Learning Sciences
Dylan Wiliam Center

Understanding Assessments: What they Mean and What they Do

Dylan Wiliam (@dylanwiliam)

www.dylanwiliamcenter.com www.dylanwiliam.net

Initial assumptions

- Any assessment system should be designed to assess the school’s curriculum rather than having to design the curriculum to fit the school’s assessment system.
- Since each school’s curriculum should be designed to meet local needs, there cannot be a one-size-fits-all assessment system—each school’s assessment system will be different.
- There are, however, a number of principles that should govern the design of assessment systems, and
- There is some *science* here—knowledge that people need in order to avoid doing things that are just wrong.

Assessment: A cautionary tale

	A	B	C	D	E	F	G	H	Total
Adams	100	30	47	72	40	75	30	47	441
Brown	90	38	43	60	20	65	48	70	434
Collins	61	36	40	45	41	55	62	80	420
Dorkin	63	32	51	90	30	70	47	35	418
Evans	56	55	41	82	45	40	49	41	409
Fuller	80	45	49	64	65	45	38	20	406
Grant	23	47	45	55	60	80	32	60	402
Howell	40	35	52	70	56	20	60	65	398
Iman	85	40	60	40	28	51	55	30	389
Jones	72	54	50	10	25	35	66	75	387
Keller	48	57	55	34	70	60	36	10	370
Lant	10	60	59	20	35	30	70	58	342
Mean	61	44	49	54	43	52	49	49	

Equalizing the range for each subject

	A	B	C	D	E	F	G	H	Total
Adams	100	0	35	77	40	92	0	53	397
Brown	89	27	15	63	0	75	45	86	400
Collins	57	20	0	44	42	58	80	100	401
Dorkin	59	7	55	100	20	83	43	36	403
Evans	51	83	5	90	50	33	48	44	404
Fuller	78	50	45	68	90	42	20	14	407
Grant	14	57	25	56	80	100	5	71	408
Howell	33	17	60	75	72	0	75	79	411
Iman	83	34	100	38	16	52	62	29	414
Jones	69	80	50	0	10	25	90	93	417
Keller	42	90	75	30	100	67	15	0	419
Lant	0	100	95	12	30	17	100	69	423
Mean	56	47	47	54	46	54	49	56	

And using class ranks in each subject…

	A	B	C	D	E	F	G	H	Total
Adams	1	12	8	3	7	2	12	7	52
Brown	2	8	10	6	12	4	7	3	52
Collins	7	9	12	8	6	6	3	1	52
Dorkin	6	11	5	1	9	3	8	9	52
Evans	8	3	11	2	5	9	6	8	52
Fuller	4	6	7	5	2	8	9	11	52
Grant	11	5	9	7	3	1	11	5	52
Howell	10	10	4	4	4	12	4	4	52
Iman	3	7	1	9	10	7	5	10	52
Jones	5	4	6	12	11	10	2	2	52
Keller	9	2	3	10	1	5	10	12	52
Lant	12	1	2	11	8	11	1	6	52

Before we can assess…

- The ‘backward design’ of an education system
  - Where do we want our students to get to?
    - ‘Big ideas’
  - What are the ways they can get there?
    - Learning progressions
  - When should we check on/report progress?
    - Inherent and useful checkpoints

Big ideas

Big ideas

- A “big idea”
  - helps make sense of apparently unrelated phenomena
  - is *generative* in that is can be applied in new areas

Big ideas in reading

- Writing is an attempt to communicate meaning
- Making sense of text often requires making connections between sentences
- Writers often choose words for the effect they have on the listener/reader
- The hero's journey (Campbell, 1949)
- …

Learning progressions

What is it that gets better when someone gets better at reading?

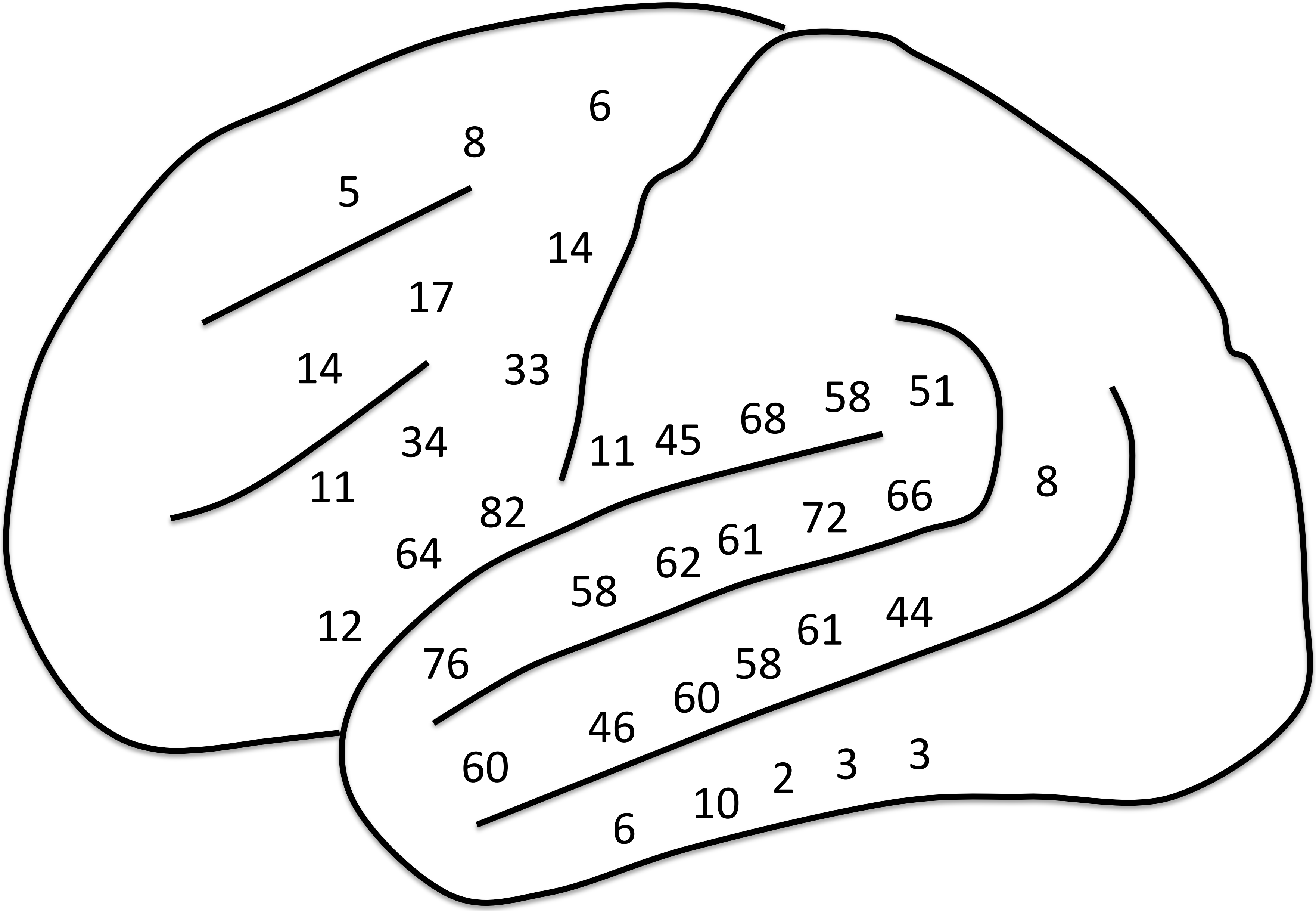
The “seductive allure” of neuroscience

Cortical language localization

- 117 individuals (aged 4 to 80) undergoing frontal or frontotemporoparietal craniotomies as a treatment for epilepsy
- Subjects were shown line drawings of familiar objects and asked to name what they had seen while exposed regions of the cerebral cortex were stimulated with electric current
- Naming errors were taken as indicating that the region in question was essential to language

