Formative self-assessment using multiple true–false questions on the Internet: feedback according to confidence about correct knowledge

KHALID S. KHAN, DAVID A. DAVIES & JANESH K. GUPTA
Academic Department of Obstetrics and Gynaecology and Physiology, University of Birmingham, UK

SUMMARY In undergraduate clinical courses, students are often dispersed over several teaching sites. Traditional curricula do not have mechanisms that allow monitoring of an individual student’s educational progress or that of a small group of students at a teaching site relative to that of the whole group. To address these issues, we have developed a web-based formative assessment system that consists of knowledge tests based on multiple true–false questions. During the test, in addition to marking true or false against each question, students indicate their level of confidence (doubt or certainty) about each answer. The feedback consists of whether the answer is correct or incorrect and the confidence with which the student had responded. Feedback to students assesses their own performance in relation to that of their peers. Feedback to tutors provides anonymized information about the level of achievement of students at their teaching site relative to that of students at other sites. This system has the tools that students can use to direct their learning and tutors can use to tailor their teaching in the light of the instantaneously available comparative feedback.

Introduction
Self-assessment, evaluation of one’s own learning in order to improve one’s overall capability, is used frequently in medical education. It is a valuable tool helping learners to become aware of their strengths and weaknesses and to undertake directed study. Multiple-choice tests of knowledge are perhaps the most common form of self-assessment because they efficiently sample the educational content, they have high degree of reliability and they are easy to score. However, their value in medical education has been questioned because they are seen as tests of factual recall, which does not particularly assess application of knowledge for problem solving. This is partly because of the way in which questions are constructed and also partly because of the way in which they are marked (Pickering, 1979; Hunt, 1982). We have devised a system of self-assessment using the Internet for an undergraduate module in obstetrics and gynaecology that overcomes several of these problems and is expected to help students improve their quality of knowledge.

Curriculum context
The undergraduate curriculum in obstetrics and gynaecology at our medical school was revised in 1998–99 to bring it in line with the recommendations of the General Medical Council (GMC, 1993). In organizational terms, each final-year block consists of approximately 50 students who are distributed between seven to eight teaching sites (hospitals). In the old curriculum, students had no formative self-assessment. Hence, they had no formal feedback to aid self-directed learning. Moreover, they did not have a mechanism to monitor their educational progress relative to that of students at other sites.

The new curriculum in obstetrics and gynaecology consists of an eight-week module in the final year called Development I. The aims, objectives and learning outcomes of this module are given in Table 1. The course is divided into seven educational themes, each with its specific objectives and learning outcomes. These are provided to both students and tutors in a course handbook that serves as a study guide. The teaching, learning and assessment strategy of the module is directed towards helping students acquire competence in the learning outcomes. The teaching and learning programme is based on independent study, lectures, small-group sessions and practical work.

Students’ assessment is based on both formative and summative assessments. Formative assessment consists of self-assessment using knowledge tests. The tests are conducted in an electronic form using the World Wide Web. The tests are administered weekly to assess each one of the educational themes in sequence as described below.

Web-based self-assessment
The web-based self-assessment is an assessment tool for knowledge, i.e. information stored in the student’s mind. Multiple true–false questions are reputed to be able to assess knowledge and its interpretation (Lennox, 1974). The web-based formative assessment of our course is based on seven tests of knowledge that consist of multiple true–false questions. The questions are grouped according to the educational themes described above and they are timed-released on the Monday of each educational theme week. Once released they remain available to the students for self-assessment until the summative assessment in week eight.

Students can take the assessments at their own convenience either at their teaching sites or on their own
In this paper, the web-based self-assessment system will be described according to the following three stages of test development:

- the question;
- the response;
- the feedback.

### The question

The multiple true–false questions included in the web-based tests were originally written by the tutors of the old curriculum and were then used for summative assessment. The results of tests conducted over the last 2 years were stored electronically, which provided the opportunity to analyse the questions and items in depth. This analysis has led to the current weekly self-assessment knowledge tests. The tests were developed using the following strategy.

First, we focused on content validity, which addresses a simple qualitative question: do the items appear (on the face of it) to be measuring what is intended to be measured? What is intended to be measured in educational terms is described in the learning outcomes of our course. We mapped each item against the learning outcomes determined at the beginning of the new course’s development by expert judgement. These learning outcomes are provided to all students and teachers in the course handbook describing each educational theme and its teaching and learning strategy in detail. Questions that did not map against the learning outcomes were discarded. If learning outcomes were not covered by the questions, we developed new questions and included them for analysis in the new tests.

