Certainty-Based Marking at UCL and Imperial College

Tony Gardner-Medwin, UCL
& Nancy Curtin, Imperial College London

Physiological Society Teaching Workshop
Main Meeting (UCL) - 4 July 2006
Proc Physiol Soc series 3, WA4

"Computer assessment just tests factual knowledge"

It certainly doesn't have to!

But laziness can make it degenerate that way

The following values were obtained in a 30 year old patient of normal build:
Plasma creatinine concentration = 0.2 mmol/l,
Urine creatinine concentration = 4 mmol/l,
Urine flow rate = 2.5 ml/min.
State whether each of the following statements is true or false.

Qu. 1: The creatinine clearance is 50 ml/min. **T/F?**
Qu. 2: The clearance of inulin would be a similar (but not identical) value to that of creatinine. **T/F?**
Qu. 3: The creatinine clearance is within normal limits. **T/F?**
Qu. 4: Urine flow rate does not reach such levels in normal subjects. **T/F?**
Qu. 5: Plasma creatinine concentration would be expected to increase if creatinine clearance is increased. **T/F?**

Don't treat physiology like anatomy:

Concerning the abdominal aorta :
Qu. 1: It ends just below the level of the umbilicus T/F?
Qu. 2: It is directly related to the liver T/F?
Qu. 3: It gives off lumbar arteries T/F?
Qu. 4: It passes behind the median arcuate ligament of the diaphragm T/F?
Qu. 5: It gives inferior phrenic branches T/F?

"Computer assessment is second best to essays / problems for assessment"

Yes when you want to test if students can write essays, organise and express ideas, or dissect problems.

No when you want to challenge knowledge and application.

It's absurd to use essays to ensure coverage of the curriculum.

Assessment should be a combination of different forms with different objectives

- *Essays / prose answers / organised expression / analysis are **really important** parts of what a student should be able to do.*
- *They are things many students do not do well*
- *- use them to test if they can, and to give feedback to help them*
- *Mediocre essays are often largely regurgitation and hard to assess for understanding - you may test understanding **better** with computerised assessment.*

"You shold use 'modern' question formats like single-best-answer or extended matching questions - not 'outdated' True/False questions"

We are really mystified why! Is there any basis for this prejudice?

Case & Swanson (http://www.nbme.org/PDF/2001iwg.pdf) - advice on formats for USA Medical Examiners - is often quoted, but there is no good argument there.

- ***Best-option Qs encourage answering simply on the basis of recognition.***
- ***Multiple options can reduce to 1 or 2 with very little knowledge, just by eliminating the obvious***
- ***Care generating multiple distractors produces a single response - equivalent effort with T/F Qs would give several responses with more info***
- ***'Best-option' answers may be just as debatable as T/F ones***
- ***A 'wrong' answer to a best-option Q doesn't indicate whether the student thought the 'right' answer almost as good, or something definitely wrong***

Best-Option Format
In which one of the following circumstances
would a dose of 75 mg aspirin per day be most appropriate?
(a) history of migraine attacks
(b) history of stomach ulcers
(c) history of thrombosis
(d) history of osteo-arthritis
(e) following a dental operation

True/False Format
Aspirin ...
(a) is most commonly sold (for adult use) in tablets of about 75 mg (T/F?)
(b) may cause stroke in excessive doses (T/F?)
(c) may cause stomach disorders in excessive doses (T/F?)
(d) is a steroidal anti-inflammatory drug (T/F?)
(e) is used in prophylaxis against thrombosis (T/F?)

"You should include a 'Don't Know' option with T/F or Best-Option Qs"

(equivalent to explicitly omitting an answer)

The implication is that students should use this option if their uncertainty is above some level.

This is practically never sensible!

It can only be rational for the student to omit a guess if the average penalty for a guess is worse than for a blank reply. This level of fixed negative marking is seldom deemed acceptable, and would simply put stress on candidates.