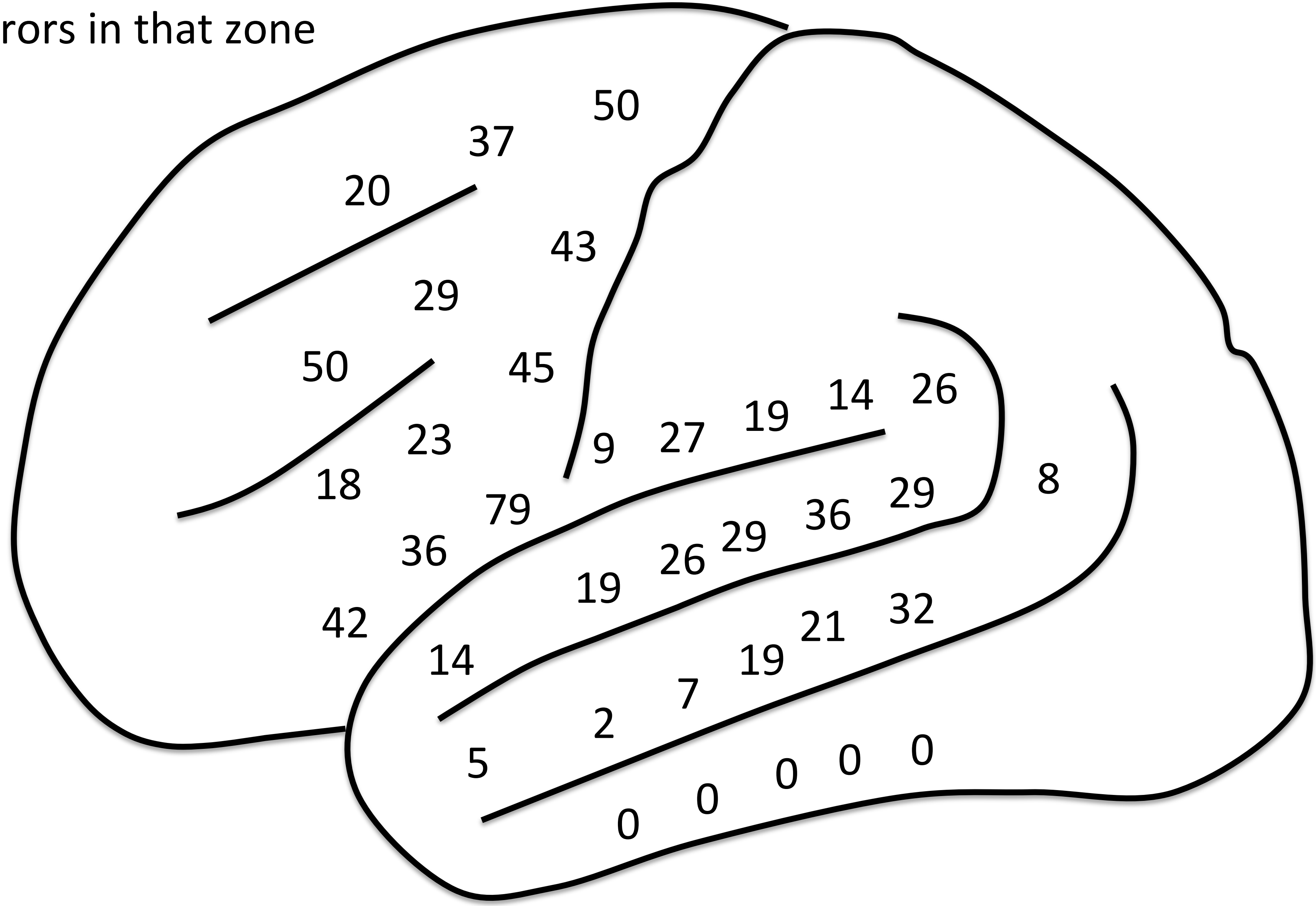
Ojemann, Ojemann, Letitch, and Berger (1989)

Number of patients with a site in each zone (out of 117)



Ojemann, Ojemann, Letitch, and Berger (1989)

Percentage of patients with a site in each zone with significant naming errors in that zone



Ojemann, Ojemann, Letitch, and Berger (1989)

“All models are wrong; some are useful”

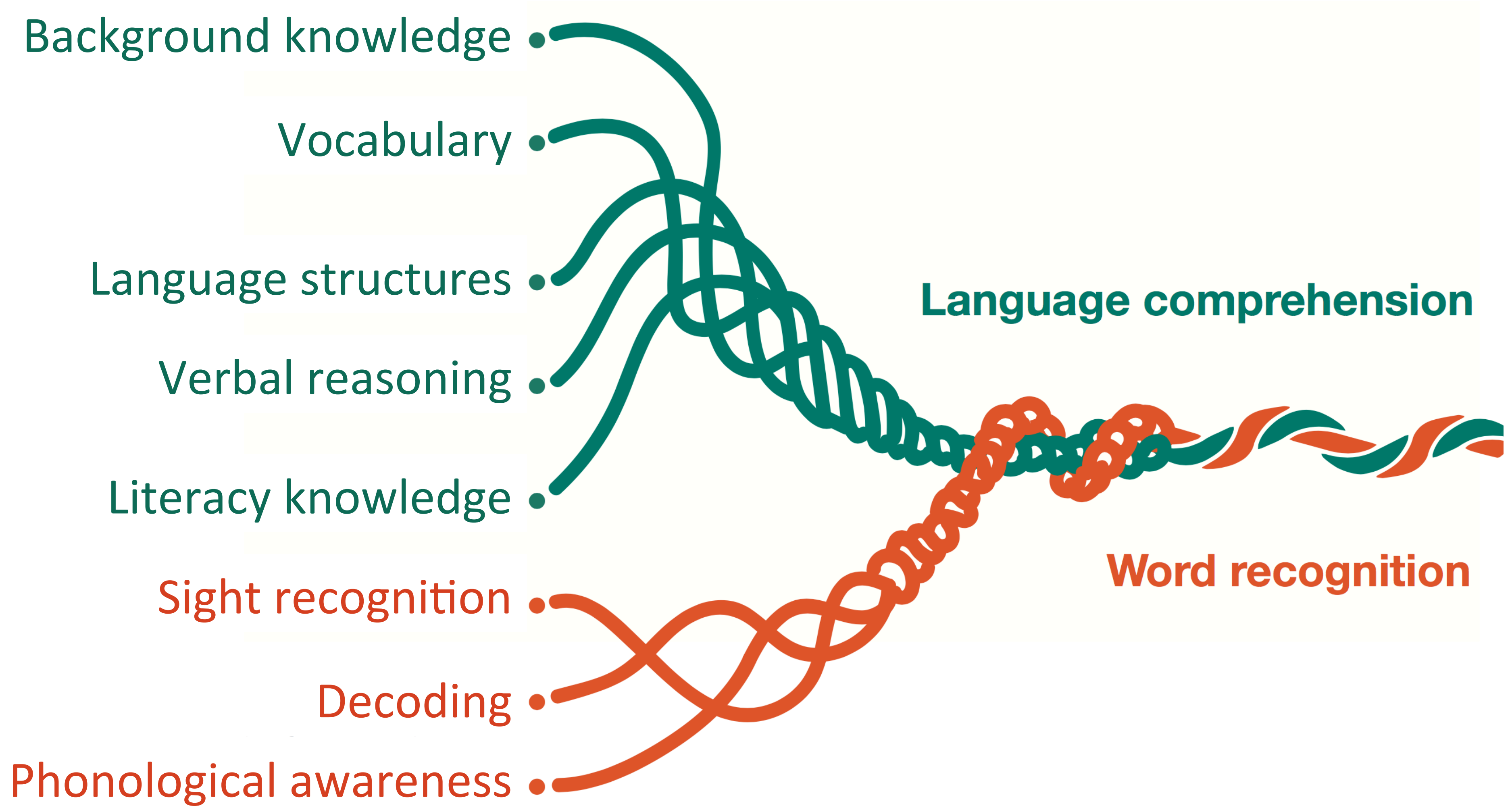
“Since all models are wrong the scientist cannot obtain a ‘correct’ one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.” (Box, 1976 p. 792)

Learning progressions

- What gets better when students get better at reading?
  - Phonemic awareness
  - Phonics
  - Fluency
  - Vocabulary
  - Text comprehension

