The meanings and consequences of educational assessments

Dylan Wiliam
Uses of educational assessments

Formative: Supporting learning
Summative: Certifying individuals
Evaluative: Evaluating educational provision
Classical test theory

Fundamental model:

\[ X = T + E \]

- **X**: Observed score
- **T**: ‘True’ score
- **E**: ‘Error’ (ie residual)

Spread of errors = \( \sqrt{1-r} \times \) spread of scores
Reliability

With an 85% reliable test, in a class of 30:

Half the candidates will get a mark within 4% of their ‘true score’

Three-quarters of the candidates will get a mark within 7% of their true score

One candidate will get a mark differing from their true score by more than 12%
Grades versus scores

Scores suffer from spurious precision
Grades suffer from spurious accuracy

When a test of reliability of 85% is used to assign candidates to one of eight grades only 60% of the candidates get the ‘right’ grade
## The reliability of national tests

<table>
<thead>
<tr>
<th>reliability of test</th>
<th>0.60</th>
<th>0.65</th>
<th>0.70</th>
<th>0.75</th>
<th>0.80</th>
<th>0.85</th>
<th>0.90</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of students</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>misclassified at KS1</td>
<td>27%</td>
<td>25%</td>
<td>23%</td>
<td>21%</td>
<td>19%</td>
<td>17%</td>
<td>14%</td>
<td>10%</td>
</tr>
<tr>
<td>% of students</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>misclassified at KS2</td>
<td>44%</td>
<td>42%</td>
<td>40%</td>
<td>36%</td>
<td>32%</td>
<td>27%</td>
<td>23%</td>
<td>16%</td>
</tr>
<tr>
<td>% of students</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>misclassified at KS3</td>
<td>55%</td>
<td>53%</td>
<td>50%</td>
<td>46%</td>
<td>43%</td>
<td>38%</td>
<td>32%</td>
<td>24%</td>
</tr>
</tbody>
</table>
The reliability of examinations

On the evidence available, somewhere between a fifth and two-fifths of students will get the ‘wrong’ level or grade in national curriculum tests, and GCSE and A-level examinations.

‘Tiering’ increases the reliability of the assessments, but means that many students cannot get the grades to which they aspire, *no matter how well they do.*
Making tests more reliable

Doubling the length of a test would increase the proportion of students correctly classified at key stage 2 by 9%.

To ensure that 90% of students are correctly classified at key stage 2 would require 30 hours of testing per subject.
Speech acts

Perlocutionary speech acts are statements about what was, is or will be (eg Michael knows his number bonds to 10)

Illocutionary speech acts are performative: they create social facts (eg “I now pronounce you husband and wife”)
Social facts

Interviewer: Did you call them the way you saw them, or did you call them the way they were?

Umpire: The way I called them was the way they were.
Validity

A property of:
  a test
  the responses to a test
  the inferences drawn from a test

“One validates not a test, but an interpretation of data arising from a specified procedure”
(Cronbach, 1971 p447, original emphasis)
Reliability and validity

Reliability: the extent to which inferences about the items actually assessed are warranted

Content validity: the extent to which inferences about the domain from which the items were selected are warranted

Predictive and concurrent validity: the extent to which inferences beyond the domain from which the items were selected are warranted
Predictive validity
Using tests for setting

<table>
<thead>
<tr>
<th>students placed in</th>
<th>should actually be in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>set 1</td>
</tr>
<tr>
<td>set 1</td>
<td>23</td>
</tr>
<tr>
<td>set 2</td>
<td>9</td>
</tr>
<tr>
<td>set 3</td>
<td>3</td>
</tr>
<tr>
<td>set 4</td>
<td>3</td>
</tr>
</tbody>
</table>
Value-added

A-level to first degree
GCSE to A-level
Key stage 2 to GCSE
Key stage 1 to key stage 2
Baseline testing to key stage 1
?
Summative assessment

Norm-referenced
Cohort-referenced
Criterion-referenced
Construct-referenced
Ipsative
Criterion-referenced assessment

John where Paul had had had had had had had had had had had had had had had had a clearer meaning

John, where Paul had had “had”, had had “had had”. “Had had” had had a clearer meaning.
Construct-referenced assessment

‘Criteria’ do not define but exemplify grades
‘Standards’ are shared by the community of practice
‘Standards’ are implicit and evolve
“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 p50).

“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 p64).
Validating assessment

Summative function
» is validated by widely shared meanings
» requires teachers to form a community of practice

Formative function
» is validated by appropriate consequences (ie learning)
» requires teachers to interpret performance in terms of learning needs
» requires learners to be members of the community of practice
Lake Wobegon

Scores

Time

X
The Macnamara Fallacy

The first step is to measure whatever can be easily measured. This is OK as far as it goes.

The second step is to disregard that which can’t easily be measured or to give it an arbitrary quantitative value. This is artificial and misleading.

The third step is to presume that what can’t be measured easily really isn’t important. This is blindness.

The fourth step is to say that what can’t be easily measured really doesn’t exist. This is suicide. (Charles Handy, The empty raincoat, 1994 p219).
Goodhart’s law

All performance indicators lose their meaning when adopted as policy targets

- Inflation and money supply
- Railtrack’s performance targets
- National Health Service waiting lists
- National curriculum 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
Reliability of coursework

<table>
<thead>
<tr>
<th>Reliability</th>
<th>0.60</th>
<th>0.70</th>
<th>0.80</th>
<th>0.90</th>
<th>0.95</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>0.40</td>
<td>0.45</td>
<td>0.52</td>
<td>0.65</td>
<td>0.75</td>
<td>0.90</td>
</tr>
<tr>
<td>consistency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluative assessment

- Light sampling approach
- Large number of tasks, allocated to students at random
- Low individual reliability, but high reliability at group level
- Broad coverage
- Minimal negative backwash