MS for Ch. 12 for:
Innovative Assessment in Higher Education edited by **Cordelia Bryan and Karen Clegg**
2nd edition, published 2019, Taylor & Francis

Certainty-Based Marking: stimulating thinking and improving objective tests

Tony Gardner-Medwin
Prof. Emeritus at University College London

Introduction

'Certainty-Based Marking' (CBM) is what was originally described at University College and Imperial College in London, and for the first edition of this book (Gardner-Medwin, 2006), as '*Confidence*-Based Marking'. Its aims and principles have not changed, but the change of name tells a story. Too often, people misunderstood CBM as something designed to boost or reward self-confidence. Its aim is much more nuanced - to reward the identification of *uncertainty* as well as genuinely reliable conclusions. It may even diminish unwarranted self-confidence, though knowledge of what one does and doesn't understand is ultimately the basis for confident decision making.

Overconfidence or its opposite, excessive hesitation, can vary with personality, upbringing and gender. A properly designed CBM strategy gives feedback that with practice can serve to correct such biasses. But the principal aim is to enhance learning by stimulating thinking about how different aspects of a student's knowledge are usefully related. Typically, the immediate response to a question - especially the sort of question good teachers ask to promote understanding - deserves extra scrutiny. CBM always asks "*How sure are you?*", challenging the student to look for justifications and reservations - aspects of knowledge seldom explored in objective testing. Such thinking promotes deeper understanding and is rewarded as we shall see by enhanced CBM scores, even if the answer remains unchanged or starts to seem less reliable.

Strong students can often do well by relying only on superficial associations, with little incentive (on conventional right/wrong mark schemes) to reflect on the reliability of their thinking unless really taxed. Weaker students try to emulate this with diligent rote learning, rejecting deeper learning as unnecessarily challenging. This may get them through a test, but can be disastrous as a basis for future learning. It becomes stressful, because if facts are learned independently there are many more to learn than if they can be deduced and checked one against another. Knowledge, especially along with understanding, is more like a network of relationships than a set of facts, and is much more efficiently stored that way.

The value of thinking about certainty, for enhancing learning and consolidation, has been researched extensively, but mostly before computer aided assessment was practical on much of a scale (see, for example, Hevner 1932, Ahlgren 1969, Good 1979). The development of CBM in London [1] has helped stimulate wider application. CBM is now included in the open-source learning management system 'Moodle' [2] and use of CBM variants have been reported from several institutions worldwide (e.g. Hassmen and Hunt (1994), Davies 2002, Rosewell 2011, Schoendorfer & Emmett 2012, Yuen-Reed & Reed 2015, Foster 2016). This chapter will focus on the London medical school experience since 1995 (ca. 1.8 million self-test sessions and 1.4 million exam answers). Students seem to pick up the logic of CBM instinctively through use - more readily than through exposition and discussion. After all, as a successful animal species we have evolved to learn to handle situations with uncertainties, risks and rewards: in childhood we call them games, while as adults they can determine our survival (Gardner-Medwin, 2018). Readers are encouraged to try CBM exercises themselves [1].

How does CBM work?

The scheme instigated at UCL in 1994 (Gardner-Medwin, 1995), which still seems a good choice, is simpler than most that had been used in previous research. It uses just three certainty levels, identified by numbers (C=1, 2, 3) and by neutral terms (low, mid, high) rather than by descriptors ("guess", "hunch", "sure", "definite", etc.) that can have idiosyncratic interpretations. Table 1 shows the marks (or 'points') awarded at each C level for correct and incorrect answers. It is these credits and penalties for answers marked right or wrong that determine how best to use the levels.

Certainty Level	Mark if Correct	Penalty if wrong
C=3 (high)	3	-6
C=2 (mid)	2	-2
C=1 (low)	1	0
No Reply	0	0

Table 1: The Certainty-Based Marking scheme

The table is easily remembered and is enough to guide students in use of CBM. The concept of a double penalty (-6) sets the criterion fairly high for choosing C=3, and with minimal knowledge it is clearly worth entering an answer with C=1 rather than 'no reply'. Students don't normally report thinking quantitatively about CBM, but nevertheless manage with practice to use C levels in a near optimal way (Gardner-Medwin 2006, 2013). They often say they think about C=1 and C=3, then if undecided they opt for C=2, which is simple and rational. The implications emerge in Fig. 1.

