Confidence-Based Marking - towards deeper learning and better exams
A.R. Gardner-Medwin
Dept. Physiology, University College London, London WC1E 6BT

[Draft for Chapter 12 in : Bryan C and Clegg K (eds) (2006) Innovative Assessment in Higher Education, Routledge, Taylor and Francis Group Ltd, London]

Introduction

This chapter looks at experience with confidence-based marking (CBM) at UCL over the last 10 years. The CBM strategy was initially introduced to improve formative self-assessment and to encourage students to think more carefully about questions in objective tests. It became known as LAPT (London Agreed Protocol for Teaching: www.ucl.ac.uk/lapt) through collaboration between a number of medical schools now mainly subsumed within UCL and Imperial College London. We have recently developed web-based tools for dissemination and evaluation in other institutions and new disciplines, and since 2001 we have been using CBM at UCL for summative (year 1,2) medical exams. CBM is seen by our students as simple, fair, readily understood and beneficial. They are motivated to reflect and justify reasons either for confidence or reservation about each answer, and they gain by expressing true confidence, whether high or low.

Experience with students suggests that many of the stronger ones find they can do well by relying on superficial associations, with little incentive (on conventional right/wrong mark schemes) to think rigorously or understand the issues thoroughly: the first thing they think of is usually good enough. At the same time, weaker students try to emulate this with diligent rote learning, rejecting deeper learning as unnecessarily challenging. A properly designed scheme for CBM ensures that in order to get the best marks students must discriminate between responses based on sound knowledge or understanding, and those where there is a significant risk of error. The essence of such a motivating scheme is that confident answers gain more marks if correct, but at the risk of significant penalty if wrong; low confidence benefits the student when there are reasons for reservation, because the penalties are proportionately less or absent. We shall see later how these constraints on the design of a proper marking scheme can be understood in graphical terms (Fig. 1), but the essential feature is that the students who benefit are those who can identify the basis for justification or reservation, not those who are consistently confident or unconfident.

This chapter will refer to previous publications where specific points are covered there in more detail (Gardner-Medwin, 1995, Issroff & Gardner-Medwin, 1998, Gardner-Medwin & Gahan, 2003). Readers are strongly encouraged, before thinking far about the issues, to try out confidence-based marking for themselves on the website (www.ucl.ac.uk/lapt) using any of a range of exercises in different fields. Experience shows that students approaching CBM as learners seem to understand its logic instinctively through application, much more readily than through exposition and discussion. The website also provides access to prior publications and authoring tools[1].

Confidence-based marking has been researched quite extensively, but mostly before computer aided assessment was practical on much of a scale (see, for example, Ahlgren, 1969; Good, 1979). Our experience is, we believe, the largest application to routine coursework, and differs from most research studies in that students have had extensive online practice before use in formal tests. We use a 3-point confidence scale (C=1,2 or 3); whenever the answer is correct, the mark (M) is equal to its associated confidence level: M=1,2 or 3. If the answer is wrong, then at the higher confidence levels there are increasing penalties: -2 marks at C=2 and -6 marks at

C=3. This scheme is set out in Table 1. For example, a student asked whether Russia has a Baltic coast would need either sound knowledge or a convincing argument to risk C=3.

Confidence levels are deliberately identified by numbers (C=1,2 or 3) or neutral descriptors (low, mid, high) rather than by descriptive terms ("certain", "very sure", "unsure", "guess", etc.), because descriptions mean different things to different people and in different contexts. A transparent mark scheme must be defined by rewards, penalties and explicit risks, not by subjective norms.

TABLE CH13.1 NEAR HERE

The rationale of CBM: the student's perspective

Several qualitative features of pedagogic importance are immediately clear to a student when thinking about answering a question with CBM. These relate directly to the second and fifth 'principles of good feedback practice' (reflection and motivation) set out in Chapter 5.

1. To get full credit for a correct answer you must be able to justify the answer, to the point that you are prepared to take the risk that - if wrong - you will lose marks. This makes it harder to rely on rote learned facts, and encourages attempts to relate the answer to other knowledge.