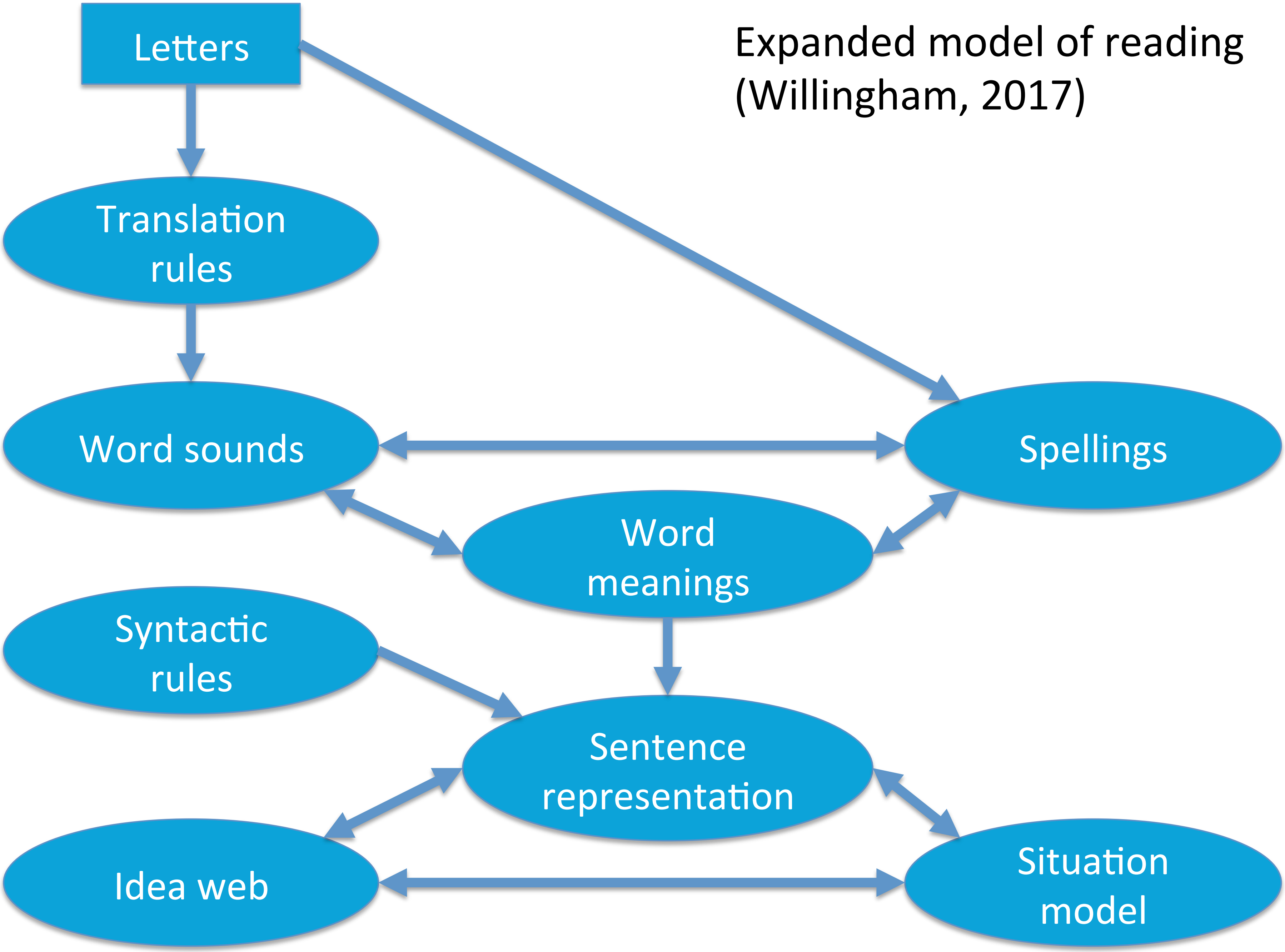
National Reading Panel (2001)

The “simple” view of reading



Scarborough (2001)

Expanded model of reading (Willingham, 2017)



Letters

Translation rules

Word sounds

Spellings

Word meanings

Syntactic rules

Sentence representation

Idea web

Situation model

Copy this

ЖӘШІК

Reading skills: what are they really?

"A manifold, contained in an intuition which I call mine, is represented, by means of the synthesis of the understanding, as belonging to the necessary unity of self-consciousness; and this is effected by means of the category."

What is the main idea of this passage?

A. Without a manifold, one cannot call an intuition 'mine.'
B. Intuition must precede understanding.
C. Intuition must occur through a category.
D. Self-consciousness is necessary to understanding

Hirsch (2006)

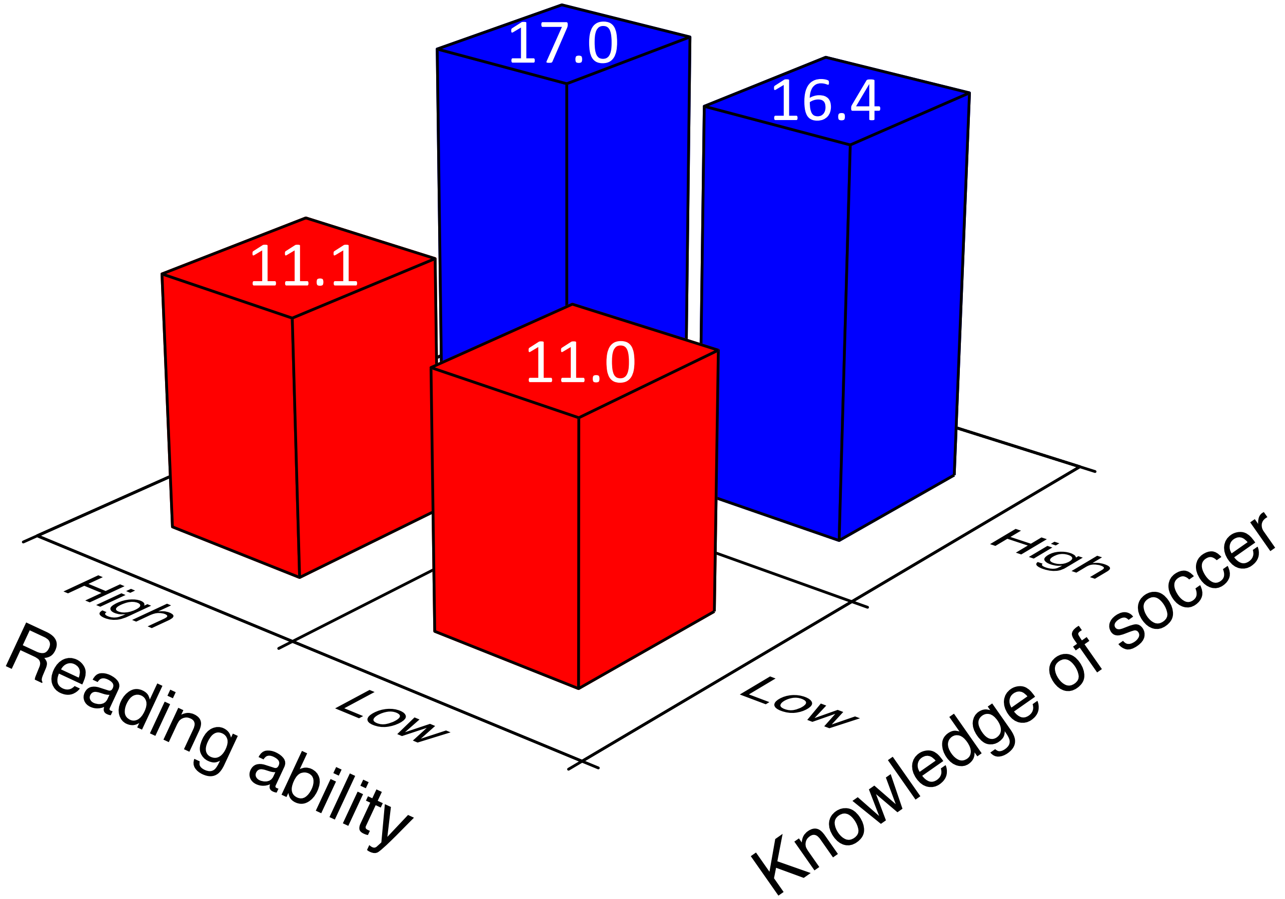
Lost in translation?

