FORMATIVE AND SUMMATIVE CONFIDENCE-BASED ASSESSMENT

AR Gardner-Medwin and M Gahan

Formative and Summative Confidence-Based Assessment

A.R. Gardner-Medwin and M. Gahan
Department of Physiology
University College London
London WC1E 6BT

Abstract

Confidence-based assessment, in which a student's rating of his/her confidence in an answer is taken into account in the marking of the answer, has several substantial merits. It has been in use at UCL with medical and biomedical students for several years, primarily for computer-based formative assessment and study, using several answer formats. For two years we have used it in summative exams with multiple true/false questions. To encourage more widespread evaluation of our system and simpler application to other disciplines we have set up a browser-based version of the software: <http://www.ucl.ac.uk/lapt/laptlite>. This paper addresses some key issues: the rationale for our simple marking scheme (1,2 or 3 marks for correct answers and 0,-2,-6 marks for wrong answers according to confidence level), student reaction and performance, gender and personality issues, comparison with other marking schemes in relation to motivation for accurate reporting of confidence, and issues of reliability and validity for summative assessment.

Keywords: assessment, confidence, probability, reliability

Rationale

To measure knowledge, we must measure a person's degree of belief. Though one could take this as the starting point for a learned debate in epistemology or the application of probability theory, the simple point is perhaps best made by considering some words we use to characterise different states. A student, with different degrees of belief about a statement that is in fact true, may be said to have one of the following:

- knowledge
- uncertainty
- ignorance
- misconception
- delusion

The assigned probabilities for the truth of the statement would range from 1 for true knowledge, through 0.5 for acknowledged ignorance to zero for an extreme delusion, i.e. totally confident belief in something that is false. Ignorance (i.e. the lack of any basis for preferring true (T) or false (F)) is far from the worst state to be in.

The original reason for introducing confidence-based testing at UCL was to help students think about and identify where they lie on the scale above, in relation to any and every issue that arises in their studies (Gardner-Medwin, 1995). Misconception (uncertain bias towards a wrong answer) about basic issues in a subject can be a huge obstacle when it comes to trying to build higher levels of knowledge, and of course the more confidently the misconceptions are held the worse this can be. So the original rationale was to improve students' study habits - to encourage an awareness that uncertain but correct answers, or lucky guesses, are not the same as knowledge, and that confident wrong answers deserve special attention: consideration of why the student assigned high confidence and how their thinking about the issue can be adjusted for greater reliability. Reflection strengthens links between different strands of knowledge, both before and after feedback - checking an answer or viewing it from different perspectives before placing what is essentially a bet under the confidence-based marking scheme. It strengthens the ability to *justify* an answer, one of the essential elements in an Aristotelian definition of knowledge (as *justified true belief*) that is often missing in students who prefer rote-learning to understanding.

This rationale for confidence-based marking has been amply justified by the enthusiasm with which students have embraced the scheme, the benefits they report in terms of identifying areas where they are weak or kidding themselves that they have adequate knowledge, and the degree to which they voluntarily think about their confidence and reflect on different approaches to the checking of an answer (Gardner-Medwin '95, Gardner-Medwin & Curtin '96, Issroff & Gardner-Medwin '98). Partly in response to suggestions from students, we have since 2001 used confidence-based marking for the computer-marked component of summative exams for 1st & 2nd year medical students (approx. 40% of the total assessment; multiple T/F Qs; optical mark reader technology). As shown later, confidence-based marks improved the statistical reliability of the exam data as a measure of student performance, compared with conventional marking.

The UCL Scheme for Confidence-Based Assessment

The UCL scheme was devised to satisfy three primary requirements:
A. *Simplicity:* understood easily with little or no practice
B. *Motivation:* students must always benefit by honest reporting of their true confidence in an answer, whether high or low.
C. *Flexibility:* applicable without modification to answers in any format that can be marked definitively as correct or incorrect.
D. *Validity:* maintaining reasonable correspondence to knowledge measures backed by the mathematical theory of information.
It is primarily implemented in software for Microsoft Windows (LAPT: *London Agreed Protocol for Teaching*: Gardner-Medwin, 2003), following an initiative in several London medical schools that are now mainly amalgamated into University College and Imperial College London (UCL, ICL). To encourage dissemination and experience with confidence assessment in other institutions and disciplines we now have a web-based version of this software (LAPT-lite: Gardner-Medwin & Gahan, 2003).

