Gardner-Medwin AR (2006). Analysis of exams using certainty-based marking.
Physiological Society, University College London 2006: Proc Physiol Soc 3, PC64

Analysis of Exams using Certainty-Based Marking

Tony Gardner-Medwin
Physiology, UCL
www.ucl.ac.uk/lapt

UCL

Knowledge, like a building, must have a firm foundation. Without this, it is not knowledge – merely words.

A student's ability to identify the reliability of an answer is integral to the assessment of knowledge and reasoning.

In formative tests, CBM aids learning by encouraging reflection about the basis and relationships of ideas.

In exams, it rewards students who can distinguish uncertain and reliable answers.

CBM should become a major plank of educational testing, especially with automated marking.

Background

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

What is CBM? CBM marks each answer according to the student's degree of certainty that their answer is correct.

How did CBM in London begin? In 1994, through collaboration of physiology depts, to improve online self-assessment (LAPT: London Agreed Protocol for Teaching).

Automated marking is forced on us through pressures on time. CBM was an attempt to improve it: make it more like face-to-face assessment, gain more information, and stimulate deeper learning.

CBM in exams? In 2001, students & staff at UCL opted for CBM for Yr 1,2 medical exams as more fair and motivating than 'number correct' or ±1 negative marking.

Fairness: a lucky hunch is not the same as a justified correct answer. Confident errors are worse than acknowledgement of uncertainty. In 2006, UCL medical students voted 52%:30% to retain CBM.

Is there experience elsewhere? Much research (mainly <1970) has shown the value of related (though usually more complex) mark schemes. UCL is probably the largest current user of CBM.

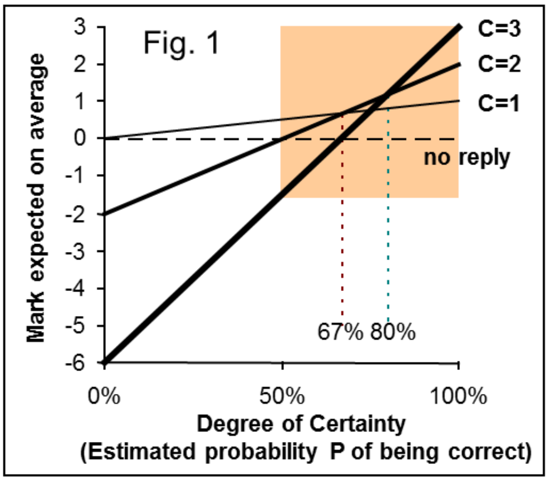
CBM increases both retention of test material and reliability of exam data. Possible reasons for poor uptake are inertia, poor comprehension and vested interest in retaining existing commercial systems.

How does CBM work?

Fig. 1 shows how the average expected mark for an answer depends on (i) the degree of certainty of being correct, and (ii) the choice of C level.

Key point: the scheme motivates accurate reporting of certainty or uncertainty. C=1 is the best choice (top graph) when unsure (P<0.67). C=3 is best when nearly sure (P>0.8), and C=2 in between. A student cannot gain by any strategy other than careful evaluation of the probability that their answer is firmly based.

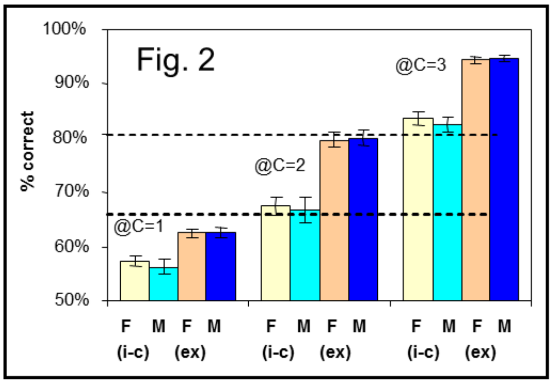
For True/False questions (used in our exams) only the shaded portion of the graph is relevant, since preferred answers cannot have an estimated P(correct)<50%.



