Optimisation of Certainty-Based Assessment Scores

A.R. Gardner-Medwin, Physiology (NPP), UCL, London

Certainty-Based Marking (CBM)

Aim : To optimise the presentation of

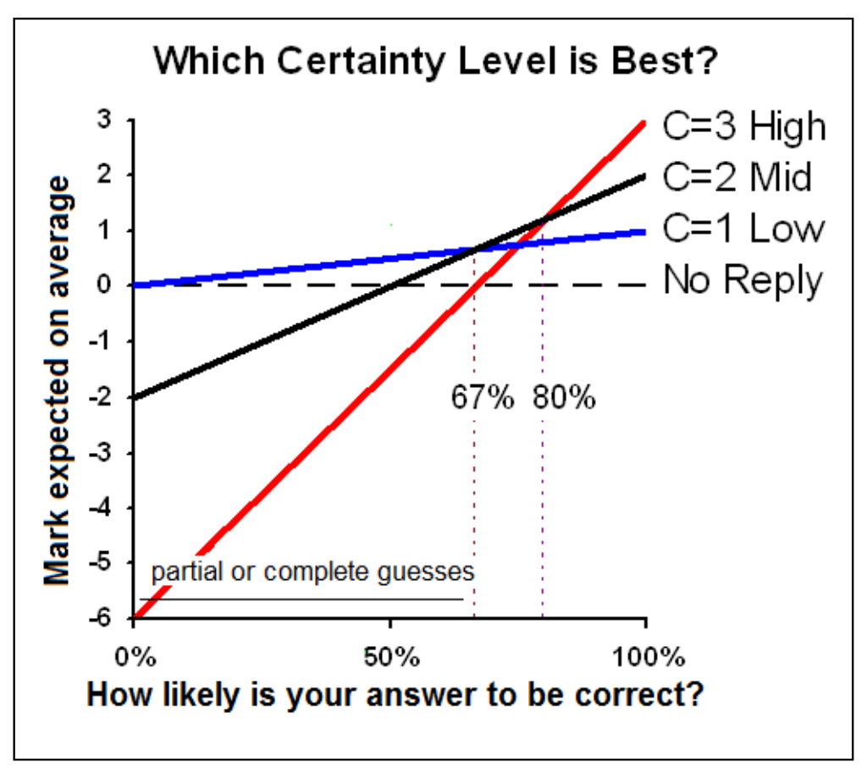
- Self-test Qs so as to challenge students and enhance study
- Exam Qs so as to increase the realism and predictive value of assessment data

Background :

- Medicine and Physiology often require integration of knowledge from different perspectives to be sure of an answer.
- Thinking about the reliability of knowledge and inference is a key academic skill, with particularly dire consequences in Medicine when it fails.
- Valid measures of knowledge or ignorance must take account of uncertainty.
- Explicit certainty judgment has been shown in many psychological experiments to enhance learning and retention.

CBM is a *proper* mark scheme in the sense that a student is always motivated and rewarded for distinguishing and identifying honestly those answers that are uncertain and well justified. It is well founded in information theory (see THEORY, to the right).

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
Probability Correct:	<67%	67-80%	>80%	–



Student perspective:

- Always motivated to be honest
- Rewarded for identifying weaknesses
- Rewarded for sound justifications
- Encouraged to reflect & link info
- Misconceptions highlighted
- Simple and transparent scheme
- Perceive it as realistic & fair

Staff perspective:

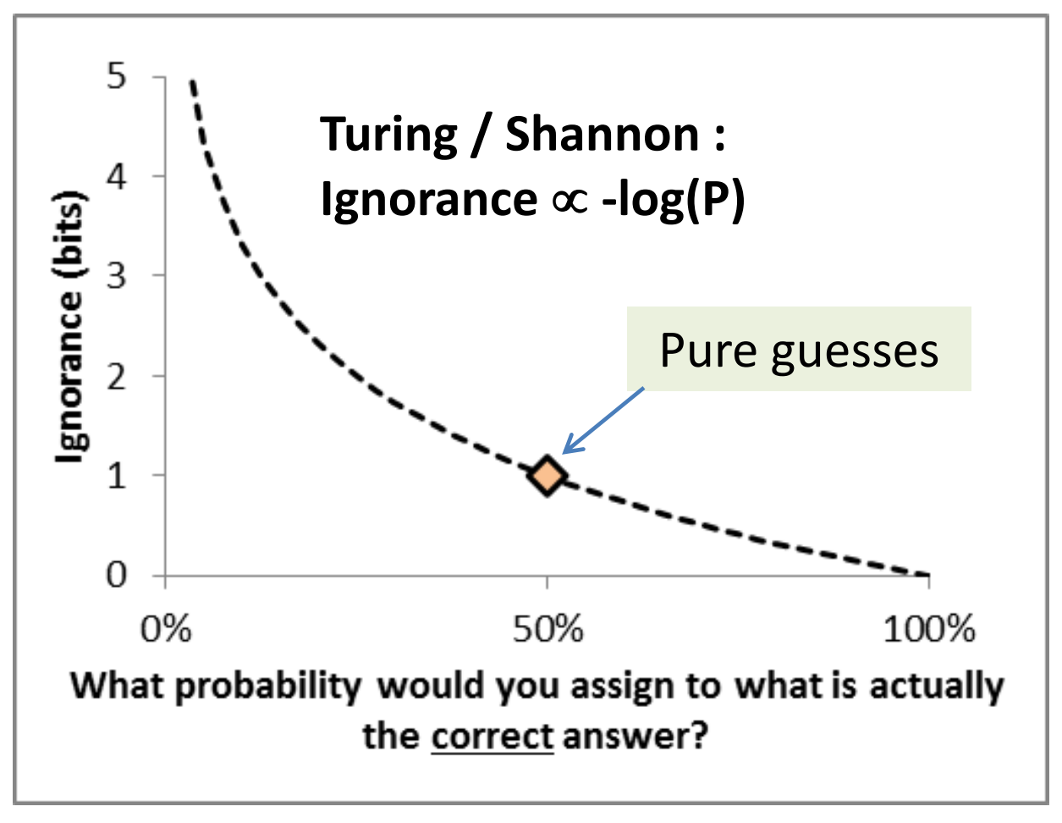
- Doesn't require new or different Qs
- Enhanced feedback about content
- Enhanced reliability & validity in exams
- Better student learning experience