2. Equally, if you can justify reasons for reservation about your answer you also gain credit, because with a higher probability of error you will gain on average by lowering your confidence. This is the *motivating* characteristic of the mark scheme (Good, 1979).

3. A lucky guess is not the same as knowledge. Students recognize the fairness and value of a system that gives less credit to a correct answer based on uncertain knowledge than to one that is soundly justified and argued. Teachers should recognize this too.

4. A confident wrong answer is a wake-up call deserving penalty. When studying, this triggers reflection about the reasons for error, and particular attention to an explanation. In exams, it merits greater penalty than a wrong answer that is acknowledged as partly guesswork.

5. To quote comments from an evaluation study (Issroff & Gardner-Medwin, 1998) Error: Reference source not found : "It .. stops you making rush answers.", "You can assess how well you really understand a topic.", "It makes one think .. it can be quite a shock to get a -6 .. you are forced to concentrate".

These points encapsulate the initial reasons for introducing CBM. Unreliable knowledge of the basics in a subject, or - worse - lack of awareness of which parts of one's knowledge are sound and which not, can be a huge handicap to further learning (Gardner-Medwin, 1995). By failing to think critically and identify points of weakness, students lose the opportunity to embed their learning deeply and to find connections between different elements of their knowledge. It is distressing to see students with good grades in GCSE mathematics struggling two years later to apply half-remembered rules to issues that should be embedded as common-sense understanding - such as the percentage change when a 20% fall follows a 50% rise.

Students benefit by learning that there are different ways to solve problems, and that efficient learning and rigorous knowledge involves the skill of always testing one idea against another. Only then can one be said to have 'understanding' of a subject. What is more, the ability to indicate confidence or reservation about opinions to others, explicitly or through body language, is an essential skill in every discipline and walk of life. These skills may nevertheless remain

largely untaught and untested in assessments before final undergraduate or graduate tests, when they become crucial in viva situations and in demanding forms of critical writing.

Is there a correct CBM mark scheme?

A student's best choice of confidence level (C=1, 2, or 3) is governed by two factors: confidence (degree of belief, or subjective probability) that the answer will be correct, and the rewards (or penalties) for right and wrong answers. The average marks obtained with our scheme (Table 1) are plotted in Fig. 1 for each confidence level against the probability of being correct.

Fig. CH13.1 NEAR HERE

One always stands to gain the best overall score by choosing a confidence level (C=1,2 or 3) for each answer that corresponds to whichever line is highest on the graph, above the point showing one's estimated probability of being correct. It is best to opt for C=1 if this probability is less than 67%, C=2 if it is 67-80%, and C=3 if it is greater than 80%. It doesn't pay to misrepresent one's confidence. This is the motivating characteristic of the mark scheme, rewarding the student's ability to judge the reliability of an answer, not their self-confidence or diffidence. CBM is quite unlike games such as poker, in which misrepresentation of confidence can gain advantage.

At UCL we have used this scheme mainly for questions with True/False answers. With just two possible answers, the estimated probability of being correct can never be less than 50%, since if it were, then one would obviously switch choices. The three confidence levels cover the possible range of probabilities (50-100%) fairly evenly. For questions with more options (a typical MCQ question, or one requiring a numeric or text entry) a preferred answer may be given with an estimated probability of being correct that is lower than 50%. For such questions we have experimented with a scheme similar to Table 1, but with lower penalties: -1 at C=2 and -4 at C=3. The graphs for this scheme (Fig. 2a) show it also to be properly motivating, with incentives to use C=1 when confidence is low (<50%) and C=3 when it is high (>75%). This gives more uniform coverage of the full probability range, from 0-100%. Data show that students can adapt confidence judgments with multiple-choice questions to whichever mark scheme is employed[2], so the better strategy in the future may actually be to avoid complexity by keeping to a single mark scheme (Table 1) for all question types.