How well do students discriminate certainty?

For both in-course (i-c) and exam data (ex) the % correct at each C level is within the optimal band for both sexes (F,M). Bars shows mean ± 95% confidence limits; student cohort: n=331.

There are no significant gender differences in the data. Psychological studies with *naïve* subjects have shown gender differences in confidence estimation, but if this is present in our students when naïve, it must disappear rapidly with practice. Both sexes are more cautious (risk-averse) in exams, using C=2 and C=3 more sparingly.



Exam Analysis - basics

17 summative True/False exams were analysed (40% of summative assessments for year 1,2 medical students at UCL). All students were well practised in formative tests and through online self-assessment.

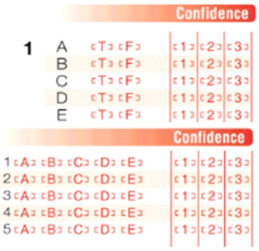
Scripts: 5706 ***Responses:*** 1.67 million (250-300 Qs and >300 students.)

Usage of C levels: **C=3:** 44% **C=2:** 18% **C=1:** 37% (**Blanks** 0.3%).

Accuracy (% correct) at each level: **C=3:** 94% **C=2:** 77% **C=1:** 61%

Overall % correct: 78.2% ± 7.4% SD (n=5706)

Raw CBM Score *(% of maximum attainable marks):* 50.3% ± 13.1% SD



Speedwell Optical Mark Reader styles for CBM

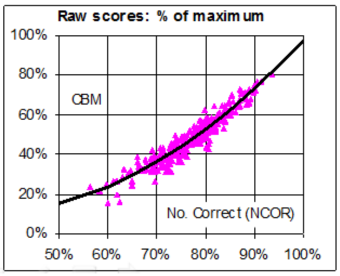
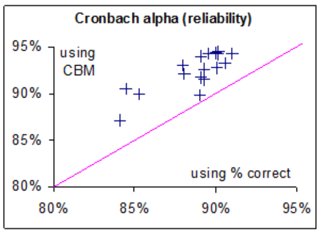


Fig. 3 shows, for one typical exam, the raw **CBM** scores for each of 342 students plotted against the number of correct answers (**NCOR**). Each is expressed as a % of maximum. Without knowledge, guessing would give NCOR=50% and CBM = 17%, while complete knowledge would give 100% on both scales. The curved line is a quadratic trendline for the data, which on extrapolation passes approximately through these points, though the ultimate in knowledge or ignorance was not quite attained by any of the students.

Fig. 4: Reliability

Cronbach Alpha is the standard measure of how reliably an exam score reflects a single characteristic ('ability') for each student in the face of random factors due to varied questions and luck.

The CBM data are more 'reliable' than % correct data, for each of the 17 exams. Alpha depends on the number of questions in a test. To raise it by the amounts shown here would require on average a 58% increase of test length with conventional marking.



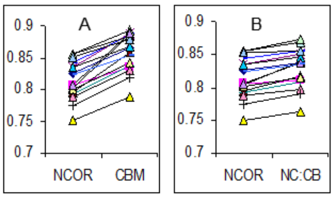


Fig. 5: *Correlation coef. (r) for scores on odd/even Qs.*

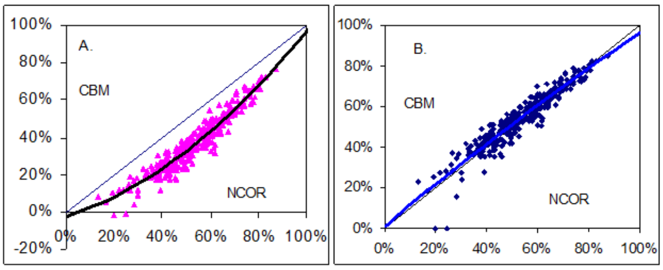
Which score best predicts performance on other questions?