AN EXAMPLE

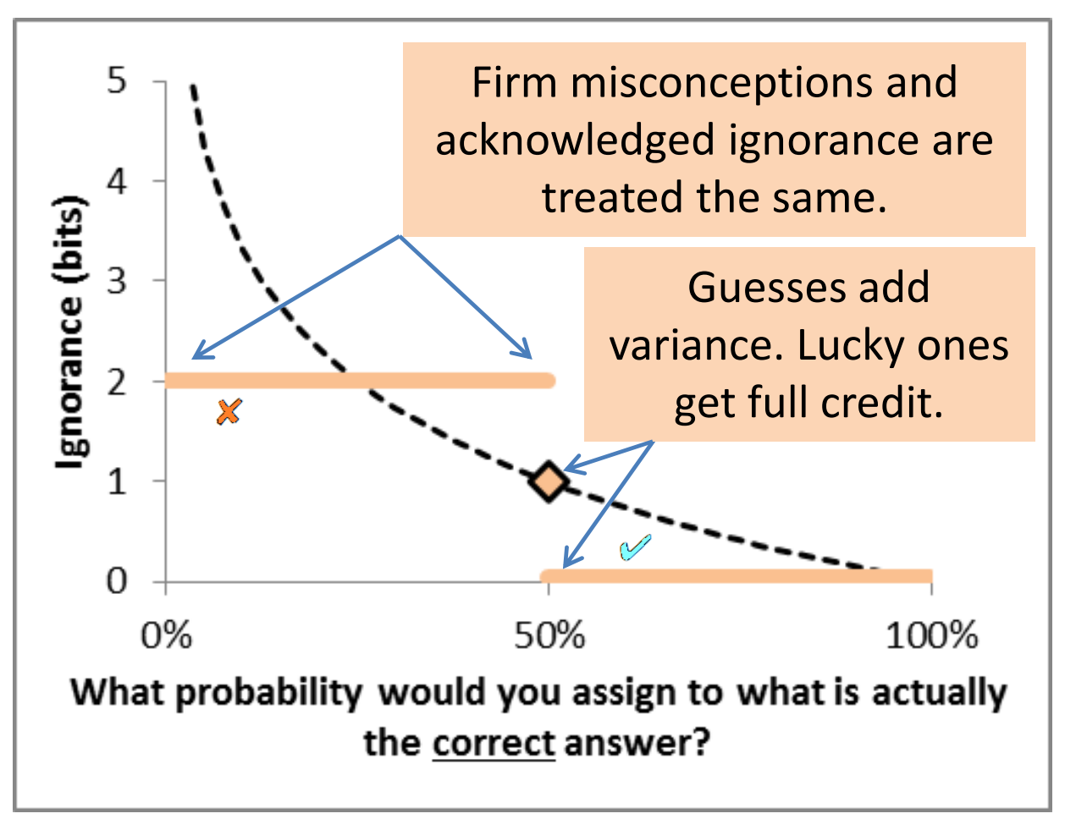
Insulin injection raises blood glucose concentration. True/False ?

IGNORANCE THEORY

Ignorance has an unambiguous definition :