Fig. CH13.2 NEAR HERE

Not all CBM schemes in the literature have been properly motivating schemes (see discussion in Gardner-Medwin & Gahan, 2003). Fig. 2b shows a superficially similar scheme (actually incorrectly attributed to LAPT: Davies, 2002) with penalties (-1, -2, -3) instead of (0, -2, -6). This scheme has the merit of simplicity. However, inspection of the graph shows that it is always best either to opt for high confidence (C=3) or not to answer at all. This shows the importance, if one is devising a new CBM scheme, of plotting such graphs. Students who use confidence levels 1 or 2 on this scheme may be following their teachers' advice, but would be quite disadvantaged. They would gain lower scores than students who were brash, or who saw that this strategy was never sensible. Though such a scheme could have some of the benefits of CBM by encouraging students to reflect, and by reducing the weighting of unconfident answers, it could never be seen as fair, given that it benefits students who only use C=3. It could not survive once the best strategy became known.

The schemes used in LAPT at UCL are not the only ones that are properly motivating, but they are simple and easily remembered and understood. They were also chosen because the scores with T/F questions (Table 1) correspond about as closely as can be achieved with a 3-point

scale to the correct measure of knowledge as a function of subjective probability that derives from information theory (Fig. 1 in Gardner-Medwin, 1995). A more complex scheme was devised and used by Hassmen and Hunt (1994) with 5 confidence levels and marks for correct answers (20, 54, 74, 94, 100) and for wrong answers (10, -8, -32, -64, -120). This scheme is in principle motivating (Gardner-Medwin & Gahan, 2003) but it is hard to remember and understand. Perhaps because of this, it has sometimes been used without the student being aware of the marks associated with different confidence levels[3]. Again, it may encourage reflection, but lacks the simplicity and transparency that seem critical if one is to obtain full engagement of students in a system designed to improve study habits and assessment.

Students rarely describe their choice of confidence level in terms of explicit probabilities, even after the principles are explained. Watching students (and particularly staff) in a first encounter with CBM, it is common to find some who initially regard anything less than C=3 as a diminution of their ego. In group working, confidence is often determined by one person, until others start to realize that a little thought can do better than a forward personality! Brashness does not long survive a few negative marks, nor diffidence the sense of lost opportunities. Despite their intuitive approach, students on average come to use the confidence bands in a nearly optimal fashion to maximize their scores. In exams, few students have percentages correct in the three bands that are outside the correct probability ranges (Gardner-Medwin & Gahan, 2003). This is consistent with the general finding that though people are poor at handling the concept of probability correctly, they make good judgments when information is evident as clear risks, outcomes and frequencies (Gigerenzer, 2003). This is perhaps the chief benefit of a simple, transparent scheme. However, our data do suggest that students initially vary more widely in their ability to calibrate judgments according to the mark scheme. Formative practice with feedback is therefore essential. Our system ensures that in addition to immediate feedback after each answer to assist reflective learning, students also receive a breakdown of % correct at each confidence level whenever they complete an exercise.

Concerns about CBM: Why don't more people use it?

Despite the clear rationale for CBM, its student popularity and statistical benefits in exams (see below), CBM has surprisingly little uptake nationally in the UK or globally. In our dissemination project (www.ucl.ac.uk/lapt) we provide tools for new institutions and teachers to experiment with their own materials and students, and to interface with their own VLE. An interesting feature of dissemination within UCL has been the stimulus to uptake within medical science courses that has come from the students themselves - often a much more potent force for change in a university than discussion amongst staff. However, it is worth addressing misconceptions that sometimes emerge in discussion.

It is sometimes thought that CBM might unfairly favour or encourage particular personality traits. In particular, it is suggested (though usually vigorously rejected by students) that CBM may lead to gender bias, usually based on the notion that it might disadvantage diffident or risk-averse personalities - supposedly more common among females. Several points can be made about this (discussed at greater length by Gardner-Medwin & Gahan, 2003). Careful analysis of data from exams and in-course use of CBM at UCL has shown no gender differences, despite clear differences (in both sexes) between summative and formative use (more cautious use of high confidence levels in exams, with consequently higher % correct when using these levels : Fig. 3). We have correlated exam data also to ethnic status, and again there is no evidence for significant ethnic variation among practised students. Gender and ethnic factors may be present on first encounter with CBM; such transient effects would not have shown up in our analysis. But if, quite plausibly, individuals or groups do have initial tendencies to be under- or over-confident, then this is an objective problem that they should be aware of. CBM offers suitable feedback and training. This is not to say that outwardly diffident or confident personalities are

undesirable or unattractive, but rather that in decision-rich occupations such as medicine, mis-calibration of reliability is a serious handicap.

