CEM
Centre for Evaluation & Monitoring

#CoeCam
@ProfCoe

The validity of England's accountability data

Robert Coe
Cambridge Assessment , 1st October 2015

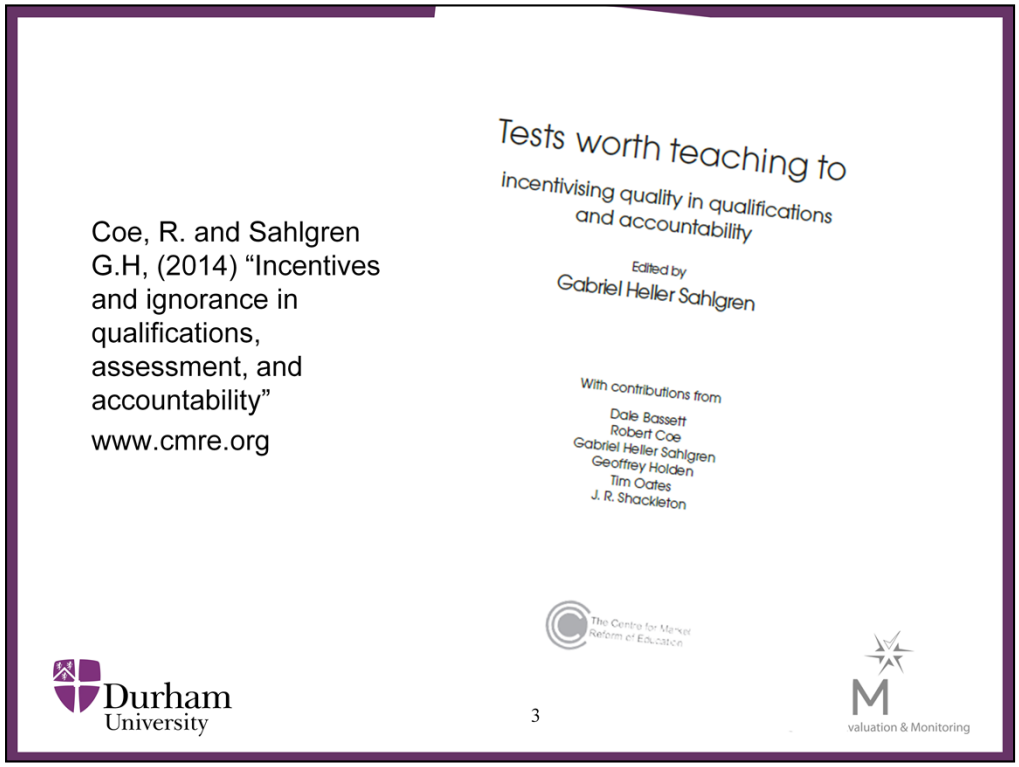
Durham University

Outline

- What is accountability?
- In what ways do accountability systems vary?
- Research evidence on accountability
- Validity: exams; Ofsted
- The Laws of Accountability
- Some suggestions

Durham University

CEM Centre for Evaluation & Monitoring



Coe, R. and Sahlgren G.H, (2014) "Incentives and ignorance in qualifications, assessment, and accountability". In G.H. Sahlgren (ed.) *Tests worth teaching to: incentivising quality in qualifications and accountability*. Centre for Market Reform of Education
http://www.cmre.org.uk/sites/default/files/Tests%20worth%20teaching%20to_web%20text.pdf

What is accountability?

- Garn & Cobb (2001) identify three types
  - Bureaucratic accountability – compliance with required rules, procedures
  - Performance accountability – measurement of outcomes (poss incl incentives)
  - Market accountability – consumer choice drives system improvement
- Accountability = Evaluation + Incentives

Dimensions of accountability
'Hard' vs 'Soft'

