Certainty-Based Marking (CBM) for Reflective Learning and Proper Knowledge Assessment

Tony Gardner-Medwin [1] & **Nancy Curtin** [2]
1 University College London, a.gardner-medwin@ucl.ac.uk,
2 Imperial College London, n.curtin@imperial.ac.uk
www.ucl.ac.uk/lapt

OVERVIEW

Certainty Based Marking (CBM) involves asking students not only the answer to an objective question, but also how certain they are that their answer is correct. The mark scheme rewards accurate reporting of certainty and good discrimination between more and less reliable answers. This encourages reflection about justification and soundness of relevant knowledge and skills, and probes weaknesses more deeply. It is easily implemented with existing test material, popular with students, grounded firmly in information theory and proven to enhance the quality of exam data. We report our experience with CBM and raise questions about constructive, fair and efficient assessment.

Keywords
Certainty, Confidence, Marking Scheme, Objective Questions, Reliability, Reflection

WHAT IS CBM?

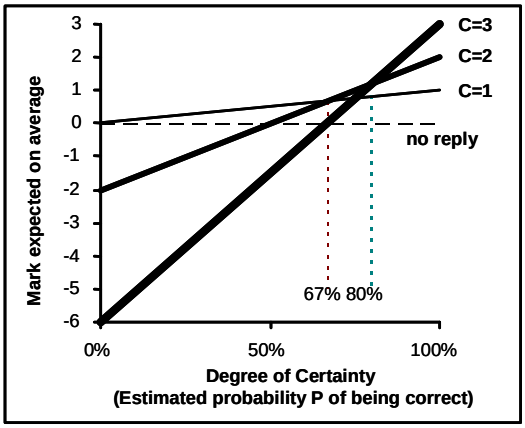
After each answer, a student indicates a degree of certainty (C) that the answer will be marked as correct, on a 3-point scale: 1 (low), 2 (mid) or 3 (high). We deliberately do not use words like 'sure' or 'very sure' because these mean different things to different people.

The best choice of C is defined by the mark scheme, which is designed so that the student is always motivated to report his/her level of certainty correctly: indicating a low level of certainty when uncertain, and vice versa (Fig. 1).

Figure 1a. Mark scheme for Certainty Based Marking

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

Figure 1b. Degree of Certainty and average expected mark



In Figure 1b, the best C level is the one that is highest at the point corresponding to your estimate of how likely you are to be correct. Each line, one for each C level, shows how your expected mark depends on your estimate of the probability that you will be marked correct. The critical transition points, to merit using C=2 or C=3, are 67% and 80%.

INFORMATION ABOUT OUR CBM USE

Our biomedical and medical students at UCL and Imperial (earlier at Charing Cross & Westminster Medical School) have been using CBM extensively for more than 10 years, to promote critical awareness and self-assessment while revising. The main current software is browser-based (www.ucl.ac.uk/lapt) with exercises on the web, on CD, or downloaded. Material is openly available in several disciplines as well as medicine, to encourage dissemination and new trials.

We run compulsory formative exercises with CBM online or using Optical Mark Reader (OMR) cards: *Speedwell Computing Services*. Most use is voluntary, however, much of it on home computers. Compulsory exercises (including Maths for medical students at UCL) are initiated within WebCT, with grades returned and recorded in WebCT. Otherwise, submission of work is merely encouraged for statistical purposes. Marking employs Javascript, run on the student's computer, so the server does not know about performance unless results are submitted. In total, submissions amount to about 1.5 million answers per year, including access from over 30 UK universities and from applicants practising for the Biomedical Admissions Test (BMAT).

UCL has used CBM for five years in medical exams, with 500-600 True/False (TF) questions contributing 40% of end-of year summative marks in years 1&2. At Imperial, in compulsory formative tests with CBM, we have been able to compare performance using TF and best-of-5 question styles.

DESCRIPTION OF PROCEDURES

Students at UCL first encounter CBM in the context of compulsory maths exercises, which they can practice as often as necessary but which they must pass eventually. This seems a good introduction, because maths is an area where students are often slapdash at first, but can learn to be more aware of when they are doing things reliably, and to check calculations or reasoning carefully. Mathematical ability also varies greatly between students. Weak or unconfident students learn to identify and build on areas where they do understand the material, identifying others where they need to seek help or think more

carefully. Some of the most appreciative initial responses actually come from those who are self-confident and able, but rapidly realise how easy it is to lose out by being careless.

