Confidence assessment

in the teaching of basic science

A. R. Gardner-Medwin

Department of Physiology, University College, London

A scheme is described for including information about confidence in the computer-based assessment of students. After each answer, students declare a confidence level of 1, 2, or 3. If the answer is correct, then this is the mark awarded. If not, marks of 0, −2, or −6 are awarded. Students do well on this scheme if they can discriminate between when they are sure of correct answers and when they are partly guessing. In self-assessment, students are trained to reflect on their reasoning, and to develop the skills of correct confidence judgement. The task of writing tests is simplified, since it becomes less important to ask complex questions. Simple direct questions discriminate better between students than they do with ordinary marking. Good students answer correctly with high confidence, while weak students moderate their confidence level if they know they are uncertain, or else lose heavily when they make mistakes. Preliminary data are presented from self-assessment trials amongst medical students.

Introduction

Automated assessment suffers from two problems that are considered here. Firstly, it seldom makes use of information about how confident a student is in the answer given, which is part of what we take into account in assessing students person-to-person. Secondly, it often involves the construction of complex questions to ensure that students cannot get good marks by a combination of partial knowledge and guesswork. Such questions can be ambiguous and open to different levels of interpretation, so the creation of satisfactory tests is time-consuming.

This paper is primarily about confidence assessment. However, the use of confidence assessment also simplifies the construction of tests, since it means that knowledge can be tested more thoroughly with simple direct questions. With confidence assessment, a student cannot do well by employing guesswork. The marking scheme is such that he/she must either declare the low confidence that is appropriate if guesswork is employed, or else lose heavily whenever a mistake is made. A student who declares confidence in wrong answers will expect to do particularly badly.

The importance of confidence judgements in the study of basics

It is important that students should have confident knowledge of the basics in their subject. If they are uncertain about things, for example the meanings of words, they are handicapped in future study. They need, of course, to be able to produce correct answers to a good fraction of the relevant questions one might ask, as tested in conventional marking schemes. However, it is equally important, perhaps even more important, that they should be able to identify when they are likely to be getting the answers right and when not. Confident belief in answers that are wrong is far worse than recognition that one simply does not know the answer.

Confidence judgements are well recognized as relevant to knowledge assessment (see, for example, De Finetti, 1965; Good, 1979; and literature reviewed in Hutchinson, 1991). Confidence is a form of probability forecast (Dawid, 1986): an estimate of the probability that an answer will turn out to be correct. Of course one can have high or low confidence about things that are matters of fact, not of probability: for example, you might judge your confidence that Canberra is the capital of Australia to be 0.3, 0.7 or 0.95, etc. What is estimated here is the probability that a proposition will turn out to be true when, in your own brain, it is subject to whatever degree of uncertainty of recall and confusion or unreliability of reasoning applies to the present issue. If your confidence judgement is good, you will be correct on about 0.7 of those occasions when you judge your confidence to be 0.7. There is a substantial psychological literature on the extent to which people make errors of such confidence judgements (see, for example, Erev *et al.*, 1994). The main point for the present is simply that students benefit from good confidence judgement. It is an essential skill for efficient study and work practice; for example, it is obviously necessary for selective and efficient use of a dictionary, or for deciding when it is appropriate to take time to check a calculation or the expression of a formula. We expect good students to show good confidence judgement, yet we seldom teach it explicitly or give them explicit practice.

A scheme for combined assessment of accuracy and confidence

Students are asked to state with each answer their level of confidence (C = 1, 2 or 3) in the correctness of their decision. If the answer is correct, then this is the score awarded. A wrong answer leads to a score of zero for level 1, and –2 or –6 for levels 2 and 3 respectively. Level 2 gives equally weighted negative marking for wrong answers, as in many standard assessment schemes including that used at University College London for medical examinations. Students are told to choose level 2 unless they are very confident (>80% chance of being right: odds 4:1), when they should choose level 3, or rather hesitant (<67% chance of being correct: odds 2:1), when level 1 is appropriate. This strategy is optimal to maximize their expected scores.