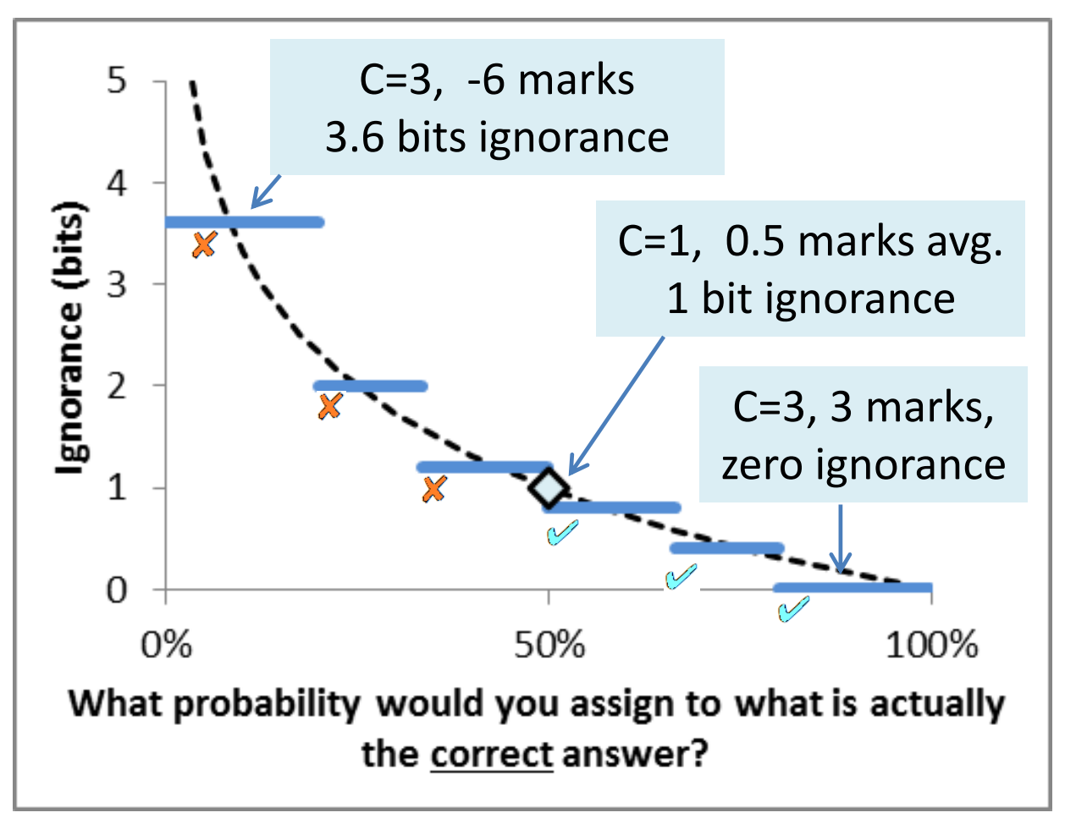


Ignorance inferred from binary (right / wrong) marks has serious problems:



(NB Negative marking generally make things worse: entering "Don't Know" instead of an uncertain answer will reduce variance, but will on average lose the student marks, unless the –ve marking is unusually severe)

CBM gives a good measure of ignorance
(= 0.4 x marks lost, relative to correct at C=3) :

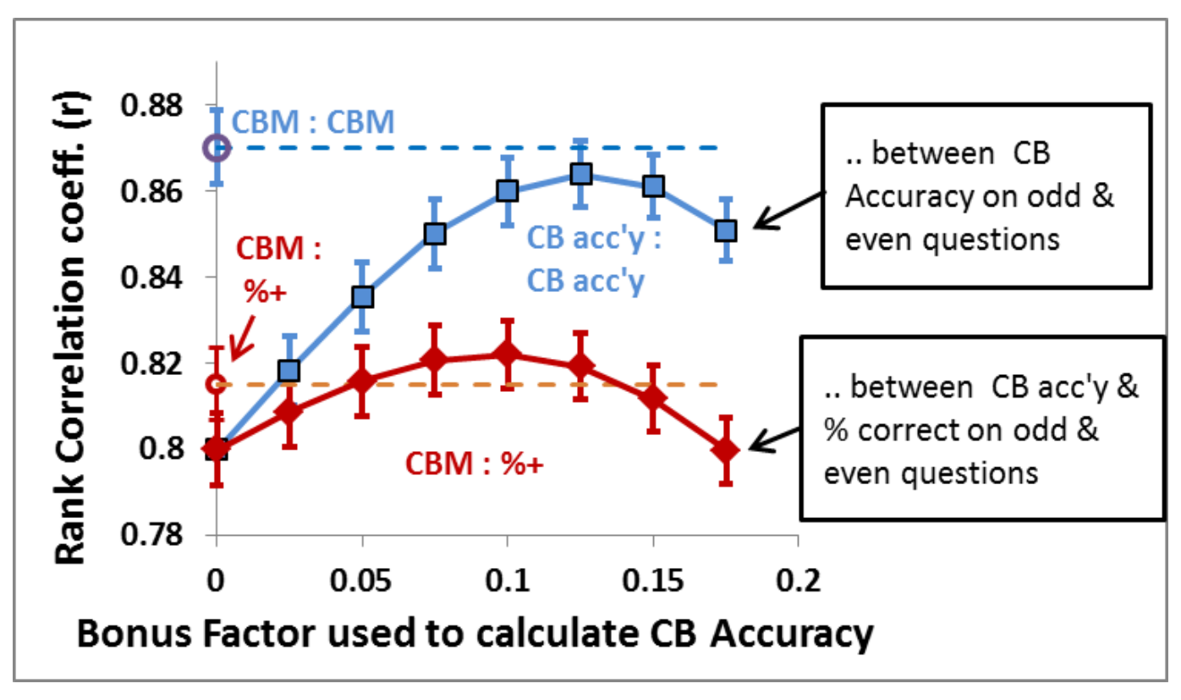


NB with Single-Best-Answer (SBA) Qs the situation is more complex, but the illustrated problems with conventional marking are more extreme.

