Confidence-Based Marking - towards deeper learning and better exams

A.R. Gardner-Medwin

Dept. Physiology, University College London, London WC1E 6BT

Abstract

Experiences with confidence-based marking for formative self-assessment over 10 years, and for exams over 3 years, are surveyed from the perspectives of pedagogy, motivation and statistics. Recent developments to encourage dissemination and collaboration are described. The system employed, in which students rate confidence in each answer on a 3 point scale by computer, via the web or on an optical mark reader card, is simple, popular, fair and readily understood. In order to gain the most marks, students must reflect and attempt to justify confidence in an answer, and report their confidence honestly and accurately. A critical point is that they benefit either by finding reasons to place greater reliance on an answer or by seeing reasons for reservation. This places a premium on careful thinking, and on checks and the tying together of different facets of knowledge, thereby encouraging deeper learning. In exams it generates higher quality data than conventional scores, with greater statistical reliability and validity as a measure of knowledge, and less contamination from chance factors associated with weak and uncertain knowledge. The puzzle remains, why this seemingly sensible strategy for objectively marked tests is so readily embraced by students and yet so little used by teachers.

Introduction

Confidence-based marking (CBM) in which a confidence rating is taken into account in the marking of each answer, was introduced at UCL in 1994 (Gardner-Medwin, 1995). It was set up to help improve students' study habits, in a protocol devised for several medical schools - now mainly subsumed within UCL and Imperial College. It became known as LAPT: London Agreed Protocol for Teaching - www.ucl.ac.uk/lapt. The idea was to make students think more carefully about how they arrive at an answer to a question, and the reliability of the various elements, instead of responding on the basis of superficial knowledge or rote learning. Strong students often find that they can get most answers right by relying just on superficial associations, with therefore little incentive (on conventional right/wrong mark schemes) to think rigorously or understand the issues thoroughly. At the same time, weaker students try to emulate this with rote learning and may reject deeper learning as unnecessarily challenging. With CBM it is not enough to be right most of the time to get good marks. The student must be able to discriminate between those responses based on sound knowledge or understanding, and those where there is a significant risk of error.

A properly designed scheme for CBM ensures that students benefit either by establishing sound reasons why an answer is likely to be correct, or by identifying issues that cast doubt on the answer - justifying reservation and low confidence. A student must always benefit by expressing their true confidence, whether this is high or low. The essence of such a motivating scheme is that a confident answer gains more marks if correct, but risks significant penalty if wrong; low confidence is preferred when there are reasons for reservation, because the penalties are proportionately less or absent. We shall see later how the constraints on the design of a proper marking scheme can be easily understood in graphical terms, 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 article surveys some pedagogic, motivational and statistical issues underlying practice at UCL. It refers to previous publications when points are covered there in more detail (Gardner-Medwin, 1995, Issroff & Gardner-Medwin, 1998, Gardner-Medwin & Gahan, 2003). Readers are 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.

For study and revision purposes we employ both dedicated Windows software (LAPT-pc) and a browser-based resource (LAPT-lite) written in Javascript for more ready access. The browser system now offers equivalent flexibility and more material, in more subject areas and for a variety of levels of education. It will be the main platform for future development and collaborations, and can be integrated with grade handling within a virtual learning environment (VLE). Since 2001, we have (partly encouraged by student preference) used confidence-based marking for parts of the summative exams in the medical curriculum at UCL, using the same marking scheme but with optical mark reader (OMR) technology. Facilities for running trial CBM tests using OMR sheets, without the initial costs of investing in hardware and software, are available through collaboration with Speedwell Computing Services (www.speedwell.co.uk), via the LAPT website (www.ucl.ac.uk/lapt).

