



Recognising the error of our ways

Dr Paul E. Newton

Presentation to the Cambridge Assessment Forum for New Developments in Educational Assessment. Downing College, Cambridge. 10 December 2008.

HOW MANY STATISTICIANS DOES IT TAKE TO CHANGE A LIGHT BULB?

ONE, PLUS OR MINUS THREE!

Other valid responses:

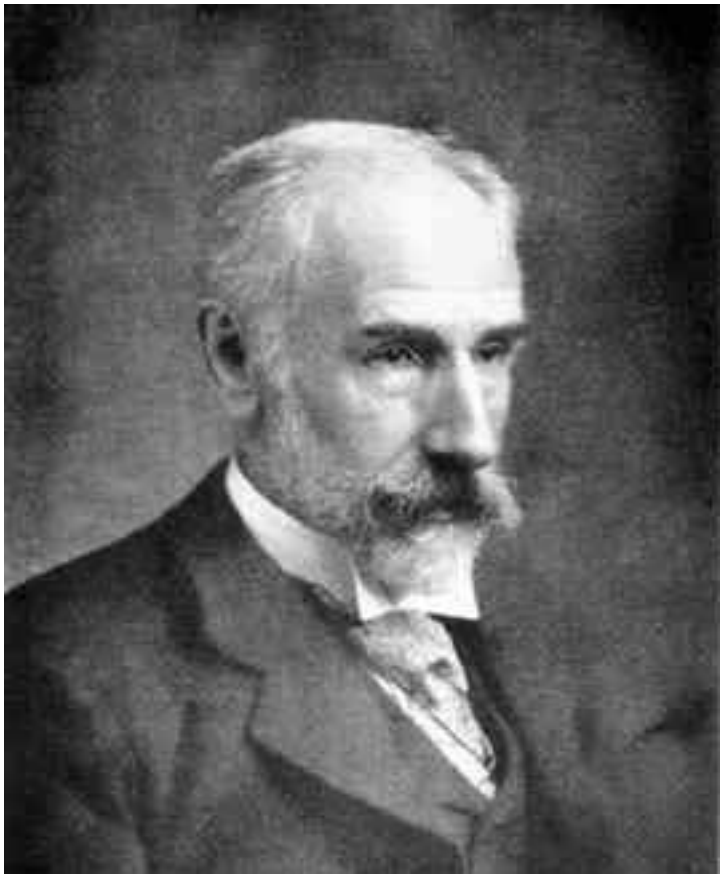
- How many did it take this time last year?
- 3.9967 (after six iterations).
- 75% of the population believe less than four.
- What kind of number did you have in mind?
- Don't bother. Nothing can be inferred from a single light bulb.
- You'd need to use a nonparametric procedure – statisticians are not normal.
- 1-n to change the bulb and n-1 to test its replacement.
- It depends whether the bulb is - vely or + vely screwed.

ONE!

HOW MANY PSYCHICS DOES IT TAKE TO CHANGE A LIGHT BULB?

Francis Ysidro Edgeworth

That examination is a very rough, yet **not wholly inefficient**, test of merit is generally accepted.



Part 1

What do we mean by ‘error’?

Variability

Whatever precautions have been taken to secure unity of standard, there will occur a certain divergence between the verdicts of competent examiners. Say full marks are thirty; then if one examiner marks 20, another might mark 21, another 19. If we tabulate the marks given by the different examiners, they will tend to be disposed after the fashion of a gend'arme's hat.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, LI, 599-635.

A gendarme’s hat?

Chapeau de Gendarme

Measurement ‘truth’

This central figure which is, or may be supposed to be, assigned by the greatest number of equally competent judges, is to be regarded as the true value of the Latin prose; just as the true weight of a body is determined by taking the mean of several discrepant measurements.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, LI, 599-635.

Measurement ‘error’

I think it is intelligible to speak of the mean judgment of competent critics as the true judgment; and deviations from that mean as errors.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, LI, 599-635.

Reliability and replication

Reliability is about quantifying the luck of the draw.