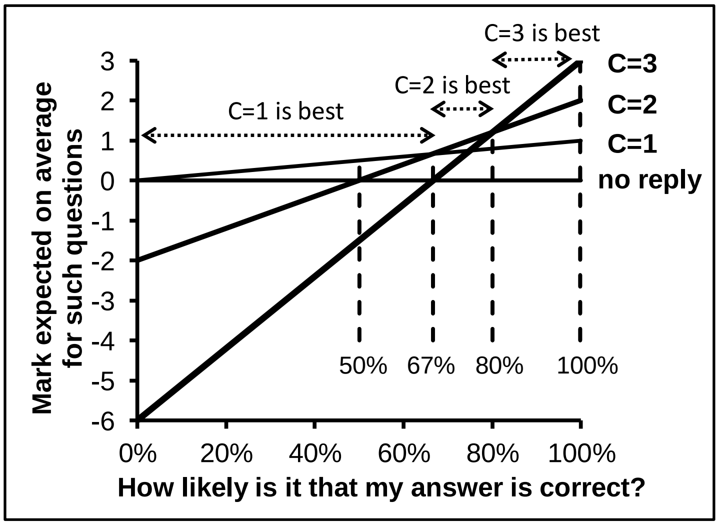


Fig. 1. Rationale for choosing certainty levels. Depending on how likely you think your answer is to be correct, you expect to gain by choosing whichever level (C=1, 2 or 3) gives the highest average expected mark: C=3 for >80%, C=1 for <67%, or C=2 in between. It is always better to enter a reply with C=1 than to omit a reply.

The figure shows that for any probability of being correct, there is a C level for which the graph is highest, meaning you expect (on average with such questions) the best reward. Above a threshold of 80% C=3 is best, while below 67% C=1 is best. Whatever your certainty, you cannot expect to gain by misrepresenting it - in other words by trying to "*game the system*". The system motivates the reporting of an honest judgment, using what statisticians call a 'proper' reward scheme for estimating probabilities (Good 1979, Dawid 1986). It is always worth sketching the equivalent of Fig. 1 to check whether a particular marking scheme properly motivates what is intended (Gardner-Medwin & Gahan 2003).

Contrast CBM with fixed negative marking schemes, which offer the option 'no reply' (mark=0) to avoid risk of a fixed penalty for wrong answers. The intention is to encourage students to omit uncertain answers, especially guesses, that increase variability in final scores. However, the consequence for the student of such omissions can be illusory and iniquitous. Penalties are commonly set to the minimum that ensures guesses do not on average improve one's score. For example, with 5 option MCQs the penalty would be -0.25 times the credit for a correct answer. Even slight partial knowledge then means the student should expect to gain on average from answering, while guesses would on average be neutral. So students are encouraged to disadvantage themselves by omitting answers, based on advice and risk aversion rather than rationality. For example, limited knowledge can often narrow MCQ options down from 5 to around 2, with a 50% chance of a final guess being correct - which would lead on average to 40% of full credit. A system that encourages students to disadvantage themselves by omitting answers in such circumstances should be illegal. CBM, by contrast, straightforwardly steers the student to enter their uncertain response with C=1, yielding on average in the example 17% of the full credit for a confident correct answer based on thorough knowledge.

The student's perspective using CBM

There are several ways in which a student's perception of CBM embodies sound principles of good learning (Cornwell & Gardner-Medwin 2008, Gardner-Medwin, 2018).

1. CBM rewards thinking about how to justify an answer, thereby developing relationships between different nuggets of knowledge far better than the rote learning of facts.
2. It rewards identification of uncertainties and inconsistencies in one's thinking, leading to more effective study.
3. Lucky guesses are not the same as knowledge. Students recognize that they should not get the same credit. Teachers and examiners should recognize this too.
4. Confident misconceptions are serious, even dangerous. When studying, a penalty is a wake-up call - triggering reflection and attention to explanations. We learn through mistakes, especially bad ones.
5. Quoting student comments from an early evaluation study (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".