The scheme has 3 confidence levels: C=1, C=2 and C=3. If the student's answer is correct, then this is the number of marks awarded (1,2 or 3). If the answer is wrong, then the marks awarded at these confidence levels are 0, -2, or -6. For the upper two confidence levels the scheme employs negative marking, but in a *graded* manner with the relative cost of a wrong answer increasing at higher confidence levels. This gradation is critical, because it ensures that the scoring scheme is properly *motivating*.

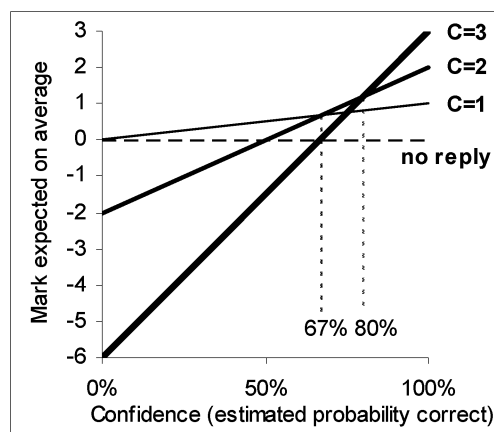| UCL Confidence-based scoring scheme			
Confidence Level	1	2	3
Score if Correct	1	2	3
Score if incorrect	0	-2	-6
Probability correct	< 67%	>67%	>80%
Odds	< 2:1	>2:1	>4:1

Fig. 1. The UCL scoring scheme, showing the ranges of probability of being correct over which each confidence level (C=1,2 or 3) is optimal in the sense that it gives the highest expectation of average score based on the probability.

The graph in Fig. 1 shows how, for each possible confidence level, the average score to be expected on a question depends on the probability of getting it right. If confidence is high, (>80%) then C=3 is the best choice. If it is low (<67%) then C=1 is best, and for intermediate estimates of the probability of being correct, C=2 is best. On this scoring scheme it is never best to give no reply, since an answer at C=1 carries the possibility of gaining a mark, with no risk of losing anything. Though this analysis of what is optimal behaviour seems rather mathematical, students easily arrive at near optimal behaviour, as shown later, on the basis of an intuitive understanding of the risks and benefits. They are shown the table of the ranges of probability or odds for which each confidence level is best (Fig. 1), but they rarely report thinking explicitly in terms of probabilities when deciding on their confidence level. The levels are always described in terms of the marks awarded (C=1,2,3) rather than in language terms such as 'very sure', 'uncertain', etc., which may mean different things to different people.

This marking scheme is appropriate for any type of answer that can be marked as definitely either right or wrong. In formative exercises we use it for answers that are T/F, multiple choice, extended matching sets, text, numbers or quantities, though in summative exams we have only at present used it with T/F answers. Each time an answer is entered, this is followed up with a request for the confidence level. It is important for formative use (study and revision) that all questions be marked individually one at a time (i.e. not in batches), with immediate presentation of feedback and explanations: this ensures that the feedback arrives while the student still has in mind the reasons for selecting an answer. This is especially important when high

confidence has been expressed for a wrong answer. A mark of -6 stings, even though it should be expected on up to 20% of the occasions when a student takes the risk of entering C=3. It stimulates attention and - an incidental spin-off from the introduction of confidence assessment - it encourages students to enter comments explaining their logic or pointing out errors (real or imagined) and ambiguities in questions. The entry and tracking of such comments (with full contextual information) is an integral and valuable feature of the LAPT and LAPT-lite software, helping to improve exercises and inform teachers of students' misconceptions.