What if the…

- **candidate** happened to have been in a different state of mind?
- **exam** happened to have comprised a different set of questions?
- **script** happened to have been marked by a different marker?
- **cut-scores** happened to have been set by a different panel?
- etc.

… would the same grade have been awarded?

Part 2

What do we know about error?

The public perception of error?

Only limited data have been published about the reliability of national curriculum tests, although it is likely that the reliability of national curriculum tests is around 0.80 – perhaps slightly higher for mathematics and science.

Black, P. & Wiliam, D. (2006). The reliability of assessments. In J. Gardner (Ed.). *Assessment and learning*. London: Sage.

Test consistency

		Target levels	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Key Stage 1 Tests	Spelling	2,3	n.a.	0.94	0.97	0.95	0.97	0.95	0.94	0.89	0.92	0.92	-	0.92
	Reading	2	n.a.	0.87	0.92	0.91	0.91	0.91	0.87	0.90	0.90	0.87	-	0.89
	Reading	3	n.a.	0.77	0.84	0.75	0.82	0.84	0.78	0.80	0.79	0.82	-	0.76
	Mathematics	2,3	n.a.	0.88	0.88	0.88	0.89	0.90	0.90	-	-	-	-	-
	Mathematics	2	-	-	-	-	-	-	-	0.88	0.88	0.83	-	0.85
	Mathematics	3	-	-	-	-	-	-	-	0.83	0.83	0.84	-	0.85
Key Stage 2 Tests	Reading	3,4,5	0.85	0.86	0.92	0.89	0.88	0.88	0.90	0.87	0.87	0.87	0.91	0.89
	Writing	3,4,5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Spelling	3,4,5	0.91	0.90	0.92	0.92	0.91	0.89	0.90	0.90	0.90	0.91	0.91	0.89
	Handwriting	3,4,5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Mathematics A	3,4,5	0.88	0.87	0.91	0.90	0.90	0.89	0.89	0.92	0.93	0.91	0.93	0.92
	Mathematics B	3,4,5	0.89	0.88	0.83	0.90	0.87	0.89	0.89	0.93	0.92	0.92	0.93	0.92
	Mental mathematics	3,4,5	-	-	0.90	0.88	0.85	0.88	0.89	0.88	0.89	0.87	0.87	0.89
	Overall	3,4,5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.97	0.97	0.97	0.97	0.97
	Science A	3,4,5	0.83	0.86	0.85	0.87	0.87	0.86	0.88	0.86	0.87	0.86	0.87	0.86
	Science B	3,4,5	0.82	0.87	0.86	0.87	0.87	0.87	0.88	0.85	0.86	0.86	0.87	0.82
	Overall	3,4,5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.92	0.93	0.92	0.93	0.91
Key Stage 3 Tests	Reading	3,4,5,6,7	0.71	0.88	0.94	0.90	0.89	0.89	0.88	0.84	0.84	0.81	0.85	0.85
	Writing	3,4,5,6,7	0.91	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Shakespeare	3,4,5,6,7	-	-	-	-	-	-	-	n.a.	n.a.	n.a.	n.a.	n.a.
	Mathematics 1	3,4,5	0.88	0.89	0.88	0.90	0.92	0.91	0.90	0.89	0.91	0.90	0.89	0.91
	Mathematics 2	3,4,5	0.88	0.94	0.90	0.89	0.92	0.92	0.88	0.91	0.91	0.91	0.90	0.90
	Mathematics 1	4,5,6	0.86	0.81	0.84	0.86	0.85	0.85	0.87	0.84	0.86	0.88	0.86	0.88
	Mathematics 2	4,5,6	0.84	0.91	0.82	0.82	0.87	0.89	0.88	0.85	0.88	0.87	0.86	0.87
	Mathematics 1	5,6,7	0.86	0.90	0.84	0.84	0.88	0.88	0.86	0.87	0.85	0.90	0.90	0.88
	Mathematics 2	5,6,7	0.88	0.87	0.85	0.83	0.88	0.91	0.88	0.88	0.88	0.89	0.90	0.87
	Mathematics 1	6,7,8	0.85	0.68	0.82	0.85	0.89	0.90	0.92	0.88	0.88	0.89	0.90	0.88
	Mathematics 2	6,7,8	0.87	0.81	0.80	0.83	0.90	0.92	0.90	0.89	0.91	0.89	0.90	0.91
	Mental mathematics A	4,5,6,7,8	-	-	0.89	0.87	0.88	0.88	0.86	0.87	0.89	0.90	0.89	0.88
	Mental mathematics B	4,5,6,7,8	-	-	0.88	0.90	0.88	0.80	0.86	0.85	0.89	0.88	0.86	0.89
	Mental mathematics C	3,4,5	-	-	0.83	0.81	0.83	0.87	0.83	0.83	0.82	0.85	0.86	0.85
	Science 1	3,4,5,6	0.88	0.90	0.91	0.90	0.93	0.94	0.90	0.94	0.91	0.92	0.93	0.92
	Science 2	3,4,5,6	0.88	0.89	0.89	0.88	0.92	0.94	0.90	0.93	0.92	0.93	0.93	0.91
	Overall	3,4,5,6	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.96	0.96	0.96	0.96	0.96
	Science 1	5,6,7	0.85	0.84	0.86	0.82	0.88	0.87	0.87	0.87	0.92	0.88	0.88	0.88
	Science 2	5,6,7	0.85	0.85	0.86	0.88	0.87	0.86	0.87	0.88	0.90	0.90	0.90	0.91
	Overall	5,6,7	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0.93	0.95	0.94	0.94	0.95