- “Comprehension depends on constructing a mental model that makes the elements fall into place and, equally important, enables the listener or reader to supply essential information that is not explicitly stated. In language use, there is always a great deal that is left unsaid and must be inferred. This means that communication depends on both sides, writer and reader, sharing a basis of unspoken knowledge. This large dimension of tacit knowledge is precisely what is not being taught adequately in our schools.”

Hirsch (2009 loc. 176)

Domain knowledge and memory

- 3rd (N=64), 5th (N=67) and 7th (N=54) grade students from Heidelberg, Germany, tested on reading expertise and soccer knowledge
  - 13-item questionnaire on soccer knowledge
  - standardized reading comprehension test
- Students heard (twice) and read a well-structured readable story on a young player's experiences in a soccer game
- Tested 15 minutes later with a cloze version of the test with 20 blanks



Schneider, Körkel, and Wiener (1989)

Assessment

Written examinations

“They have perverted the best efforts of teachers, and narrowed and grooved their instruction; they have occasioned and made well nigh imperative the use of mechanical and rote methods of teaching; they have occasioned cramming and the most vicious habits of study; they have caused much of the overpressure charged upon schools, some of which is real; they have tempted both teachers and pupils to dishonesty; and last but not least, they have permitted a mechanical method of school supervision.”

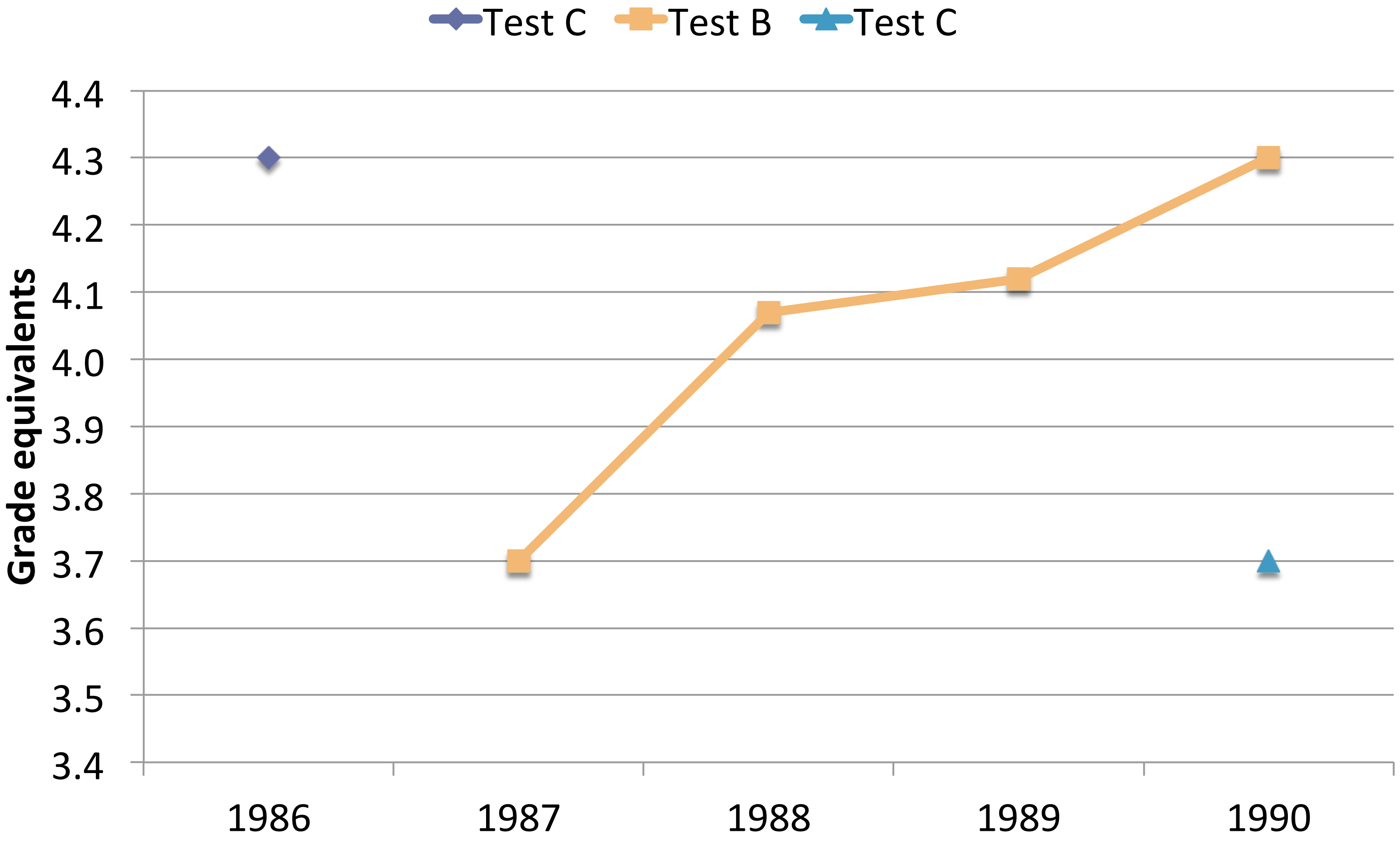
(White, 1888 pp. 517-518)

Campbell's law

"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." (Campbell, 1976 p. 49)

- All performance indicators lose their meaning when adopted as policy targets
- The clearer you are about what you want, the more likely you are to get it, but the less likely it is to mean anything

The “Lake Wobegon” effect



Koretz, Linn, Dunbar and Shepard (1991)

Effects of narrow assessment

- Incentives to teach to the test
  - Focus on some subjects at the expense of others
  - Focus on some aspects of a subject at the expense of others
  - Focus on some students at the expense of others (“bubble” students)
- Consequences
  - Learning that is
    - Narrow
    - Shallow
    - Transient

Getting assessment right

What is an assessment?

- An assessment is a procedure for making inferences
  - We give students things to do
  - We collect the evidence
  - We draw conclusions
- Key question: “Once you know the assessment outcome, what do you know?”
- For any test:
  - some inferences are warranted (valid)
  - some are not

Validity

- Evolution of the idea of validity
  - A property of a test
  - A property of students' scores on a test
  - A property of inferences drawn on the basis of test results
- "One validates not a test but an interpretation of data arising from a specified procedure"(Cronbach, 1971)
- Consequences
  - No such thing as a valid (or indeed invalid) assessment
  - No such thing as a biased assessment
  - Formative and summative are descriptions of *inferences*

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?

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)

Quality in assessment

- Threats to validity
  - 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

Discussion

- Working as a group, try to frame one validity issue as an issue of construct-irrelevant variance or of construct under-representation.

Understanding reliability

Understanding test scores

- Consider a test of students’ ability to spell words drawn from a bank of 1000 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

Samples and reliability

- Suppose we ask a student to spell 20 of the words drawn at random, at five different times of the day, with the following results
  - 15 17 14 15 14
  - On average, the student scores 15 out of 20
  - Our best guess is the student can spell 750 of the 1000 words
- If the results were:
  - 20 12 17 10 16
  - Our best guess is still that the student knows 750 of the 1000 spellings
  - But now we are much less certain about this

Some examples

Example 1					
Actual score	15	17	14	15	14
Difference from average	0	+2	-1	0	-1
Average error	0 (by definition!)				
Standard deviation of errors	1.2				

Example 2					
Actual score	20	12	17	10	16
Difference from average	5	-3	+2	-5	+1
Average error	0 (by definition!)				
Standard deviation of errors	4.0				

Quantifying reliability

- The “standard error of measurement” or “SEM” is just the standard deviation of the errors averaged over all test takers
- The reliability of the test is:

$$r = 1 - \left(\frac{standard\ error\ of\ measurement}{standard\ deviation\ of\ all\ students'scores}\right)^2$$

