14th Annual

Hawker Brownlow Thinking & Learning Conference

PROFESSOR DYLAN WILIAM

SATURDAY 20 MAY

Assessment Literacy: The Meaning and Consequences of Educational Assessments

Session 3

MELBOURNE

Hawker Brownlow EDUCATION

03 8558 2444

03 8558 2400

www.hbconf.com.au

conferences@hbe.com.au

PROFESSOR DYLAN WILIAM

Professor Dylan Wiliam is emeritus professor of educational assessment at University College London. In a varied career, he has taught in urban public schools, directed a large-scale testing program, served a number of roles in university administration, authored numerous books, and pursued a research program focused on supporting teachers to develop their use of assessment in support of learning. As one of the United Kingdom's leading experts on assessment, Dylan has an extensive history of research and consultation in this area.

A message from Hawker Brownlow Education

We hope that you have found these conference papers and the accompanying sessions useful. Please be aware that the contents of these papers are the intellectual property of the speaker and no reproduction for any purpose is authorised. We urge you to take care of this booklet. Replacement copies will not be made available either during or after this conference.

Published in Australia by

Hawker Brownlow EDUCATION

Hawker Brownlow Education
P.O. Box 580, Moorabbin, Victoria 3189, Australia
Phone: (03) 8558 2444 Fax: (03) 8558 2400
Toll Free Ph: 1800 334 603 Fax: 1800 150 445
Website: www.hbe.com.au
Email: orders@hbe.com.au

Printed in Australia

CODE: MELDWM0303
0517

Assessment Literacy: The Meaning and Consequences of Educational Assessments

Dylan Wiliam (@dylanwiliam)

Quality in assessment

What is an assessment?

3

- An assessment is a process for making inferences
- Key question: "Once you know the assessment outcome, what do you know?"
- Evolution of the idea of validity
  - A property of a test
  - A property of students' results on a test
  - A property of the inferences drawn on the basis of test results

Validity

4

- For any test:
  - some inferences are warranted
  - some are not
- "One validates not a test but *an interpretation of data arising from a specified procedure*" (Cronbach, 1971; emphasis in original)
- Consequences
  - No such thing as a valid (or indeed invalid) assessment
  - No such thing as a biased assessment

Threats to validity

5

- Construct-irrelevant variance
  - Systematic: good performance on the assessment requires abilities not related to the construct of interest
  - Random: good performance is related to chance factors, such as luck (effectively poor reliability)
- Construct under-representation
  - Good performance on the assessment can be achieved without demonstrating all aspects of the construct of interest

Validity revisited

- "Validity is an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." (Messick, 1989 p. 13)
- Social consequences:
  - "Right concern, wrong concept" (Popham, 1997)
  - No such thing as "consequential validity"

Understanding reliability

Understanding test scores

- Consider a test of students' ability to spell words drawn from a bank of 100 words.
- What we can conclude depends on:
  - The size of the sample
  - The way the sample was drawn
  - Students' knowledge of the sample
  - The amount of notice given

Reliability and sample size

9

- What can you conclude about a student who:
  - correctly spelled 1 out of 2 words
  - correctly spelled 5 out of 10 words
  - correctly spelled 10 out of 20 words
  - correctly spelled 50 out of 100 words?
- If you're sampling, conclusions about the unsampled items will be subject to error
- Assessment literacy requires knowing how big the error is

The standard error of measurement

- The "standard error of measurement" (SEM) is just the standard deviation of the errors, so, on any given testing occasion
  - 68% of students score within 1 SEM of their "true score"
  - 96% of students score within 2 SEM of their "true score"

Reliability of test scores

11

- Basic assumption of classical test theory:

$$X = T + E$$

  - In other words, the score a student gets on any occasion is their "true score" plus or minus some error
  - This does *not* mean assuming that ability is fixed
  - The "true score" is just the long-run average score over numerous testing occasions with similar tests, so
    - The average error is zero
    - For a reliable test, the errors are bunched closely
    - For an unreliable test, the errors are spread out
    - The more spread out the errors, the less reliable the test