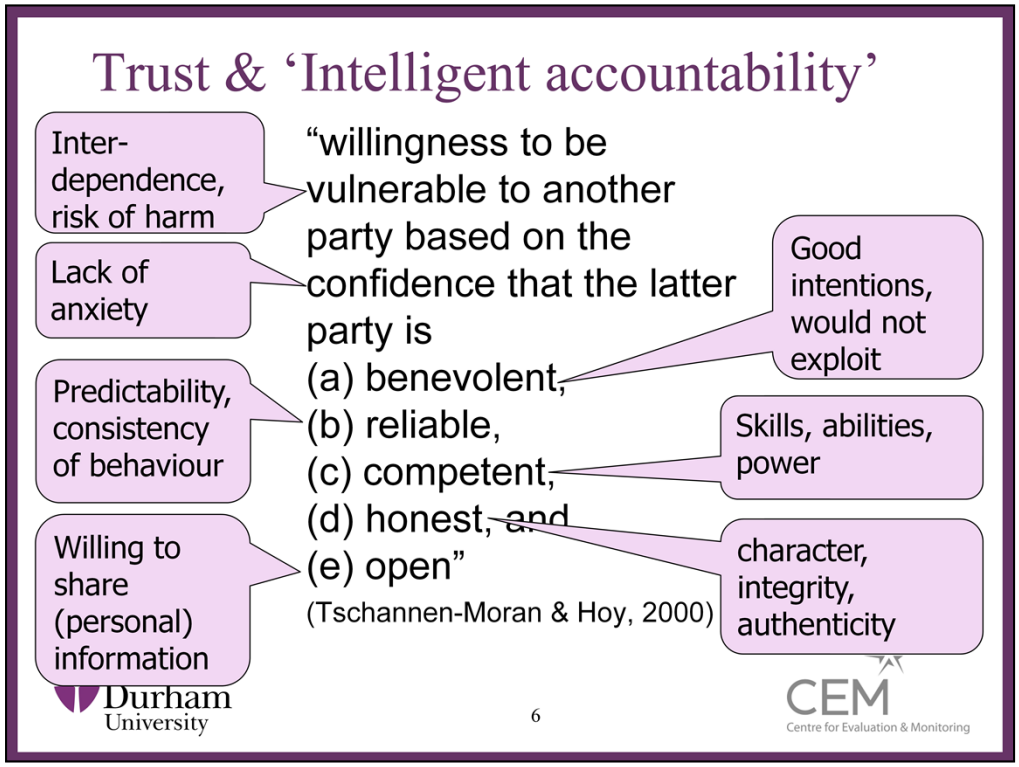
(Coe & Sahlgren, 2014)

- Links to incentives
  - Direct, explicit, vs implicit, intrinsic
- Public openness
  - Published vs confidential
- Objectivity/judgement
  - Direct measure vs interpreted indicator
- Improvement mechanism
  - Consequences vs feedback
- Prioritised actors
  - Consumers vs professionals