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). Thinking critically and identifying points of weakness is an opportunity to consolidate connections between different elements of knowledge. It is distressing to see students with good GCSE grades struggling two years later to apply half-remembered rules to issues that should just be embedded as common-sense understanding - for example how to combine successive percentage changes in a quantity.

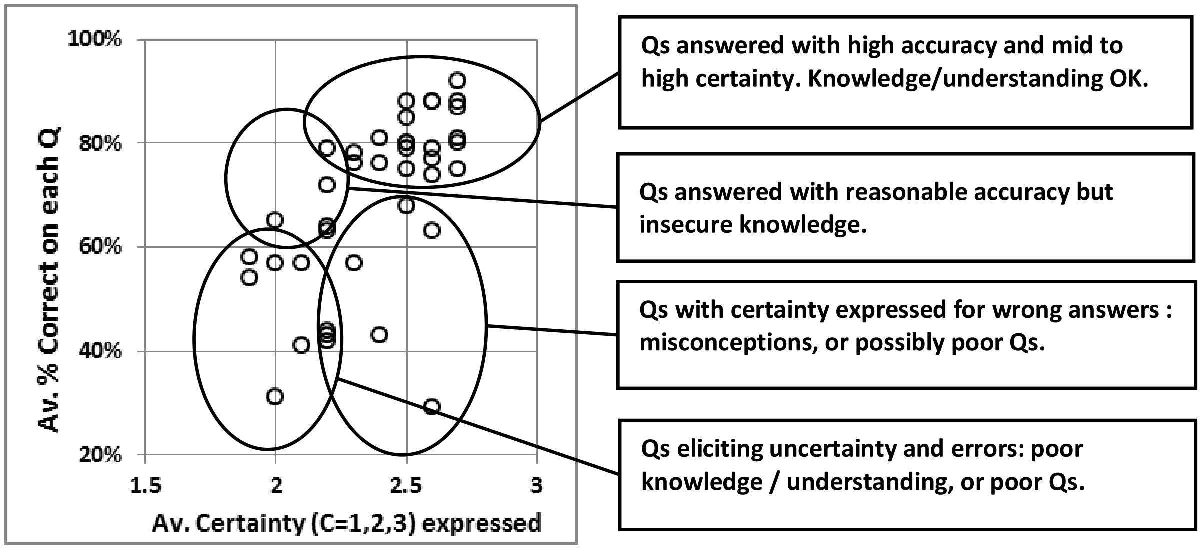
There are different ways to solve problems and retrieve facts. Communicating the reliability of rival ideas is a key part of inter-personal communication in every walk of life, either explicitly or through body-language, and it deserves emphasis in education. Such skills however can remain largely untaught and untested in assessments until final exams, when they are often expected in demanding forms of critical writing and in viva situations. CBM can be a constant stimulus and reminder of their importance.

An interesting evaluation study (Foster, 2016) has used a form of CBM in school mathematics. Pupils aged 11-14 were given numerical exercises and asked to rate how confident they were in each answer on a scale 0-10. They were told their total score would be the sum of their ratings for correct answers minus the sum for incorrect answers. Completely new to the idea, their reaction was encouraging: a ratio 106:28 positive to negative comments. They found it challenging, but constructive. The scheme might need revision in continued use since it is not a properly motivating scheme for reporting certainty: a pupil should expect to do best by rating each answer 10 if the the chance of being right was judged >50%. One negative comment from a pupil: "*I don't like this marking scheme as it is partly based on your confidence in yourself*." seems perceptive: students who were confident and not averse to risk would likely have performed in a more nearly optimal way, which may have contributed to a gender difference reported: despite the fact that girls in the trial achieved greater accuracy than boys, they had similar average confidence ratings.

Benefits to teachers from using CBM

Self-tests can help students learn effectively, whether or not they use CBM. There is a danger, however, that setting up such exercises can be seen by students as a form of assessment rather than as a challenge to assist their study. Several guideline principles should apply to self-tests in my opinion. Students should at least optionally be able to keep selftest marks private (not visible to teachers) to avoid fear of humiliation. After all, the more mistakes they make the more they learn - especially if feedback is immediate and accompanied by explanations. They should be able to control the questions they choose to answer and be able to be marked out of the subset they choose; this enables them to focus on challenging themselves in areas of weakness or interest. They can be encouraged to work together to stimulate discussion, with anonymous comment facilities shared with other students and staff, to facilitate discussion of questions and improvement of content. While it is useful for teachers to know who does and doesn't use self-tests, submission of scores can be voluntary and anonymised without sacrificing much of its value as feedback for teaching. Online self-tests can be programmed to operate entirely within the student's computer once initiated, so a central computer need not handle performance data at all unless the student chooses to submit it on completion.