Defining reliability

12

- To find out how spread out the errors are, we can't use the average (because it's zero!), so
- We square the errors (to get rid of the minus signs) and *then* average (this is called the *variance*)
- The reliability of a test is then defined as:

$$r = 1 - \frac{variance\ of\ errors}{variance\ of\ observed\ scores}$$

- When the variance of the errors:
  - is zero, the test is perfectly reliable (reliability = 1)
  - equals that of the observed scores, the test score is just noise (reliability = 0)

The standard error of measurement

13

- The "standard error of measurement" (SEM) is just the standard deviation of the errors, so, on any given testing occasion
  - 68% of students score within 1 SEM of their true score
  - 96% of students score within 2 SEM of their true score
- Re-arranging the defining equation for reliability:

$$SEM = Score\ SD \times \sqrt{1 - r}$$

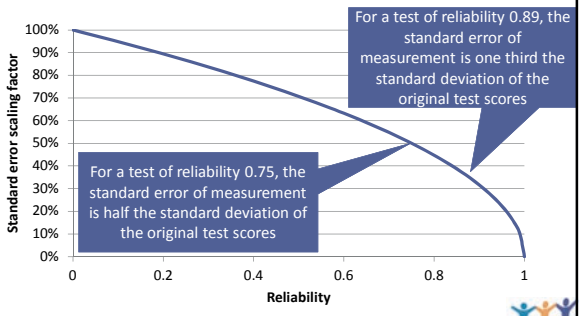
Relationship of reliability and error

For a typical test (average score 50, standard deviation 15), a student who should have scored 50 will actually score:

Reliability	SEM	Two-thirds of the time (68%)	Almost always (96%)
0.70	8.2	42 to 58	34 to 66
0.75	7.5	43 to 58	35 to 65
0.80	6.7	43 to 57	37 to 63
0.85	5.8	44 to 56	38 to 62
0.90	4.7	45 to 55	41 to 59
0.95	3.4	47 to 53	43 to 57

Reliability and standard error

15

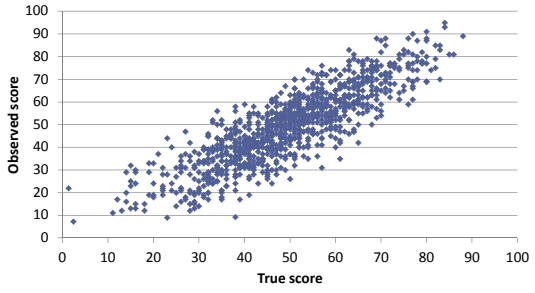


Sources of unreliability

- Rater variability
  - e.g., mark—re-mark
- Student variability
  - e.g., good days and bad days
- Student x item variability
  - e.g., effects of item choice or student choice