Confidence-based marking has been researched quite extensively, mostly before computer aided assessment was readily available (see, for example, Ahlgren, 1969). Experience at UCL and ICL is probably, however, the largest project where it has been used for routine teaching, learning and assessment. Our experience is different from some research studies in that our students have had extensive online practice with the system before use in tests. The LAPT scheme uses a 3-point confidence scale (C=1,2 or 3), with a judgment made after each answer in a test. When the answer is correct the mark (M) is equal to the confidence level: M=1,2 or 3. If the answer is wrong, then M=0, -2 or -6 according to confidence level. It is clear to the student that however low one's confidence, it is always best to enter an answer at C=1 rather than to refrain from answering: there is always a chance of gaining a mark. With C=2 and C=3 there are progressively greater rewards for correct answers, but with increasing risk if wrong. This makes it only worthwhile to opt for the higher levels if one believes that the probability of being right is greater than 67% (for C=2) and 80% (for C=3), as illustrated later in Fig.1. The mark scheme is set out in Table 1. Confidence levels are deliberately identified by numbers (C=1,2 or 3) or just neutral descriptors (low, mid, high) rather than descriptive terms such as "certain", "very sure", "unsure", "unconfident", "guess", etc., because descriptive terms mean different things to different people and in different contexts. It is important that the mark scheme is transparent and defined by rewards, penalties and explicit risks, not by subjective linguistic norms.

Confidence level:	C=1 (low)	C=2 (mid)	C=3 (high)	No Reply
Mark if correct:	**1**	**2**	**3**	**(0)**
Penalty if wrong (T/F Q)	**0**	**- 2**	**- 6**	**(0)**

Table 1: The normal LAPT Confidence-Based Mark scheme

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. The fundamental points are as follows:

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 recognise the fairness and value of a system that rewards a correct answer based on uncertain knowledge less than one that is soundly justified and argued. Teachers should recognise 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 where 67% of students rated CBM useful or very useful (Issroff & Gardner-Medwin, 1998): "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" (full transcripts available on the web site).

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 50% drop is followed by a 20% rise. Students need to learn that there are different ways to solve a problem and to justify a solution, and that efficient learning and rigorous knowledge involve the habits and skills 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 an opinion to others, either explicitly or through body language, is a valued and necessary skill in every academic discipline and in every respected walk of life. These are skills that nevertheless remain largely untaught and untested in many forms of assessment before final undergraduate or graduate years, 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 at each level. The average marks obtained with our scheme (Table 1) are plotted in Fig. 1 for each confidence level against the probability of being correct.

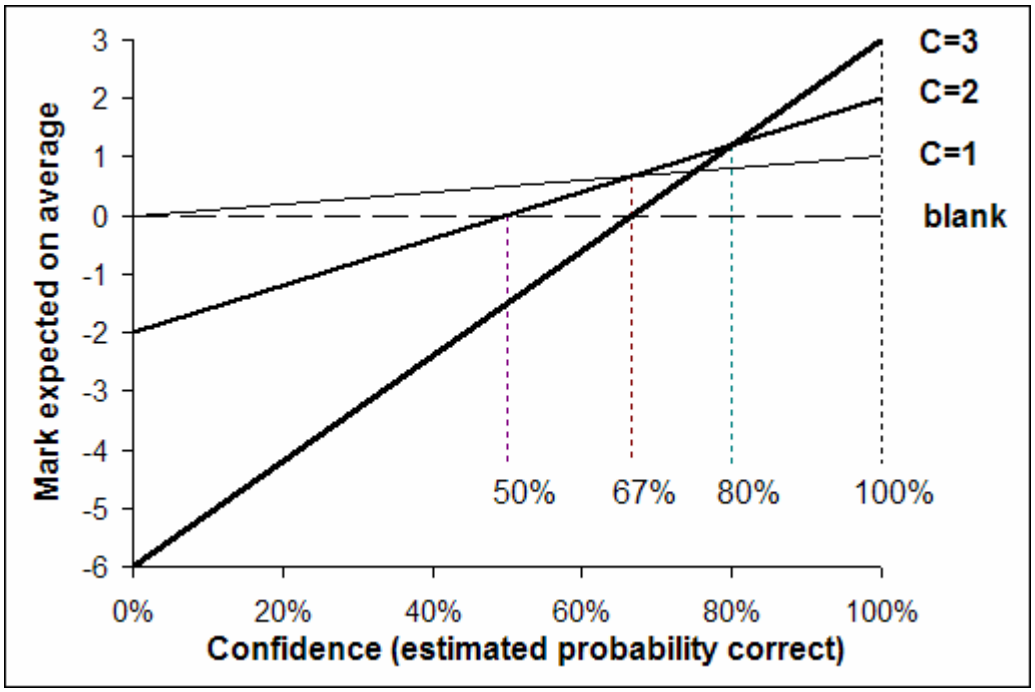

Fig. 1. The average mark expected on the basis of a student's estimated probability of an answer being correct, for each of the three confidence levels and for a blank reply with the mark scheme in Table 1. The best confidence level to choose for a particular estimated probability of being correct is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