The most extensive use of CBM is for formative tests and pre-exam revision. To encourage self-assessment earlier in the year alongside coursework, we use follow-up tests that are closely tied to specific practicals or classes, where performance is not recorded unless voluntarily submitted. These are well appreciated, and save staff time on marking of follow-up exercises.

Students have access to 'help' links while working with CBM, explaining the mark scheme and giving a breakdown of percentage correct achieved at different certainty levels. Students obviously need practice before exams, but it has never proved necessary to explain or discuss the mark scheme in any detail, since it is transparent and easy to remember, and the risks and benefits of opting for different C levels are at least qualitatively very clear. Issues of poor calibration in the use of C levels are discussed below. Students generally regard the procedure as helpful and fair, and both at Charing Cross & Westminster Medical School and at UCL many students suggested in evaluation surveys that they would prefer CBM in exams.

RATIONALE IN TERMS OF EDUCATIONAL IDEAS

The rationale for our use of CBM, its relation to proper measures of knowledge, and details of new developments and data analysis are published and available on the website (Gardner-Medwin, 1995, Gardner-Medwin & Gahan, 2003; Gardner-Medwin, 2006a). In this article the approach will be to pose questions raised by our experience and interactions with students and staff, paralleling to some extent a recent presentation to a Physiological Society teaching workshop (Gardner-Medwin & Curtin, 2006). We hope this will provoke more discussion. Points 1-8 below, concerning general issues about objective testing, are offered provocatively and without argument. Readers may either react to them from their own perspective or (1-4) look at our slides from the workshop (www.ucl.ac.uk/lapt/UCL06_tw.pdf) to read our views. Subsequent points, specifically about CBM, are presented here in more detail. We start by considering the general rationale of objective testing and CBM.

Of course we all want student learning to be more effective and less extravagant in staff time. Part of a strategy for this can involve self-assessment tasks alongside teaching material, wherever possible challenging deeper knowledge than simply factual or associative learning. Indeed in this sense, self-assessment material is teaching material. 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 can be less effective at probing weaknesses than face-to-face confrontation or feedback on student scripts. Students who get an answer right often think they knew the answer, when in fact all they did was plump for the most 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. Worse than this, we think it encourages sloppy habits of thought in students.

CBM differentiates between different students who 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 it rewards reflection that leads to the conclusion that an answer is less certain than initially thought. The approach has a basis in probabilistic decision theory, but students find it intuitively easy to use, and cannot cheat by misrepresenting their certainty. Brains have evolved to make decisions under uncertainty, in the context of potential risks and benefits. This is an important, intuitive task in intellectual as well as everyday endeavours. Accurate

expression of reliability is therefore recognised as a fundamental part of discourse in every discipline.

We certainly don't advocate computer-marked tests, even with CBM, as an ideal or sole form of assessment. But in large classes, especially where there is critical core material as in medicine, there is no option but to use them as a substantial component of assessment, and particularly of self-assessment to support learning. We must use them in the best possible way. Other assessments can be more probing, but unless carried out on an extravagant scale they are bound to be based on small samples of student knowledge and are therefore limited in reliability. This is no reason to omit such assessments: they stimulate deeper learning by the fact that students need to prepare thoroughly for them. But computer-marked tests are necessary to cover the range of a syllabus efficiently. Scepticism and inertia are rife in universities, so we encounter many proffered reasons (or perhaps excuses) for continuing familiar practices rather than experimenting with objective testing or CBM. We start with some general conclusions we have arrived at, which we know will be provocative to some people:

1. ***Objective testing need NOT simply test factual knowledge and encourage rote learning.***
2. ***Objective testing is for some (not all) purposes BETTER assessment than essays or problems.***
3. ***The notion that you should use 'modern' question formats like single-best-answer or extended matching questions rather than 'outdated' True/False questions is often generalised far beyond any valid supporting evidence we know of. T/F questions are often BEST PRACTICE.***
4. ***It is (common) BAD PRACTICE to include a 'Don't Know' option with T/F or Best-Option Qs.***