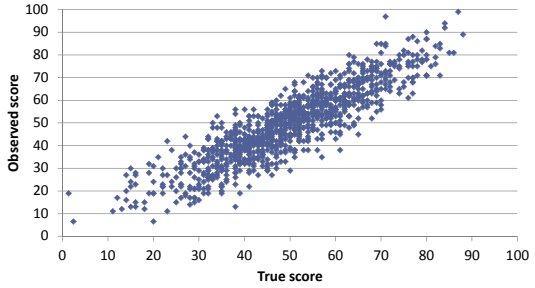
Reliability: 0.75

17



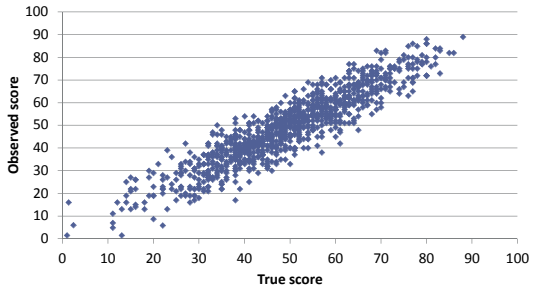
Reliability: 0.80

18



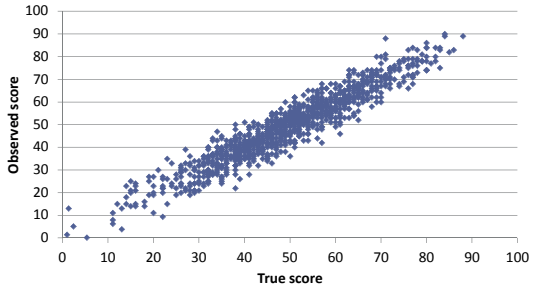
Reliability: 0.85

19



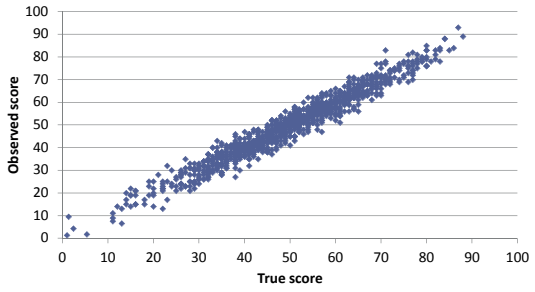
Reliability: 0.90

20



Reliability: 0.95

21



Reliability

- No measurement is perfectly reliable
- Educational measurements are no exception
  - With an 85% reliable test, in a group of 30:
    - Half the students will get a score within 4% of their 'true score'
    - Three-quarters of the students will get a score within 7% of their true score
    - One student will get a score differing from their true score by more than 12%
- And the only way to make a test more reliable is:
  - to make it longer, or
  - to make it narrower (same thing really)

Teacher assessment is essential

23

- The only way to improve the reliability of a test is to make it longer:
  - Increase testing time
  - Use information from teachers
- Teachers' involvement is not optional but essential
- However, teacher assessment brings problems of its own
  - Standardization
  - Random variation
  - Bias

Discussion question

24

Discussion

- If you had to ask one question now, what would it be?

Understanding what this means in practice

Grouping students by ability

Using tests for grouping students by ability

Using a test with a reliability of 0.9, and with a predictive validity of 0.7, to group 100 students into four ability groups or "sets":

		should be in set			
		set 1	set 2	set 3	set 4
students placed in	set 1	23	9	3	
	set 2	9	12	6	3
	set 3	3	6	7	4
	set 4		3	4	8

Only 50% of the students are in the "right" set

Diagnostic testing

The limits of diagnostic testing

29

- 120-item multiple choice test for teacher licensure
  - Four major subject areas
    - language arts/reading
    - mathematics
    - social studies
    - science
  - 30 items per subject area
  - Sub-score reliabilities range from 0.71 to 0.83

How reliable are 10-item subtest scores?

30

- Items for each subject area ranked in order of difficulty (i.e., 1 to 30)
- Three parallel 10-item forms created in each subject area:
  - Form A: items 1, 4, 7, … 28
  - Form B: items 2, 5, 8, … 29
  - Form C: items 3, 6, 9, … 30
- Sub-score reliabilities in the range 0.40 to 0.60
- On form A, 271 examinees scored 7 in mathematics and 3 in science

Scores of 271 students on form B

Math subscore \ Science subscore	1	2	3	4	5	6	7	8	9	10
1	0	0	0	1	1	1	0	0	0	0
2	0	0	0	1	3	1	2	0	0	0
3	1	0	0	1	2	4	3	1	1	1
4	0	0	2	7	7	6	4	0	1	0
5	0	1	1	4	10	14	8	5	1	1
6	0	0	0	3	0	11	15	8	1	1
7	0	0	0	2	4	11	10	7	4	0
8	0	0	0	1	2	13	7	5	4	0
9	0	0	0	1	6	3	7	4	3	0
10	0	0	0	1	1	2	1	1	0	0

110 out of 271 (41%) examinees got a better form B score in science than mathematics

Sinharay, Gautam and Halberman (2010)

32

- A student scoring 7 on mathematics and 3 on science would probably want to improve the latter
- But 110 of the 271 examinees got a better score in science than mathematics on Form B
- Correlation of science subscores on Forms A and B is 0.48
- Correlation of science subscore on Form A with total score on Form B is 0.63
- In other words, the total score on a test is a better guide to the score on a sub-test than another score on the same sub-test