Fig. CH13.3 NEAR HERE

A second misconception is that the aim of CBM is somehow to boost self-confidence. Of course self-confidence should ultimately be boosted by any effective learning tool. Students often say, in evaluation questionnaires, that the LAPT system with CBM has improved their confidence in their knowledge (Issroff & Gardner-Medwin, 1998)[4]. But they also say, and this seems pedagogically more important, that it forces them to think more, and reveals points of weakness. CBM places a premium on understanding: on the ability to link and cross-check pieces of information, and therefore to distinguish sound and weak conclusions. The net effect may be to undermine confidence as students come to realize that sound knowledge cannot be based on hunches. But this realization is itself a step towards academic success and the building of self-confidence.

Some people suggest that CBM may be useful for formative assessment but not for exams, supposedly because what matters are correct answers, not students' confidence in their answers. From an epistemological standpoint, this seems simply wrong: a lucky guess is not knowledge, and a firm misconception is far worse than acknowledged ignorance. My view is that we fail our students if we don't acknowledge this. But we can also regard the issue simply as a psychometric one. Our exam data shows improved reliability using CBM[5], consistent with research data with other forms of CBM (Ahlgren, 1969). Some of this improvement is due to involvement of an extra skill that varies between students: ability to handle CBM. But even if one adopts number correct as the criterion of performance, our data showed that CBM scores are the better predictor of this on a separate set of questions, and are therefore both more valid and more reliable as a measure of knowledge.

Confidence-based marking places a premium on careful thinking and on checks that can help tie together different facets of knowledge. It thereby encourages deeper understanding and learning. It is popular with students, and helps them develop valuable skills. In exams it produces higher quality data for their assessment. A puzzle remains, why this seemingly sensible strategy for objectively marked tests, which has been known and researched over many decades, is so little employed by teachers or vendors of software.

Acknowledgements

Supported by the Higher Education Funding Council for England, under the Fund for the Development of Teaching and Learning, Phase 4. Some of the software was written by M. Gahan. Data for analysis were provided by D. Bender at UCL and N. Curtin at Imperial College.

References

Ahlgren A. (1969) Reliability, predictive validity, and personality bias of confidence-weighted scores. <www.p-mmm.com/founders/AhlgrenBody.htm>

Davies P. (2002) There's no confidence in Multiple-Choice Testing, ... *Proceedings of the 6th International CAA Conference*, Loughborough, pp 119-130

Gardner-Medwin AR (1995) Confidence assessment in the teaching of basic science. *Association for Learning Technology Journal* 3:80-85 1995

Gardner-Medwin A.R., Gahan M. (2003) Formative and Summative Confidence-Based Assessment. *Proceedings of the 7th International CAA Conference*, Loughborough University, UK, pp. 147-155 (www.caaconference.com)

Gigerenzer G. (2003) *Reckoning with Risk*. Penguin Books, London UK, 310 pp.

Good I.J. (1979) "Proper Fees" in multiple choice examinations. *Journal of Statistical and Computational Simulation* 9,164-165

Hassmen P, Hunt DP (1994) Human self-assessment in multiple-choice testing. *Journal of Educational Measurement* 31, 149-160.

Issroff K., Gardner-Medwin A.R. (1998) Evaluation of confidence assessment within optional coursework. In : Oliver, M. (Ed) *Innovation in the Evaluation of Learning Technology*, University of North London: London, pp 169-179

Confidence level :	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)