Durham University

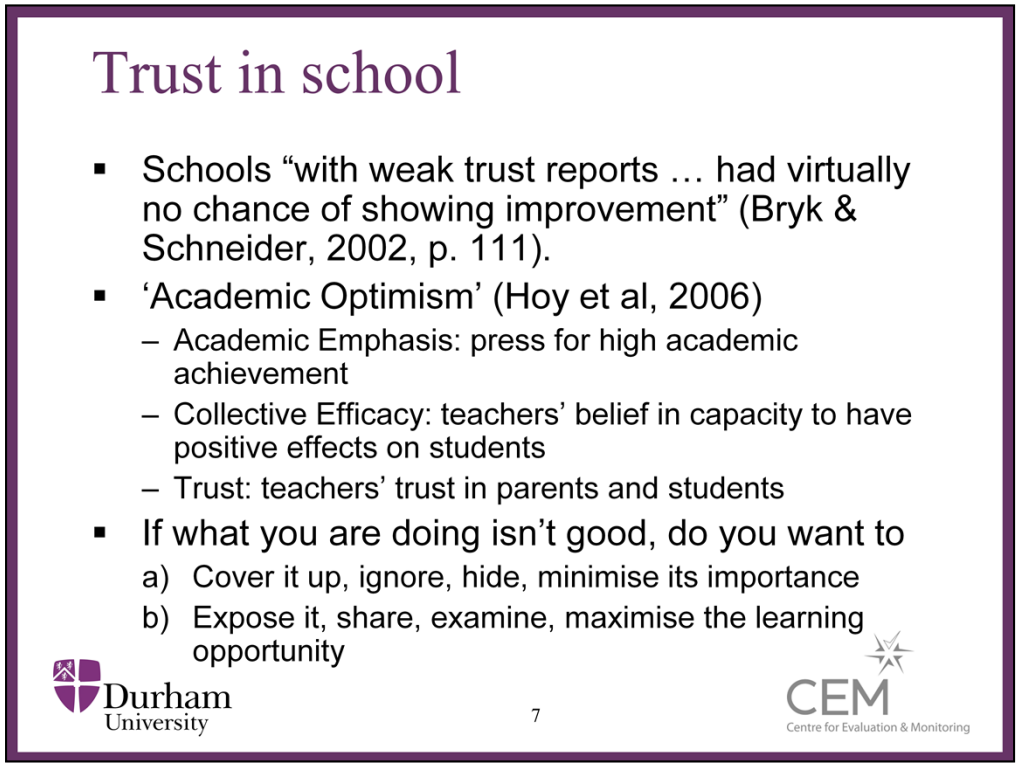
5

CEM Centre for Evaluation & Monitoring



Tschannen-Moran, M., & Hoy, W. K. (2000). A multidisciplinary analysis of the nature,

meaning, and measurement of trust. *Review of Educational Research*, 70(4), 547-593.



Bryk, A., & Schneider, B. (2002). Trust in schools. New York: Russell Sage.

Hoy, W. K., Tarter, C. J., & Hoy, A. W. (2006). Academic optimism of schools: A force for student achievement. *American educational research journal*, *43*(3), 425-446.