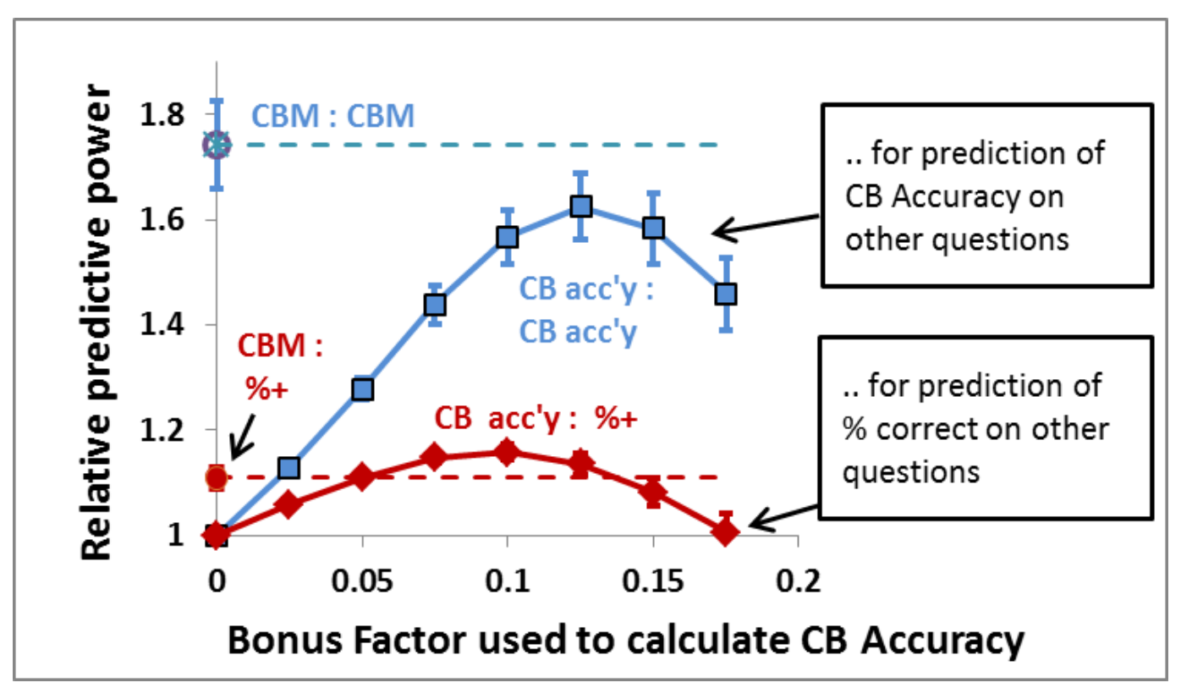
Psychometric optimisation of CB Accuracy

A good measure of the quality of an assessment is how well the score or ranking based on one half of the test (e.g. odd numbered Qs) correlates with that based on the other half. This correlation is substantially enhanced with CBM. With "CB Accuracy" the scale of the "Bonus" added to Accuracy is a variable (<0.125 to ensure CBA<=100%). Data from 17 UCL Yr 1&2 medical exams are used here to assess the validity and reliabilty of CB Accuracy and to optimise the bonus factor (0.1 in the graphs presented in the left panel).

Mean correlation coeff ± s.e.m. for rankings based on odd & even no. Qs , using CBM totals or CB accuracy , paired with either the same score type or with % correct.



Enhancement of predictive power of answers (increase of r/(1-r) compared with right/wrong marking) based on the above correlations. A factor of 1.5 means that the benefit is approximately equivalent to a 50% increase in Q numbers in a conventional test.



A bonus factor of 0.1 is clearly a good choice, giving nearly as good psychometric reliability as the CBM mark itself, and better prediction of accuracy and rank based on accuracy on the complementary Q set.

CBM Implementation: www.ucl.ac.uk/LAPT

All you need to implement CBM for self-tests in your institution (following a model developed for Imperial & Kings, London), are:

- A server site where students & staff authenticate with a local userid
- Links to that site, specifying each self-test, in your VLE

You can use open exercise files, or private ones sited either on your server or at UCL. Editing is simple. Contact me at ucgbarg@ucl.ac.uk for more information. Wholly self-contained software packages are under development, but server loading for new users is almost negligible because computation nearly all takes place on the student's computer.

SUMMARY

CBM makes sense!
Doesn't require special Questions
Always motivates students to give careful honest judgment

SELF-TESTS

More sound and fair measure
↑ reflection & linking of Info
↑ realism about uncertainty
Highlights misconceptions
Students like it!

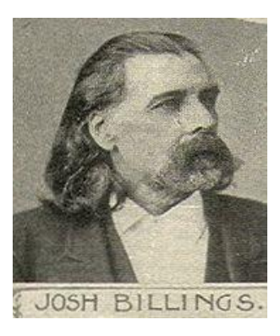
EXAMS

↑ psychometric reliability
↑ psychometric validity
↓ question numbers
No loss of conventional exam info

Some Distinguished Tweets !

"When you know a thing, to hold that you know it,
when you do not know a thing, to allow that you do not know it – this is knowledge."
Confucius

"... there are known knowns;
... there are known unknowns;
... But there are also unknown unknowns
Donald Rumsfeld



"It's not ignorance does so much damage;
- it's knowin' so derned much that ain't so."
attr.: Josh Billings