How does CBM help? Firstly, of course there are the benefits discussed earlier, from a student's perspective. CBM can prompt useful discussion, because questions like "*Why are you so sure/unsure*" can be very constructive. A pleasant surprise to me, starting to write questions for self-tests (about physiology and maths), was that CBM makes life easier: you shouldn't always worry about pitching questions at an appropriate level of difficulty. Students have a range of different strengths and weaknesses. CBM makes a diverse mixture of questions work well. When a student thinks a question is really easy (which without CBM they might even see as demeaning), they think "*OK, definitely a C=3 here*". Weak students are encouraged by identifying their strengths as well as weaknesses. Feedback to teachers can be striking: a surprising number of correct answers to 'easy' questions may be accompanied by low C ratings, indicating an insecurity in understanding. I have come to hesitate using terms like 'easy' and 'difficult' for questions, because these depend so much on how individual students have approached the subject. Teachers get useful feedback about question quality, helping the elimination of ambiguities and improvement of explanations. An unforeseen bonus arising from the use of -6 penalties with CBM has been the readiness with which students will try to justify in comments why their particular slant on a question was completely reasonable: not always true, but helpful in improving exercises and explanations!



Qs answered with high accuracy and mid to high certainty. Knowledge/understanding OK.

Qs answered with reasonable accuracy but insecure knowledge.

Qs with certainty expressed for wrong answers : misconceptions, or possibly poor Qs.

Qs eliciting uncertainty and errors: poor knowledge / understanding, or poor Qs.

Figure 2 Distribution of responses to a set of 40 questions (mostly single option MCQ and numerical) written as self-test practice for an exam. Likely issues about individual questions are highlighted, and can prompt consideration alongside question text and wrong answer choices. Thanks to N.Curtin (Imperial College) for anonymised data.

Teachers get a new dimension of feedback about their questions with CBM: average certainty ratings as well as accuracy. Fig. 2 shows one plotted against the other for each question in an exercise. With self-tests the data can help with improvements and course planning, while equivalent data in exams can flag questions that may have been widely misunderstood and may warrant exclusion from assessment.

Assessment with CBM

CBM, at least in London, has been used more for self-tests than formal assessment. There are both benefits and obstacles to its use in assessment. The UCL medical school ran 1st and 2nd year objective exam components (using true/false questions) with CBM for 5 years 2001-2006. The result was an enhancement of reliability (equivalent to using over 50% more questions with conventional marking) and strong student support for its continued use (Gardner-Medwin 2006, 2013). However, a review discontinued its use in exams (with also a switch to MCQ and extended matching (EMQ) question styles), apparently because CBM was out of line with exam practice elsewhere and it was thought it could lead to confusion over standard setting.

Comparison of standards with and without CBM is an interesting challenge. Fig. 3A shows a typical distribution of CBM exam scores (average marks per question) plotted against conventional accuracy. A student with 80% accuracy typically gets around 50% of the maximum possible CBM score. This divergence is inevitable, since both percentages could only be equal if every correct answer was entered with C=3 and all the others with C=1 or 'no reply'. Given that students will often have partial knowledge, raising the probability of being correct above chance but not to 100%, this can't happen. However, CBM percentages substantially lower than conventional scores can seem a bit demoralising to a student, and confusing to examiners.

I have tried three approaches to this problem. Simplest, to boost student morale, is just to scale CBM scores so 100% corresponds to all correct at C=2; the maximum is then 150% and the median (typically 70-75%) roughly comparable to median accuracy. For exam assessments, a more complex non-linear scaling of CBM grades [Gardner-Medwin & Curtin, 2007] can retain CBM grades within a 0-100% range and ensure approximate equivalence on average between CBM and accuracy at each level of ranking. The same mark criteria for accuracy and CBM will then pass about the same number of students, though students with better identification of reliable and unreliable answers will rank higher with CBM. Though somewhat complex, this does simplify standard comparisons.