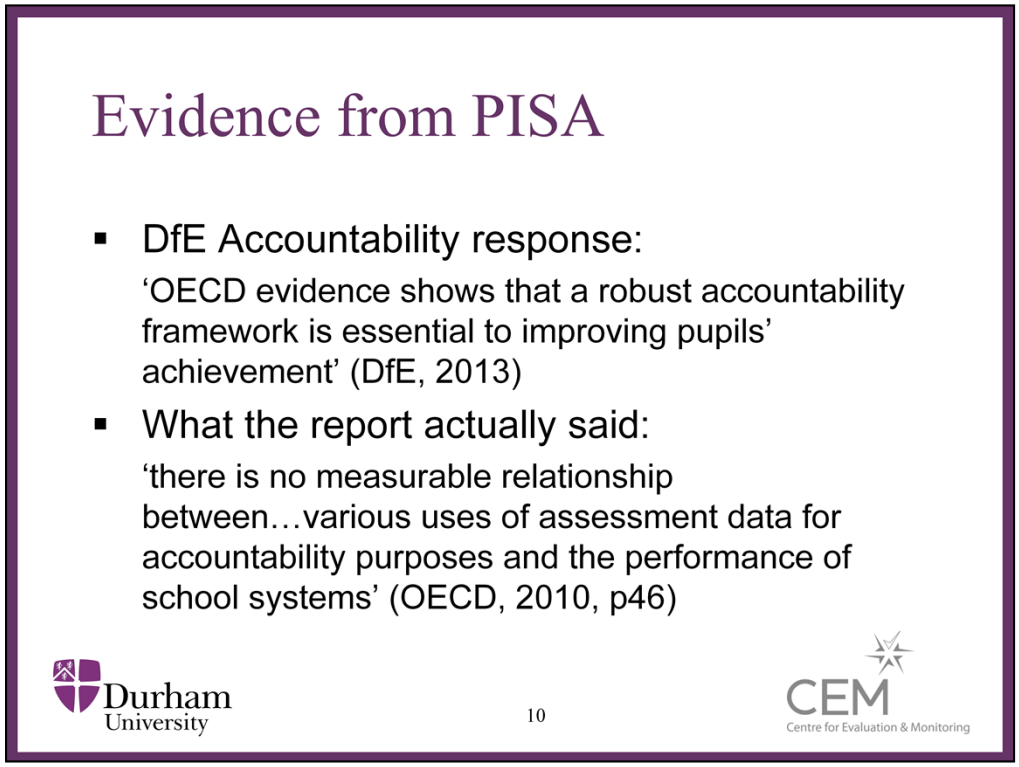
CEM
Centre for Evaluation & Monitoring

Evidence on impact of accountability

Durham University

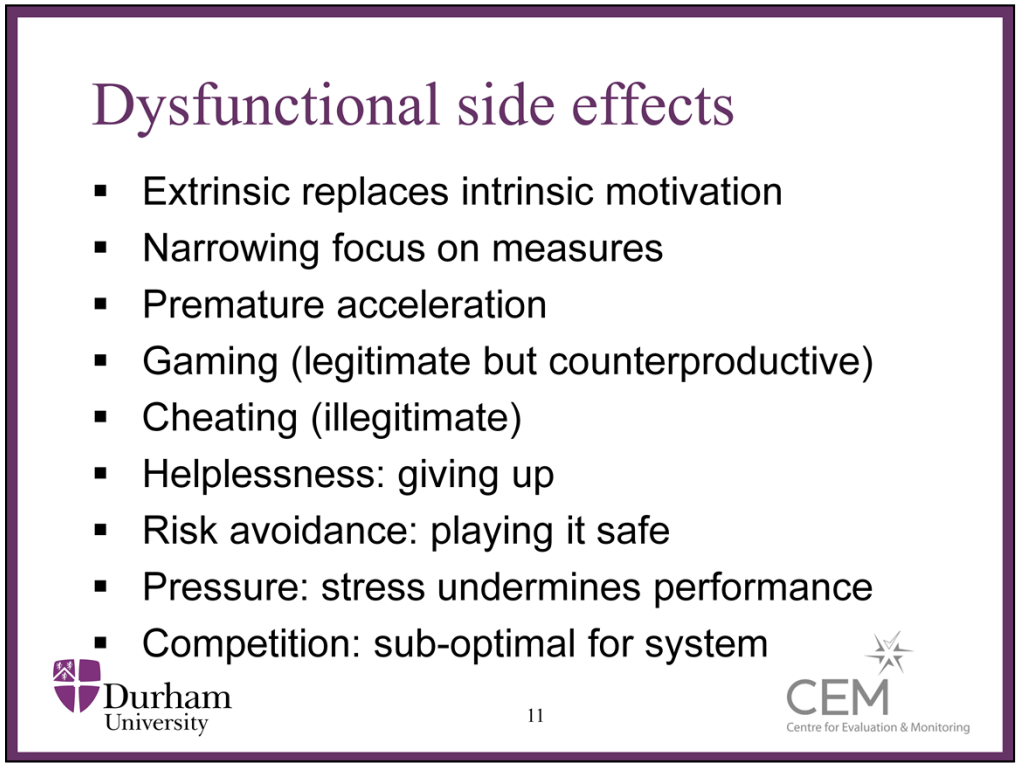
Research evidence

- Meta-analysis of US studies by Lee (2008)
  - Small positive effects on attainment (ES=0.08)
- Impact of publishing league tables (England vs Wales) (Burgess et al 2013)
  - Overall small positive effect (ES=0.09)
  - Reduces rich/poor gap
  - No impact on school segregation
- Other reviews: mostly agree, but mixed findings
- Lack of evidence about long-term, important outcomes
- Not clear how far impact on targeted outcomes transfers to other measures



Department for Education (DfE) (2013), 'Reforming the Accountability System for Secondary Schools: Government Response to the February to May 2013 Consultation on Secondary School Accountability'. Report, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/249893/Consultation_response_Secondary_School_Accountability_Consultation_14-Oct-13_v3.pdf (accessed 15 June 2014).

OECD (2010), 'PISA 2009 Results: What Makes a School Successful? Resources, policies and practices, Volume IV', http://www.oecd.org/pisa/pisaproducts/48852721.pdf (accessed 14 June 2014).