Marker consistency

Agreement between markers (n = 9) and Lead Chief Marker	English 100 marks N, 3, 4, 5, 6, 7	Reading 50 marks B4, 4, 5, 6, 7	Writing 50 marks B4, 4, 5, 6, 7
Mean coefficient of correlation (marks)	0.92	0.94	0.80
Percentage exact agreement (levels)	59 %	61 %	52 %

Level setting consistency

3 to 6 tier		Confidence Interval		
	Tucker Linear	Lower	Upper	Final
Level 3	45	42	48	42
Level 4	72	70	74	69
Level 5	105	103	106	104
Level 6	135	133	136	134

Dylan Wiliam on error

[…] it is likely that the proportion of students awarded a level higher or lower than they should be because of the unreliability of the tests is at least 30% at key stage 2

Wiliam, D. (2001). *Level best?* London: ATL.

Overall reliability (parallel forms)

Agreement between performance across test forms	English 100 marks B3, 3, 4, 5	Reading 50 marks B3, 3, 4, 5	Writing 50 marks B3, 3, 4, 5
Classification consistency (two forms)	73 %	73 %	67 %
Classification accuracy – rough!! (one form)	84 %	84 %	79 %

Part 3

What do we say about error?

Sometimes we dodge questions

The Qualifications and Curriculum Authority said the test was carefully trialled and pre-tested to make sure it was appropriate and stimulating for the age group.

Ward, H. (2002). Children exhausted by 'too wordy' reading challenge. *The TES*, 24 May.

A QCA spokesman said that all the questions cited were consistent with national curriculum requirements.

Shaw, M. (2002). A gender-bending question. *The TES*, 17 May.

Sometimes we downplay error

A Qualifications and Curriculum Authority spokeswoman said: “We are confident that the quality of the marking of tests is robust.”

Mansell, W. (2003). Row over test marks at 14. *The TES*, 11 July.

Occasionally ‘inevitable’

“It was a proof-reading error on our part.” said a spokesman for the authority. “We make no excuses and this error should not have happened, but we have made sure no students suffer as a result.”

Mistakes are inevitable in an examinations system which deals with 18 million papers a year, says the QCA.

Hook, S. (2002). Anger at blunder in key skills paper. *The TES*, 24 May.

Occasionally ‘unacceptable’

However, any level of error has to be unacceptable – even just one candidate getting the wrong grade is entirely unacceptable for both the individual student and the system.

QCA. (2003). A level of preparation. TES Insert. *The TES*, 4 April.

Sometimes we're just not sure

Dr Boston: Error exists. As I said before, this a process of judgment. Error exists, and error needs to be identified and rectified where it occurs. I am surprised at the figure of 30%. We have been looking at the range of tests and examinations for some time. We think that is a very high figure, but whatever it is it needs to be capable of being identified and corrected.