Fig. 3 illustrates the third, probably best, approach in which any benefit the student has derived from effective use of CBM is separated and added as a bonus to conventional accuracy, yielding *'CB accuracy'*. The benefit (Fig. 3A) is the average CBM mark minus what it would have been if the student had not distinguished reliable and unreliable answers, using an identical C level (appropriate for the overall accuracy) throughout. Such 'benefits' can be negative, seldom in exams but more commonly in self-tests as students work with wrong ideas. The bonus added to accuracy (Fig. 3B) is one tenth of this calculated benefit, using a factor optimised empirically for improvement of statistical reliability with CB accuracy (Gardner-Medwin, 2013).

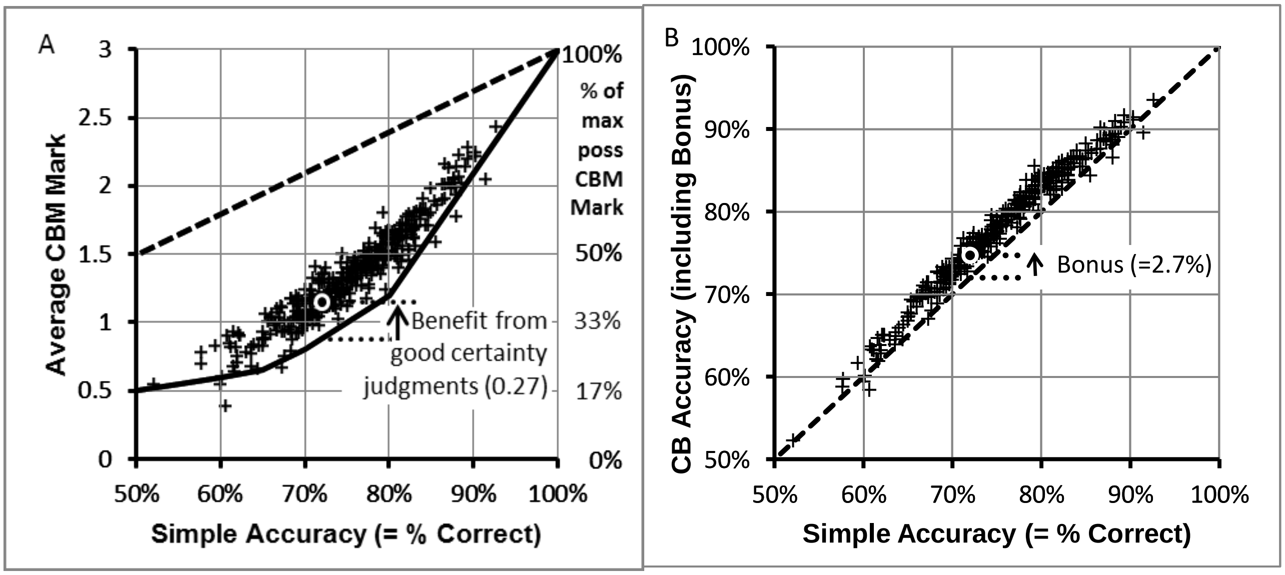


Figure 3. Exam scores (320 medical students, 300 T/F questions). ***A****. Average CBM vs. accuracy, for each student. Dashed line: unattainable equality between CBM (expressed as % of maximum) and accuracy. CBM scores are lower, but mostly above the full line showing CBM scores for a student who doesn't discriminate reliability but has the same accuracy (credited with a fixed C level appropriate for the accuracy).* ***B****. CB accuracy expressed as conventional accuracy plus bonuses calculated from the benefits illustrated in A. Thanks to D.Bender (UCL) for anonymised data.*

Conclusion

Certainty-based marking addresses some of the key elements of knowledge and understanding that we try to impart through education. These days it is easy to check facts online. Many professions have shifted towards a culture of collaboration, making it easier to seek help addressing uncertainties. The important thing is always to be aware whether your ideas are reliable or need checking or consulting. This is the core concept of CBM: rewarding accurate judgment of reliability. The necessary thought processes are a bit like a self-consultation, linking an issue to the rest of your knowledge.