*It actually **disadvantages** able students - because they have more insight into which answers are uncertain, and could be persuaded to omit these, though they would on average gain by answering.*

"Frequently Missed Opportunities" with Computerised Assessment

• Immediate feedback	in formative assessment
• Integrated comment system	
• Challenge to students to find good reasons for certainty or doubt	in both formative & summative (exam) assessment

We need : "Assessment for Learning"

-- applied to computerised assessment

cf. **Paul Black** (King's) for more general application
- Google him !

LAPT - "London Agreed Protocol for Teaching"

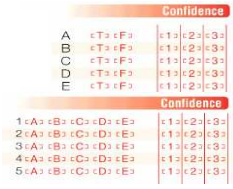
-- incorporates these principles, not yet understood by e.g. **WebCT**, **QuestionMark**

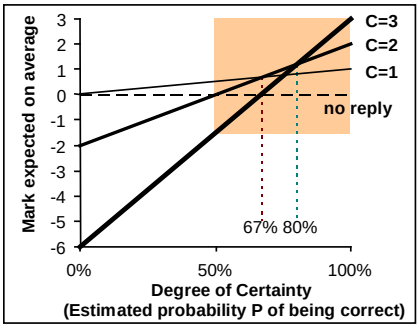
Certainty-Based Marking to motivate & reward deeper thinking

www.ucl.ac.uk/LAPT

Degree of Certainty :	C=1 (low)	C=2 (mid)	C=3 (high)	No Reply
Mark if correct:	1	2	3	0
Penalty if wrong:	0	- 2	- 6	0

OMR forms (from UCL or Speedwell)





CBM - main current usage

Learning through online Self-Assessment (open access)
- UCL, Imperial + students at >30 universities
- Follow-up exercises for lectures, practicals
- Revision, with past exams
- Student-written exercises

Formal online tests (using link to WebCT)
- Maths & key skills tests at UCL
- Formative module tests at Imperial

Paper (OMR) tests
- Formative module tests at UCL & Imperial
- Yr 1,2 exams at UCL

Lecture/ seminar context
- Junior Doctor sessions at Imperial [Sara Marshall]

"Are you prepared to act on your answer?"

Poster Communication: Analysis of UCL exams

(Sara Marshall - Imperial College)

Qu. 7: Follow up management of the unconscious patient

- Following treatment, Mrs Brown feels much improved and asks when she can go home, where her daughter will stay with her. Which of the following management plans are appropriate? (best of 5)

A. Allow her to go home, advising that she return if she has further hypoglycaemic symptoms.

B. Allow her to go home, but switch her to a shorter acting sulphonylurea with GP follow up in 48 hours.

C. Admit her for observation for at least 24 hours

D. Admit her for dextrose and insulin sliding scale for 24 hours

Level of confidence / Choose correct answer	1 I am guessing, but I think this is the correct answer	2 I am pretty sure I am correct but need advice before proceeding	3 I am happy to proceed
A			
B			
C			
D			

(Sara Marshall - Imperial College)

$$\text{Quack coefficient} = \frac{\text{number of -6 scores}}{7} \times 100$$

Lessons from experience with CBM

Practice is needed before use in exams
- not really a problem, since its objective is to encourage better thinking

Exams should re-use questions from an open database only very sparingly
- CBM places a premium on answers that the student has good reason to believe are correct

Students can lose out through excessive risk-aversion or over-confidence
- this lesson is learned in practice.
- adjustment can be made in exams to compensate for poor calibration**

Limited experience with Best-Option Qs suggests that students need extra practice if they are not to exaggerate confidence in their answers to these
- further research required
- students may benefit from learning that such Qs may need more thought

Standard setting
- the CBM mark range is unfamiiar, but scaling can align it with the familiar**

**** see Poster**

Positive Features :
- students like CBM, consider it helps them study and is more fair
- they have voted at UCL to retain it in Yr 1,2 exams
- it is more reliable and valid than conventional marking in exams
- it is more closely related to what we mean by 'knowledge'