Measuring progress

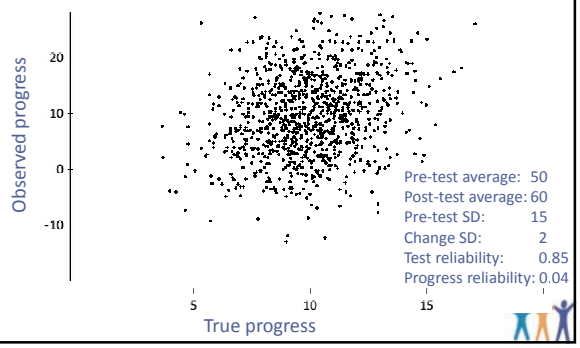
Reliability, standard errors, and progress

34

Grade	Reliability	SEM as a percentage of annual progress
1	0.89	26%
2	0.85	56%
3	0.82	76%
4	0.83	39%
5	0.83	55%
6	0.89	46%
Average	0.85	49%

In other words, the standard error of measurement of this reading test is equal to six months' progress by a typical student

True and observed growth scores



Fortunately…

36

- While progress measures for individuals are rather unreliable, progress measures for groups are much more reliable.
- The standard error for the average score of a group of individuals is the standard error for individuals, divided by the square root of the group size, so
  - if the standard error of individual progress is 10 marks
  - the standard error for the average progress of a class of 25 is just 2 marks

If you must measure progress…

37

- As rules of thumb:
  - For individual students, progress measures are meaningful only if the progress is more than *twice* the standard error of measurement of the test being used to measure progress
  - For a class of 25 students, progress measures are meaningful if the progress is more than *half* the standard error of measurement of the test being used to measure progress.

Discussion question

38

Discussion

- From what you have heard so far, what are the key challenges regarding the design of an assessment system for your school?

Meanings and consequences of assessment

- Evidential basis
  - What does the assessment result mean?
- Consequential basis
  - What does the assessment result do?
- Assessment literacy (Stiggins, 1991)
  - Do you know what this assessment result means?
  - Does it have utility for its intended use?
  - What message does this assessment send to students (and other stakeholders) about the achievement outcomes we value?
  - What is likely to be the effect of this assessment on students?

Recommendations: recording and reporting

40

- Records should be kept at the finest manageable level
- Where profiles of scores are aggregated this should be to a relatively fine scale
- When assessment outcomes are reported, they are always accompanied by a margin of error, such as:
  - Standard error (SEM)
  - Probable error (0.675 x SEM)

Discussion question

41

Discussion

- Pulling together what you have heard today, what are the most important principles for the design of an assessment *system* for your school?

Force-field analysis (Lewin, 1954)

42

What are the forces that will support or drive the adoption of good assessment practices in your school/authority?	What are the forces that will constrain or prevent the adoption of good assessment practices in your school/authority?
+	—

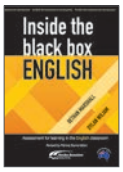
RELATED RESOURCES

Hawker Brownlow

Available from Hawker Brownlow Education

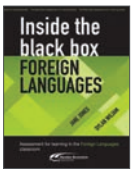
Qty	Code	Title	Price
	SOT8281	Ahead of the Curve: The Power of Assessment to Transform Teaching & Learning	$50.00
	SOT2446	Content, Then Process DVD	$150.00
	SOT8112	Embedded Formative Assessment	$35.95
	SAT8899	Embedding Formative Assessment Professional Development Pack	$544.50
	LSM0546	Embedding Formative Assessment Quick Reference Guide	$15.00
	LSM4971	Embedding Formative Assessment: Practical Techniques for F–12 Classrooms	$35.95
	GLA1284	Inside The Black Box	$10.95
	GLA1369	Inside The Black Box of Assessment	$10.95
	GLA1280	Inside the Black Box Series Set of 11	$110.00
	GLA1383	Inside The Black Box: Design and Digital Technologies	$10.95
	GLA1314	Inside The Black Box: English	$10.95
	GLA1376	Inside The Black Box: Foreign Languages	$10.95
	GLA1345	Inside The Black Box: Geography	$10.95
	GLA1352	Inside The Black Box: ICT	$10.95
	GLA1321	Inside The Black Box: Maths	$10.95
	GLA1307	Inside The Black Box: Primary Years	$10.95
	GLA1338	Inside The Black Box: Science	$10.95
	LSM8306	Leadership for Teacher Learning: Creating a Culture Where All Teachers Improve so That All Students Succeed	$39.95
	SAT5085	Redesigning Schooling Series Complete Set	$90.00
	SAT5190	Redesigning Schooling: Principled assessment design	$15.95
	SAT5107	Redesigning Schooling: Principled curriculum design	$15.95
	GLA1291	Working Inside The Black Box	$10.95
		Total (plus freight) $	

