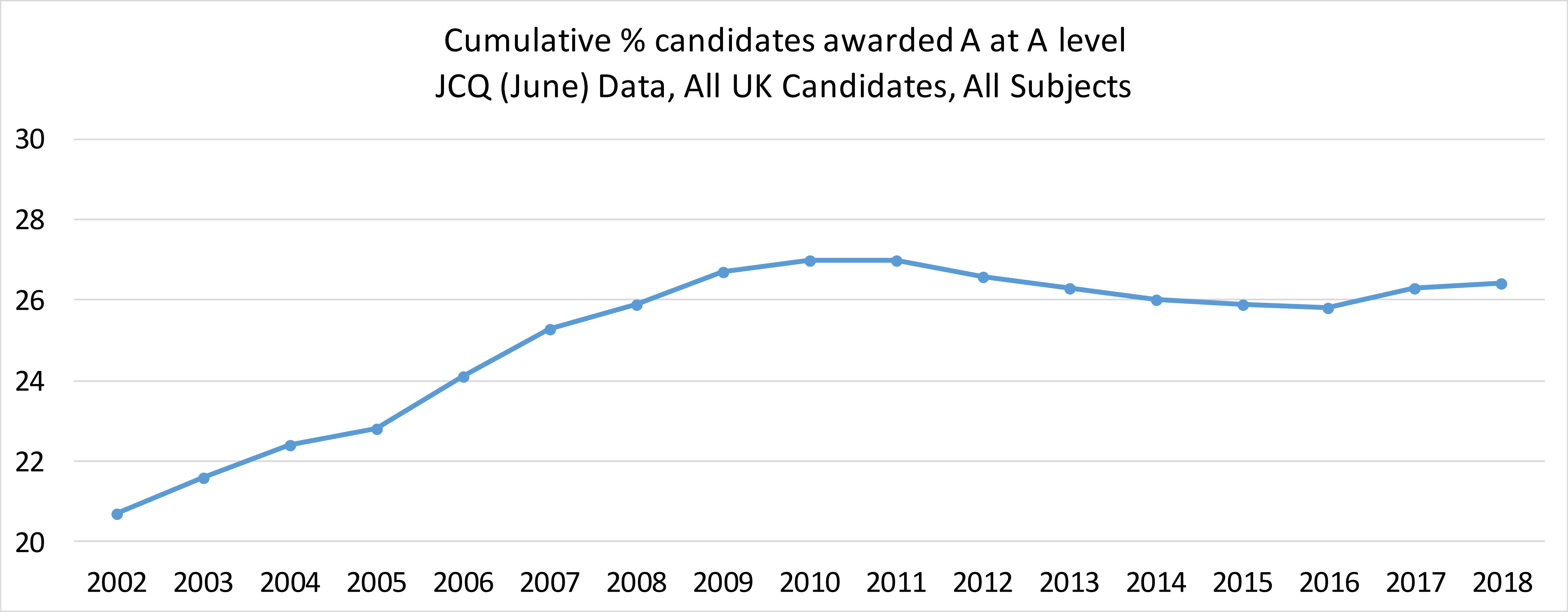
This year's A-level results prove that the exam is not fit for purpose.