ABSTRACT: Certainty-Based Marking at UCL and Imperial College Proc Physiol Soc, 2006

A.R. Gardner-Medwin, Dept. Physiology, UCL, London WC1E 6BT
N.A. Curtin, Div. Biomedical Science, Imperial College London, SW7 2AZ

Certainty-based marking (CBM), or Confidence-Based Marking as it was initially termed, was set up in London 12 years ago as a project involving several Physiology Departments. The initial acronym LAPTTOP (London Agreed Protocol for Teaching and Testing of Physiology) soon lost its 'TOP' as usage spread to other areas of the medical curriculum, mainly at UCL and Charing Cross. Current versions of LAPT are on open access (at www.ucl.ac.uk/lapt) and accessed by students from campus computers at more than 30 UK universities. However, despite usually enthusiastic reaction to the concept at conferences, staff at other universities have been slow to engage.

We all want student learning to be more effective and less extravagant in staff time. Part of the strategy can involve self-assessment tasks alongside teaching material, wherever possible challenging deeper knowledge than simply factual or associative learning. A strength of this approach is that staff time can pay off many times over with new student cohorts, but a weakness is that self-assessment is less effective than face-to-face confrontation at probing weaknesses: students who get an answer right often think they knew the answer, when in fact all they did was plump for the more likely answer and strike lucky. A lucky guess is not knowledge, and it is incorrect and inefficient (in statistical terms, adding variance) to mark an assessment as if it were. CBM differentiates between students who may all give the same answers in a test: it rewards those who can distinguish their more reliable and less reliable answers. It places a premium on being able to think through a thorough justification for an answer, and indeed it rewards reflection that leads to a conclusion that the answer is less certain than initially thought. Students find it intuitively easy to use, and cannot cheat by misrepresenting their certainty: correct reporting of one's degree of certainty is always the best strategy to maximise expected score. CBM avoids irrational dilemmas that are often created for students about whether to answer or not, in exercises with fixed negative marking schemes. If you are not familiar with CBM, go to the website (above), try it out, and read the papers linked from the site.

CBM has been used in several contexts at UCL and Imperial. Greatest activity involves online self-assessment exercises designed for teaching and revision (some using past exam papers). Invigilated formative tests with Optical Mark Reader cards (Speedwell Computing Services: www.speedwell.co.uk) have been run with True/False, Single Best Answer (SBA) and Extended Matching (EMQ) question types. Compulsory summative online maths tests (with numerical answers) are employed at UCL, with opportunity for those having difficulty to repeat tests indefinitely with randomised questions and parameters. Online tests are used to establish comprehension of key points following practical work in place of assessed write-ups. In the Imperial clinical course, questions are presented in seminars with students generating self-assessed answers using CBM, with certainty levels linked to qualitative concepts like 'would be willing to proceed on this basis' or 'need to confirm', with enthusiasm from junior doctors about the stimulus it gives to their judgment of clinical knowledge and issues. CBM is used in summative UCL exams with clear evidence of improved statistical reliability and assessment validity with True/False questions (poster communication at this meeting).

Though CBM has been popular and successful, some cautions emerge. Students need practice to use CBM to best advantage, and indeed this is part of the benefit: they learn to judge how reliable is their knowledge. Re-use of published questions in exams becomes a more serious problem: CBM places a premium on good reasons to be sure of an answer, amongst the best of which is of course to have seen the question and answer before. With SBA and EMQ in formative tests, students have tended to be overconfident in their answers. This may result from less practice with these question types, which are less common in our institutions. An alternative hypothesis is that students may narrow options to just 2 or 3 through limited knowledge, and then exaggerate confidence in the final choice because they are correctly confident about some of their rejections. We are keen to discuss evidence bearing on the relative merits and costs of the use of these different question types.