Table CH13.1: The normal LAPT Confidence-Based Mark scheme

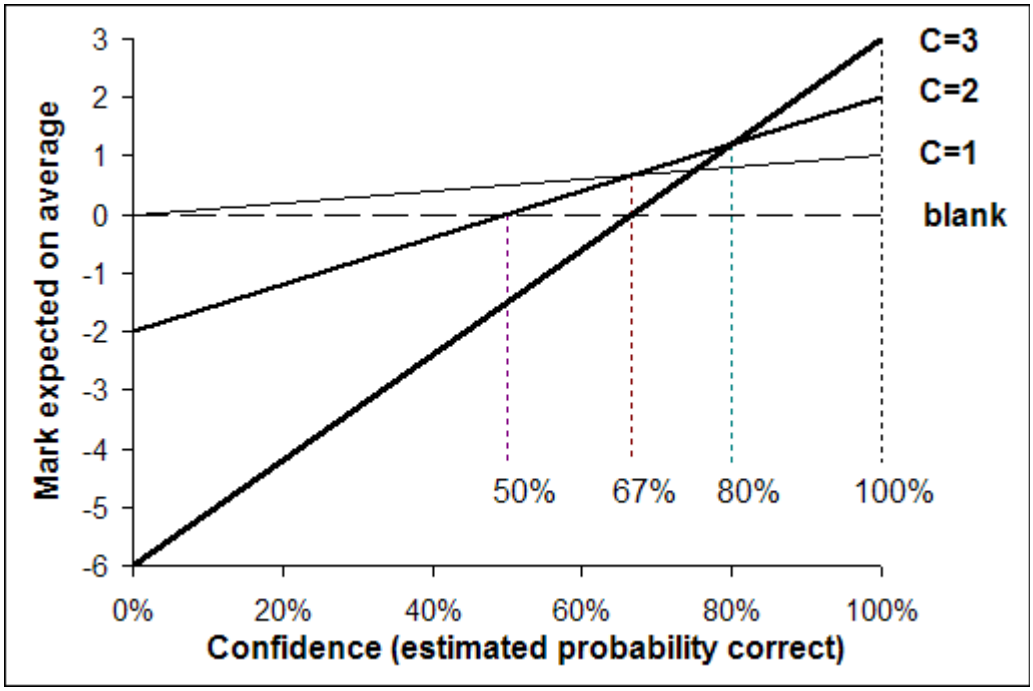
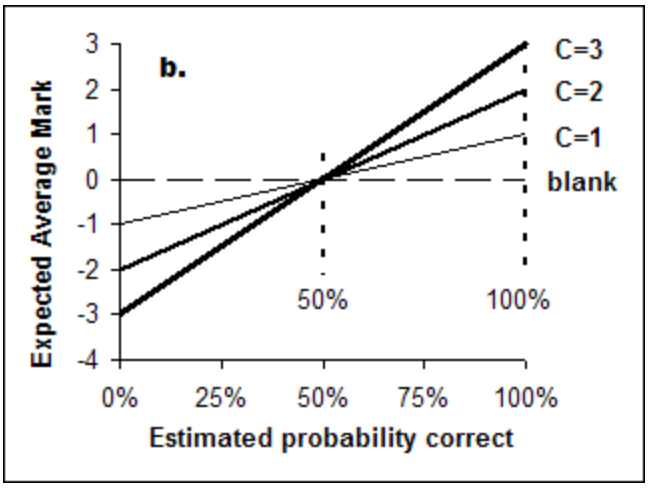


Fig. CH13.1. Rationale for a choice of confidence level. The average mark, expected on the basis of a student's estimated probability of an answer being correct, is shown for each confidence level and for a blank reply, with the mark scheme in Table 1. The best choice for a particular estimated probability of being correct is the one with the highest graph.



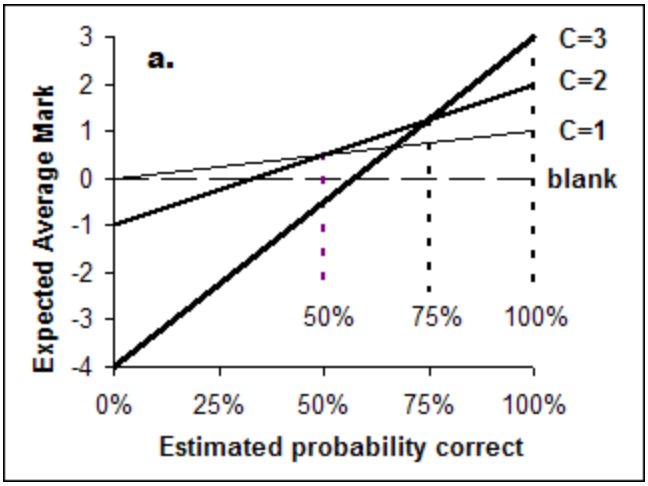


Fig. CH13.2. Characteristic graphs for different CBM schemes. These graphs are equivalent to Fig. 1, but for a scheme used on a trial basis in LAPT for questions with more than 2 possible answers and for a non-motivating scheme used elsewhere (for details, see text).