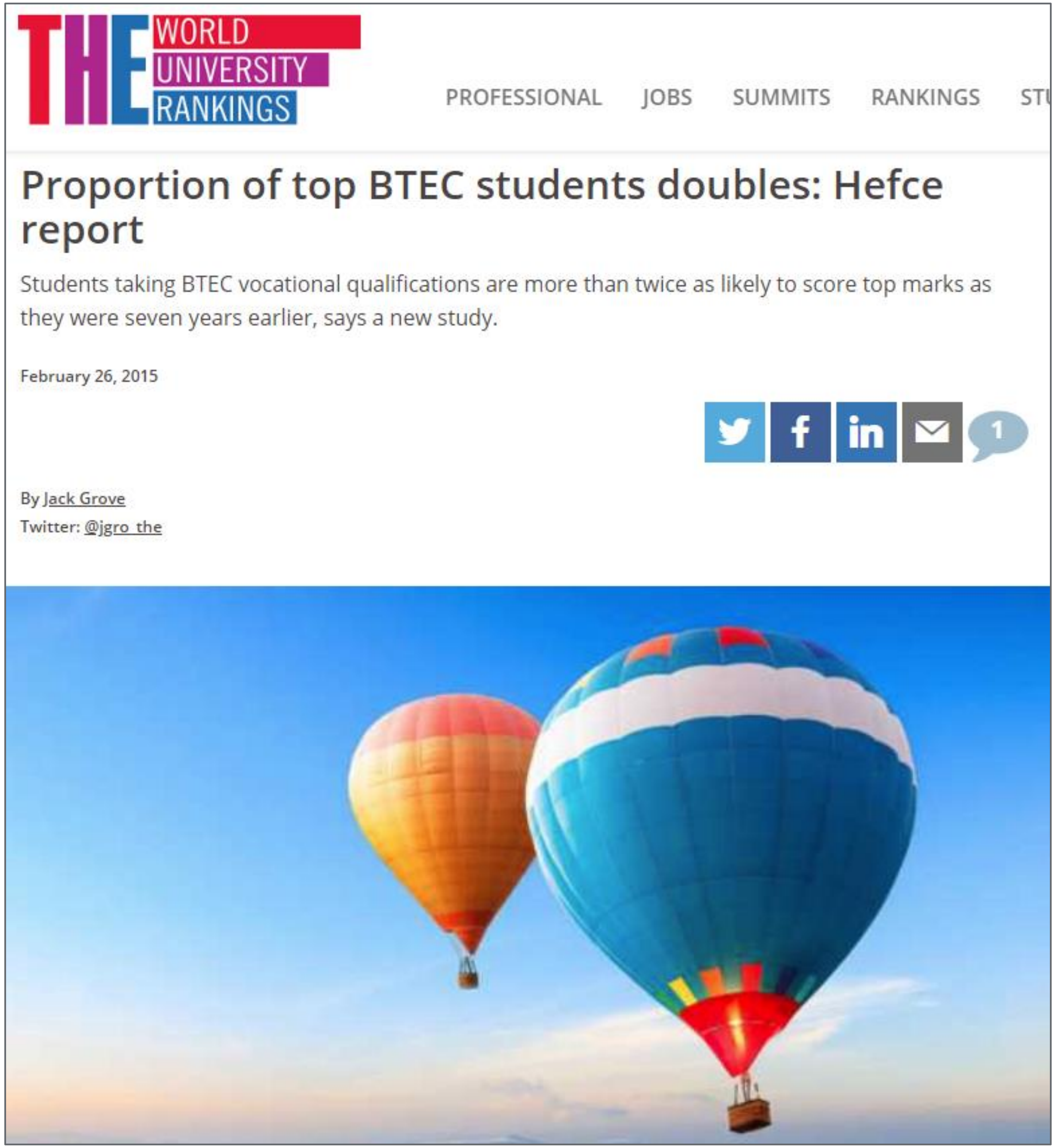
7:51PM BST 20 Aug 2009

50 Comments

At current rates of academic "improvement", within nine years no one sitting an A-level will actually fail the exam, while over the same period the number of A grades will rise to more than a third of all entries (it is already over a quarter). Such are the wondrous effects of the grade inflation that has become endemic in public examinations, and yesterday resulted in an improvement in A-level results for the 27th year in a row.

No longer any evidence of grade inflation at A level





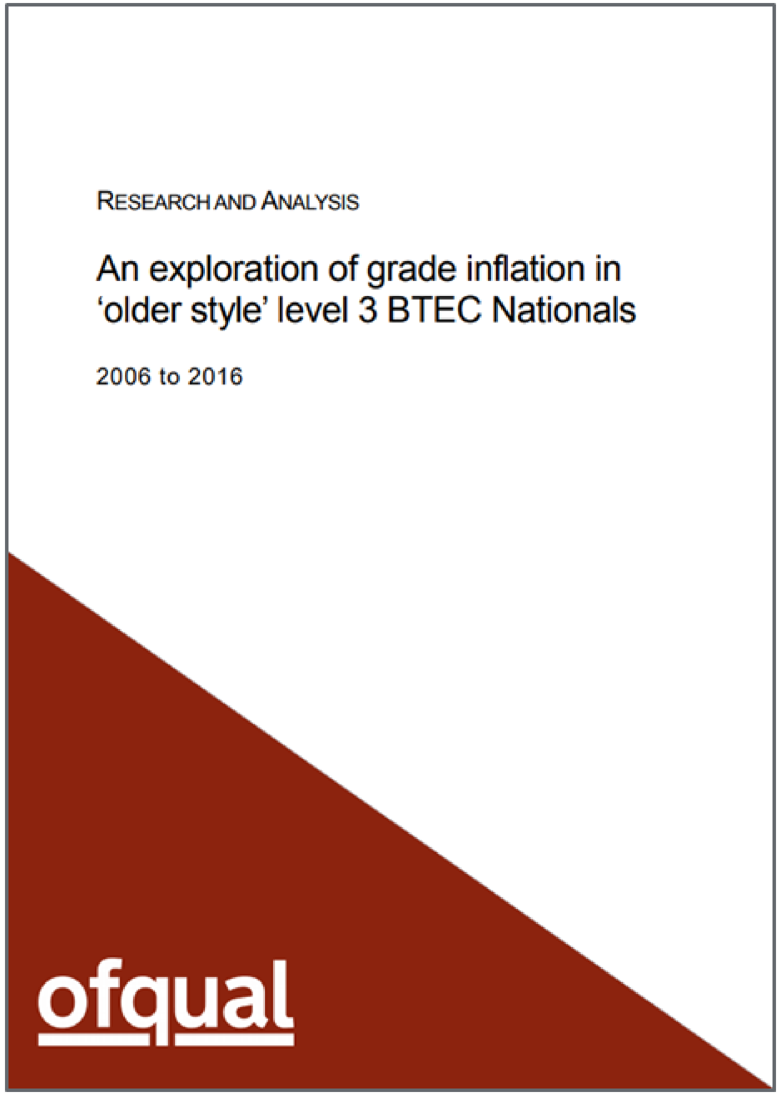
Proportion of top BTEC students doubles: Hefce report

Students taking BTEC vocational qualifications are more than twice as likely to score top marks as they were seven years earlier, says a new study.

February 26, 2015

By Jack Grove
Twitter: @jgro_the

https://www.gov.uk/government/publications/strengthening-vocational-and-technical-qualifications