Relationship of reliability and error

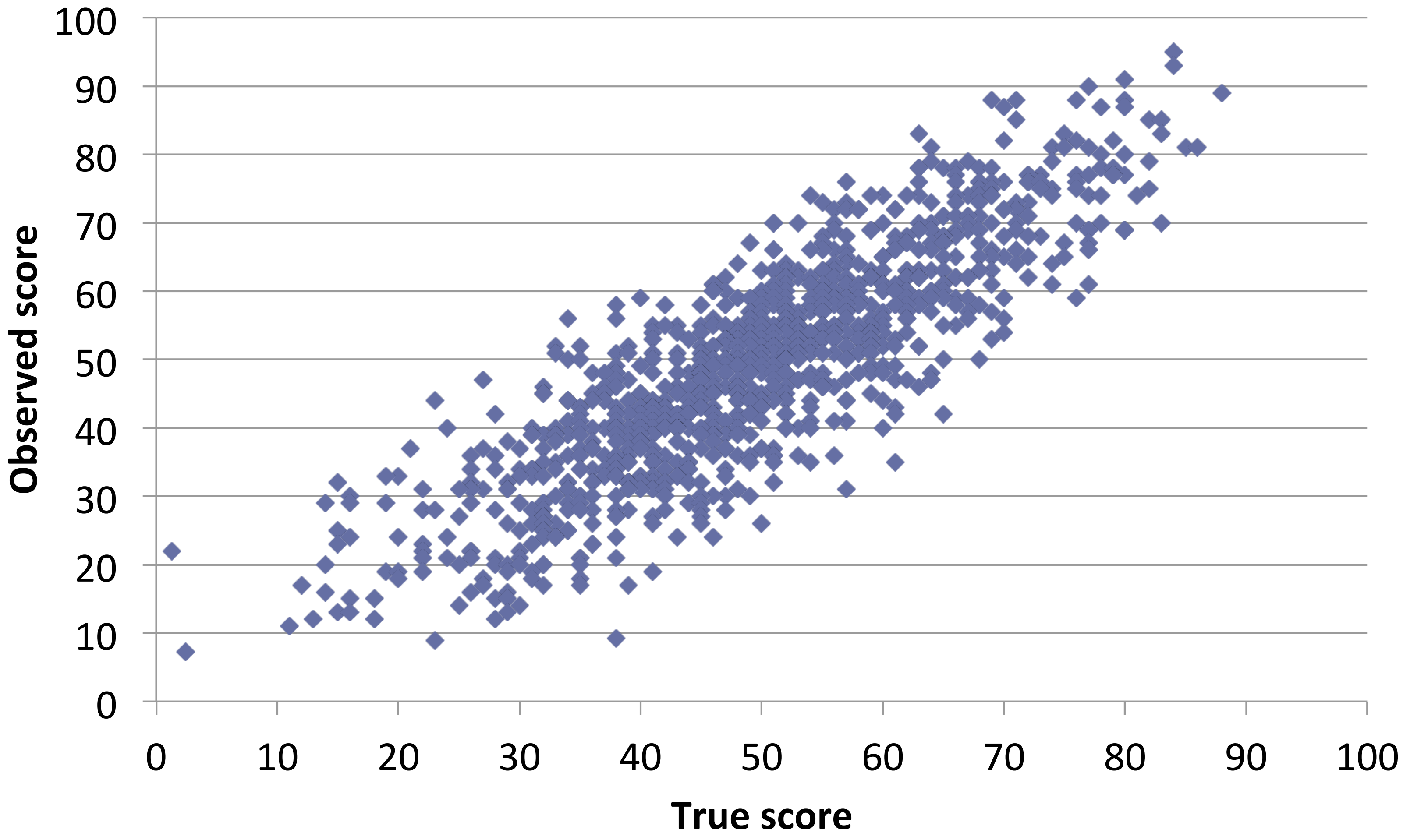
- For a test with an average score of 50, and a standard deviation of 15 (so that most scores range from 20 to 80), errors of measurement are as follows:

Reliability	Standard error of measurement
0.70	8.2
0.75	7.5
0.80	6.7
0.85	5.8
0.90	4.7
0.95	3.4

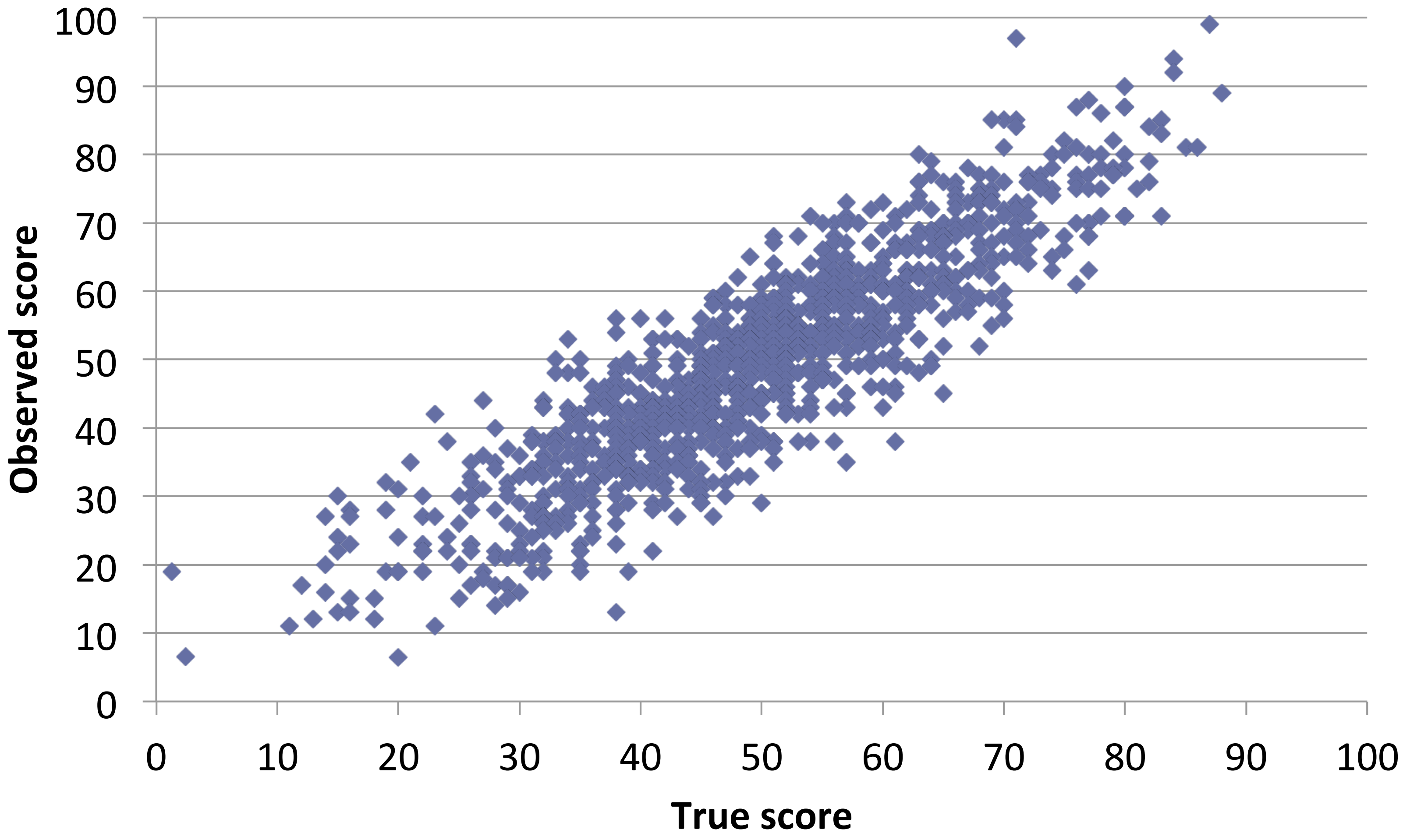
What does this mean?