A more intuitive way of looking at reliability is the correlation between odd and even numbered questions on an exam. CBM scores were not only better than NCOR at predicting equivalent scores on interleaved questions (A), but better predictors even of NCOR on the interleaved questions (B). For stats, see abstract.

In psychometric terms, CBM is both a more ***valid*** and more ***reliable*** index of performance. Even if one were to regard NCOR as the gold standard of performance, CBM is a better estimator of this than NCOR.

Fig. 6: Standard setting

Raw CBM data (Fig. 3) has its own characteristic range of values, not linearly related to % correct. To make pass-marks similar on different schemes, for comparison and standard setting, it is best first to scale each score so that total guessing gives 0% (A). The regression relation is then linearised by raising CBM to the power 0.6 (B), as established consistently in both on-line and exam data.

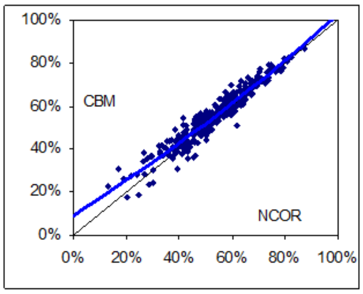


Fairness: students with poor calibration

Though practised students are generally well calibrated, some (1.3%) got <80% correct at C=3 (over-confident) and more (16%) got >67% correct at C=1 (under-confident). Under-confidence was commoner in high-scoring students and over-confidence in weak students. Where the % correct at one C level is not in the optimal range in this way, adjustment can be made by reallocating all the Qs answered at this C level to the optimal level. This never reduces a student's score.

Adjustments averaged 1.2%, being zero for 40% of scripts, <1% for 25% and >10% for 0.6% of scripts. It can be argued that adjustment is indulgent to students who are confident while performing badly (for example, two students in Fig. 6a who obtained negative CBM scores, worse than chance). However, it eliminates any possible claim that failure in such cases was due to poor CBM experience rather than to lack of knowledge.

Fig. 7 Adjusted CBM scores. *As Fig. 6B, but with C allocated to optimal levels when there was poor calibration.*



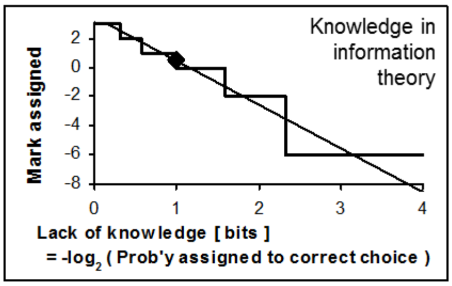
Conclusions

- CBM has worked well in True/False summative exams, improving both the reliability and validity of the resulting data.
- CBM motivates students to evaluate reservations or justifications for answers, and to report correctly how reliable they think each answer is.
- CBM is both well founded in information theory and readily accepted by students as a more fair assessment than conventional marking.
- CBM in exams requires that students are familiar with the scheme through practice in formative and/or online self-assessment.
- Adjustments can be made for students who disadvantage themselves with non-optimal bias of their choice of CBM levels, in either direction.
- CBM requires no special skills in question setting
- CBM scores can be made readily comparable with conventional scores, for standard setting.
- CBM is readily implemented through software available from UCL; commercial vendors should be pressured to offer CBM.

What is knowledge anyway ?

✓ knowledge
✓ uncertainty
0 ignorance
× misconception
✘ delusion

decreasing certainty about something true, increasing certainty about something false