RESEARCH AND ANALYSIS

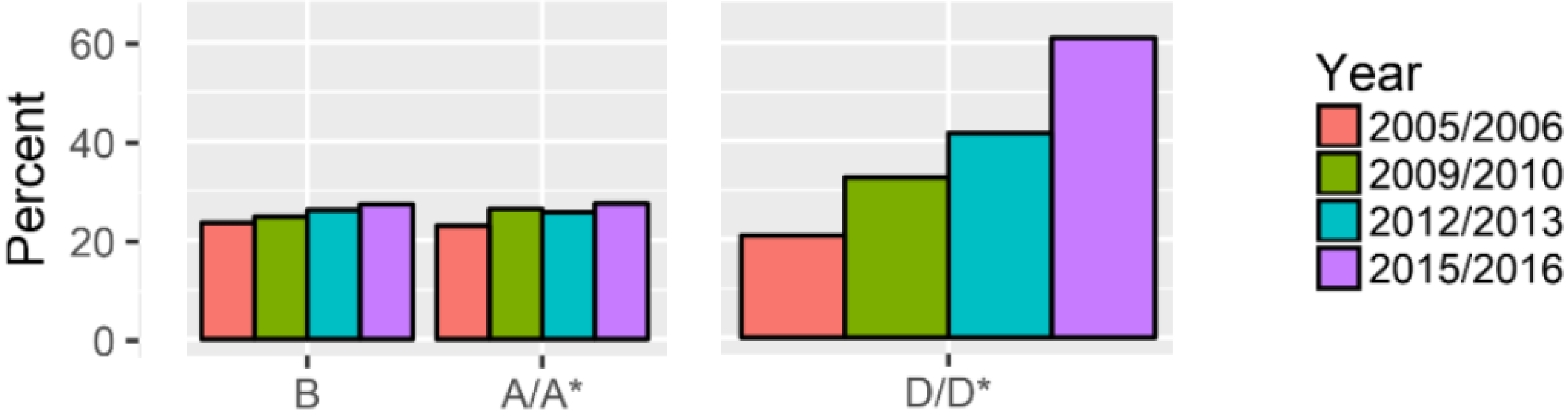
An exploration of grade inflation in 'older style' level 3 BTEC Nationals

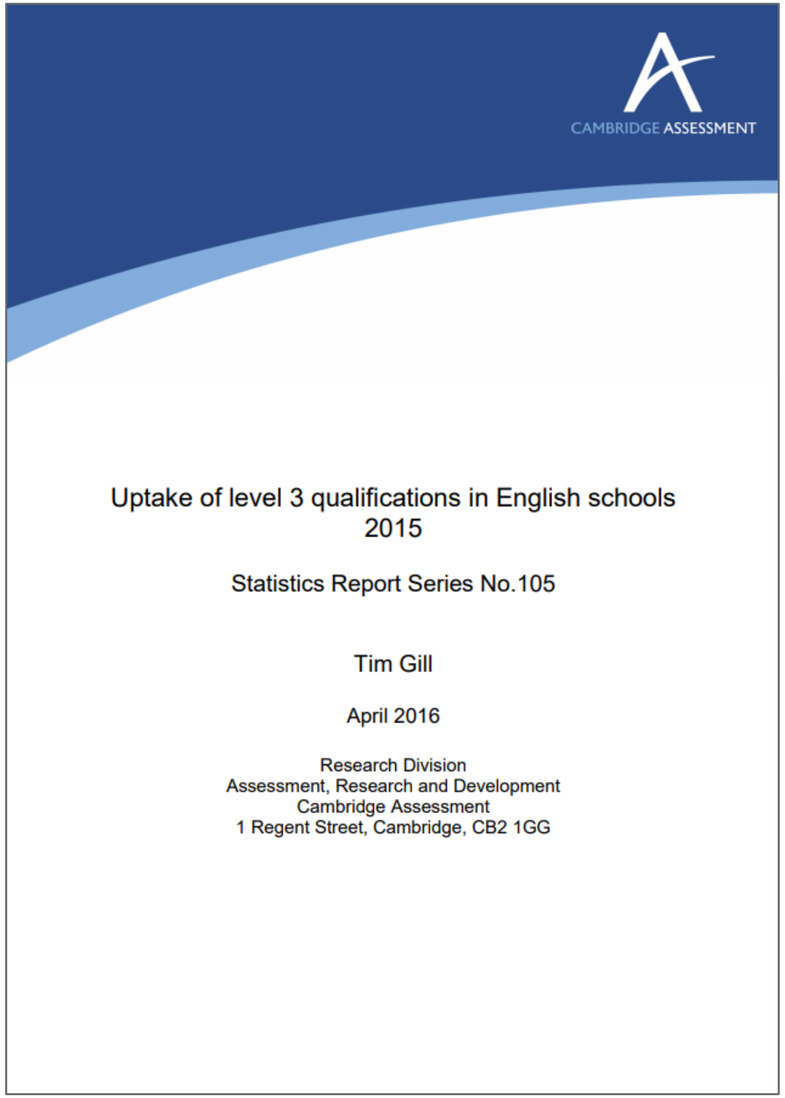
2006 to 2016

ofqual

Ben Cuff,
Nadir Zanini,
Beth Black

A Level **BTEC (Subsidiary Diploma)**





CAMBRIDGE ASSESSMENT

Uptake of level 3 qualifications in English schools
2015

Statistics Report Series No.105

Tim Gill

April 2016

Research Division
Assessment, Research and Development
Cambridge Assessment
1 Regent Street, Cambridge, CB2 1GG