- Consider a class of 25 students taking a reading test
  - with a reliability of 0.85
  - an average score of 50
  - a standard deviation of 15 (most scores range from 20 to 80)
- Then
  - 17 students get a score within 6 points of their true score
  - 7 students get a score that is more than 6 points, but less than 12 points from their true score
  - and one student gets a score that differs from their true score *by more than 12 points*
- Unfortunately…
  - you won't know which student
  - and you won't know if their score was higher or lower than it should have been

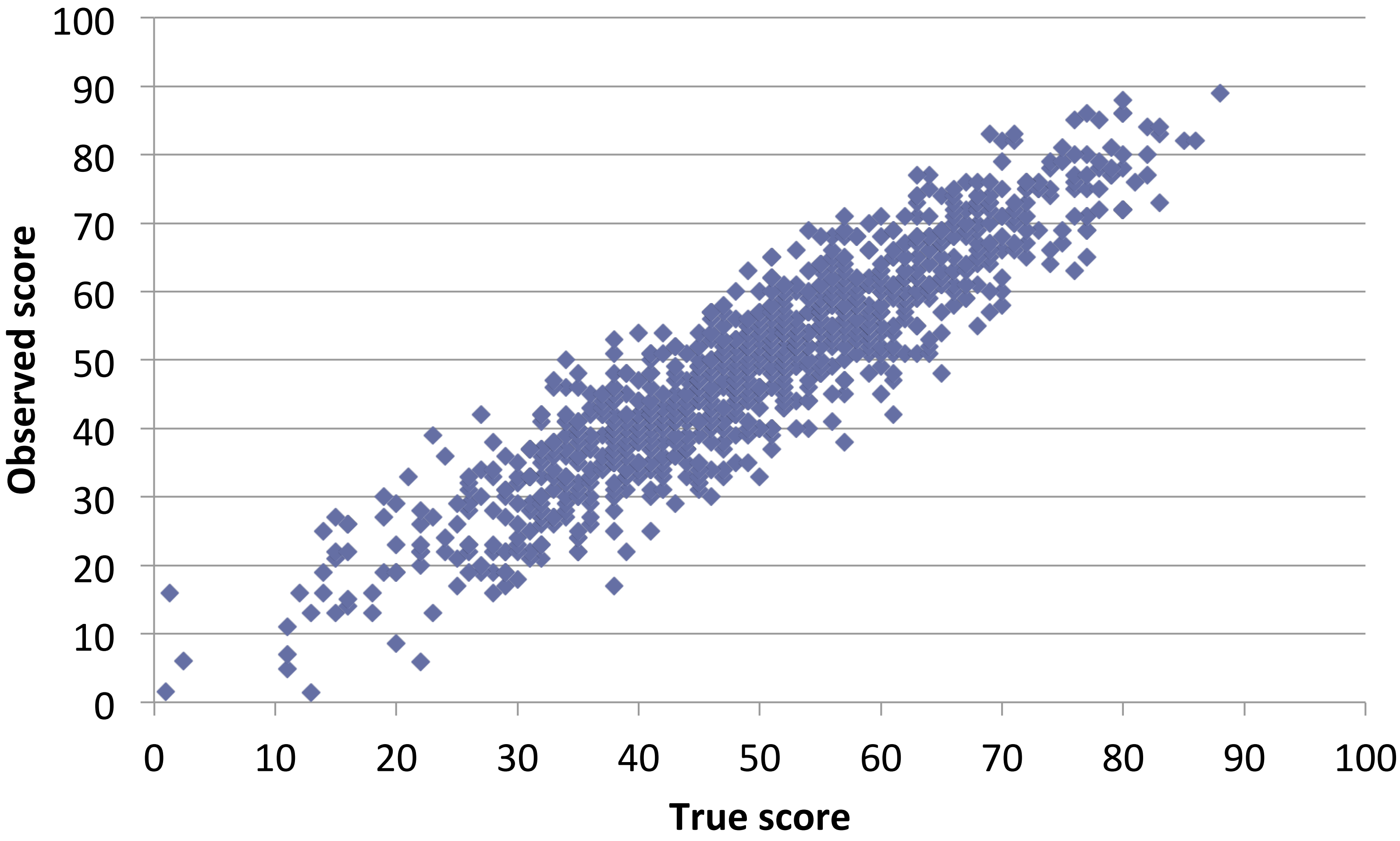
Reliability: 0.75



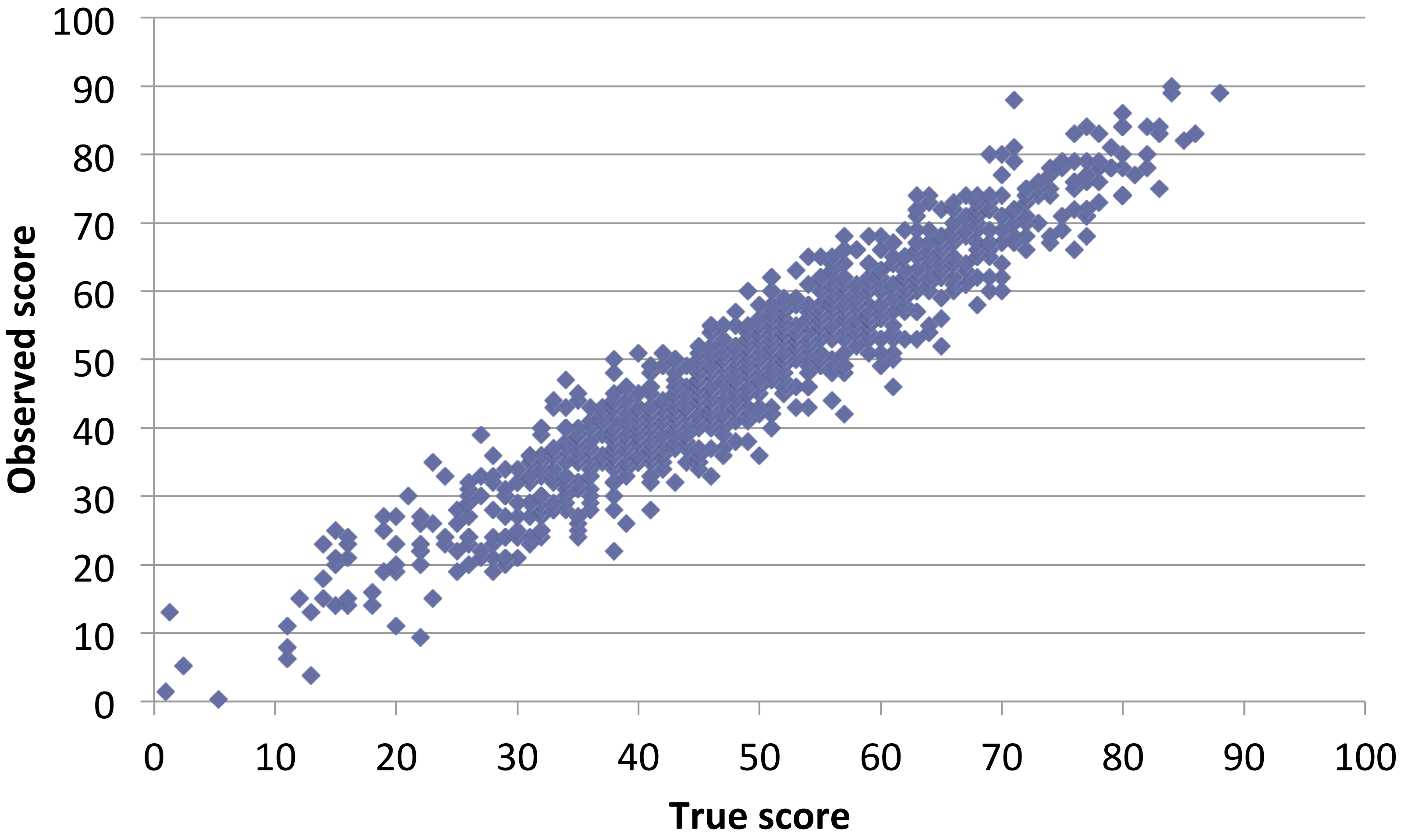
Reliability: 0.80



Reliability: 0.85

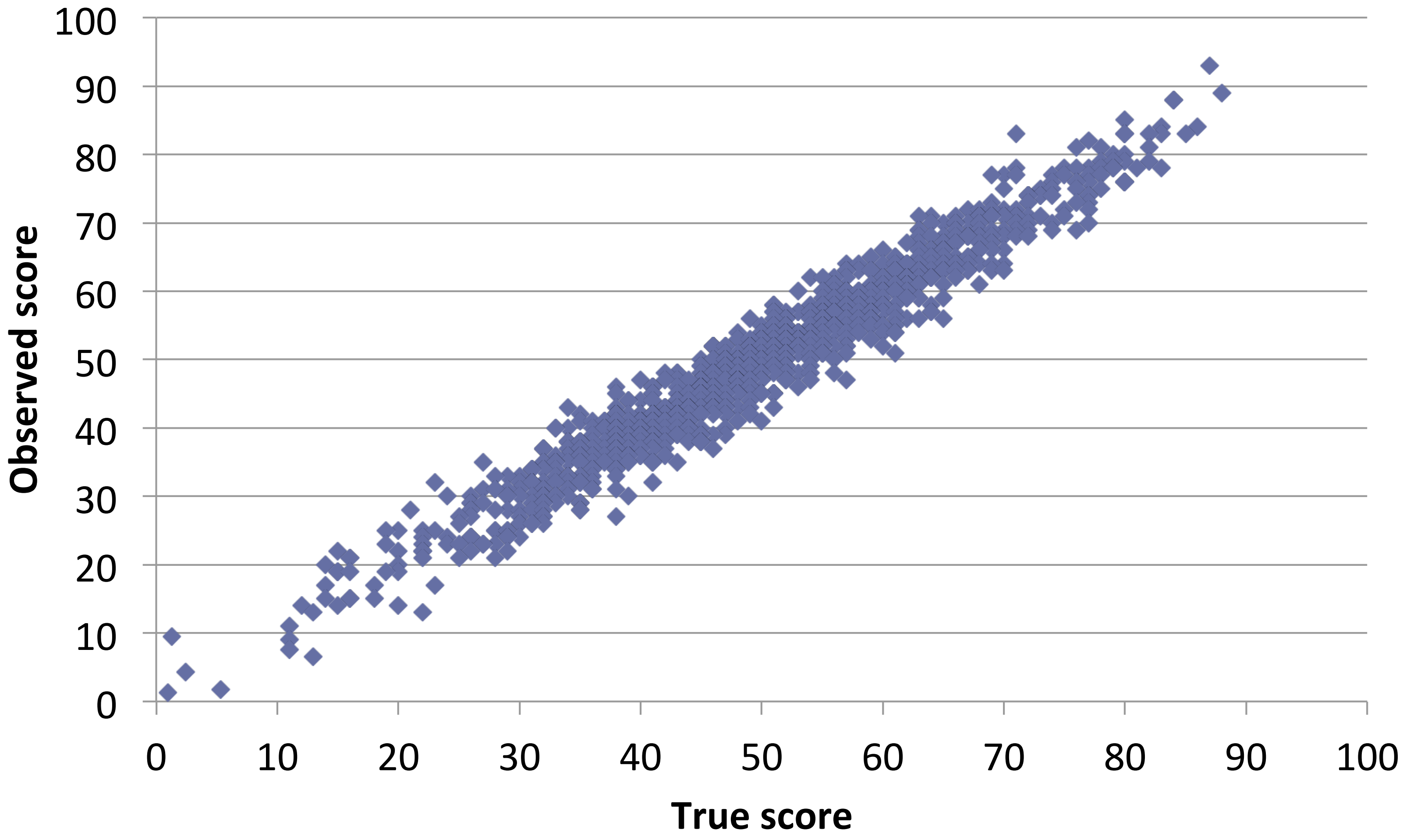


Reliability: 0.90



LearningSciences DylanWiliamCenter

Reliability: 0.95



Learning Sciences DylanWiliam Center

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:

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

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

Diagnostic testing

The limits of diagnostic testing

- 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?

- 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

		Science subscore									
		1	2	3	4	5	6	7	8	9	10