House of Commons Children, Schools and Families Committee. (2008). *Testing and Assessment.* Third Report of Session 2007–08. Volume II. Oral and written evidence. HC 169-II. London: TSO Limited.

The cloak of secrecy

Boards seem to have strong objections to revealing their mysteries to 'outsiders' […] There have undoubtedly been cases of inquiries […] where publication would have been in the interests of education, and would have helped to prevent the spread of 'horror-stories' about such things as lack of equivalence which is an inevitable concomitant of the present cloak of secrecy.

Wiseman, S. (1961). The efficiency of examinations. In S. Wiseman (Ed.). *Examinations in education.* Manchester: MUP.

A new dawn of openness

In presenting this booklet to the public […] we in the GCE boards have found ourselves in a dilemma. If we merely state that comparability exercises are regularly conducted and do not show our hand, we appear to have something to hide. If we try to explain them, their complexities and limitations invite misunderstanding and misrepresentation. On balance, the preferable alternative seemed to be to 'publish and be damned'. We have, and probably shall be.

(A. Robin Davis, in) Bardell, G.S., Forrest, G.M. & Shoesmith, D.J. (1978). *Comparability in GCE*. Manchester: JMB.

Recent Cambridge work

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, 38(2), 247-264.

Crisp, V. & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*, 33, 6, 943-961.

Greatorex, J. and Bell, J.F. (2004). Does the gender of examiners influence their marking? *Research in Education*, 71, 25-36.

Suto, W.M.I. & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, 34, 2, 213-233.

Suto, W.M.I. & Greatorex, J. (2008). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy & Practice*, 15, 1, 73-89.

Part 4

What do we still need to know about error?

The current situation

Little **systematic and sustained** effort to evaluate the reliability of results from national tests and exams.

Evaluation work to date has been:

1. relatively **isolated** – not part of routine monitoring
2. **partial** – only certain facets, small no. tests and exams
3. **under-theorised** – little debate over interpretation

Part 5

Should we be saying more about error?

Reasons NOT to be more open

1. **The error narrative isn't watertight**
   - We still don't properly understand truth and error
   - We still don't have good error statistics
2. **The error narrative is complicated**
   - It's hard to explain error statistics meaningfully
   - The media won't allow us sufficient space to do so
3. **The error narrative is misleading**
   - Results are best estimates, regardless of error margins
4. **The error narrative is damaging**
   - We can't afford to threaten public confidence

The ‘true value’ of a script?

The (hypothetical) **average mark** awarded to the script, were it to have be marked many times by:

1. the (actual) **principal examiner**
2. each member of the (actual) population of **examiners** trained to mark the examination by the (actual) **principal examiner**
3. each member of multiple (hypothetical) populations of trained **examiners**, each population having been trained by a different **principal examiner**

… or something else entirely?

The traditional conception

Measurement error = inconsistency across replications

Cronbach's alpha 'expects'

1. all items measure a single construct
   - ➢ each item is a mini-replication
2. perfect consistency across replications
   - ➢ perfect inter-item correlation

Yet, for curriculum-based exams, items intentionally assess different aspects of a construct

A broader conception

Assessment error =

1. **inconsistency across replications**
   - random error (reliability)

   plus

2. ***some* of the consistency across replications**
   - systematic error (validity)

An even broader conception

What inferences might be drawn from a grade C in A level sociology?

- **a level of attainment at the point of taking the exam**
- **a level of attainment some time after taking the exam**
- **a potential to succeed in that domain in the future**
- **a potential to succeed in a different domain in the future**
- **etc.**

It’s hard to explain error stats

In terms of a five-point grading scale... all that can properly be said about a candidate awarded a grade 3 is that his ‘true’ grade could be as high as a grade 2 or as low as a grade 4.

Willmott, A.S. & Nuttall, D.L. (1975). *The reliability of examinations at 16+*. Basingstoke: Macmillan Education Ltd.

[…] results on a six or seven point grading scale are accurate to about one grade either side of that awarded.

Schools Council. (1980). *Focus on examinations*. Pamphlet 5. London: Schools Council.