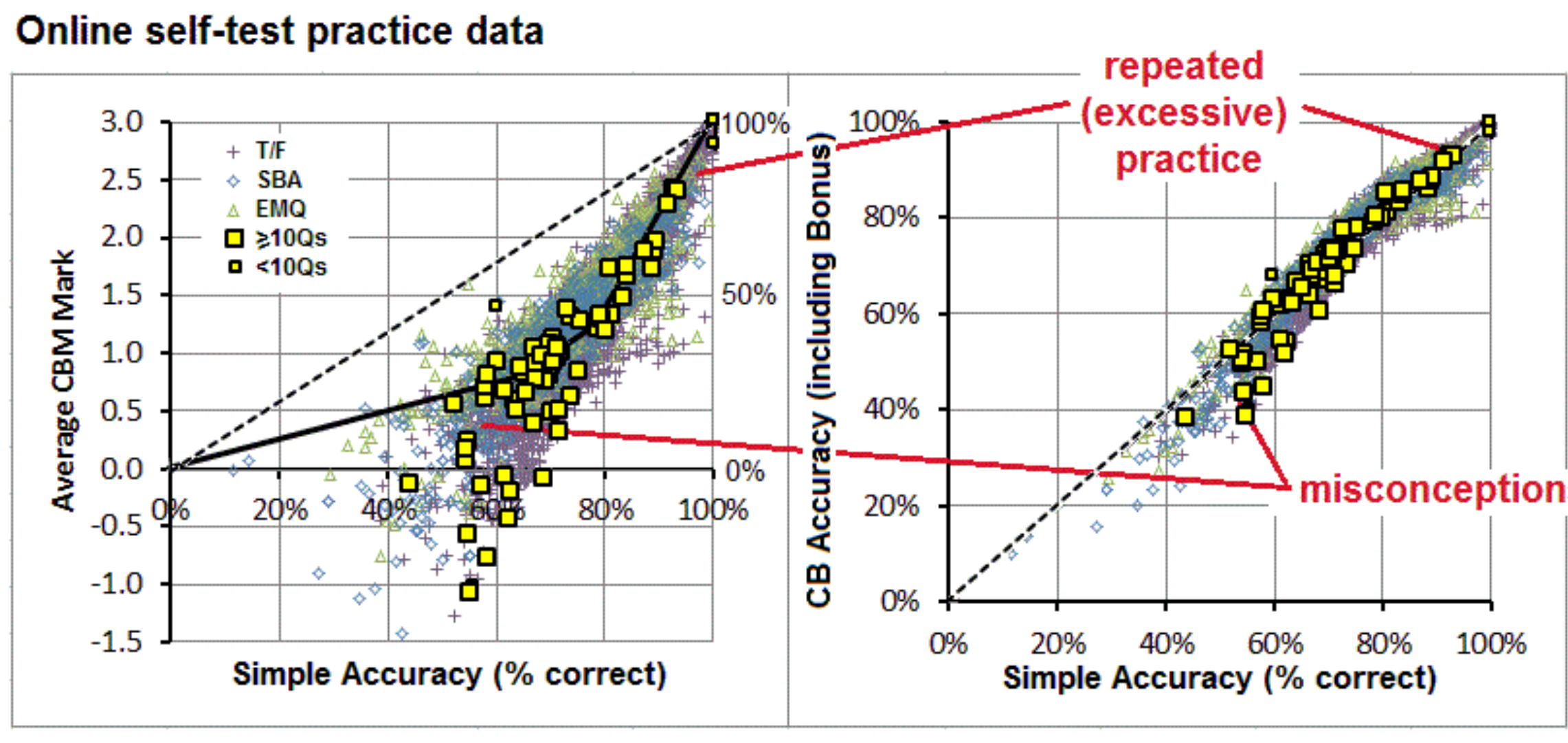
"A lucky guess is not knowledge.
A firm misconception is worse than acknowledged ignorance.
So why do we mark students as if these things weren't true?"
TG-M

HOW BEST TO PRESENT CBM SCORES

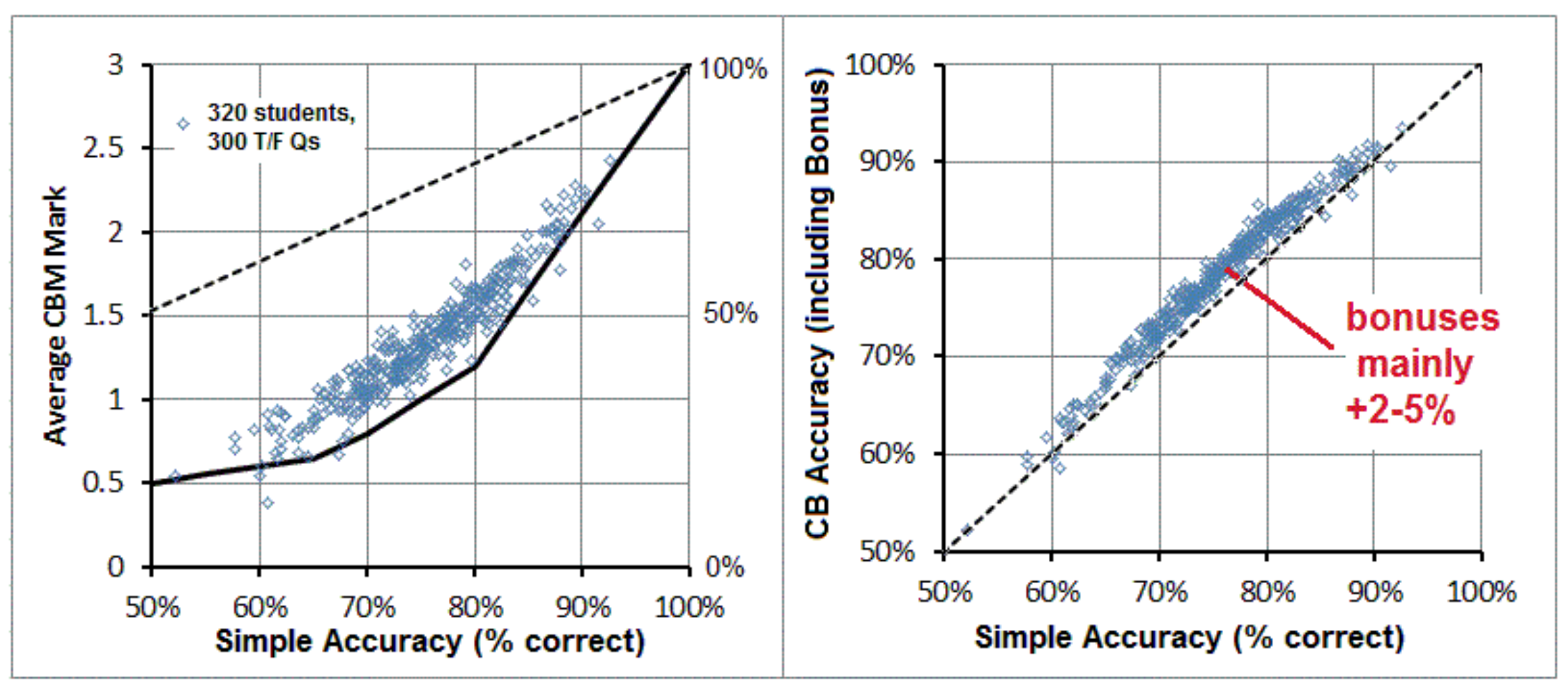
CBM motivates a student to reflect and identify uncertain vs. reliable answers. This is how you maximise your score. But no student can realistically expect to get an average CB mark as great (expressed as a % of maximum) as their accuracy, or % correct on a test. This would only be attainable if every correct answer was given at C=3 and every error at C=1, which in most tests is unrealistic. The graphs below show how (both in self-tests and exams) 80% accuracy is typically associated with an average CBM mark = 1.5, only 50% of maximum.

