# The Development of a New Method of Knowledge Assessment: Tailoring a Test to a Doctor's Area of Practice

Jane Dacre, MBBS, MD, Henry W.W. Potts, MSc, PhD, CStat, David Sales, MBBS, Hilary Spencer, and Alison Sturrock, MBChB

## Abstract

The practice of clinical medicine is becoming increasingly specialized, and this change has increased the challenge of developing fair, valid, and reliable tests of knowledge, particularly for single candidates or small groups of candidates. The problem is particularly relevant to the UK's General Medical Council's Fitness to Practice procedures, which investigate individual doctors. In such cases, there is a need for an alternative to the conventional approach to reliability estimation that will still allow the delivery of reproducible and standardized tests. This report describes the three-year process (starting in 2005) of developing a knowledge test that can be tailored for individual doctors practicing in narrowly specialized fields or at various stages in their training. The process of test development for this study consisted of five stages: item writing, to create individual questions; blueprinting, to establish the content and context that each item might test; standard setting, to calculate for each question a theoretical probability that a doctor of just-adequate capability would answer the question correctly; reference data collection, to determine for each item the distribution of scores to be expected from a large population of doctors in good standing; and test assembly, to select sets of questions that together formed complete and balanced tests. Tailored testing is a valid, feasible, and reproducible method of assessing the knowledge of one doctor or small groups of doctors who are practicing in narrow or subspecialty areas.

Acad Med. 2009; 84:1003–1007.

In the past decade, there have been significant developments in the United Kingdom in the way that clinical medicine is delivered and that doctors are assessed. Previously, most doctors practiced as generalists, and only a minority specialized in particular areas of medicine. With recent advances in medical science and the demands for high-quality specialized care, subspecialization of doctors and the number of recognized subspecialties have increased. Clinicians have therefore developed narrower fields of knowledge but have explored those fields to much greater depths.

This development poses a problem for efforts to devise an assessment of those doctors who practice in subspecialties or who have a very restricted field of practice. More attention has been focused on this problem by the proposed introduction of revalidation, which would require doctors to demonstrate that they are up-to-date in the relevant clinical knowledge in their area of practice.

Subspecialization also causes difficulty in the assessment of individual doctors who have been referred to the General Medical Council (GMC). Since 1997, the GMC has assessed the performance of individual doctors whose fitness to practice has been called into question; this assessment uses workplace-based peer review and a test of competence called the Fitness to Practice (FtP) procedures. Tests have been developed for each specialty.[1] The peer review components were the predecessors of the now well-established workplace-based assessments in use in the National Health Service of the United Kingdom. The current FtP test of competence includes a knowledge test and an objective structured clinical examination (OSCE); each specialty has its own test. However, with the subspecialization of doctors, it has become increasingly difficult to assess a doctor's overall knowledge and abilities solely on the basis of his or her performance on a standard test of competence.

Therefore, during the three years starting in 2005, we developed a tailored test of knowledge to be used in the assessment of individual doctors as part of the GMC FtP procedures. In this tailored test, each assessment is unique to the doctor under investigation; thus, we had to develop new ways to assess these doctors. Using the same principles and this model, smaller specialties could develop their own assessments.

## Development of the Innovation

In this study, the overall process of test development consisted of five stages:

- item writing, to create individual questions;
- blueprinting, to establish the content and context that each question might test;
- standard setting, to calculate for each question a theoretical probability that a

**Prof. Dacre** is professor of medical education, honorary consultant physician, vice dean, Faculty of Biomedical Sciences, and director, Division of Medical Education, University College London Medical School, London, United Kingdom.

**Mr. Potts** is a chartered statistician, psychologist, and lecturer, Centre for Health Informatics and Multiprofessional Education, University College London Medical School, London, United Kingdom.

**Dr. Sales** is a general practitioner and consultant in medical education, based in southeast England.

**Ms. Spencer** is a research assistant, Division of Medical Education, University College London Medical School, London, United Kingdom.

**Dr. Sturrock** is clinical lecturer in medical education and honorary consultant physician, University College London Medical School, London, United Kingdom.