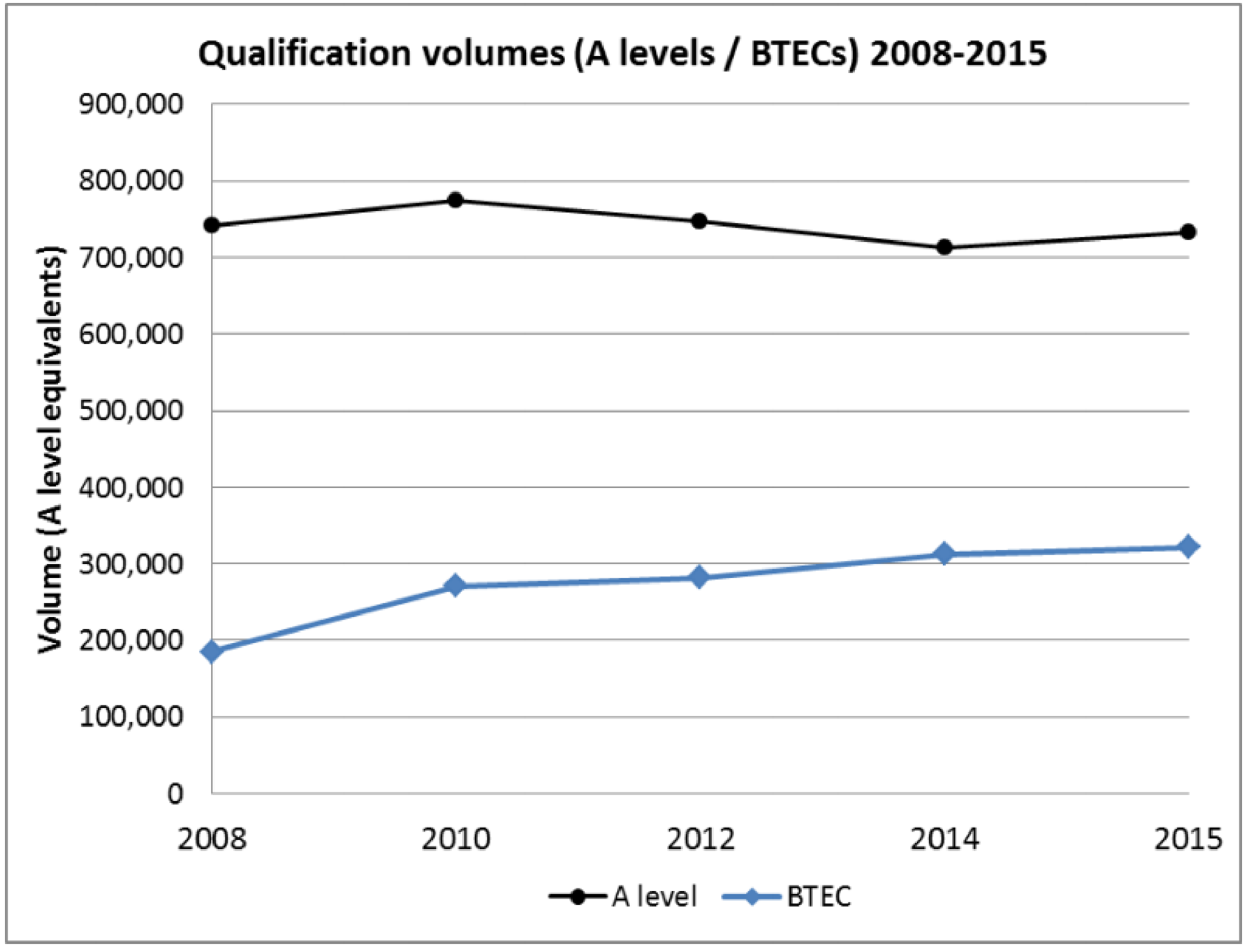
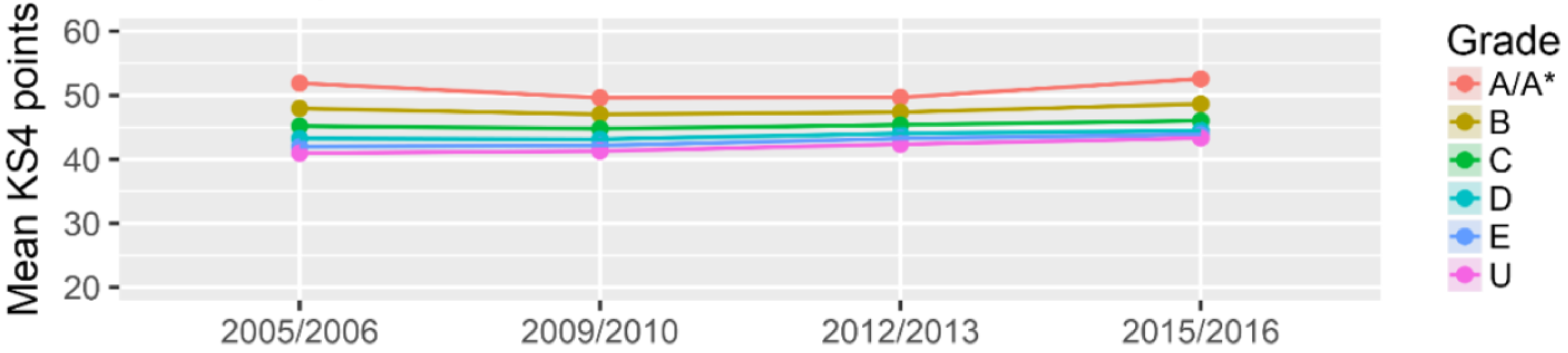
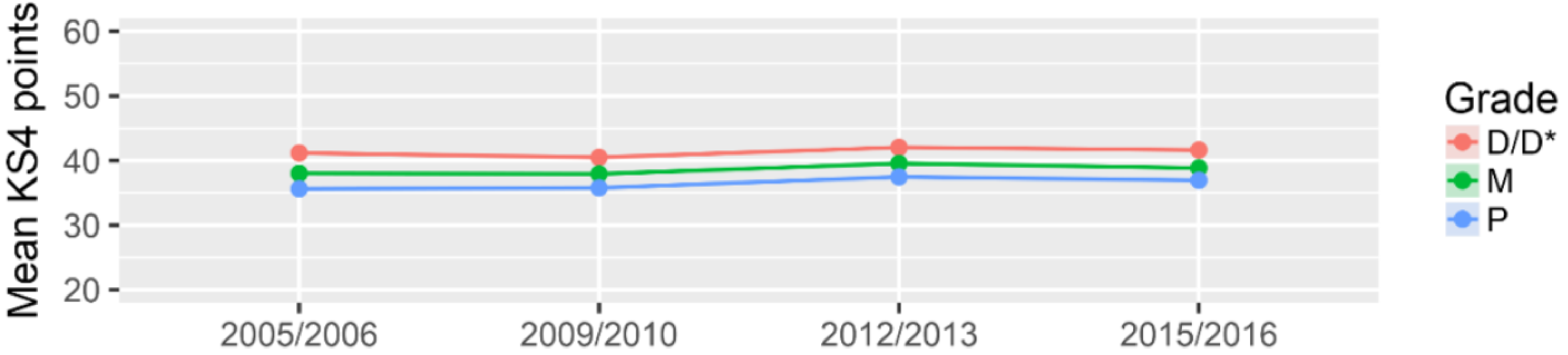