THE PROBLEM! This simple comparison, even for students above average at judging uncertainty, can be demoralising and counterproductive. There is nothing wrong with CBM scores; but they are fundamentally different from (and psychometrically superior to) accuracy measures. The problem of presentation is tackled here by generating a **"CB Accuracy"** by adding a **BONUS** to the simple accuracy as a measure of how well the student categorises responses as uncertain or reliable. The bonus is positive or negative, proportional to the amount the average CBM mark is above (or below) the average that would be obtained (shown by heavy black lines below) if the student had used the same optimal C level for all his/her answers. Negative bonuses are common in self-tests when students often have misconceptions (confident errors), but as is evident in exam data, students can aspire to gaining positive bonuses of 2-5%.

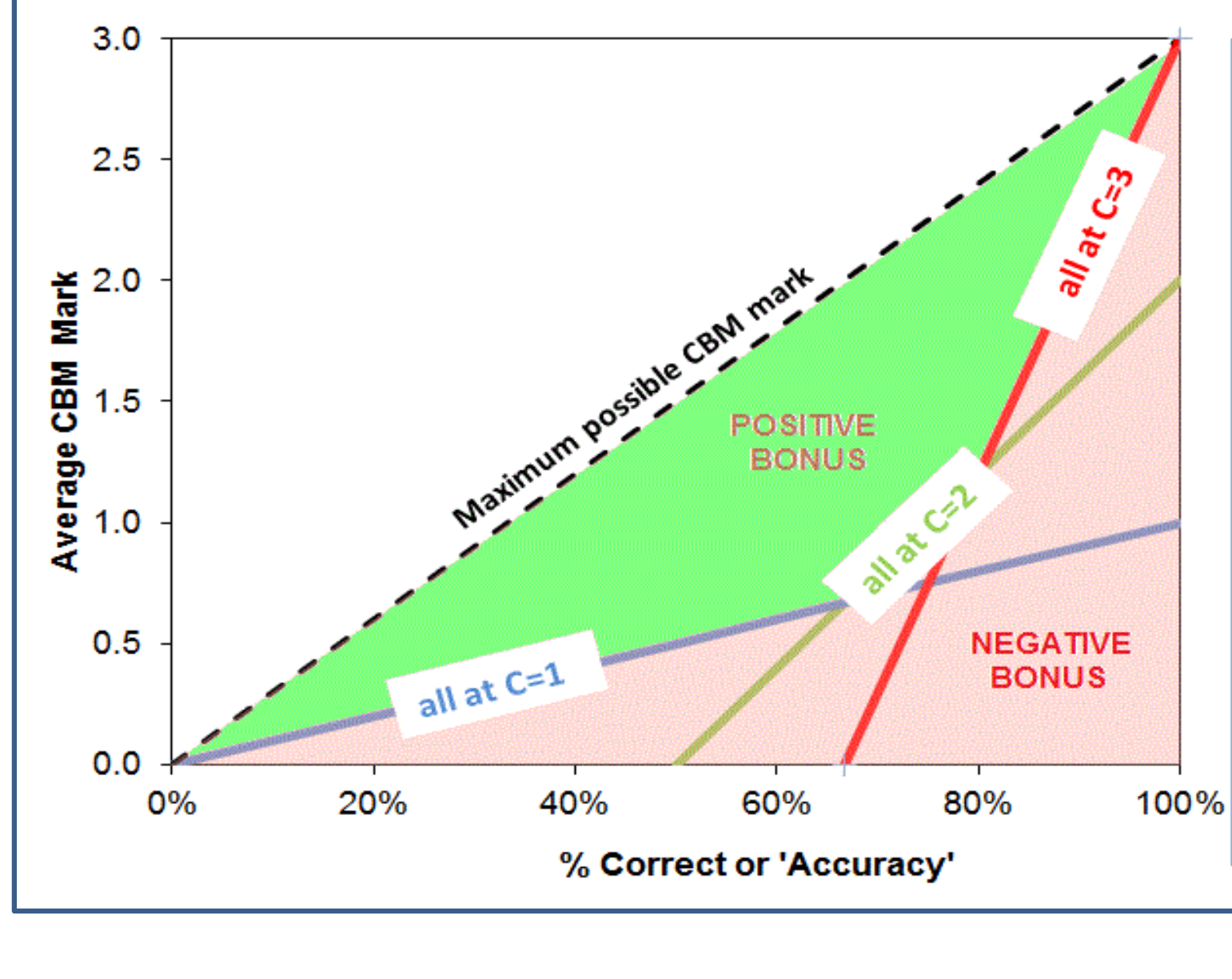
Online self-test practice data



Exam Data



Yellow squares show students' scores on self-test physiology T/F Qs, as displayed in feedback to students and staff using LAPT (www.ucl.ac.uk/LAPT). They are superimposed for comparison on a background of historic data from ca. 9000 sessions with various Q types. Exam data is from a 1st yr medical exam at UCL after students had had substantial experience using CBM in LAPT.

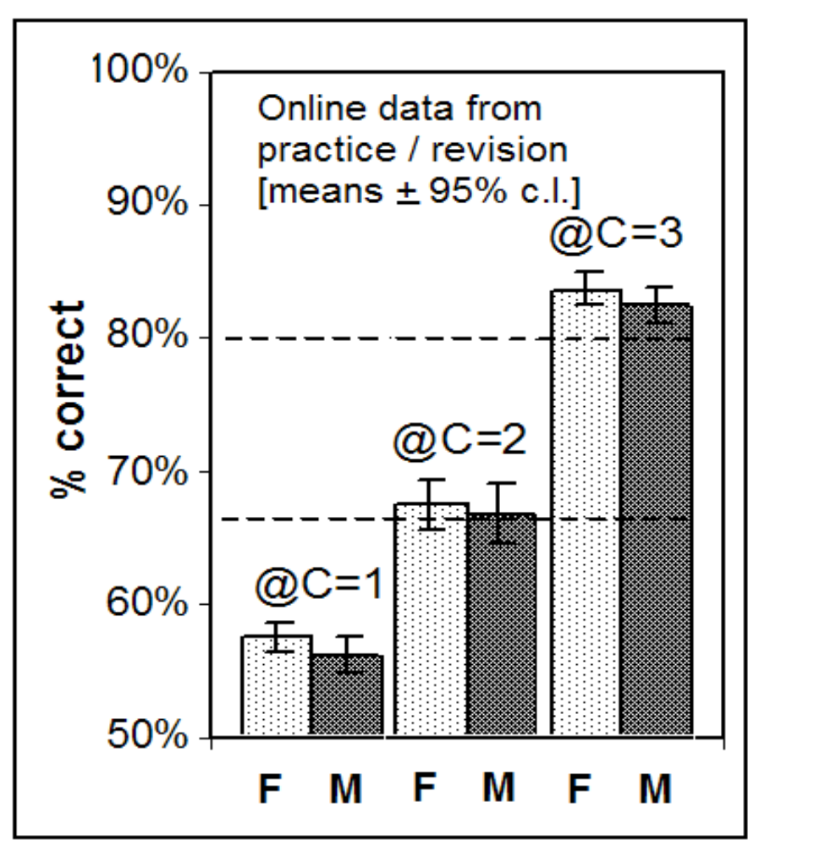


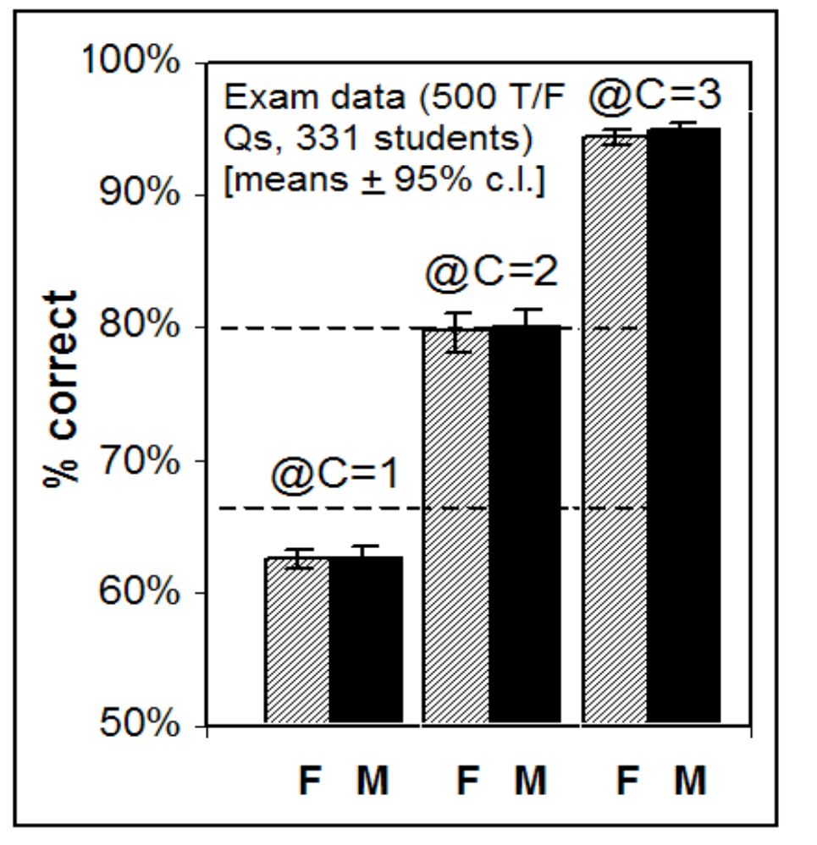
CBM Bonus =
10% x (av. CBM mark*
– mark expected with
no discrimination)

CB Accuracy =
Accuracy + CBM Bonus

** This is the "Bonus Factor" studied in the panel at top right.*

How well do students discriminate reliability?