Correspondence should be addressed to Prof. Dacre, Division of Medical Education, UCL Medical School, Holborn Union Building, Highgate Hill, London, UK N19 5LW; e-mail: (j.dacre@medsch.ucl.ac.uk).

doctor of just-adequate capability would answer the question correctly;

- reference data collection, to determine for comparison the distribution of scores for each item to be expected from a large population of doctors who are in good standing and are similar in specialty and grade to the intended candidate; and
- test assembly, to select sets of questions that together formed complete and balanced tests.

Each of these stages is described below.

**Item writing**

We recruited, from all of the UK Medical Royal Colleges, question writers with expertise in writing single-best-answer (out of five choices) and extended matching questions. Specific invitations were sent to individuals known to write similar items for medical schools and the Professional and Linguistic Assessments Board test (a validated test used for the assessment of international medical graduates wishing to practice in the United Kingdom). General practitioners were targeted because they make up the largest specialty group in the United Kingdom and thus constitute the highest volume of referrals to the GMC; we therefore initially used general practitioners to develop this test method.

Specialty-specific and cross-specialty item-writing days were then convened, each consisting of a training session, which was followed by a facilitated writing activity to a defined template and content and then peer review of questions. The training involved a plenary session to familiarize the writers with the question style and to provide them with practice in constructing questions in a group. This session also familiarized the writers with areas of the blueprint in which additional questions were required for the bank. The facilitated writing involved writers who worked in small groups, with support from item-writing experts. We provided the participants with guidance on the house style and with a template to use in completing each question, which ensured the correct wording outline and layout of each item. Toward the end of the session, the peer group as a whole reviewed completed questions, and the group provided constructive feedback on the content of the question. We then referred the questions generated to an editing group composed of three content experts, who reviewed each question for content, context, and validity.

All questions were blueprinted to a grid designed specifically for that purpose.[2] We categorized them according to the headings of Good Medical Practice[3] (the code of practice published by the GMC) and an agreed-upon list of specialty-specific common (frequent—such as acute coronary syndrome) and key (rare, but serious if missed—such as Wegeners granulomatosis) medical conditions. For example, one question might be classed as "upper gastrointestinal bleeding, diagnosis"; another might relate to ethical concerns in contraception. This process enabled the construction of well-balanced tests appropriate to the skills and experience of the doctor concerned.

Speedwell Computing Services, a company specializing in the development of software for processing examination and assessment data, provided management of the item bank. Questions were stored within the blueprint, in a question bank designed for this purpose.

**Standard setting**

During the three-year period, we held eight standard-setting days to determine a theoretical standard for each question. The standard for each question was set on one occasion only. Standard-setters were recruited from a pool of doctors who were familiar with standard setting and who knew the level of knowledge expected of the doctor in question. Each question was reviewed individually by 8 to 12 standard-setters who used the Angoff method.[4] This method is a process used internationally to set standards in tests of competence in medicine. First, the group discusses and agrees on the nature of a just-adequate doctor, and then conceptualizes a group of such borderline doctors. Each standard-setter then estimates the percentage of the group of borderline doctors who would correctly answer each question. Percentages assigned by each standard-setter are averaged for each question and then summed to arrive at an overall minimum acceptable score.

**Creation of the reference group**

To ensure that the new knowledge items were reliable, fair, and valid, we tested each item on a group of volunteer doctors (the reference group). Since October 2006, we held 47 specialty-specific days to pilot-test these questions. To recruit volunteers, advertisements were placed on the UK Medical Royal College Web sites, on doctors.net, and in the *BMJ*, *GMC Today*, and *Hospital Doctor*; in addition, invitations were sent to the education departments at all UK postgraduate deaneries. The advertisement contained a statement making it clear that if the performance of an individual doctor were particularly poor, the investigators would be obligated to refer that individual to the GMC.

We recruited doctors from all over the United Kingdom. The only restriction was that doctors had to be fully registered with the GMC; the group included all levels of experience. Before accepting a doctor as a volunteer, we checked each doctor's GMC registration number to ensure that he or she was currently registered to practice and was not under active investigation by the GMC.

Volunteers visited the GMC offices in London to take a knowledge test (using single-best-answer questions, which means that candidates select the most correct answer from (1) a list of five options, (2) a list of extended matching items (in which candidates match a scenario to an option from a list of >10 possible correct answers), and (3) a 12-station OSCE under examination conditions, including time restraints. Each volunteer received a small honorarium, and all received formal feedback about their performance in relation to that of the other doctors taking the same test.