Figure 1: Volumes of A levels and BTECs taken by students 2008-2015

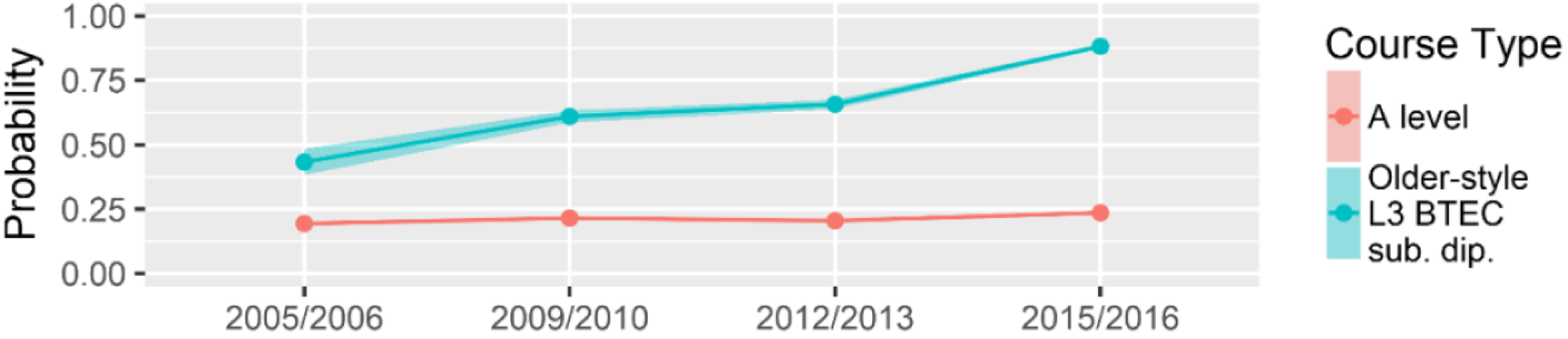
A level - prior attainment



Older-style L3 BTEC subsidiary diploma - prior attainment



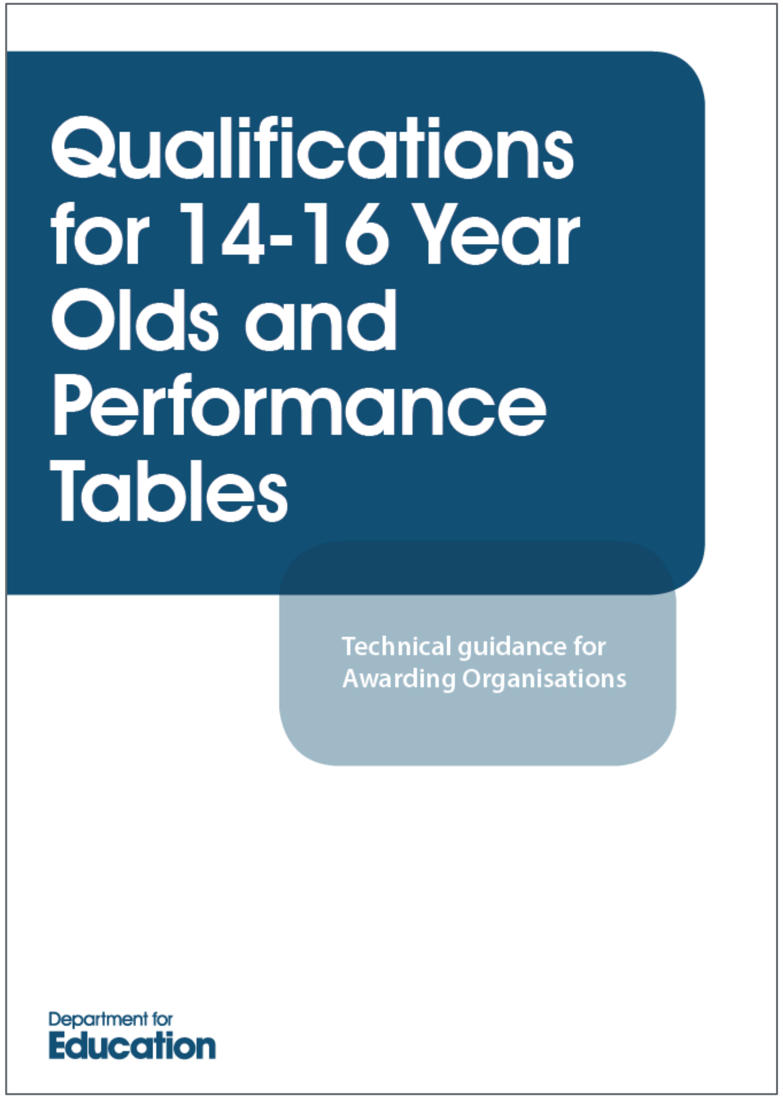
Predicted probability of achieving a top grade: Older-style L3 BTEC subsidiary diploma and A level



We know how hard challenges like standardisation/comparability can be

- But is it particularly challenging to maintain a firm grip on standards within Competence-Based Qualifications, like ‘older style’ BTECs?

Becoming ‘GQ-like’?



Qualifications for 14-16 Year Olds and Performance Tables

Technical guidance for Awarding Organisations

Department for Education

- “Professor Wolf’s report is very clear that assessment methods for many vocational qualifications need to be strengthened […] This **helps to ensure** that vocational qualifications offer **a comparable level of challenge** to academic qualifications and are seen to do so. External assessment also **provides an additional check that standards are consistent** across centres.”