GLA1291

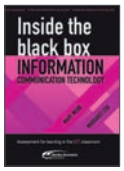


GLA1314

GLA1369



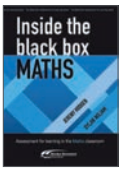
GLA1376



GLA1352

GLA1307

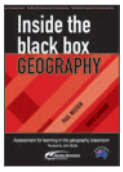
GLA1338



GLA1321

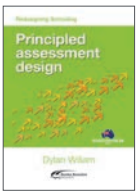
GLA1284

GLA1383

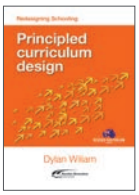


GLA1345

GLA1280

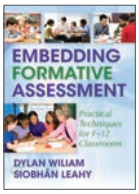


SAT5190

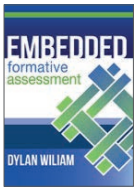


SAT5107

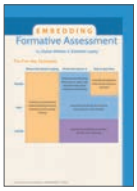
SAT5085



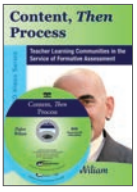
LSM4971



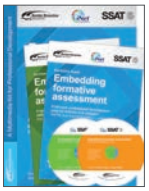
SOT8112



LSM0546

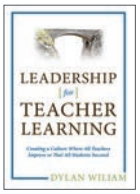


SOT2446



SAT889

SOT8281



LSM8306

Attention ..... Order Number .....

Name of School .....

Address .....

..... State .....P/Code .....

Country .....

Email: .....

☐ Yes, I would like to receive emails from Hawker Brownlow Education about future workshops, conferences and the latest publications.

Terms of Trade

- Prices are quoted in Australian dollars ($AUD) and include GST
- All prices are subject to change without notice.
- For New Zealand customers, at the time of invoice, we will convert the amount into New Zealand dollars ($NZD) so that you can pay by cheque or credit card in New Zealand dollars ($NZD).
- Full money-back guarantee.
- We do realise it is difficult to order sight unseen. To assist you in your selection, please visit our website <www.hbe.com.au>. Go to 'Browse Books' and most titles will give you the option to view the first few pages of the book. Click 'View Contents' on your selected book page.
- We will supply our books on approval, and if they do not suit your requirements we will accept undamaged returns for full credit or refund. Posters are for firm sale only and will not be sent on approval. Please be aware that delivery and return postage is the responsibility of the customer.
- Freight costs are determined at Australia Post rates, with a minimum delivery charge of $9.50 within Australia and $15.00 for New Zealand for each order.
- Please provide your street address for delivery purposes.

To place an order, request a catalogue or find out more about our resources:

Call
1800 334 603
(03) 8558 2444

Fax
1800 150 445
(03) 8558 2400

Online
www.hbe.com.au

Mail
Hawker Brownlow Education
PO Box 580,
Moorabbin, VIC 3189

Do you want to know all about the latest professional development events in your area? Be the first to find out about new releases from world-renowned and local authors with the HBE e-newsletter! Upcoming titles will feature authentic assessment and digital media, along with a strong focus on success in mathematics and literacy. Sign up to our FREE e-newsletter at www.hbe.com.au.

Online 'On Account' ordering now available!
If you have a pre-existing account with Hawker Brownlow Education, you can now order online and pay using that account.