The media won’t let us explain

Peter Smith, general secretary of the Association of Teachers and Lecturers, said: “The tests are riddled with fundamental flaws. We are not against testing, but we're utterly opposed to half-baked interpretation of the results.”

A Department for Education and Skills spokesman said: “The tests provide an effective and reliable means of assessing pupils' progress at key points in their education.”

Henry, J. (2001). Professor calls for end to 'bogus' tests. *Times Educational Supplement*, 30 November.

The error narrative misleads

Regardless of the margin of error to which it is subject, a candidate's observed grade is the best estimate of his or her true grade. Because of this, selectors will, in the long run, make the most correct selection decisions by taking observed grades at their face value.

Cresswell, M.J. (1986). Examination grades: how many should there be? *British Educational Research Journal*, 12(1), 37-54.

The error narrative damages

Christina Townsend

Chief Executive, Edexcel (2000)

Ron Tuck

Chief Executive, SQA (2000)

Wiliam Stubbs

Chairman, QCA (2002)

Estelle Morris

Education Secretary, DfES (2002)

Reasons FOR BEING more open

1. **Ignorance is no excuse**
   - We've had 120 years to get the error narrative right
2. **The error narrative is illuminating**
   - For students, parents, researchers, policy makers
3. **The 'myth of perfection' is damaging**
   - We can't afford to condone public misperceptions
4. **A new world order is emerging**
   - Freedom of Information, Accountability, Regulation

Ignorance is no excuse

Edgeworth (1888)

- true and error values
- sources of random *and* systematic error
- classification accuracy (and result reporting)
- borderlining
- multiple marking
- standards over time
- examiner, and mark distribution, scaling
- methods of aggregation

The error narrative illuminates

For students and teachers

- **maybe you're better than your grades suggest**
- **maybe you're worse than your grades suggest**

The error narrative illuminates

For **employers** and **selectors**

- maybe such fine distinctions shouldn't be drawn
- maybe other information should be taken into account

The error narrative illuminates

For **parents**

- maybe that difference in value added is insignificant
- maybe those kinds of inference cannot be drawn

Holy Trinity (CVA=102) versus All Saints (CVA=98)

- taught the **core subjects** better (‘02 to ‘08)?
- taught the **national curriculum** better (‘02 to ‘08)?
- **educated** better (‘02 to ‘08)?
- will **educate** better (in the future)?
- will educate **my child** better?

The error narrative illuminates

For **policy makers**

- maybe that proposed use of results is illegitimate
- maybe that policy change will compromise accuracy

The myth of perfection damages

The assessment profession disempowers itself

- **less talk**
  - less action
  - less thought

The myth of perfection damages

The assessment profession disempowers itself

- **less education**
  - less public understanding
  - less constructive debate

The myth of perfection damages

30 per cent of pupils may be given the wrong test level, a finding which ministers have never disproved.

Mansell, W. (2006). Persistent professor returns. *Times Educational Supplement*, 18 August.

The myth of perfection damages

Q298 Annette Brooke: I did ask the Qualifications and Curriculum Authority what it was doing to investigate the matter. I was not very satisfied that it was checking out the figure. I think that it is important to check it out. Perhaps you could ask them to do it, and then you can put your hands up and say that there is not a 30% error rate.

House of Commons Children, Schools and Families Committee. (2008b). *Testing and Assessment.* Third Report of Session 2007–08. Volume II. Oral and written evidence. HC 169-II. London: The Stationery Office Limited.

A new world order

Newspapers and activists invoke a supposed public 'right to know'. Freedom of information has become an admired ideal, and freedom of the press is still going strong. […] It seems that openness and transparency are set to replace traditions of secrecy and deference, at least in public life.

Onora O'Neill (2002). *A question of trust*. BBC Reith Lecture 4.

So, should we be more open?

AGAINST	FOR
The error narrative isn't **watertight**	Ignorance is **no excuse**
The error narrative is **complicated**	
The error narrative is **misleading**	The error narrative is **illuminating**
The error narrative is **damaging**	The 'myth of perfection' is **damaging**
	A **new world order** is emerging