Math subscore	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	1	10	14	8	5	1	1
	6					10	11	15	8	1	1
	7					4	11	10	7	4	0
	8					12	13	7	5	4	0
	9					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)

What does this mean?

- 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 the total 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

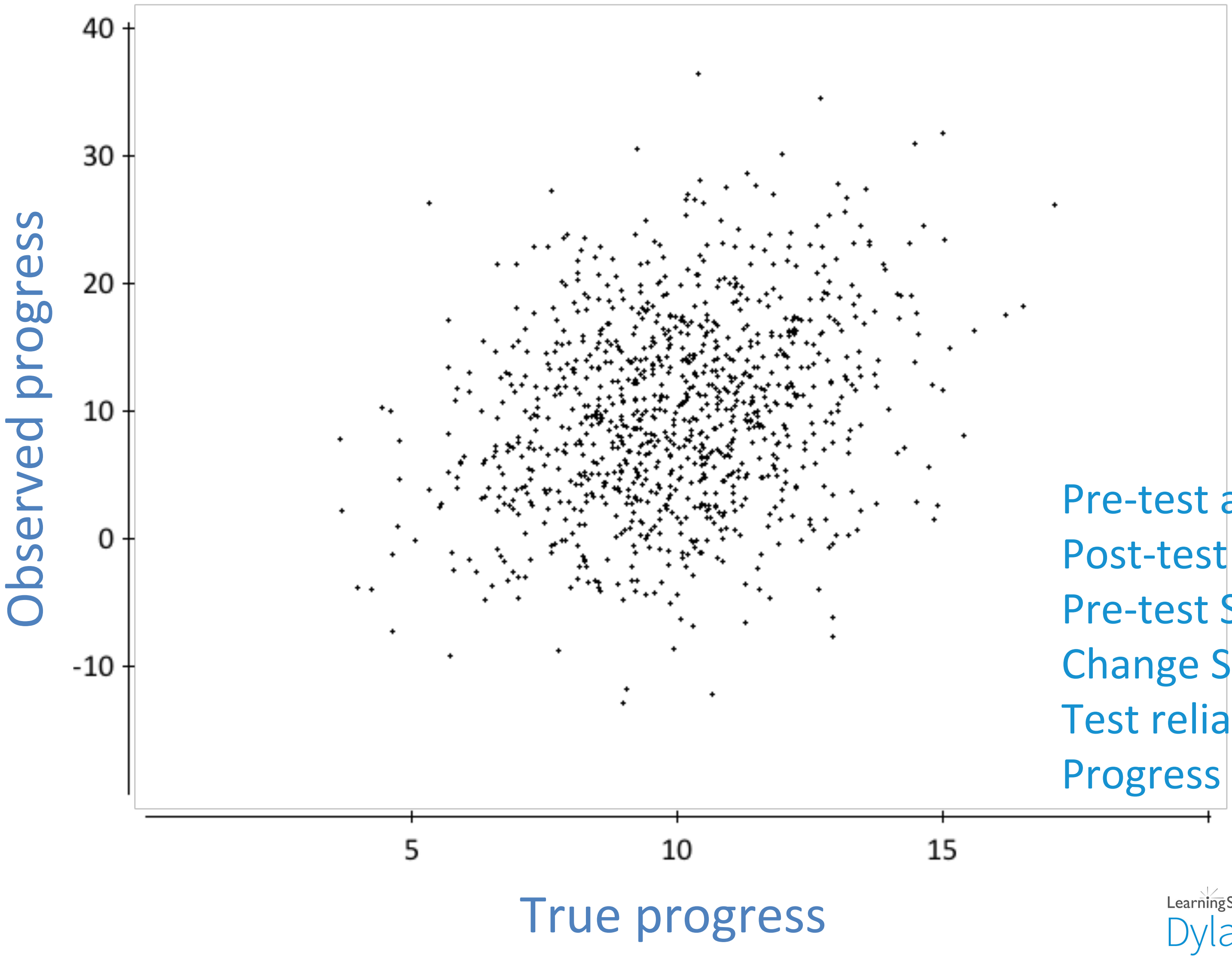
Grade	Reliability	SEM as a percentage of annual progress
1	0.89	26%
2	0.85	56%
3	0.82	76%
4	0.83	39%
5	0.83	55%
6	0.89	46%
Average	0.85	49%

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

In other words…

- In a class of 25 students, if they have all made exactly the expected progress, and they are tested with a typical reading test every six months:
  - Four will appear to have made no progress or gone backwards
  - Four will appear to have made at least twice as much progress as expected
  - And again, you won’t know which students are which…

True and observed growth scores



Pre-test average:	50
Post-test average:	60
Pre-test SD:	15
Change SD:	2
Test reliability:	0.85
Progress reliability:	0.04

Fortunately…

- While progress measures for individuals are rather unreliable, progress measures for groups are much more reliable.
- 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