Eg Hutchins M. (2014) Exam Factories? The impact of accountability measures on children and young people . NUT https://www.teachers.org.uk/files/exam-factories.pdf

Meadows, M. (2015) Teacher ethics in summative assessment, paper presented at Oxford University Centre for Educational Assessment http://www.slideshare.net/ofqual/ofqual-ethicssymposiummichellemeadows

Limitations of existing research on accountability

- Conflicting claims & evidence – hard to summarise and reconcile
- Most evidence is quite weak
  - Conceptual: what is accountability anyway? Which aspects? Separating intended and unintended
  - Values: is a narrow focus good? Creative & exploratory work
  - Representativeness: often selective/anecdotal
  - Outcomes: perceptions of actors, not directly & independently observed
  - Attribution: cause and effect? Genuine change? Eg anxiety over exams; teacher turnover

Durham University

CEM Centre for Evaluation & Monitoring

Overall evidence-based conclusions

- Easy to cherry-pick

  '[E]ducational policy makers and practitioners should be cautioned against relying exclusively on research that is consistent with their ideological positions to support or criticize the current high-stakes testing policy movement' (Lee, 2008, p. 639)
- Direct incentives do drive people's behaviour; current evidence suggests accountability has (small) positive effects on targeted attainment
- Accountability systems always seem to have some undesirable side-effects
- Balance of positive & negative effects likely to depend on a range of factors; current knowledge does not allow us to predict confidently

Durham University

CEM Centre for Evaluation & Monitoring

CEM
Centre for Evaluation & Monitoring

Validity:
Uses, interpretations and decisions regarding exams and inspection

Durham University

Uses of exams (1): "Subject"

Use	*Interpretation*
The use of a B in GCSE maths as a filter for A level study in maths. The requirement for candidates who do not achieve a particular threshold in KS2 maths or English to repeat the test in Y7	The grade/score indicates specific competences within the subject domain that the candidate is likely to be able reproduce in the future.

Uses of exams (2): "Ability"

Use	*Interpretation*
The use of '5A*-C at GCSE' as a filter for proceeding to A level study in any subjects. Universities offering places conditional on candidates achieving particular grades in A level subjects that have no direct content overlap with the intended subject of degree	The grade indicates competences transferable to other academic study in other disciplines that the candidate is likely to be able reproduce in the future.

Uses of exams (3): "Teaching"

Use	*Interpretation*
Use of GCSE results in league tables to judge schools (eg to inform parent choice) Use of GCSE results to inform decisions about 'support' for schools (eg special measures or interventions triggered by a 'floor standard').	Average grades for a class or school (especially if referenced against prior attainment) indicate the impact (and hence quality) of the teaching experienced.

Validation

- Who should determine whether it is OK to use exams in this way?
  - Provider (exam boards)
  - User (universities, employers)
  - Regulator (Ofqual)
- What evidence would warrant these uses?

"Assessment developers should be required by the regulator (Ofqual) to state explicitly what interpretations, uses, and decisions their assessment outcomes are intended to support, and which are not appropriate. Evidence should be provided to show how well the assessments support the intended purposes."

Accountability measures (1)

(Coe & Sahlgren, 2014)

1. Do the measures represent valued outcomes?
2. Are there important outcomes not captured by the measures?
3. Is what is measured sensitive to changes in the desired behaviours (e.g. improvements in instruction or greater effort)?
4. Could performance on the measures reflect irrelevant or misleading confounds?
5. What are the limits of precision, misclassification, or consistency (reliability) of the measures?

Durham University

CEM Centre for Evaluation & Monitoring

Accountability measures (2)

(Coe & Sahlgren, 2014)

6. Are the measures fair to all subgroups, including individuals with disabilities, different language, cultural, or social backgrounds, or to schools that serve different kinds of communities?
7. Could it be possible to improve performance on the measures without any real improvements in valued outcomes?
8. Could it be possible to make significant improvements in some important educational outcome without that being reflected in any improvement in the measures?
9. Do the measures depend on judgements made by people with a potential conflict of interest?

Durham University

CEM Centre for Evaluation & Monitoring

CEM
Centre for Evaluation & Monitoring

Ofsted

“I understand that would be a long journey for [Ofsted] because there is a big cultural change there,” he concedes. “But it is relatively easy. They just need to decide to do their job properly.”

TES 24.4.15

Durham University

scheme has been funded despite an "outrageous" lack of evidence as to whether it is effective, it has been claimed. Serious concern has also been raised about insufficient research into the effectiveness of academies.