We then reviewed every item (individual question) included on that day in terms of feedback from the volunteers; we also reviewed item statistics (facility indices and point-biserial correlation coefficients, which show how difficult and how discriminating each item was during that assessment). Questions that received poor feedback (questions that volunteers felt were inappropriate or were badly worded) or had poor performance statistics (low or negative

point-biserial correlation coefficients, which suggested that the items were not discriminating) were rejected and either returned to an item-writing day for revision or deleted. We entered the remaining questions and their associated statistics into a pool from which a specific test-of-competence paper can be created.

### Test assembly

Unlike other high-stakes medical assessments designed for large volumes of candidates, each GMC assessment of an individual doctor is tailored to his or her practice. Using the doctor's completed portfolio as a guide, the test writer selects each item according to the item statistics obtained by volunteer doctors of similar specialty and grade and the doctor's current range of practice. Each test can include knowledge and skills that may be generic for all doctors or specific to their specialty or, when appropriate, very specific subspecialty material.

The written papers in these tests comprise 200 items, predominantly in the single-best-answer format, to enable wide sampling by both content and context. Items typically are vignette based, and they require reasoning skills rather than straightforward recall of facts. All tests have a three-hour time limit.

## Test Implementation

In the GMC FtP assessments to date, the doctor's knowledge test results have been reported as numerical scores, with no pass/fail mark. The individual doctor's score is compared with the performance of a subset of doctors in the reference group. An example of such a comparison is shown in Figure 1. Previously, the doctor under investigation would have answered the same set of questions as the doctors in the reference group. The figure shows the histogram of the scores of 188 established GPs in an extended-matching-set knowledge test. A best-fitting normal distribution is superimposed.

With the tailored test, no member of the reference group will have taken the same test as the doctors in this study, but each of the questions will have been answered during a number of different volunteer (validation) events and in a variety of test configurations. In this way, we learn how each individual question performs.

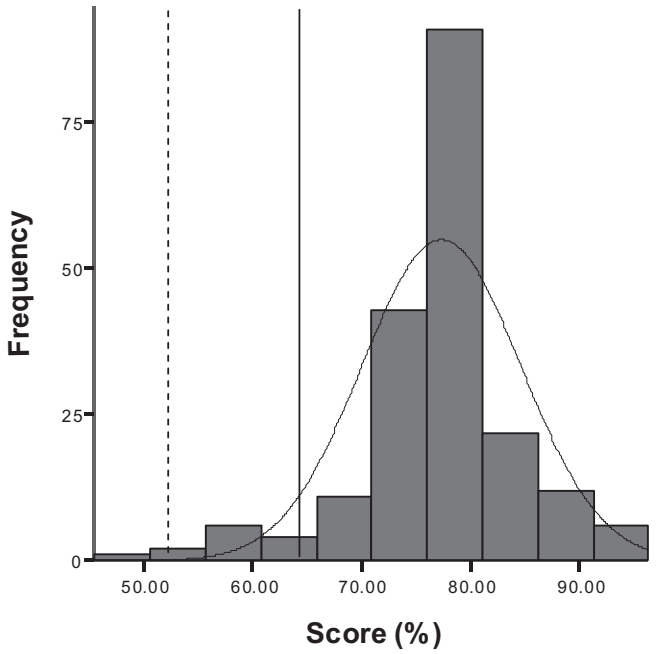


**Figure 1** Performance of 188 established doctors shown in a histogram. A best-fit normal curve is superimposed. The dotted vertical line shows the score achieved by the test subject (61%). The solid vertical line marks the Angoff score (68.5%).

For example, we know that question 1 has been answered correctly 82% of the time, across its appearance in different test configurations and validation events; question 2 has been answered 73% of the time; question 3 has been answered 75% of the time, and so on. The expected performance on the tailored test is the mean performance by the reference group on each of the question items.

Using this model, we can calculate the expected "typical" performance on the tailored test; however, we also need to know the range of scores that would be expected. Mathematical modeling was used to calculate this range (see the boxed item entitled "Details of the Modeling Procedure"). By using this mathematical model, we can predict the total scores that the reference group would have achieved. These predictions are based on the reference group's performance on individual items under similar conditions and on tests of comparable difficulty. The reference group's performance is expected to follow a normal distribution, with a mean equal to the mean

**Details of the Modeling Procedure**

In each tailored test, there are $N$ questions. Performance of the reference group on question i follows a Bernoulli distribution with a probability of success, $P_i$. We estimate the performance of the question from the data available, using only the questions in which the sample size is sufficiently large to ensure a good approximation to the population value. Thus, the mean for each item is $P_i$, and the variance is $P_i\,(1 - P_i)$.