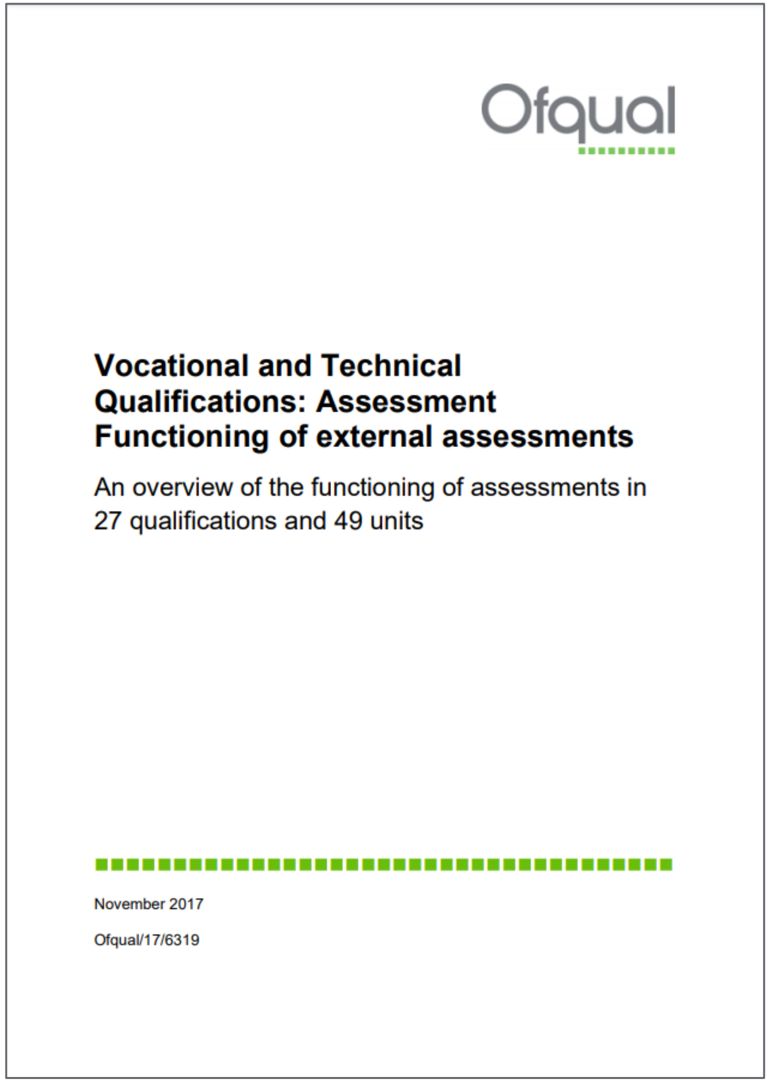
We know what a good GQ test looks like

- But what does a good VTQ test look like?

What makes a good test?

1. Questions (and test overall) should:
   a) be of an appropriate level of difficulty
   b) differentiate between learners, but (only) on the basis of their proficiency
2. Test overall should:
   a) deliver reliable results
   b) embody standards that are comparable with comparable tests

https://www.gov.uk/government/publications/assessment-functioning-of-external-assessments



Ofqual

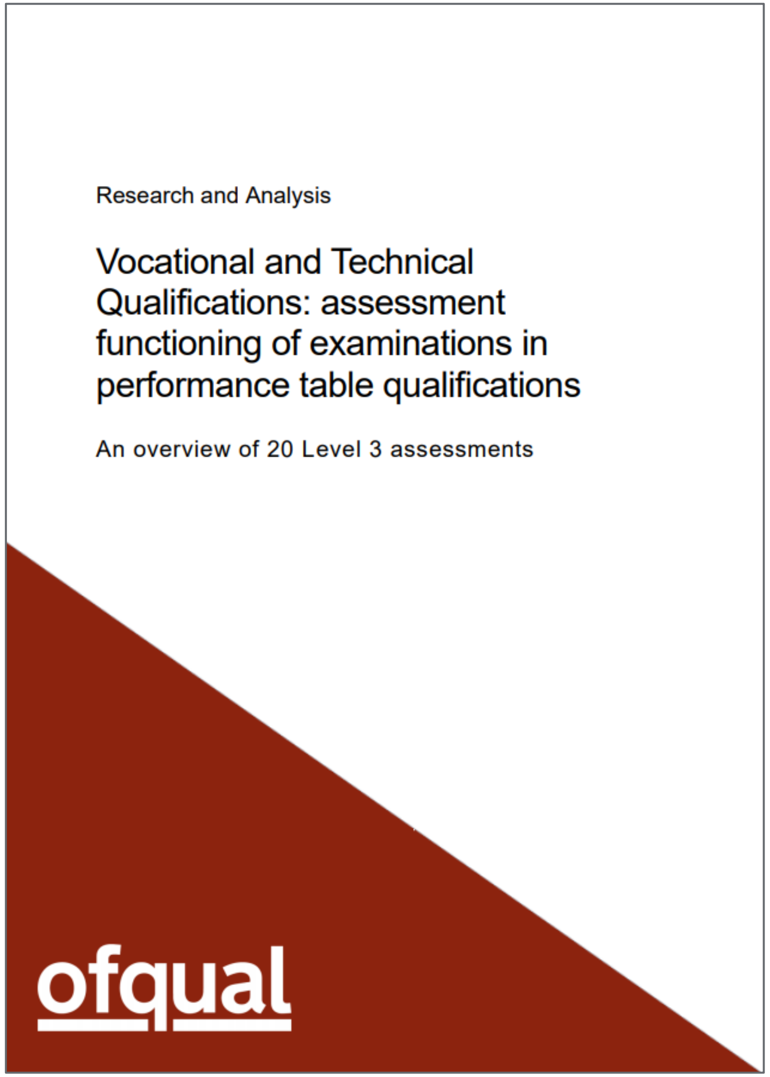
Vocational and Technical Qualifications: Assessment Functioning of external assessments