Professor Robert Coe, whose work on school standards has been cited by education secretary Michael Gove, told a major conference on research in education that practice in schools needed to be more closely linked to academic analysis. Ofsted, he said, was "part of the problem". "It is not research based or evidence based," the director of Durham University's Centre for Evaluation and Monitoring said.

According to Professor Coe, there is no proof that the watchdog's inspections and lesson observations lead to "valid" judgements. "What is the evidence about people making those kinds of judgements? Do we know that inspection creates benefits to the system?" he said. "Some studies suggested that, actually, schools take a long time to recover from inspections and they don't do any good, and yet we are spending I don't know how many millions on Ofsted … and the whole point of it is to raise standards. So let's see some evidence."

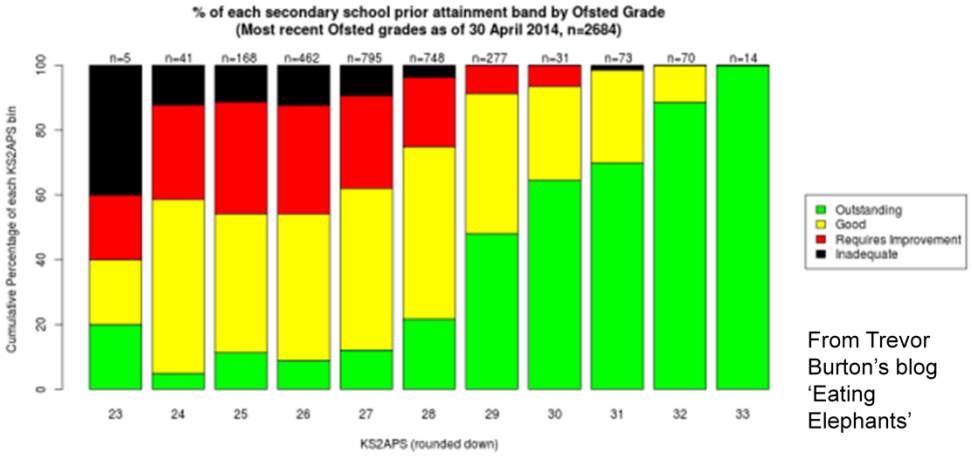
In a TES interview, Ofsted chief inspector Sir Michael Wilshaw described the claims as "tosh and nonsense". He said new figures released this week, showing a 9 percentage point rise in the proportion of schools judged to be good or outstanding (see panel, page 9), proved that the watchdog's tougher inspection regime had "galvanised the system".

But speaking at the ResearchED conference in London last weekend, Professor Coe questioned the basis of the watchdog's verdicts and said it needed to demonstrate that its lesson observations were valid. Classroom observation in general was the "next Brain Gym" because there was no scientific evidence to show that it led to better learning, he said.

But Sir Michael said: "I don't know of any headteacher who doesn't believe that classroom observation isn't anything other than a help. The fact that we are an inspectorate and we do make judgements has made a huge amount of difference."

TES Sept 2013

What's the easiest way to a secondary Ofsted Outstanding?



From Trevor Burton's blog 'Eating Elephants'

'Ofsted has not disputed the figures but insists that its inspectors pay "close attention" to prior pupil attainment and take a broad view of schools.' (TES)