By strengthening awareness of how facts and ideas relate, CBM assists learning and understanding and leads to more fair assessment. Knowledge is not a binary thing ("*you know it or you don't*"). CBM gives graded credit for partial knowledge; lucky guesses are not treated like knowledge. Negative knowledge is penalized (i.e. firm misconceptions, potentially hazardous, and worse than a baseline of acknowledged ignorance) . Knowledge has many facets. You may struggle to retrieve facts, yet recognise them with certainty if presented (as in many MCQ tests). You may know something but not understand it. Training with CBM helps develop the connections and strategies on which retrieval and understanding depend. Always paramount are the key processes of education like explanation, inspiration, encouragement and involvement, but CBM sits well alongside them and can go some way to dispelling the negative image that often attaches to testing in education.

Notes

1. More information, examples and links are at: https://tmedwin.net/cbm/selftests . The software can use existing sets of questions (True/False, MCQ, EMQ, text, numerical), with or without CBM, provided answers can be categorically identified as right or wrong. It has sophisticated options for randomisation, alternative answers, tailored feedback, anonymised comments, etc. and is available free for use with exercises hosted on the site or elsewhere. Queries, suggestions and involvement in further developments are welcome. Contact: a.gardner-medwin@ucl.ac.uk .

2. CBM in Moodle: https://docs.moodle.org/en/Using_certainty-based_marking .

References

Ahlgren A (1969) Reliability, predictive validity, and personality bias of confidence-weighted scores. Paper at the American Educational Research Association Convention. https://eric.ed.gov/?id=ED033384

Cornwell R & Gardner-Medwin T (2008) "Perspective on Certainty-Based Marking: An Interview with Tony Gardner-Medwin," *Innovate: Journal of Online Education*: Vol. 4 (3),7 Online: http://nsuworks.nova.edu/innovate/vol4/iss3/7 Transcript: http://www.tmedwin.net/~ucgbarg/tea/Innovate/

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

Dawid AP (1986), 'Probability forecasting' in Kotz, S., Johnson, N.L. & Reid, C.B. (eds), *Encyclopedia of Statistical Sciences,* 7, 210-1

Foster C (2016) Confidence and competence with mathematical procedures. Educ Stud Math. 91:271–288

Gardner-Medwin AR (1995) Confidence assessment in the teaching of basic science. Research in Learning Technology (formerly ALT-J) 3:80-85

-------- (2006) 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, pp 141-149

-------- (2013) Optimisation of Certainty-Based Assessment Scores. Proc 37th IUPS, PCA167. Online: http://www.physoc.org/proceedings/abstract/Proc%2037th%20IUPSPCA167, http://tmedwin.net/~ucgbarg/tea/IUPS_2013a.pdf

-------- (2018) The value of self-tests and the acknowledgement of uncertainty. In Enhancing Learning and Teaching with Technology - What the research says. Ed. Luckin R, UCL IOE Press

Gardner-Medwin AR & Curtin, N (2007) Certainty-Based Marking (CBM) for Reflective Learning and Proper Knowledge Assessment. In: REAP Int. Online Conf. on Assessment Design for Learner Responsibility. http://www.tmedwin.net/~ucgbarg/tea/REAP/REAP_CBM.htm

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

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

Good IJ (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.

Hevner K (1932) Method for correcting for guessing and empirical evidence to support. J. Soc. Psych. 3, 359-362.

Issroff K & Gardner-Medwin AR (1998) Evaluation of confidence assessment within optional coursework. In : Oliver, M. (Ed) Innovation in the Evaluation of Learning Technology, Univ. N. London: London, pp 169-179

Rosewell JP (2011). Opening up multiple-choice: assessing with confidence. In: 2011 International Computer Assisted Assessment (CAA) Conference: Research into e-Assessment, 5/6 Jul 2011, Southampton, UK.

Schoendorfer N & Emmett D (2012) Use of certainty-based marking in a second-year medical student cohort: a pilot study. Adv Med Educ Pract. 3,139–143

Yuen-Reed G & Reed KB (2015) Engineering Student Self-Assessment Through Confidence-Based Scoring. Advances in Engineering Education 4 (4), 8