Next are some more specific opinions about objective testing that seem very strange to us, couched in a form that we do *NOT* agree with, though we can't claim much experience or evidence for our scepticism. Again we would welcome discussion, and relevant evidence:

5. ***All forms of negative marking are de-motivating to students. You must use carrots, not sticks.***
6. ***Objective testing has no place in subjects like social science or psychology***
7. ***True/False questions are harder to write than Multi-Choice questions***
8. ***A uniform question type in exams should be used, to avoid confusing students***

Now we get to specific reactions to CBM. Staff at conferences usually react with enthusiasm to the concept of CBM. But when you ask why they don't use it, here are some of the answers. We respond briefly to each of the assertions we challenge.

9. ***CBM assesses something different from knowledge, more about personality than ability***

   CBM would indeed disadvantage students who tend always to be either confident or diffident about answers, regardless of how well they can justify them. Reliability judgment is however part of any sensible knowledge measure. Analysis of students' judgments shows that, after formative practice, very few are poorly calibrated, with no evidence for gender differences (Gardner-Medwin & Gahan, 2003). In exams (when students tend to be slightly less confident than online) we can to some extent compensate for poor calibration with automatic adjustments (Gardner-Medwin & Gahan, 2003). Overly self-confident or diffident students need to become more self-aware if they are not to be handicapped in academic work, and practice with CBM should help with this.

10. CBM encourages students to 'play the system'

A somewhat inscrutable comment, but common! CBM is indeed a system, designed so that success requires the ability to distinguish reliable from unreliable answers. Gambling without knowledge simply does not pay, which is a lesson that students rapidly learn after expressing confidence for a few uncertain answers.

11. CBM seems appropriate in medicine, but less so in other fields

It is easy to argue that judgment of the reliability of one's conclusions is a matter of life and death in medicine. But the same is true in many fields, from engineering to politics, and accurate judgment is crucial to success in any field.

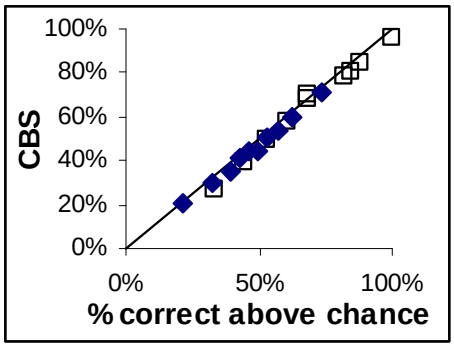
12. CBM might make standard setting more difficult, and this already causes us enough grief!

Standard setting can be problematic with any marking scheme. Much care has gone into the design of a Certainty-Based Score (CBS) for tests using CBM, so that standards set with conventional mark schemes correspond directly to CBS equivalents. The principles for this are:

(1) to scale both conventional and CBM scores so that 0% = the level expected by chance guessing and 100% = totally correct, confident performance,
(2) to linearise the relation between CBS and conventional scores (applying an empirically determined power-law scaling of CBS : www.ucl.ac.uk/lapt/laptlite/sys/lpscoring.htm) so that average marks over the full range (0% - 100%) are equivalent (Fig. 2).

At Imperial College we tested this with formative exercises combining TF and Best-of-5 question styles. On average the two types of score on each component were very close (Fig. 2), though of course individual students with the same % correct will have gained better or worse than average CBS scores, depending on how well they distinguished reliable and uncertain answers.

Figure 2. Equivalence of CBS and conventional scores for standard setting



In Figure 2. for the True/False (filled symbols) and best-of-5 (open symbols) components of a formative test, 345 students were ranked by conventional scores. Then for each decile, mean CBS scores are plotted against % correct above chance. The line corresponds to equality between the two scores.

13. CBM is not available on my favourite tool (e.g. Questionmark, Blackboard/WebCT or Moodle)

It would be great if these vendors or developers would incorporate it! At present you can run CBM exercises from within a VLE by using links to the external software at UCL. In Moodle and WebCT, most question types can be exported automatically to the format required to run in this way (see www.ucl.ac.uk/lapt/laptlite). Grades can be accessed by authenticated students or staff outside the VLE or, where facilities exist, they can be uploaded to the VLE. Full adaptation of VLE interfaces to embed CBM would require intimate knowledge of the individual program structures. We believe that the benefits of CBM are such that e-learning tools that fail to incorporate CBM will eventually lose out.