An overview of the functioning of assessments in 27 qualifications and 49 units

November 2017

Ofqual/17/6319

Beth Black,
Qingping He,
Stephen Holmes,
Caroline Morin



Research and Analysis

Vocational and Technical Qualifications: assessment functioning of examinations in performance table qualifications

An overview of 20 Level 3 assessments

ofqual

VTQs on the ‘approved list’ for DfE’s 16-19 performance tables

2016 series

- 49 tests (27 qualifications)
- mainly L1 and L2
- health & social care, carpentry, hospitality, digital media, applied science, mathematics

2017 series

- 20 tests (15 qualifications)
- L3 only (Applied Generals/Tech Levels)
- applied science, business, digital media, engineering, health & social care, IT/computing, sport

Item functioning statistics – computed separately for each question

- Facility indices
  - how easy/hard is the question?
- Discrimination indices
  - do candidates who tend to perform poorly/well on the question also tend to perform poorly/well on the test overall?

Test functioning statistics – computed for the test overall

- Mean of marks
  - □ how easy/hard is the test?
- Standard deviation of marks
  - □ how widely spread are the marks (across the mark range)?
- Reliability coefficient
  - □ an estimate of the degree to which results are likely to be replicable

We need to make certain assumptions

- Each test is intended to provide an **overall estimate** of proficiency
  - it is not intended to certify the attainment of a specified set of AC for the unit
- Each test is intended to **differentiate** between candidates
  - between gradations of proficiency (pass, merit, distinction)
- All candidates were adequately **prepared** for their tests
  - they were enrolled on an appropriate course (at the right level)
  - they were taught the subject content appropriately
  - they were given appropriate experience of test-taking

2016 tests – Facility indices

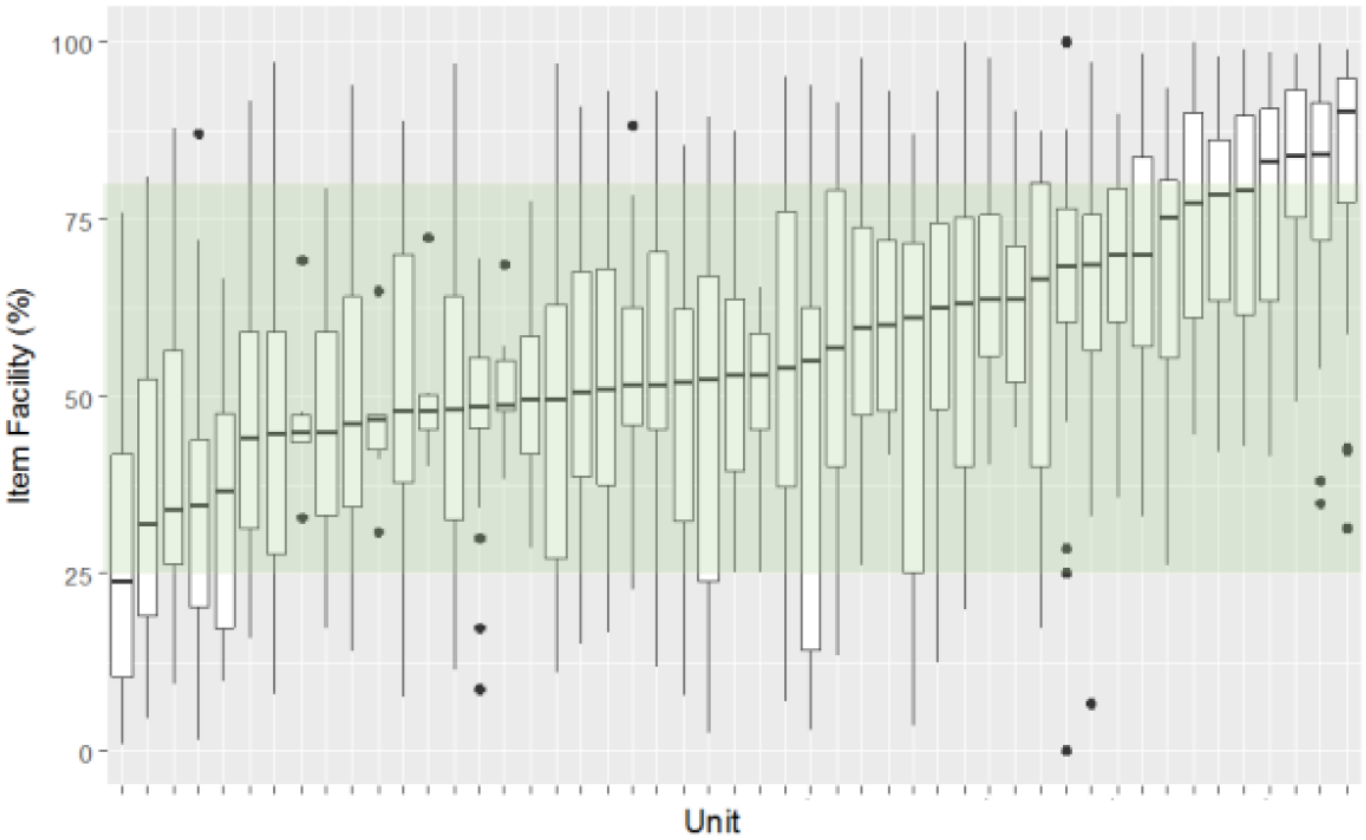
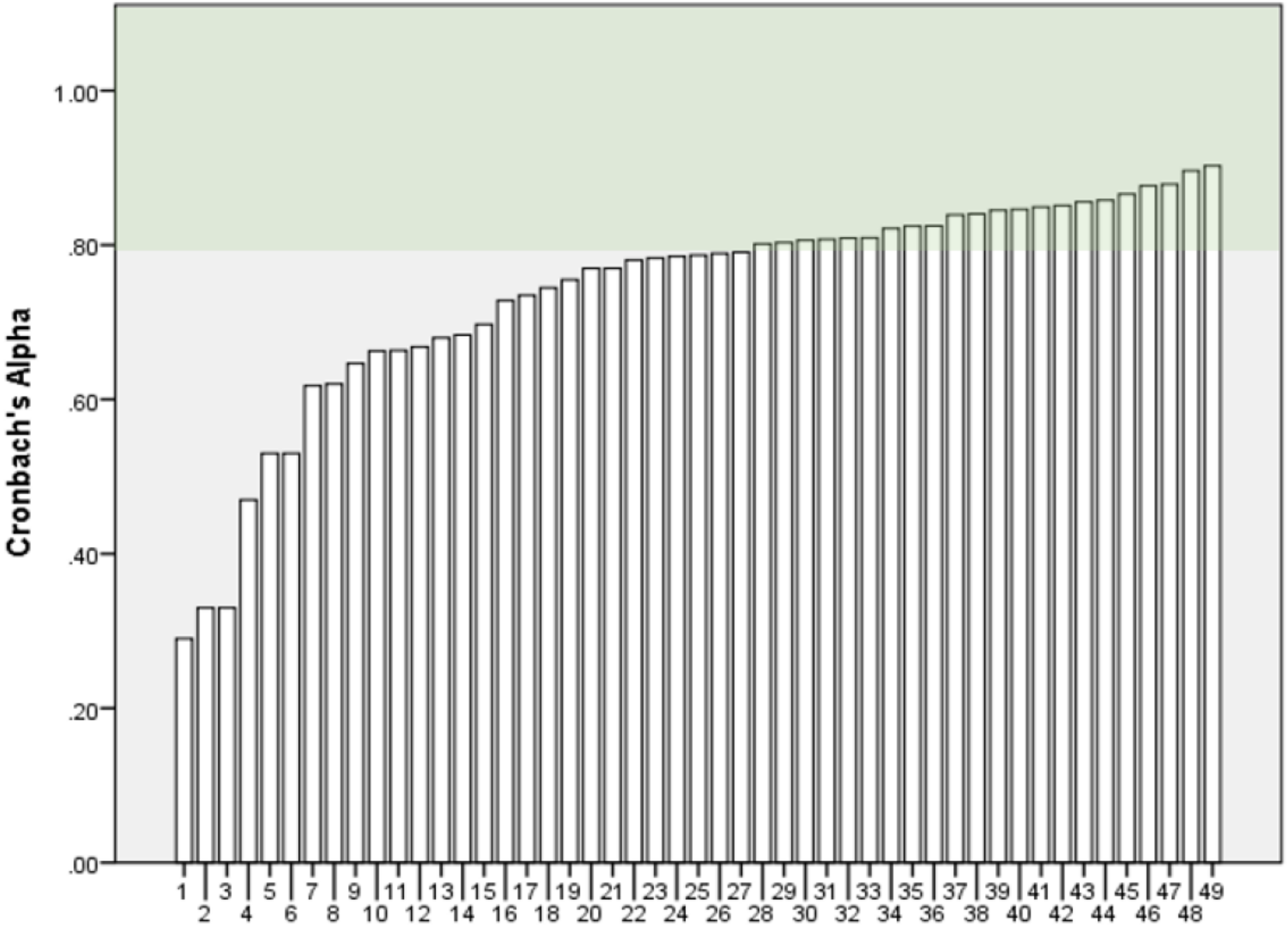