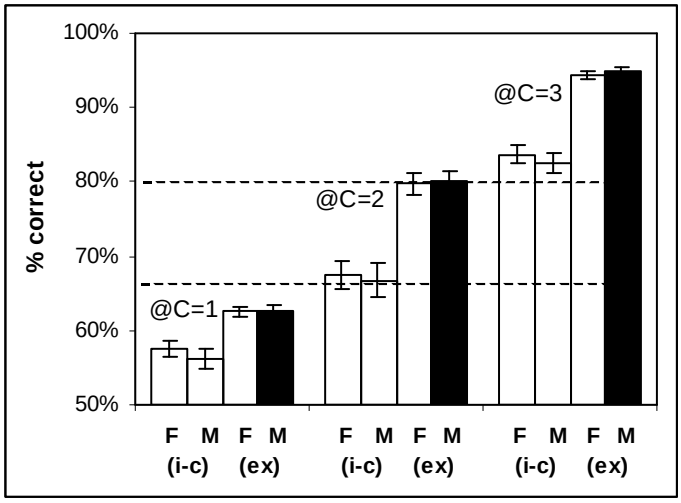


Fig. CH13.3. Performance broken down by confidence, context and gender. Accuracy (mean % correct) is shown at each confidence level for T/F answers entered in voluntary in-course (i-c) exercises (mean 1005 Qs) and in exams (ex) at the end of year (500 Qs), separated by gender (190F, 141M). Bars are 95% confidence limits for the means. Differences between exams and in-course work are significant at each confidence level (P<0.01%) but gender differences are not significant (Gardner-Medwin & Gahan, 2003).

[1] For study and revision purposes we employ either dedicated Windows software or a browser-based resource written in Javascript. The browser system now offers equivalent flexibility and more material, in more subject areas; it will be the main platform for future development and collaborations, and can be integrated with grade management in a virtual learning environment (VLE). Other institutions may use the UCL software to run exercises placed on their own or a UCL site. For exams we use optical mark reader (OMR) technology, and for trial purposes UCL currently offers a processing service, saving initial investment in hardware and software, through collaboration with Speedwell Computing Services (www.speedwell.co.uk).

[2] Data from 10,000 (mostly multiple-choice) answers to practice questions for a Biomedical Admissions Test (see www.ucl.ac.uk/lapt), carried out mainly by inexperienced CBM users, gave mean percentages at each confidence level that were within the appropriate optimal bands (Figs. 1 or 2a) when the exercises were made available at different times using the two schemes.

[3] E.g. SACAT: Self Assessment Computer Analyzed Testing, online at www.hpeusa.com

[4] 67% of students rated CBM 'useful' or 'very useful'. 40% said they sometimes changed their answers when thinking about their confidence. Additional comments: ‘If [score is] high - gives more confidence’; '[I] revise areas where confidence is weak’; ‘Useful guide to level of knowledge and real ability’ (Issroff & Gardner-Medwin, 1998).

[5] In data from six medical exams, each with ca. 350 students and 250-300 true/false questions (in groups of five, on related topics), we calculated reliability indices for both CBM scores and numbers correct, using the standard (Cronbach Alpha) measure. These were 0.925 ± 0.007 for CBM (mean ± SEM, n=6) and 0.873 ± 0.012 for numbers correct. The significant improvement ($P<0.001$, paired t-test) corresponded to a reduction of the chance element in the variance of exam scores from 14.6% of the student variance to 8.1%. We also correlated the two types of score on half of the questions (odd numbers) for their ability to predict the number correct on the remainder. CBM scores were substantially better predictors (Gardner-Medwin & Gahan, 2003).