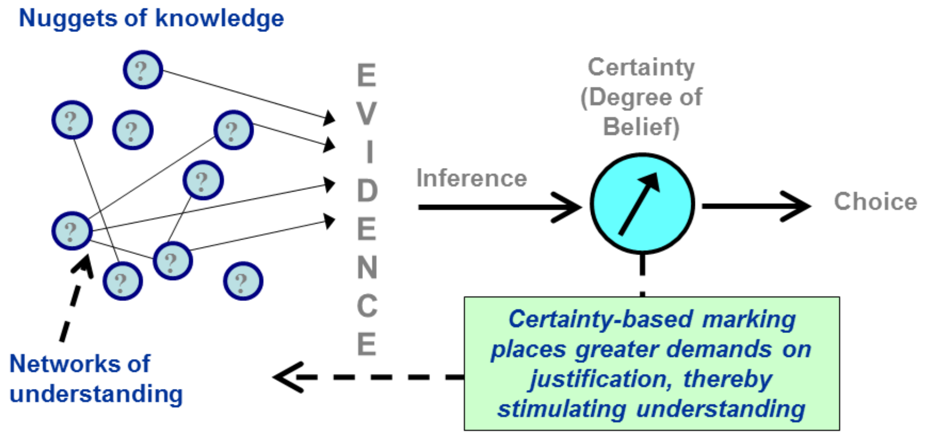


Knowledge is justified true belief. Proper justification requires understanding.

What is understanding?

To understand = to link correctly the facts that bear on an issue.

[This is how you can (usually) tell a student from a parrot!]



Certainty-based marking places greater demands on justification, thereby stimulating understanding

ABSTRACT:

Analysis of Exams using Certainty-Based Marking
A.R. Gardner-Medwin, Physiology, UCL, London WC1E 6BT

Certainty-Based Marking (CBM) has been used at UCL in 17 summative medical exams (years 1&2), each with 250-300 True/False questions and >300 students. Students enter answers on OMR sheets (Speedwell Computing Services) with an index of certainty or confidence that each one is correct. The 3-point scale (C=1,2,3) corresponds to marks given for correct answers, with penalties 0,-2,-6 for errors. This mark scheme is proper, in the sense that students gain by indicating low C when their probability of error is low and high C when it is high. Optimal threshold probabilities are 0.67 and 0.8 for C=2,3. Students were well practised through self-assessments (www.ucl.ac.uk/lapt) and formative tests with detailed feedback. The aim is to encourage care in justification of answers and to improve exam data.

CBM and conventional (number-correct: NCOR) scores were both scaled so 0%=chance performance (at C=1) and 100%= maximum. CBM scores were linearised (raised to the power 0.6), so that the regression of CBM vs NCOR is typically close to the line of equality. Mean scores were CBM=55.0%±12.6% SD and NCOR=53.3% ±12.8% SD. A measure of exam reliability is Cronbach Alpha, indicating how well the combined data reflect a single variable ('ability') characteristic of the student. This was higher for CBM scores than for NCOR (92.4% vs 88.7%, difference 3.7%± 0.31% SEM, n=17, P<0.001%).

A more intuitive way to view reliability is in terms of the correlation between scores from alternate questions: sets with odd and even numbers. If the data are reliable, then the score on one set is a good predictor of the score on the other. The mean correlation coefficient (r) for CBM was 0.859±0.030 SD, significantly greater than for NCOR (0.814±0.030; difference 0.045±0.0042 SEM, P<0.001%). CBM scores were not only better predictors of CBM on the alternate set, but also better predictors of NCOR (CBM vs NCOR: r=0.829±0.030 SD, greater than NCOR vs NCOR by 0.015±0.0021 SEM, P<0.001%). Improvements were largest for the bottom third of each class, critical for standard setting and pass/fail decisions: NCOR alone r=0.428, CBM 0.560 (P<0.001%), NCOR vs CBM 0.460 (P<0.1%).

Most students achieve percentages correct in the optimal ranges for each C level. Where students were over- or under-confident (too low or high a % correct with a given level), upward score adjustments (averaging 1.2%) were used in the above analysis, calculated by re-assigning C to the optimal level. The proportion of papers where this adjustment exceeded 2% was just 3.1% for over-confidence and 18% for under-confidence. Though such compensation is perhaps generous, it ensures that no student can argue that a fail mark was simply due to poor calibration of confidence. Weak students benefit if they correctly identify reliable answers, but do not lose out if they fail to do this correctly.

Acknowledgements: Supported by HEFCE FDTL4. D.Bender & C.Farguson kindly provided the exam data.