After determining validity, we assessed questions for quality measuring the degree of difficulty of questions and their ability to discriminate between students of differing abilities. These qualities of multiple true–false questions can be measured mathematically in a number of different ways (Lennox, 1974). We used the following approaches: The degree of difficulty was measured in terms of percentage of correct responses to the question (percentage correct), called the facility index. The discrimination index was based on comparison of the proportion of correct responses between high performers (upper quartile) and low performers (lower quartile). For all questions that represented a valid assessment of the learning outcomes, we included them in the new tests as long as the facility index was > 0.7 regardless of discrimination index; and facility index 0.3–0.7 with discrimination index > 0.3. We used this approach because when multiple true–false questions are used for formative assessment, discrimination is less important. If the facility and discrimination were < 0.3, we either amended the question and included it in the new test for reanalysis or we discarded the question completely.

In this manner, we assembled a database of test questions and categorized them according to educational themes. Prior to each Development I block, we randomly select 15–20 questions for each one of the educational themes for use by students in that block. When students ask to take a test from the self-assessment website, the

---

### Table 1. The aim, objective and learning outcomes of Development 1 module at University of Birmingham Medical School.

<table>
<thead>
<tr>
<th>Aim</th>
<th>To teach core gynaecology and obstetrics (and related genetics and neonatology)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective</td>
<td>To prepare students to investigate, diagnose and manage patients with general gynaecologic and obstetrics problems (and related genetics and neonatology)</td>
</tr>
<tr>
<td>Learning outcomes*</td>
<td>By the end of the course (divided into seven educational themes) students should be able to:</td>
</tr>
</tbody>
</table>

* Theme week 1:
  - Investigate, diagnose and manage patients with benign menstrual and menopausal problems

* Theme week 2:
  - Investigate, diagnose and manage patients with incontinence, prolapse, abnormal smears, postmenopausal bleeding and adnexal mass

* Theme week 3:
  - Investigate, diagnose and manage patients with infertility, contraceptive issues and sexually transmitted diseases

* Theme week 4:
  - Investigate for genetic problems and screen patients in pregnancy

* Theme week 5:
  - Investigate, diagnose and manage patients with bleeding and medical problems in pregnancy

* Theme week 6:
  - Investigate, diagnose and manage patients with normal and abnormal labour

* Theme week 7:
  - Investigate, diagnose and manage puerperal and neonatal problems

* Theme week 8:
  - Pass the summative assessment

Note: * Each educational theme has specific objectives and learning outcomes which are provided in a course handbook that serves as a study guide.

home computer or centrally at the medical school computer cluster. When a student logs on to the website and wishes to take a self-assessment test, the student is asked to choose an educational theme. When the educational theme is selected, the student is presented with 15–20 questions on that theme each with 1–4 true–false items. Their responses are collated and feedback is provided immediately, which serves as an educational tool. The results of these tests are purely for self-assessment, and they do not contribute towards summative assessment. In addition, the questions used in formative assessment are excluded from the end-of-course summative assessment.
questions for the specified educational theme are presented to them for response.

**The response**

Indicating true or false against each item of a question is the commonest way of responding to multiple true–false questions. These true or false responses are then used to determine whether a student can answer the questions correctly, and form the basis of a test score. If the student gives a correct answer no distinction is made between whether the student is sure or unsure of the correct answer; it is simply assumed that the student possesses the specific knowledge. Similarly if the student gives a wrong answer, no distinction is made between whether the student was sure or unsure of its correctness; a wrong answer is simply taken to mean that the student has not yet learned the correct knowledge. From an educational viewpoint it is important to identify the degree to which a student is sure or confident about the correctness of his or her responses (Hunt, 1982; Hassmen et al., 1996).

The practical importance of knowledge is that doctors make decisions about the care of their patients based on knowledge about which they are relatively sure (including knowledge about consequences of their decisions). The knowledge about which a doctor is relatively sure may be correct or wrong. In either case the doctor can use correct or incorrect knowledge as a basis for decision making. If the knowledge is correct then the decision making will be well informed. However, if the knowledge is wrong, the decision making will be misinformed. Between these two extremes lie the grey areas where the doctor’s knowledge is not usable for decision making. The relationship between correctness of knowledge, the confidence with which the doctor believes the knowledge to be correct, and the usability of knowledge in decision making is shown diagrammatically in Figure 1.