Fig. 2. The relationship between marks assigned (3,2,1,0,-2,-6) and the appropriate information-theoretic measure of lack of knowledge for T/F answers, proportional to the log of the subjective probability assigned to the correct truth value for a proposition. The diamond corresponds to acknowledged ignorance.
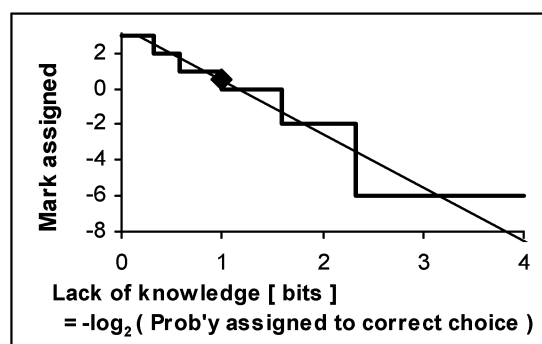
The relationship between marks awarded and the student's knowledge, or more strictly lack of knowledge, based on Shannon's theory of information is shown in Fig. 2. The relationship is only clearcut in this way for T/F answers, where confidence for a correct answer is always implicit in confidence expressed for a wrong answer. For questions with more than 2 possible answers (MCQ, text, etc.), the graph is valid for correct answers, but only shows the minimum lack of knowledge corresponding to a mark for a wrong answer. This minimum is correct only if the student's 2nd choice of answer (after being told the first choice was wrong) would be both correct and totally confident. It is a fundamental drawback of MCQs that they can fail to pick up serious misconceptions, where a student is convinced that the right answer is wrong, but unsure what would be right. The correspondence with theory in Fig. 2 is about as good as can be achieved with 3 confidence levels and 6 discrete marks. Though this mathematical nicety is probably the least important of the constraints that a confidence-based marking scheme should conform to, a wide discrepancy would be worrying.

Issues surrounding confidence-based assessment

It is important to recognise that the objective of confidence-based marking is not to reward or discourage self-confidence. The aim is to encourage reflection, self-awareness, and the expression of appropriate levels of confidence. One of the major limitations of computer-aided assessment is that it generally implements little of the subtlety of face-to-face assessment. Confidence-based assessment is one way in which it can catch up.

A commonly encountered view is that confidence-based assessment must somehow be introducing a bias into assessment that favours one or other gender, or certain personality types. A perceptive view of this issue was given by Ahlgren (1969) at a conference entitled "*Confidence on Achievement Tests - Theory, Applications*", in which he argues that the value of confidence assessment should be seen primarily in the context of education rather than psychometrics, and that the benefits of improved reliability and concerns about supposed unfairness are secondary issues deserving research but of much less significance. Because of the concern, particularly about gender, it is worth presenting data from our experience at UCL.

Despite careful scrutiny, no significant gender differences have emerged in data from over 3 million answers recorded on campus computers and in exams. Fig. 3 shows the percentages correct at different confidence levels for a single cohort of 1st year medical students in 2000/01 (331 students), , compared across the sexes and between in-course data and exams at the end of the year. The students were considerably more cautious (achieving a higher % correct) in their use of the high confidence levels in the exams than when working to aid study and revision, but both sexes behaved without any significant difference under both conditions. For most of these data the students were very familiar with the confidence assessment principle and had received much feedback about their performance, so differences in behaviour at the outset may have disappeared. But since personality traits that lead to inappropriate over-confidence or under-confidence in such tasks are undesirable, such a learning experience can only be beneficial.

