Joachim P Sturmberg
MBBS, MFM, PhD, FRACGP, is Associate Professor of General Practice, Monash University, Victoria and University of Newcastle, New South Wales. jp.sturmberg@gmail.com

Elizabeth A Farmer
BSc, MBBS, PhD, FRACGP, is Dean, Graduate School Medicine, University of Wollongong, New South Wales.

Assessing general practice knowledge base

The applied knowledge test

A multiple choice based knowledge test has been part of the examination for Fellowship of The Royal Australian College of General Practitioners (RACGP) since its inception. In line with current best practice, the format changed to a clinically based applied knowledge test in 2000. The test consists of two question formats – single best answer questions and extended matching questions. This article describes the features and characteristics of the RACGP applied knowledge test; the design and management of the question bank; the setting of the pass mark; and the performance of the test in terms of validity, reliability and candidate acceptability.

The Royal Australian College of General Practitioners (RACGP) Fellowship examination assesses competence for unsupervised clinical practice anywhere in Australia through three segments, each with a unique focus. The applied knowledge test (AKT) is a written examination that tests candidates' applied clinical knowledge. Other segments assess clinical problem solving skills[1] and ability to perform in a clinical situation. Approximately 400–500 candidates sit each administration of the RACGP examination, which is held twice yearly throughout Australia.

How applied knowledge questions are designed

The familiar multiple choice questions (MCQ) format of knowledge testing has been part of the examination since its inception in 1967. Reviews of the examination and candidate feedback consistently indicated that the MCQ was not seen as a good test of general practice knowledge. In 1999, as part of a re-design of the examination, the written testing components were modified to reflect actual clinical practice more effectively – the 'applied knowledge test'. Since then, all questions are based on a clinical scenario. This requires candidates to identify the most correct diagnosis or management strategy from a list of options. Candidates demonstrate not only their knowledge, but also their ability to apply that knowledge to the clinical situation described. Scenarios cover general practice encounters in line with BEACH data.[2]

Question formats

The AKT consists of two item types: 70 single best answer (SBA) questions and 80 extended matching questions (EMQ). Single best answer questions require the candidate to select the one correct answer from five response options (Table 1); EMQ items require candidates to select the most likely or best answer from a list of up to 26 choices. Frequently, several EMQ items are presented together with the same menu list and the same medical theme (eg. diagnosing

Table 1. Examples of question formats used in the AKT

Extended matching questions
A. Coeliac disease
B. Depression
C. Diabetes mellitus
D. Duodenal ulcer
E. Eating disorder
F. HIV/AIDS
G. Hyperthyroidism
H. Inflammatory bowel disease
I. Lymphoma
J. Malaria
K. Oesophageal stricture
L. Substance abuse
M. Tuberculosis

For the following patient with weight loss, select the MOST likely diagnosis from the options provided.

41 year old Kit Fung is a part time teacher's aide who has been underweight since her last child was born 3 years ago. She now complains of having lost another 3 kg in the past 6 months. She has had two previous miscarriages and has two children. She has had episodic bloating diarrhoea for a year but it has been worse lately. There are no other changes in her bowel habits, and she doesn't drink any alcohol.

Answer: A. Coeliac disease

Single best answer questions

8 year old Miranda Kelly presents with upper sternal pain. You note that she has a pyrexia of 40°C and swelling of the interphalangeal joints. She had streptococcal pharyngitis 2 weeks before this episode started. The MOST likely diagnosis is:
A. toxic synovitis
B. rheumatoid arthritis
C. rheumatic fever
D. eptic arthritis
E. steoarthritis

Answer: C. rheumatic fever

a respiratory condition). As in real practice, some scenarios would allow a number of possible responses (eg. a number of differential diagnoses, or management strategies). The single correct response is that which describes the most typical or likely condition, or the most appropriate management strategy.

To assist candidates to familiarise themselves with these two question formats, the RACGP offers online learning resources (www.gplearning.com.au) and pre-exam training workshops run by state faculties.

Constructing questions

Most scenarios are written by examiners who are general practitioners and RACGP Fellows. Writers come from all ages, stages of career and practice locations, and are recruited from examination panels in each Australian state. The writers, who work in groups, are asked to create scenarios and related questions that reflect general practice challenges in a range of settings including metropolitan, rural and remote, and that relate to patients from various ethnic backgrounds and age groups. Some scenarios are accompanied by images such as electrocardiographs, radiographs or clinical photographs. Questions are pilot tested among the writing groups. Items are reviewed by the national coordinator for the AKT segment before being entered into the item bank of the examination management system.

Selecting, reviewing and finalising the AKT

For each administration, the examination management system selects items from the item bank according to a number of criteria. The paper thus generated, is checked for duplication of topics and conditions, and if necessary, suitable replacement questions are inserted manually. Experienced examiners then review the draft paper for content relevance and clarity of wording. The final paper is then prepared for printing and a master scoring sheet forwarded to the data manager for scoring and analysis of candidates' responses.