A student will always expect to do best, trying to build up a good score, by choosing the confidence level for each answer that corresponds to the highest line above the axis at the point showing his/her estimated probability of being correct. It is clearly 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 never pays systematically to exaggerate or under-rate your confidence. This is the motivating characteristic of the mark scheme, making it a measure of the student's ability to judge the reliability of each answer, not of self-confidence or diffidence. CBM rewards this as a knowledge-based skill, with no incentive to misrepresent confidence in the way that occurs in games such as poker.

At UCL we have used this mark scheme mainly for questions with True/False answers. For any such question, with just two possible answers, the estimated probability of being correct can never be less than 50%, since if it were less then obviously one would prefer to switch choices. The three possible confidence levels therefore cover the possible range of probabilities (50-100%) fairly evenly. For questions where the answer has more options (as in a typical MCQ question, or one requiring a numeric or text entry) a student's preferred answer may be given with a much lower estimate of the probability of being correct. 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 are in Fig. 2a, showing it also to be properly motivating with incentives to use C=1 when confidence is low and C=3 when it is high. This scheme has a more uniform coverage of the possible probabilities of being correct, from 0-100%. This theoretical advantage may be outweighed however by the added complexity of having two mark schemes for different question types. Data from over 10,000 answers given by relatively inexperienced CBM users practising for a Biomedical Admissions Test on our website (with mostly multi-choice questions and using the scheme of Fig. 2a) suggest that naive users of a 3 point scheme, without as yet the experience of much feedback, may instinctively match confidence choices to probabiities in a way that better corresponds to optimal behaviour under the scheme of Table 1 and Fig. 1 than that of Fig. 2a.

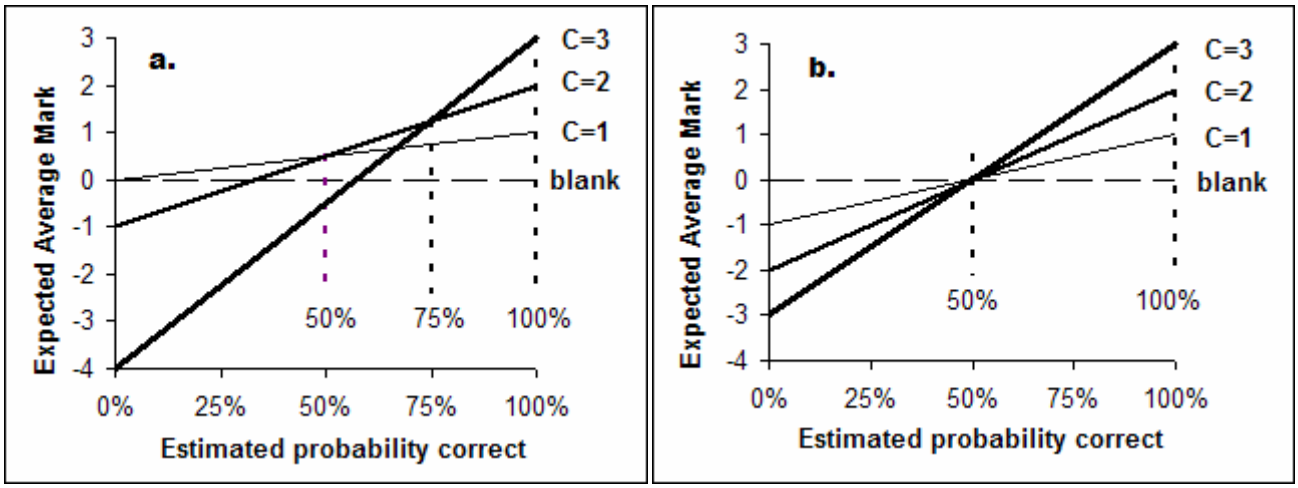

Fig. 2. Average marks expected as a function of student's estimated probability of being correct, for a scheme used on a trial basis in LAPT for questions with more than 2 possible answers and for a non-motivating scheme used elsewhere (for details, see text).