We assume independence between the $N$ Bernoulli distributions for the $N$ questions in each tailored test. Therefore, the mean performance on the tailored test is $\sum_{i=1}^{N} P_i$ and the variance is $\sum_{i=1}^{N} P_i\,(1 - P_i)$. The distributional form, as a sum of $N$ Bernoulli distributions in which $N$ is large and $P_i$ and $(1 - P_i)$ are never very small, will be approximately normal. This number can then be rescaled for expression as a percentage score.

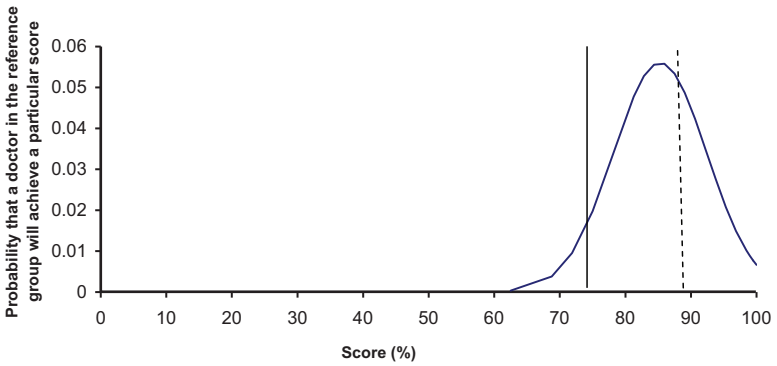


**Figure 2** Modeled performance of general practitioners in the reference group. The dotted vertical line shows the score achieved by the test subject (89%). The solid vertical line marks the Angoff score (74.6%).

performance over all of the questions asked and a variance, as described in the boxed item.

For this process to be valid, each question must perform at an equivalent level every time it is included within a specific test; that is, if 80% of doctors at an equivalent level of experience answer the question correctly in one specific test, then approximately 80% of doctors at the same level of experience should answer it correctly every time it is used. To ensure that this assumption is valid, the reference group's tests and the tailored test are designed to allow similar overall performance and are taken under comparable examination conditions.

As opposed to a histogram of actual data, the curves on the graphs shown in Figure 2 are probability density functions based on the actual data for each individual question included in the specific test. This means that the curve has been built mathematically to represent how the doctors in the reference group would have performed if they had taken the same test as the doctor under investigation.

This modeling process assumes that performance on the individual questions is statistically independent of performance on other questions—that is, it ignores intraquestion correlation. In reality, we would expect that there is some correlation between an individual's performance on the different questions, especially in the area of subspecialization. Any intraquestion correlation will reduce the variance of expected test performance. In ignoring this, we are making a conservative approximation, in that we are giving doctors who are being assessed the benefit of the doubt, because any below-average test performance will stand out more as the modeled variance is reduced. It would be impractical to model the intraquestion correlations, but, in our reference group data, they were on the order of 0.1 or lower and, thus, would have a negligible effect.

This procedure can be generalized to use in any subset of the reference group. We took care to include only questions for which the reference group was large enough that the estimate of the difficulty of each question is robust.

The tail on the left of the normal distribution curve in Figure 2 does not represent completed test scores of individual doctors in the reference group. It does, however, represent the probability, based on actual performance on individual items, that doctors in the reference group would achieve this score.

## Discussion

This study showed that it is feasible to develop a tailored knowledge test that is suitable for use for doctors with a defined area of clinical practice. Such a test was developed for the GMC FtP procedures, in which the doctor under investigation may work in a restricted clinical area. The tailored test is created by validating a large number of individual test items and modeling the results of an identical test. It addresses the difficulties seen in assessing individuals whose area of practice is narrow.

As long as the validated database of questions is sufficiently large, it is possible to tailor a test to any individual while also comparing the results with a cohort of questions previously answered by practicing colleagues. Tailored knowledge testing also could be used to revalidate doctors practicing in narrowly specialized areas or for assessments in subspecialties with small numbers of candidates.