With this background we considered it insufficient to ask students to simply identify responses to questions as correct or incorrect. As explained earlier, this is because such a dichotomy ignores the level of confidence a student has in the correctness of his or her knowledge. By ignoring this aspect of learning, when a student correctly responds

<table>
<thead>
<tr>
<th>Correctness of knowledge</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of confidence</td>
<td>Sure</td>
<td>Unsure</td>
</tr>
<tr>
<td>Usability of knowledge</td>
<td>Usable</td>
<td>Unusable</td>
</tr>
<tr>
<td>Level of decision making</td>
<td>Well informed</td>
<td>Somewhat Informed</td>
</tr>
</tbody>
</table>

*Source: Based on Hunt (1982).*

**Figure 1.** The relationship between correctness of knowledge, the confidence with which the person believes the knowledge to be correct, and the usability of knowledge in decision-making.
Confidence assessment in formative knowledge tests

Table 2. Scoring matrix for the confidence responses.

<table>
<thead>
<tr>
<th>Confidence responses</th>
<th>Very sure</th>
<th>Fairly sure</th>
<th>Neutral</th>
<th>Unsure</th>
<th>Pure guess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>+50</td>
<td>+45</td>
<td>+37</td>
<td>+27</td>
<td>+10</td>
</tr>
<tr>
<td>Wrong</td>
<td>−60</td>
<td>−32</td>
<td>−16</td>
<td>−4</td>
<td>+5</td>
</tr>
</tbody>
</table>

Source: Based on Hassmen & Hunt (1994).

to a question, it is not possible to determine whether the correctly identified knowledge is usable for decision making or not. Moreover, and more importantly, when a student incorrectly responds to a question, it is not possible to determine whether the incorrect knowledge is due to plain ignorance or, worse still, due to misinformation. Misinformation is particularly dangerous because the student strongly believes that the wrong answer is correct. Such knowledge is usable and it is highly likely to result in erroneous decision making in healthcare.

In a competence-based curriculum, the purpose of education is to prepare students to perform various tasks safely. Hence, assessment should be directed towards identifying responses that are both correct and usable; being correct alone is not sufficient. Therefore, the responses in our web-based assessment require students to indicate their level of confidence (doubt or certainty) about each answer. The level of confidence is graded as follows:

- very sure;
- fairly sure;
- neutral;
- unsure;
- pure guess.

These choices appear in a pop-up menu in a box next to the true or false lozenges against each item of a question. The student reads the question and the items simply marking them true or false and choosing their level of confidence for each answer. After submitting their responses to the questions and items in the test the student is prompted to request feedback.

The feedback

The feedback is itemized on a web page that can be printed out on paper for use in self-directed study. Feedback is given to students on whether the answer is correct or incorrect and the confidence with which they had responded. An explanation is also provided as to why the answer is correct or incorrect. The most useful information for the students is to see for how many questions they were confident of the answer but it was wrong. This means that the particular topic or fact has been misunderstood and it should be revised.

At the end of the itemised analysis there is a summary score based on the traditional scoring system: The true or false responses are used to determine whether a student can answer the questions correctly which determines a non-weighted test score presented as a percentage. A summary confidence weighted score is also provided. The confidence score depends on both the knowledge possessed and the confidence with which it is possessed.

Table 2 shows the scoring matrix for generating a weighted score for each response, which is summed up from all items to generate a summary confidence mark also expressed as a percentage (Hassmen & Hunt, 1994). For interpretation of the summary scores, the students are advised that if they are correct in their assessment of how surely they know the answer to a question, their confidence score will be roughly equal to their non-weighted score. Significant deviation between the two scores is important as it identifies whether students over- or underestimate their knowledge. For example, if a student’s non-weighted score is higher that the confidence score, he/she knows more than he/she thinks he/she does and should be reassured by that.

This feedback to students is anonymous and non-threatening, as it is not used for summative assessment. It is immediate, individualized and meaningful, directing further study. It is convenient as the students can take the tests at any time and they can repeat the tests after directed study. It is frequent in that it allows weekly testing in line with curricular progress. Above all it is comparative with students being able to assess their own performance over time as well as see their scores in relation to those of their peers. These factors encourage a large proportion of students to take formative assessments. An example of the use of formative assessment scores to track performance of individual students over time is given in Table 3. With such an assessment the students should also overcome the isolation they might have felt in the past when they had little idea of what students were learning at other teaching sites.

As far as the course organizers are concerned, the feedback allows tracking of students’ achievement during the course at the various teaching sites. The performance of students in the formative tests provides feedback on the effectiveness of instruction. An example of the performance of various teaching sites for 5 weeks during the course, where some sites are doing not as well as others, is given in Table 4. In this manner, teaching sites where students are doing worse that other sites can be identified and corrective measures can be taken in time well before the summative assessment.