Unfortunately not all schemes used in the literature on confidence based marking have been properly motivating schemes (Gardner-Medwin & Gahan, 2003). Fig. 2b shows a scheme that is superficially similar to the LAPT scheme (and actually incorrectly accredited to it: Davies, 2002) but with penalties (-1, -2, -3) instead of (0, -2, -6). This has the merit of simplicity, but 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 used confidence levels 1 or 2 on this scheme, perhaps following teachers' advice when unconfident, would be quite disadvantaged. They would gain lower scores than students who were brash, or clever enough to see that this strategy was never sensible. Though such a scheme could have some of the benefits of CBM by encouraging students to think more, 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, and 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 also in principle motivating (Gardner-Medwin & Gahan, 2003) but since the scheme is hard to remember and understand, and sometimes promoted for use without the student being aware of the marks associated with the different confidence levels (e.g. SACAT: Self Assessment Computer Analyzed Testing, available online at www.hpeusa.com), it cannot really be seen as more than a qualitative way of encouraging students to think about confidence. Ease of understanding and transparency seem critical considerations when engaging students with a system designed to improve study and assessment, and this is what we have tried to achieve.

Students rarely describe their choice of confidence level in terms of explicit probabilities, even when the principles have been explained to them. Watching students (and particularly staff) in their 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 realise 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 maximise their scores, with few showing proportions of their answers 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 rather poor at handling the abstract concept of probability correctly, they make good judgments when information is evident as clear risks, outcomes and frequencies (Gigerenzer, 2003). For this reason it seems preferable to stick to a CBM scheme that is simple and transparent in terms of the choice of confidence levels and outcomes, rather than introducing subjective criteria and probabilities prone to misinterpretation. However, since students may differ in their ability to calibrate their judgments optimally to match a particular mark scheme, it is essential in work with CBM to provide clear feedback showing how well students are succeeding in this calibration. In formative use of CBM we provide such feedback in the form of a breakdown of % correct at each confidence level at the end of an exercise, as well as the all-important immediate feedback after each answer for the benefit of the student's reflective learning.

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. As part of a dissemination project we have provided tools on the UCL website (www.ucl..ac.uk/lapt) to enable new institutions and teachers in new disciplines or areas of education to experiment with CBM in their own contexts, with their own materials and students, and even 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 or exhortation amongst staff. However, it is worth addressing misconceptions that sometimes emerge in discussions and correspondence with teachers.

It is sometimes thought that CBM might unfairly favour or encourage particular personality traits. In particular, it is often suggested (though usually vigorously rejected by students) that CBM may favour one or other sex, usually based on the notion that it might disadvantage diffident or risk-averse personalities - supposedly more common amongst females. Several points can be made about this, some discussed at greater length by Gardner-Medwin & Gahan (2003). One is empirical : careful analysis of comparative data from exams and in-course use of CBM at UCL has shown clear, statistically highly significant, differences in risk-aversion between summative and formative use of CBM (with more cautious use of C=3 in exams: Fig. 3). However, there are not even small significant gender differences within our data for either summative or formative assessments, as also evident in Fig. 3. Nor is there any evidence for ethnic differences: the statistical relationships (linear regression lines) between CBM scores and percentage correct are almost identically superimposable for the wide range of ethnic groups we have as students at UCL. Most of these data were obtained from usage when the students were already quite well practised with CBM, so there might be transient differences on first encounter. But if individuals or groups do initially have a tendency to be disadvantaged through under- or over-confidence (and we all have known students who in discussion do show such traits), then this is an objective mismatch between expectation and performance that the student should be encouraged to be aware of, and to correct through feedback such as is provided by CBM. This is not to say that outwardly diffident or confident personalities are undesirable or unattractive, but rather that it is a serious handicap, especially in decision-rich occupations such as medicine, if inferences are not correctly calibrated for reliability.