Optimisation of certainty-based assessment scores

A.R. Gardner-Medwin

Physiology (NPP), UCL, London, UK

Certainty-Based Marking (CBM) rewards students for identifying which of their answers are uncertain and which reliable. Confident answers incur negative marking if wrong, while uncertain answers receive lower credit but no penalty. The scheme at UCL is shown in Table 1 (1,2). In self-tests this encourages reflection and deeper learning, because the student gains from evaluating both doubts and justifications. CBM gives a sound measure of knowledge (1) and enhances the statistical reliability of assessments (3,4). Problems arise, however, with student perception of CBM final scores, because average marks as a fraction of the maximum are lower than the percentage of correct answers ('accuracy'). For example, students with 80% accuracy typically get only 50% of the CBM maximum (Fig.1).

A new strategy is analysed here using data from 17 medical exams with CBM at UCL (250-300 true/false Qs, 320-360 students). In this, a CB bonus (+ or -) is added to conventional accuracy to reflect how well or badly the student has distinguished uncertain from reliable answers. This bonus is a fraction of the difference between average CB mark and the average that would have resulted with a uniform certainty level, appropriate to the student's overall accuracy. The multiplying 'Bonus Factor' (BF) is shown here to be optimally ca. 0.1 to give maximum statistical reliability for the resultant 'CB Accuracy' (accuracy + bonus). Reliability was calculated as the rank correlation (r) between scores on subsets (odd and even questions) in each exam. This increased from a mean of 0.80 with BF=0 (i.e. conventional accuracy) to 0.86 with BF=0.1 (increase = 0.060 ±0.0039), almost as high as for the CBM average (0.87). This reflects an increase of 57% in predictive power (Fig. 2a), equivalent to adding 57% more questions without CBM. The power to predict conventional accuracy was increased by 16% (Fig. 2b), showing greater validity for CBM scores even with accuracy as the standard.

The CB bonuses added to accuracy in exam data are typically 2-5%, though students doing online self-tests while studying often get negative bonuses due to misconceptions (confident errors). Bonuses and CB accuracy scores are implemented online using BF=0.1 in LAPT at UCL (2), available for use by other institutions. CBM is also available in Moodle, though additional code (5) is required to implement the enhanced scoring and feedback to students.

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
Probability correct:	<67%	67-80%	>80%	-

Table 1

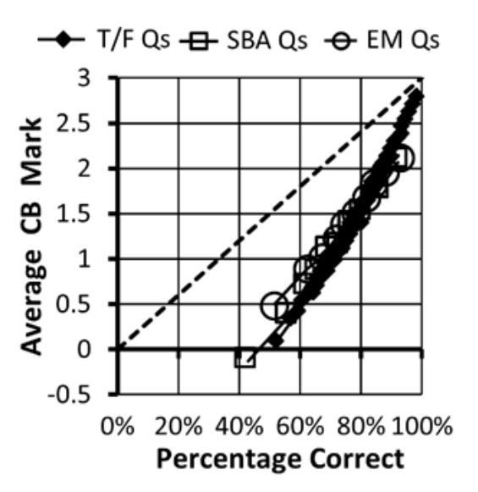


Fig. 1

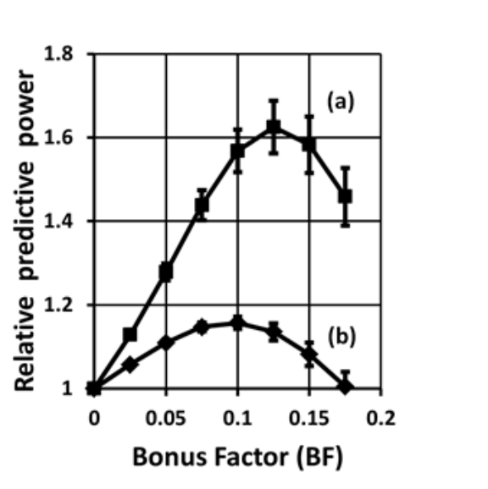


Fig. 2

Fig.1 Data from 9,000 online LAPT self-test sessions with >20 answers, showing mean CBM averages plotted against accuracy, for centiles (True/False Qs) or deciles (Single Best Answer, 4-5 options and Extended Matching Questions, 10 options) ranked by accuracy.

Fig.2 Predictive power *(=r/(1-r))* of 'CB Accuracy' scores as a function of bonus factor, calculated from *r* = Pearson rank correlation comparing (a) CB scores vs CB scores and (b) CB scores vs accuracy, for interleaved question subsets. Mean ± sem for 17 exams.

Gardner-Medwin AR (1995). *ALT-J* **3**, 80-85

London Agreed Protocol for Teaching (LAPT), http://www.ucl.ac.uk/lapt

Ebel RL (1965) *J Educ. Measurement*. **2**, 49-57

Gardner-Medwin AR & Gahan M (2003) *Proc. 7th CAA Conf. July 2003*, 147-155

Gardner-Medwin AR (2012). *CBM and Moodle*, http://www.tmedwin.net/cbm/moodle

Thanks to D.Bender for providing access to past exam data.

Where applicable, the authors confirm that the experiments described here conform with The Physiological Society ethical requirements.