Durham University

23

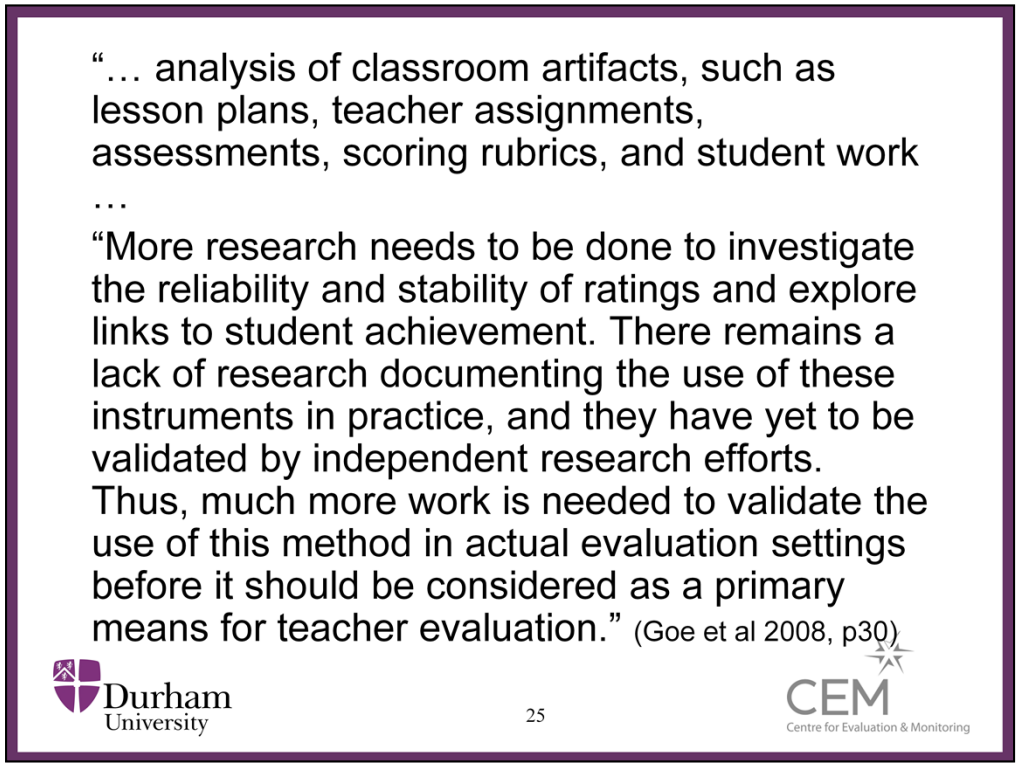
CEM Centre for Evaluation & Monitoring

What's the easiest way to a secondary Ofsted Outstanding?

https://jtbeducation.wordpress.com/2014/06/29/whats-the-easiest-way-to-a-secondary-ofsted-outstanding/

Quotation from William Stewart, TES, 22 Aug 2014, Is Ofsted's grading 'scandalous'? https://www.tes.co.uk/article.aspx?storycode=6440390

http://www.cem.org/blog/414/



Goe, L., Bell, C., & Little, O. (2008). Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. National Comprehensive Center for Teacher Quality. http://files.eric.ed.gov/fulltext/ED521228.pdf

What should Ofsted do?

- Inspectors should have to pass an exam
  - "an independently accredited professional standard, with a clear specification for what inspectors need to know, the kinds of training that should be expected and a rigorous assessment to show that the standard has been met."
- Establish the validity of inspection judgements
  - "… providing evidence of the validity of judgements shouldn't be done once in a half-baked study conducted by Ofsted, but should be done rigorously, with independent scrutiny, on an ongoing basis every time a judgement is made."

(Coe, TES, 24 Oct 2013)

Durham University

CEM Centre for Evaluation & Monitoring

26

Coe, R. (2013) "Critics should not have to prove Ofsted's judgements are wrong; it should be up to Ofsted to prove they are right". Blog for TES Opinion, 24 Oct 2013 http://news.tes.co.uk/opinion_blog/b/weblog/archive/2013/10/24/critics-should-not-have-to-prove-ofsted-s-judgements-are-wrong-it-should-be-up-to-ofsted-to-prove-they-are-right.aspx

CEM
Centre for Evaluation & Monitoring

Ways forward

Durham University

Laws of Accountability

1. Meddling with qualifications and accountability is irresistible to politicians

2. Unintended effects are always underestimated

Durham University

CEM
Centre for Evaluation & Monitoring

"Given such ignorance, a policy of dictating a single accountability structure for all schools in England can hardly be described as evidence based. A more scientific approach would be to allow a range of variation … within what is politically acceptable, and then randomly allocate different groups of schools to experience accountability systems that differ on these factors."