Conclusion

In undergraduate clinical courses, where students are dispersed over several teaching sites, our web-based formative assessment system allows individual students to monitor their educational progress. It also allows course administrators to monitor the performance of groups of students at various teaching sites relative to that of the whole group. The test analysis allows the measurement of confidently possessed correct knowledge and it identifies misinformation (incorrect knowledge possessed with confidence).
### Table 3. Use of formative assessment scores to track performance of individual students.

<table>
<thead>
<tr>
<th>Student</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5</th>
<th>Week 6 &amp; 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Attempt</td>
<td>Score</td>
<td>Attempt</td>
<td>Score</td>
<td>Attempt</td>
<td>Score</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Conf.</td>
<td>Mean</td>
<td>Conf.</td>
<td>Mean</td>
<td>Conf.</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>82%</td>
<td>2</td>
<td>77%</td>
<td>1</td>
<td>85%</td>
</tr>
<tr>
<td>B*</td>
<td>1</td>
<td>61%</td>
<td>2</td>
<td>66%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Class average</td>
<td>56</td>
<td>76%</td>
<td>62</td>
<td>76%</td>
<td>55</td>
<td>79%</td>
</tr>
</tbody>
</table>

Notes: * = Student not performing well; attempt = number of attempts; mean = average non-weighted score of all attempts; conf. = average confidence score of all attempts. See text for details. Italic lettering represents attempt and score for poorly performing students.

### Table 4. Use of formative assessment scores to track performance of different teaching sites.

<table>
<thead>
<tr>
<th>Site</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5</th>
<th>Week 6 &amp; 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Uptake</td>
<td>Score</td>
<td>Uptake</td>
<td>Score</td>
<td>Uptake</td>
<td>Score</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Conf.</td>
<td>Mean</td>
<td>Conf.</td>
<td>Mean</td>
<td>Conf.</td>
</tr>
<tr>
<td>A</td>
<td>100%</td>
<td>82%</td>
<td>100%</td>
<td>77%</td>
<td>100%</td>
<td>87%</td>
</tr>
<tr>
<td>B</td>
<td>100%</td>
<td>78%</td>
<td>75%</td>
<td>70%</td>
<td>75%</td>
<td>76%</td>
</tr>
<tr>
<td>C*</td>
<td>80%</td>
<td>66%</td>
<td>72%</td>
<td>69%</td>
<td>79%</td>
<td>80%</td>
</tr>
<tr>
<td>D</td>
<td>100%</td>
<td>80%</td>
<td>77%</td>
<td>79%</td>
<td>100%</td>
<td>80%</td>
</tr>
<tr>
<td>E</td>
<td>100%</td>
<td>85%</td>
<td>100%</td>
<td>79%</td>
<td>100%</td>
<td>75%</td>
</tr>
<tr>
<td>F</td>
<td>100%</td>
<td>82%</td>
<td>72%</td>
<td>77%</td>
<td>100%</td>
<td>76%</td>
</tr>
<tr>
<td>G*</td>
<td>50%</td>
<td>61%</td>
<td>50%</td>
<td>81%</td>
<td>75%</td>
<td>69%</td>
</tr>
<tr>
<td>All sites</td>
<td>90%</td>
<td>76%</td>
<td>86%</td>
<td>75%</td>
<td>90%</td>
<td>75%</td>
</tr>
</tbody>
</table>

Notes: * = Teaching sites not performing well; mean = average non-weighted score of students at each site; conf. = average confidence score of students at each site. See text for details. Italic lettering represents uptake and score of poorly performing teaching sites.
Notes on contributors

KHALID KHAN is a Lecturer in the Academic Department of Obstetrics and Gynaecology of the University of Birmingham, UK.

DAVID DAVIES is a Senior Lecturer in the Department of Physiology of the University of Birmingham, UK.

JANESH GUPTA is a Senior Lecturer in the Academic Department of Obstetrics and Gynaecology and Physiology of the University of Birmingham, UK.

Acknowledgement

The authors are grateful to Afran, Ahmed, Beatie, Cooper, Morgan, Gee, Ibrahim, Kelly, Lueseley, Mascarenhas, Olah, Kilby, Penketh, Shahmanesh, Mann, Whittle and others who contributed questions to the database.

References


LENNOX, B. (1974) Hints on the Setting and Evaluation of Multiple Choice Questions on the One from Five Type, Medical Education Booklet No. 3 (Dundee, Association for the Study of Medical Education).

PICKERING, G. (1979) Against multiple choice questions, Medical Teacher, 1, pp. 84–86.