14. I tried CBM in a tutorial and my students said they didn't like it

Though not our experience, we can imagine how this can happen. Firstly, experience with just a few questions may fail to overcome an inappropriate initial 'all-or-none' approach to certainty. Secondly, though group discussion about reliability of answers is very productive, group dynamics can initially produce responses that are more socially determined than thought through - potentially embarrassing to those who initiate them. The best practice with CBM is probably individual work, or in groups of 2 or 3. Our students usually experience hundreds or thousands of questions in the course of study, leading to familiarisation with certainty judgments as second nature.

15. CBM requires extra time in an assessment

Possibly, but not much. We have no specific data on this, but in an evaluation study (Issroff & Gardner-Medwin, 1998) many students said they sometimes changed their answer while thinking about their certainty, and would appear therefore to have been making good use of extra time. Reliability judgments do tend to emerge automatically alongside any constructive thinking, as a sort of gut feeling, so any slowing of performance under pressure may be minimal. Reliability analysis of exams (Gardner-Medwin, 2006b) suggests in any case that the number of questions needed for equally reliable assessment data can fall by a third or more with CBM compared with conventional marking.

16. CBM is only appropriate with True/False questions

This notion may have mistakenly arisen because at UCL and Imperial our principal experience is with True/False questions, traditionally employed in many of our medical assessments. Though we have not used other styles with CBM in summative exams, use of CBM with numerical, best-of-5 and extended matching questions has shown no special problems with these styles. Practice is of course always needed with new styles, and appropriate scaling is required (12 above) to deal with chance performance and make scores comparable.

EVALUATION

Two evaluation studies for CBM at UCL have been published (Issroff & Gardner-Medwin, 1998; Longstaffe & Bradfield, 2005), and there is an interview transcript for a JISC case study on assessment available at www.ucl.ac.uk/lapt/jisc_transcript.doc .

REFERENCES

Gardner-Medwin AR (1995) *Confidence Assessment in the Teaching of Basic Science*. ALT-J (Association for Learning Technology Journal) 3:80-85

Gardner-Medwin AR (2006a) Confidence-Based Marking - towards deeper learning and better exams In : Innovative Assessment in Higher Education. Ed.: Bryan C and Clegg K. Routledge, Taylor and Francis Group, London

Gardner-Medwin AR (2006b). Analysis of exams using certainty-based marking. *Physiological Society Main Meeting, UCL*, Proc Physiol Soc series. , 3, PC64

Gardner-Medwin AR & Curtin NA (2006). Certainty-based marking at UCL and Imperial College. *Physiological Society Teaching Workshop*, Proc Physiol Soc series. , 3, WA4

Gardner-Medwin AR & Gahan M (2003) *Formative and Summative Confidence-Based Assessment* Proc. 7th International Computer-Aided Assessment Conference, Loughborough, UK, July 2003, pp. 147-155

Issroff K & Gardner-Medwin AR (1998) *Evaluation of Confidence Assessment within Optional Coursework*. In Innovation in the evaluation of Learning Technology. Ed. M.Oliver, Univ. N. London Press, pp169-179.

Longstaffe JA & Bradfield JWB (2005) A review of factors influencing the dissemination of the London Agreed Protocol for Teaching (LAPT) - a confidence based marking system. www.ucl.ac.uk/lapt/CBM_review.doc

Please reference as:

Gardner-Medwin, A.R. & Curtin, N.A. (2007). Certainty-Based Marking (CBM) For Reflective Learning And Proper Knowledge Assessment. *From the REAP International Online Conference on Assessment Design for Learner Responsibility, 29th-31st May, 2007*. Available at http://ewds.strath.ac.uk/REAP07

Re-Engineering Assessment Practices in Scottish Higher Education (REAP) is funded by the Scottish Funding Council under its e-Learning Transformation initiative. Further information about REAP can be found at http://www.reap.ac.uk