The confidence assessment scheme is a proper marking scheme (Dawid, 1986), in the sense that a student can never expect to gain systematically by either overestimating or underestimating confidence. High marks are gained firstly, of course, by getting the

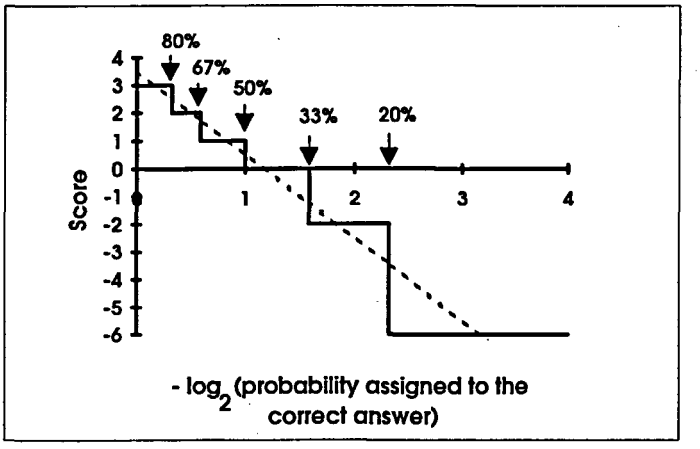

Figure 1. The relation between score and the probability assigned by a student to the right answer, in a binary test. The score is approximately linearly related to the logarithm of the probability, plotted on the abscissa. Note that an assigned probability of 20% means that the student has assigned 80% probability to the wrong answer.

answers correct. However, students will lose out unless they develop the introspective skills that lead to correct confidence judgement, or discrimination between when they are likely to be getting answers right or wrong. Since reflecting on confidence may lead to students changing their minds about the preferred answer, they are given the opportunity (by entering -1 as the confidence) to go back and change the answer. The new answer is then what counts, and confidence is requested over again. Once a valid confidence level has been entered, the student is immediately told (for formative assessment) whether the answer was correct and what score was awarded.

With binary (true/false) questions, the score on this simple scheme is approximately linearly related, over its six discrete values from 3 to -6, to the logarithm of the confidence expressed for the correct answer (Figure 1). Such a logarithmic function is one of the more commonly proposed proper scoring functions for eliciting honest confidence declarations (Good, 1979), and in fact has important theoretical merit as a measure of knowledge or ignorance when it comes to combining scores for knowledge about different aspects of a topic. Note that with binary questions, 80% confidence expressed for an answer that is wrong implies 20% confidence for the correct answer. With questions having free-format answers, this would simply be an upper bound for the confidence for a correct answer.

The delivery system

A new authoring and delivery system (LAPTTOP: London Agreed Protocol for Teaching and Testing of Physiology) has been developed to incorporate confidence assessment. This is supported by the physiology departments in most of the London medical schools, and is also being used by several other departments in University College. Although designed with the needs of physiologists in mind, it is equally applicable to many other disciplines. It runs exercises defined in ASCII text files, capable of being prepared and edited on any wordprocessor, together with diagrams prepared as .PCX or .BMP files.

The main novel feature is the incorporation of confidence assessment. Important related features, however, are the ability to store and collate student comments entered in the context of individual questions (e.g. indicating that questions seem to be ambiguous or marked incorrectly), and to total how often each question is answered correctly and incorrectly at each confidence level. Questions that are answered incorrectly with high confidence can be particularly revealing to teachers on the relevant course.

Confidence judgements are handled in the same way for binary and free-format answers. The LAPTTOP system can verify answers consisting of words, phrases, numbers within specified ranges, or quantities (numbers + units). It is used at present solely for formative assessment, i.e. self-assessment and tutorial exercises. For these, the flexibility to vary the question format in any sequence is particularly valuable. The system runs on the University College Ethernet, collating usage information in one place from any number of simultaneous users. Students can also download the files and use them for study at home, since they run on any IBM-compatible PC without stringent specification requirements.