Because we were keen to create tests of a length equivalent to tests achieving adequate reliability, and because we were constrained in the structure of the assessment day to a testing time of three hours, we elected to create a single test of knowledge rather than to create a generic and a specific test. The reliability of a test is based on the extent that a test is dependable, repeatable for individual candidates, and free from error; however, competence is highly content- and context-specific. This test was therefore designed to include 200 well-validated items to ensure a sufficiently large sample content[5]; the test has been shown to be reliable when tested on volunteer doctors.

However, a test cannot be reliable if it is not valid. In most assessments, reliability estimates depend on a sufficiently large sample through all possible sources of error, such as test items, examiners, and test occasion. This arrangement is not possible in the GMC FtP procedures, which test only one doctor at a time. However, because "reliability is not the whole story,"[6] it is relatively simple to derive a numerical score to estimate reliability, whereas validity is a conceptual term that is less easy to measure.

To ensure that this new test is valid, we have to use surrogate markers of reliability. These include ensuring that each question is written and reviewed by trained item writers, using experienced test developers to construct the test against a blueprint, getting trained content experts to perform standard setting in a rigorous manner by using an internationally accepted method, and, finally, ensuring that each item is pilot-tested by a volunteer group. The development of this concept is in keeping with current opinion suggesting that overreliance on the reliability of a test is an unnecessarily reductionist approach.[7,8] The currently popular Cronbach coefficient alpha uses just one of several analyses that may gauge the accuracy of educational measurements, and additional components of variance have been recommended.[9]

The statistical model used in the GMC test of competence assumes that each item performs independently from other items. This is a potential pitfall that is mitigated by careful blueprinting of the specific test to ensure no duplication of subject areas, and our data suggest that intraquestion correlations are low. The use of a modeled distribution does lead to the expectation of a greater amount of statistical understanding from users than does the previous approach, and it has required more training of the GMC performance assessors to increase their confidence to a point at which they can explain the method to a defense barrister at a GMC hearing.

The use of knowledge-based assessments at higher levels of clinical experience has been discussed in the context of revalidation and of the assessment of trainees who are nearing the completion of their training. This kind of assessment is being launched by the Federation of the Royal Colleges of Physicians for candidates in an increasing number of medical specialties. The numbers of practitioners working in these very specific areas are small, which reduces the rigor of conventional knowledge testing. In a programmatic, instructional-design approach to assessment, "simple" psychometric evaluation will not suffice.[10]

We therefore believe that we have developed a novel idea based on a legitimate Bayesian approach and have designed a method of robust, tailored testing that could be applied to revalidation, FtP assessments, and other high-stakes tests involving a single candidate or a small number of candidates.

## Acknowledgments

The authors are grateful to Joanne Turner and Samantha Henry for their help in the administration of these tests and to Lucinda Etheridge for reviewing this article. This project was funded by the General Medical Council.

## References

1 Southgate L, Cox J, David T, et al. The assessment of poorly performing doctors: The development of the assessment programmes for the General Medical Council's Performance Procedures. Med Educ. 2001; 35(suppl):2–8.
2 Sales D, Sturrock A, Boursicot K, Dacre J. The development of an electronic blueprint for the General Medical Council's Fitness to Practise procedures. Med Teach. (in press).
3 General Medical Council. Good Medical Practice, 3rd ed. London, United Kingdom: GMC; 2006.
4 Angoff W. Scales, norms, and equivalent scores. In: Thorndike R. ed. Educational measurement. Washington, D.C.: American Council on Education; 1971.
5 Swanson D, Norman G, Linn R. Performance-based assessment: lessons from the health profession. Educational Res. 1995; 24:5–11.
6 Schuwirth L, van der Vleuten C. How to design a useful test: The principles of assessment. Edinburgh, UK: Association for the Study of Medical Education; 2007.
7 Schuwirth L, van der Vleuten C. A plea for new psychometric models in educational assessment. Med Educ. 2006;40:296–300.
8 Schuwirth L, van der Vleuten C. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004;38:974–979.
9 Cronbach L. My current thoughts on coefficient alpha and successor procedures. Educational Psychol Meas. 2004;64:391–418.
10 van der Vleuten C, Schuwirth L. Assessing professional competence: from methods to programmes. Med Educ. 2005;39:309–317.