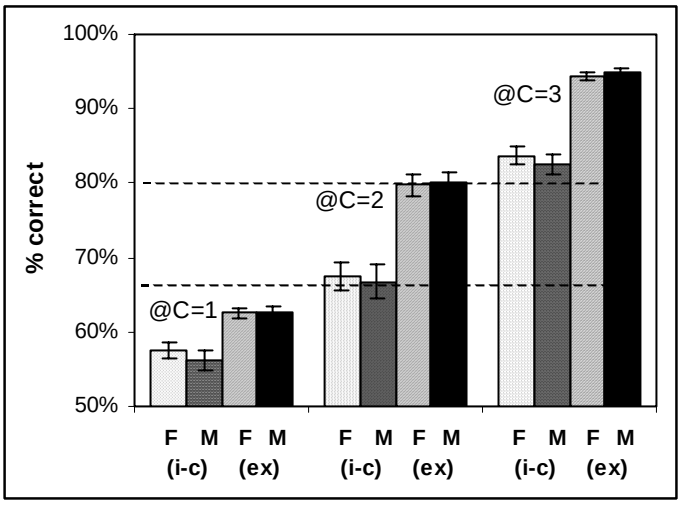

Fig. 3. Mean percentage correct at each confidence level for T/F answers entered in voluntary in-course (i-c) exercises (mean 1005 Qs) and end of year exams (500 Qs), separated by gender (190F, 141M). Bars show 95% confidence limits for the means. Differences between exams and in-course work are significant at each confidence level (P<0.01%) but gender differences are not statistically significant [from Gardner-Medwin & Gahan, 2003]

A second misconception is that the aim of CBM is somehow to boost self-confidence. Of course self-confidence is something that should improve through the use of any effective learning tool. Students often say, in evaluation questionnaires, that use of the LAPT system with CBM has helped to improve (and sometimes to diminish) their self-confidence in relation to the subject (Issroff & Gardner-Medwin, 1998). But they also say, and this is surely pedagogically more important, that it forces them to think more, and reveals points of weakness in their knowledge. CBM places a premium on understanding: on the ability to link and cross-check pieces of information, and to identify strong and weak conclusions. The net effect may be to reduce confidence as students come to realise that sound knowledge cannot be based on hunches. But this realisation is itself a step towards the building of self-confidence and academic success.

A third issue concerns the use of CBM in exams. It is sometimes suggested that what matters is whether a student produces the right answers, not whether s/he has confidence in these answers. As an issue in epistemology, this seems simply incorrect: a lucky guess is not knowledge, and a firm misconception is far worse than acknowledged ignorance. As a personal view, it seems to me that we let our students down on both counts if we mark their answers as if this were not true. But the issue can also be examined from a purely statistical psychometric point of view. In six exams, each involving ca. 350 students and 250-300 true/false questions presented in groups of five on related topics, we have looked at the reliability of CBM and percentages correct as indices of student performance. The standard measure of reliability (Cronbach Alpha) was 0.925 ± 0.007 (mean ± SEM, n=6) for CBM scores and 0.873 ± 0.012 for the percentages correct. This improvement (P<0.001, paired t-test) corresponds to a substantial reduction of the random element in the variance of exam scores from 14.6% of the student variance to 8.1%. The increased reliability of CBM scores is consistent with earlier research data with different forms of CBM (Ahlgren, 1969). Undoubtedly some of this improvement is due to differing ability to handle the component of knowledge that relates to confidence judgment. But in related analysis of the same data (Gardner-Medwin & Gahan, 2003) CBM marks on one half of the questions in an exam (odd or even numbers) were shown to be better predictors of % correct on the other half of the questions than were the % correct scores on the first half. So it

seems clear on even the most basic criterion of knowledge, at least under our exam conditions, that CBM scores offer a more reliable and valid measure of knowledge.

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. Exams were run by D. Bender.

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.* Seventh International Computer-Aided Assessment Conference Proceedings, 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