Confidence data for binary questions in physiology

Statistics are presented here from use of a bank of 376 binary questions in basic physiology over a period of seven weeks prior to exams. Students were encouraged – not required – to use the system, and data was collected anonymously. A total of 121 sessions were recorded, in 14 of which the students chose not to use confidence assessment. The remaining 107 sessions were analysed (77 male students, 30 female). 6,950 questions were presented on the screen, of which 7% were skipped without answering, 26% answered at confidence C = 1, 20% at C = 2 and 48% at C = 3. The percentages of the answers given at each confidence level that were in fact correct were 56% at C = 1, 71% at C = 2 and 83% at C = 3. These are within the limits of confidence for which it is optimal for the students to choose each of the three levels (50%–67% for C = 1, 67%–80% for C = 2, 80%–100% for C = 3). This suggests that on average these students were well calibrated in their confidence judgements, getting about the expected percentages correct at each level. Four sessions identified individuals who were significantly overestimating their confidence when using C = 3: percentages correct (out of 16–45 questions answered at this confidence level) were in each case more than three standard deviations below expectation for an 80% chance of being correct ($P<0.002$). There was no evidence of significant underestimation of confidence in any sessions.

The male students in this self-selected sample had better knowledge than the females (71% correct overall for the males, 64% for the females). The males answered a substantially higher fraction of the questions with high confidence (58% at C = 3, compared with 31% for the females). The percentages correct at each confidence level were not, however, much different for the sexes (56%, 71% and 82% for the males and 57%, 72% and 85% for the females). Thus the greater fraction of males choosing C = 3 was due largely to the difference in ability rather than to the known tendency of females to declare lower confidence than males for the same level of performance (Walkerdine, 1990). Such problems as were revealed in the calibration of confidence judgements seem to have been

largely among the males. Fourteen out of fifteen sessions in which the scores at C=3 or C=2 were more than two standard deviations below the appropriate expectation were male (including all four extreme cases mentioned above).

The benefits of confidence assessment

The students' informal reactions indicate that they appreciate confidence assessment. They see it as testing something important in relation to their knowledge, and they like the option of avoiding negative marking (with C = 1) when they are hesitant. In only 12% of sessions did they opt not to use confidence assessment, or not to vary their declared confidence level, despite the fact that their multiple-choice-question examinations do not involve confidence assessment. With some students, the principles involved have seemed complicated when first explained, but present no problem once they gain experience.

In many situations, university teachers are faced with a wide range of student backgrounds and abilities. It is important to identify and assist students early on, where knowledge is weak or absent. Because of the diversity of student problems, this is an area where automated techniques can help a great deal. One of the merits of confidence assessment is that it is possible to give students relatively easy basic tests without seeming condescending to good students who get them mostly correct. These students get a kick out of getting the answers right at maximum confidence, while the weak students can answer questions initially with low confidence, then increase confidence as they repeat exercises in an area. We use the system in this way, for example, to help with the basic mathematical skills necessary for a BSc degree.

There is little evidence from our data so far to suggest that many medical students are radically wrong in their confidence judgements overall. Where students do have such a problem, use of a confidence-assessment scheme seems a potentially valuable tool for helping with the problem. Even for students who are well calibrated in this respect, it is instructive for them to think, on the occasions when they make mistakes with high confidence, why they failed to see the risks they were taking. Such experiences should help students to develop careful habits of thought and to identify when their knowledge is tentative, or subject to serious misconceptions.

Acknowledgements

The physiology questions were written by staff at the Charing Cross and Westminster Medical School. The LAPTTOP system is available to university departments agreeing to pool material with other registered users.

References

De Finetti, B. (1965), 'Methods for discriminating levels of partial knowledge concerning a test item', *British Journal of Mathematical and Statistical Psychology*, **18**, 87–123.

Dawid A.P. (1986), 'Probability forecasting' in Kotz, S., Johnson, N.L. & Reid, C.B. (eds), *Encyclopedia of Statistical Sciences*, **7**, 210–18.

Erev, I., Wallstein, T.S. & Budescu D.V. (1994), 'Simultaneous overconfidence and underconfidence – the role of error in judgement processes', *Psychological Review*, **101**, 519–27.

Good, I.J. (1979), ' "Proper Fees" in multiple choice examinations', *Journal of Statistical and Computational Simulation*, **9**, 164–5.

Hutchinson, T.P. (1991), *Ability, Partial Information, Guessing*, Adelaide, Rumsby.

Walkerdine, V. (1990), *Schoolgirl Fictions*, London, Virago.