Fig. 3. Mean percentage correct at each confidence level for T/F answers entered in voluntary in-course exercises (i-c: mean 1005 Qs) and end of year exams (500 Qs), separated by gender (190F, 141M). Bars show 95% confidence limits for the means. The small gender differences are not statistically significant (t<1.5).
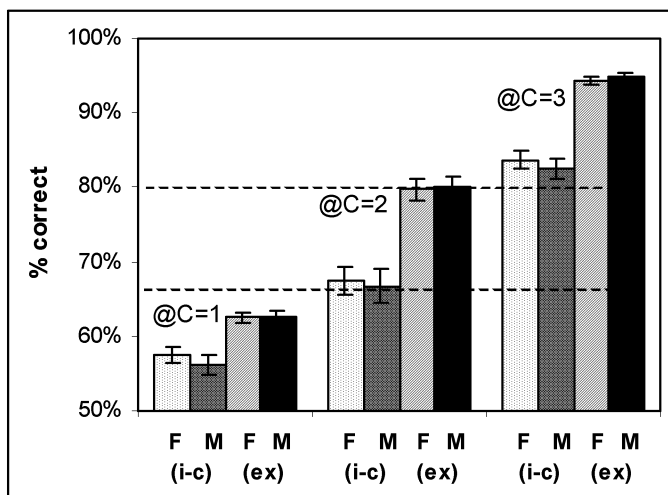
Since an individual tendency to either over-confidence or under-confidence can lead to a loss of marks with confidence assessment, it was important to examine this in the context of exam data. Overall in the exams, 41% of answers were entered at C=3 with 95% correct, 19% at C=2 with 79% correct and 40% at C=1 with 62% correct. The percentages correct were within the optimal ranges but for C=2 and C=3 were near the top of these ranges, reflecting caution or under-confidence. Only 2 students (1F, 1M; both weak) were over-confident, with percentages correct at any confidence level that were significantly below the optimal range (in each case about 60% correct at C=2). Under-confidence was more common: 8 students gained significantly

>67% for answers entered at C=1 and 43 students >80% for answers at C=2. The most extreme examples were two students (1F, 1M) with 90% correct at C=2. Educationally, the important issue is for students to learn to distinguish between confident and unconfident answers, rather than to handle a particular marking scheme with optimal calibration. A simple adjustment was made to deal with this issue fairly: each set of answers at the same confidence level was treated as if entered at the most beneficial level for the percentage correct. In practice this made little difference to the marks or rank orders.

Motivating marking schemes

A crucial feature in confidence-based assessment is the *motivating* nature of a marking scheme (Fig. 1). Without this, a system that awards higher marks for answers entered at high confidence simply rewards those students who are bold enough or perceptive enough to see that it is never advantageous to enter low confidence. One of the major learning issues in use of confidence-based marking is the realisation that you can be rewarded for acknowledging and communicating low confidence. Correct and honest expression of confidence is a valued communication skill in any arena. Computer-aided assessment offers an excellent platform for experience and practice of this, backed up by encouragement of students to apply the principle to written work also, stating when they are or are not sure of a fact or an argument. In choice of a marking scheme, it is necessary to pay careful attention to the way it depends on confidence, to ensure proper motivation.

The simplest confidence related self-awareness is the appreciation that one is guessing, or nearly guessing. Discussions often centre round the issue of discouraging students from guessing in exams, because guessing adds random noise to assessment data. Negative marking schemes are often said to discourage guessing, but unless the scheme is motivating it will not be rational for a student to behave in this way (Gardner-Medwin, 1999). For the purpose of discussion, graphs analogous to Fig. 1 are set out in Fig. 4 for six different marking schemes. Options (including the option to omit an answer) should logically be preferred by a student when the corresponding line is the highest on the graph, for whatever probability corresponds to his/her confidence.

With no negative marking (Fig. 4, top left) it is obviously never rational to omit an answer. Few students fail to appreciate this, so exam data is bound to be corrupted by random answers when students have little or no relevant knowledge. A simple negative marking scheme, with equal numbers of marks added and subtracted for right and wrong answers, motivates a student to omit a reply when confidence for being marked right is less than 50%. This is valuable for MCQ and open ended answers, but for T/F answers the confidence probability can never be less than 50%, since if it were then a student should prefer the complementary answer. Use of such a scheme for T/F questions (as at UCL a few years ago) is not only irrational but tends to act against the interests of students who follow advice to omit guesses, since even a modest degree of partial knowledge may ensure sufficient bias in favour of correct answers to make the difference between a pass and a fail