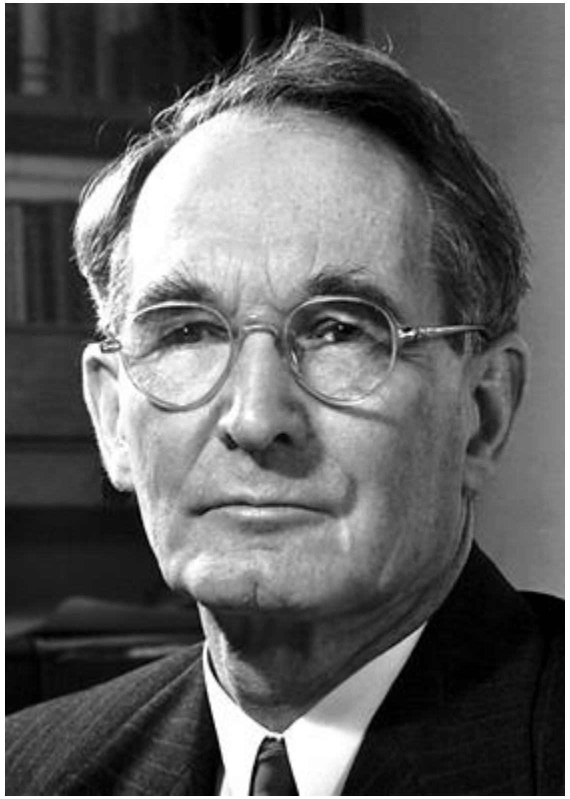
Figure 1: ***Box and whisker plots showing the distribution of item facilities for each of the 49 tests. The green area indicates the ideal range of item facilities. Tests arranged according to ascending order of median facility value (black horizontal line)***

2016 tests – Reliability coefficients



We need to engage technically with VTQs

“nothing more than doing one’s damnedest with one’s mind, no holds barred” (Bridgman, 1955)



Credit: Wikimedia Commons

- **WHY?** Because VTQ assessment is never unproblematic, so it always needs to be studied (scientifically)
  - □ validity (*includes* reliability, comparability, bias)
- **HOW?** By comparing how VTQ assessment is supposed to work (in theory) with how it actually works (in practice)
- **AND?** Statistics can be our friends!
  - □ just as long as we get to know them really well and treat them right