Coe & Sahlgren, 2014

Durham University

CEM Centre for Evaluation & Monitoring

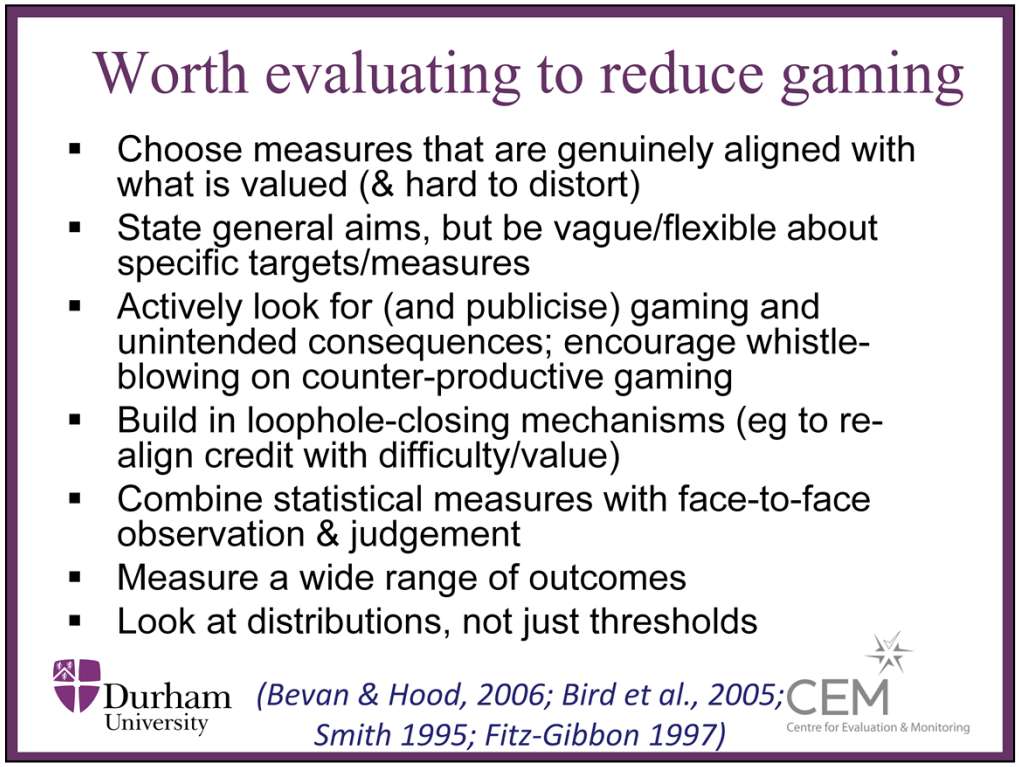
Dimensions of accountability
'Hard' vs 'Soft'

(Coe & Sahlgren, 2014)

- Links to incentives
  - Direct, explicit, vs implicit, intrinsic
- Public openness
  - Published vs confidential
- Objectivity/judgement
  - Direct measure vs interpreted indicator
- Improvement mechanism
  - Consequences vs feedback
- Prioritised actors
  - Consumers vs professionals

Durham University

CEM Centre for Evaluation & Monitoring



GWYN BEVAN & CHRISTOPHER HOOD (2006) What's measured is what matters: targets and gaming in the English public health care system. *Public Administration*, Volume 84, Issue 3, pages 517–538, August 2006

Bird S. M., Cox D., Vern T. F., Goldstein H., Holt T., Smith P. C. (2005) Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Volume 168, Issue 1, pages 1–27, January 2005

Fitz-Gibbon, C.T. (1997) *The Value Added National Project: Feasibility Studies for a National System of Value Added Indicators (Final Report)*. London: SCAA.

Smith, P. (1995) 'On the unintended consequences of publishing performance data in the public sector' *International Journal of Public Administration*, 18 (2/3) 277-310.

Rescuing high-stakes teacher assessment

- Training in assessment and moderation
- Link teacher assessed mark distribution to within-centre exam mark distribution
- Spot checks (risk targeted): can students reproduce it?
- Support whistle-blowing
- Signed declarations from teachers, headteachers and students
- Questionnaire audit of practices: 'too good to be true' triggers spot check

Durham University

CEM Centre for Evaluation & Monitoring

CEM
Centre for Evaluation & Monitoring

www.cem.org

Summary …

1. Evidence on accountability is not great, but suggests small positive impacts
2. Dysfunctional side-effects are also real
3. Exams and inspection should be improved
4. We need experiments to learn how to optimise

@ProfCoe

Robert.Coe@cem.dur.ac.uk

Durham University