score. Prejudice against fixed negative marking schemes may stem partly from explicit or implicit realisation of this point, though for MCQs with more than 2 options this scheme is preferable to marking without penalty.
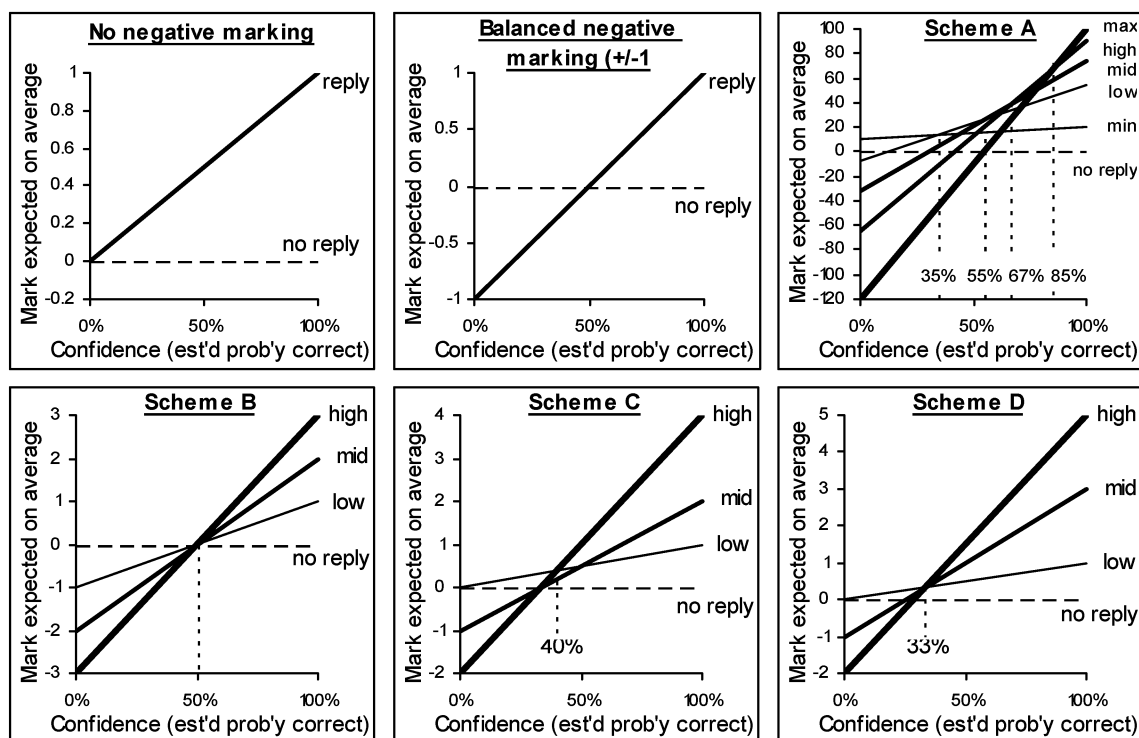

Fig. 4. Plots of the average mark expected on a question for each available option, given the student's estimate of the probability of being marked correct. For each scheme the best choice for a given probability is the one for which the graph is highest above that point on the horizontal axis. Scheme A is from Hassmen & Hunt (1994), Schemes B-D are from Davies (2002).

Scheme A in Fig. 4, used by Hassmen & Hunt (1994) and Khan et al. (2001), has five confidence levels with marks (20,54,74,90,100) for correct answers and (10,-8,-32, -64, -120) for wrong answers. It is a properly motivating scheme, in that each level has a range of probabilities for which it is optimal (with thresholds of 35%, 55%, 67% and 85%). If the implications are clearly understood by a student, then the scheme discriminates two lower levels of confidence than our simpler UCL scheme, but the lowest level option would have to be deleted for T/F answers where one cannot rationally assign P<35% even for an answer determined by the toss of a coin. The complexity and inflexibility of this scheme are rather negative factors.

Schemes B-D in Fig. 4 are examples of schemes that are not properly motivating, used by Davies (2002). Examination of the graphs will show that it is *never* rational to use the middle confidence level with any of these schemes. Scheme B was erroneously attributed to Gardner-Medwin (1995) where the UCL scheme is originally described. The best marks with Scheme B (in 3-choice MCQ questions used by Davies) would have been achieved if a student always used high confidence or (for confidence <50%) omitted replies altogether. Neither of the lower confidence levels is useful, and

students who took advice to use these levels would be disadvantaged. Schemes C,D used by Davies (2002) are only slightly different from B. In Scheme C the low confidence level is best if confidence is less than 40%, though there is relatively little margin between this and the chance probability of 33% in 3-choice MCQs. In Scheme D, high confidence is the only useful option unless the student expects to do worse than chance. The comment made by a student in this study is rather apposite: *'If a student is not a confident person, then they will not do well'*. One should strictly qualify these considerations of what is rational use of available confidence levels by saying that it assumes the student is attempting to maximise an expected score. Choice of lower confidence levels can occasionally be rational in order to minimise the risk of failing an exam rather than to maximise the expected score; but the distinction is subtle and applicable only in unusual circumstances.