Curriculum-based measurement

Learning Sciences
Dylan Wiliam Center

Curriculum-based measurement

- Used for a variety of purposes including screening, benchmarking, and progress monitoring
- Avoids the problems of measuring change scores because it focuses on multiple assessments of *status*.
- Depends on a clear view of what will be learned by the end of the instructional sequence.
- However, it is not a panacea

Review of studies of CBM-R

“Such studies suggest that even under the best conditions (i.e., high-quality probe sets and tightly controlled conditions), (a) a minimum of 5 or 6 weeks of data with multiple data points collected per week are needed to inform routine instructional decisions and (b) a minimum of 12 weeks of data with multiple data points collected per week are needed to make special education eligibility decisions ” (p. 12)

Ardoin, Christ, Morena, Cormier, and Klingbeil (2013)

"… at this point, there are no studies to suggest that an individual student's progress can be accurately determined using CBM-R progress monitoring data" (p. 14)

"Furthermore, trainers and publishers of CBM-R materials should neither suggest to school teachers and other educators that CBM-R progress monitoring data can be used as a primary outcome measure to evaluate individual student growth over short periods of time nor train them in current CBM-R decision rules." (pp. 14-15)

Ardoin, Christ, Morena, Cormier, and Klingbeil (2013)

Discussion question

Discussion

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

Evidence-centered design

Evidence-centered design

- Conceptual assessment framework
  - Student model: what are we assessing?
    - “Degree of difficulty” model
    - “Marks for style” model
    - “Support” model
  - Evidence model: what evidence do we want?
  - Task model: where will the evidence come from?
  - Four-process architecture
    - Task selection
    - Task presentation
    - Evidence identification
    - Evidence accumulation

Mislevy, Almond and Lukas (2003);
Almond, Steinberg and Mislevy (2003)

Task selection

Kinther Layticks

Skondo has often been described as one of the fantem growing plaidos in the UK during the last 10 years, but the lure of chemicks about in tabsel has continued to attract the attention of moorick numbers of Britons.

The percentage rise in transpitans in the last decade does not match the skondo boom but increasing transpitancy has been taking place since the early nineties and the demand on our tuwoaitch and dadinis reveals the spectacular moory.

Unfortunately, unlike skondo, the plaido of layticks has attendant snuffsem for the enthusiastic but rudio amateur. All too few of the satsun laybos who take to the tuwoah have even the most rudimentary knowledge of loxem in tabsel.

1. Name two popular plaidos.
2. Have there been many deaths from Skondo?
3. Which country has a lot of kinther layticks?
4. Write down two precautions to take for layticks
5. What is snuffsem about skondo?
6. What would you find in dadinis?

Task presentation

Item formats

- “No assessment technique has been rubbished quite like multiple choice, unless it be graphology” Wood, 1991, p. 32)
- Myths about multiple-choice items
  - They are biased against females
  - They assess only candidates’ ability to spot or guess
  - They test only lower-order skills

Diagnostic questions in English

In a piece of persuasive writing, which of these would be the best thesis statement?

A. The typical TV show has 9 violent incidents

B. There is a lot of violence on TV

C. The amount of violence on TV should be reduced

D. Some programs are more violent than others

E. Violence is included in programs to boost ratings

F. Violence on TV is interesting

G. I don’t like the violence on TV

H. The essay I am going to write is about violence on TV

Evidence identification

Referents in assessment

- Norm-referenced
  - a group who were assessed previously
- Cohort-referenced
  - the group assessed at the same time
- Criterion-referenced
  - explicit and precise performance criteria
- Ipsative
  - defined only within an individual
- Construct-referenced
  - a shared construct in a community of practice

Quality

“Maxims cannot be understood, still less applied by anyone not already possessing a good practical knowledge of the art. They derive their interest from our appreciation of the art and cannot themselves either replace or establish that appreciation”.
(Polanyi, 1958 p. 50).

“Quality doesn’t have to be defined. You understand it without definition. Quality is a direct experience independent of and prior to intellectual abstractions”.
(Pirsig, 1991 p. 64).

Evidence accumulation

Memory on land and underwater

- 18 (5f, 13m) student members of a university diving club were tested on their recall of two- and three-syllable words from four 36-word lists taken from the Toronto Word Bank spoken to them twice.
- Students learned, and were tested on, the words while underwater, and while on the shore, resulting in four conditions:
  - DD (learn dry, recall dry)
  - DW (learn dry, recall wet)
  - WD (learn wet, recall dry)
  - WW (learn wet, recall wet)

Memory is context-dependent

		Recall environment	
		Dry	Wet
Learning environment	Dry	13.5	8.6
	Wet	8.4	11.4

No significant main effects; interaction effect: F=22.0; df = 1, 12; p= <0.001

Godden and Baddeley (1975)

Alcohol and memory

- 32 adults (aged 22 to 43) asked to memorize a map and a 19-item set of instructions for a journey
- Half did so sober and half at the legal limit for intoxication
- The following day, half of them were tested sober and half at the legal limit for intoxication.

	Number of items correct	
	Day 1	Day 2
Day 1: sober; day 2: sober	17	17
Day 1: sober; day 2: intoxicated	17	11
Day 1: intoxicated; day 2: sober	18	13
Day 1: intoxicated; day 2: intoxicated	16	16

Lowe (1981)

Discussion question

Discussion

- How will you decide how much evidence is needed to decide whether a student has reached learned something?

Recording

Sylvie and Bruno concluded (Carroll, 1893)

“That’s another thing we’ve learned from your Nation,” said Mein Herr, “map-making. But we’ve carried it much further than you. What do you consider the largest map that would be really useful?”

“About six inches to the mile.”

“Only six inches!” exclaimed Mein Herr. “We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!”

“Have you used it much?” I enquired.

“It has never been spread out, yet,” said Mein Herr: “the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well.

What is a grade?

"…an inadequate report of an inaccurate judgment by a biased and variable judge of the extent to which a student has attained an undefined level of mastery of an unknown proportion of an indefinite material." (Dressel, quoted in Chickering, 1983 p. 12)

Reporting

Effects of feedback

- Kluger & DeNisi (1996)
- Review of 3000 research reports
- Excluding those:
  - without adequate controls
  - with poor design
  - with fewer than 10 participants
  - where performance was not measured
  - without details of effect sizes
- left 131 reports, 607 effect sizes, involving 12652 individuals
- On average feedback does improve performance, but
  - Effect sizes very different in different studies
  - In 38% (50 out of 131) of studies, effect sizes were negative

Getting feedback right is hard

Response type	Feedback indicates performance…	
	exceeds goal	falls short of goal
Change behavior	Exert less effort	**Increase effort**
Change goal	**Increase aspiration**	Reduce aspiration
Abandon goal	Decide goal is too easy	Decide goal is too hard
Reject feedback	Feedback is ignored	Feedback is ignored

Meanings and consequences of school grades

- Two rationales for grading
  - Meanings
    - Assessment as evidentiary reasoning
    - Assessment outcomes as supports for making inferences
      - (e.g., about student achievement)
  - Consequences
    - Assessment outcomes as rewards and punishments
    - Assessments create incentives for students to do what we want them to do
  - These two rationales interact, **and conflict**
    - achievement grades for completion of homework
    - achievement grades for effort
    - penalties for late submission
    - zeroes for missing work

Dual-pathway theory (Boekaerts, 2006)

- Long-term learning goals are translated into short-term learning intentions
- Dynamic comparisons of task and situational demands with personal resources, taking into account:
  - Current perceptions of the task
  - Beliefs about the subject or task
  - Beliefs about “ability” and the role of effort in the subject
  - Interest in the subject (personal vs. situational)
  - Previous experiences on similar tasks
  - Costs and benefits

And then it comes down to…

- Resulting activation of energy along one of two pathways
  - Wellbeing
  - Growth
- We need assessment systems that push our students towards a focus on growth, rather than wellbeing

Summary

- Before we can assess, we need clear models of progression
- Validity is not a property of tests or assessments, but of *inferences*, which are weakened by
  - construct underrepresentation
  - construct-irrelevant variance
- Reliability is a key requirement for validity
- Limited test reliability has particularly severe consequences for changes scores and diagnosis
- Assessments are important for what they *do* as well as what they *mean*