Reliability and validity

Full analysis of the statistics of our exam data will be published elsewhere. Key results are shown here (Fig. 5). One of the best tests of the reliability of a marking scheme is a comparison of marks for students on one set of questions with marks on a different set. Good correlation means the measurements tell you something about the student, not just about chance factors. Fig. 5 shows the correlations between data from odd numbered and even numbered questions in 6 medical exams, each with over 300 students and 250-300 T/F questions. Correlations were significantly higher for comparison of confidence-based marks than conventional marks ($P<0.01$).

Fig. 5. Mean values of r^2 for correlations between scores on interleaved sets of questions in 6 exams, using confidence-based marks and conventional marks. Error bars show $\pm$ 1 s.e.m. (n=6) for the separate means and paired differences, which are all significant at $P<0.01$.
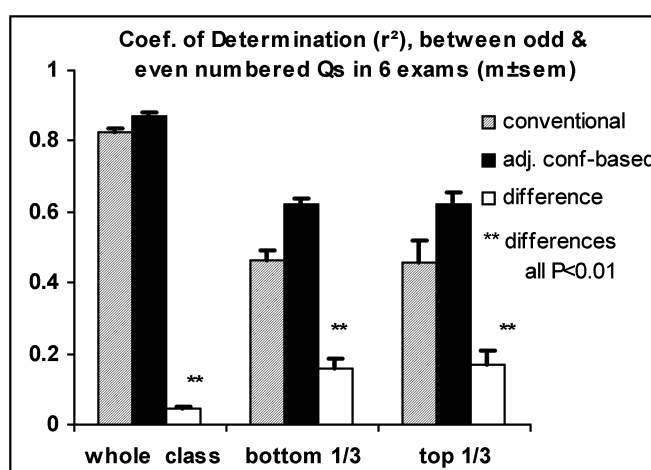
Since exam data tends to be more consequential for the weakest and the strongest students in a class, correlations were examined for the bottom and top thirds (divided according to conventional scores). For both groups the improved reliability with confidence-based scores was equally clear and significant, though coefficients were lower due to the more restricted range of abilities (Fig. 5). Such an improvement in the ability to discriminate between students means that confidence-based data can give results of equivalent statistical reliability to conventional marking with fewer questions, by a factor of 2.0 for the lowest third of our groups and 2.2 for the top third.

This improvement is not due simply to differences in ability to make good confidence judgements, because the confidence-based scores on one set of questions actually proved better than conventional scores at predicting even the *conventional* scores on the other set of questions, by factors of 1.3-1.4. Thus there is no sacrifice of validity in data from confidence-based marking. Even if one were to ignore educational arguments for confidence-based marking (to improve the way in which students handle problems), or its merits as a more complete measure of knowledge, it can - at least under our conditions - achieve substantial improvements and savings in assessment performance, even on conventional criteria.

Acknowledgements

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. (1999) *Rational and irrational marking schemes.* Journal of Physiology, 515P: 48P

Gardner-Medwin, A.R. (2003) *LAPT: London Agreed Protocol for Teaching.* <http://www.ucl.ac.uk/lapt>

Gardner-Medwin AR & Curtin NA (1996) *Confidence assessment in the teaching of Physiology.* Journal of Physiology, 494:74P

Gardner-Medwin AR & Gahan M. (2003) *LAPT-lite* <www.ucl.ac.uk/lapt/laptlite>

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. ISBN 1-85377-256-9, pp 169-179

Khan K.S., Davies D.A., Gupta J.K. (2001) *Formative self-assessment using multiple true-false questions on the Internet: feedback according to confidence about correct knowledge.* Medical Teacher, 23, 158-16