Maintaining the item bank

The item bank serves several interdependent functions. As well as storing items, it holds information for each item about the presenting complaint, domains of general practice represented according to the RACGP curriculum,[3] and age and gender characteristics of each 'patient'. Finally, it also records the usage history of each item and information about its performance, such as item total correlation and item analyses.

These data form an important part of assessing the validity of the AKT component. The item total correlation measures the consistency with which all items of a paper assess the knowledge base of general practice.

Blueprinting

Blueprinting ensures the content representativeness of each exam segment as well as the examination as a whole, and is drawn from the BEACH survey.[2] Questions for the AKT paper are selected according to this blueprint and are further stratified by age groups to ensure a similar representation of paediatric, adult and geriatric cases in each paper.

Marking

Candidate answer sheets are computer scanned and automatically scored against the answer key provided to the statistician. Candidates' aggregate and segment scores are provided for their feedback.

Validity and reliability of the AKT segment

A number of strategies are used to support the validity of the AKT segment. Validity is the primary concern when evaluating any test. According to the standards for educational and psychological

Table 2. Reliabilities of the AKT examination segment (1999.2–2006.2)

Time of administration	Cronbach's α
1999.2	0.82
2000.1	0.82
2000.2	0.83
2001.1	0.85
2001.2	0.87
2002.1	0.90
2002.2	0.88
2003.1	0.88
2003.2	0.86
2004.1	0.86
2004.2	0.88
2005.1	0.82
2005.2	0.82
2006.1	0.84
2006.2	0.80
.1 = first examination, .2 = second examination	

testing, validity '... refers to the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences'.[4] Validity is a complex construct and four types are frequently considered in assessment:

- face validity estimates how well a test appears to measure a certain criterion; it does not guarantee that the test actually measures that criterion
- content validity estimates how far your test samples from or represents all of the relevant content
- concurrent validity estimates the degree to which a test correlates with a criterion measure given at the same time
- predictive validity estimates the degree to which a score on the test predict scores on another criterion measure available in the future.

Face validity of the AKT is supported by the use of writers and reviewers who are GPs from a wide range of Australian general practice backgrounds. Content validity is achieved by selecting items according to a blueprint, which represents the profile of complaints presenting to Australian general practice.[2] Systematic review of relevant evidence ensures the validity of the answer key. However, formal testing of concurrent or predictive validity has not been done.

The AKT is characterised by a high level of internal consistency. The internal consistency of a test is a measure of the coherency of the test. Cronbach's α is commonly used to index internal consistency and ranges between 0–1; scores above 0.8 are commonly accepted as a gold standard for high stakes examinations.[5] The reliabilities obtained over the past 15 administrations of the examination from 1999.2 to 2006.2 have consistently met and surpassed that criterion (*Table 2*).

In addition, AKT scores typically correlate with those of the two other examination segments in the range of 0.55–0.65. This figure indicates that the AKT has some variance in common with the other segments. Candidate acceptance of a test format is also an important aspect of its performance as well as providing support for its validity.

Standard setting

Standard setting for each of the three RACGP examination segments was introduced in 1999. A modified Angoff method[6] has been applied throughout that period to the AKT. As part of the ongoing review of all aspects of the RACGP examination, AKT standard setting is currently the focus of attention in an attempt to improve its consistency. In general terms, each AKT paper is standard set by a group of 20 experienced GP judges. They review each AKT item, determining a required level of performance and taking into account the characteristics of each item. Judgments are combined over items and judges to create a pass mark.

Conclusion

The AKT was introduced as part of the RACGP's commitment to maintaining relevance and quality of knowledge testing in the new RACGP examination. The available evidence supports the validity of the AKT segment. Analysis of its performance over the past 7 years has demonstrated that its reliability remains at a level that satisfies international expectations for a high stakes assessment.

Maintaining content validity, blueprinting and standard setting require ongoing attention to ensure ongoing performance of the AKT segment at international standards. Newer issues involve further development of the standard setting processes and the testing of their validity and reliability. Testing of concurrent or predictive validity would enhance knowledge about the value of the AKT in high stakes general practice testing.

Conflict of interest: none declared.

References
1. Farmer E, Hinchy J. Assessing general practice clinical decision making skills. Aust Fam Physician 2005;34:1059–61.
2. Britt H, Miller G, Knox S, et al. General practice activity in Australia 2004–05. Canberra: Australian Institute of Health and Welfare (General Practice Series No 18); 2005 December 2005.
3. The RACGP. Training Program Curriculum. South Melbourne: The RACGP, 1999.
4. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington: American Psychological Association, 1999.
5. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrica 1951;6:297–334.
6. Angoff W. Scales, norms, and equivalent scores. Educational measurement. Washington, DC: American Council on Education, 1971;508–600.

AFP CORRESPONDENCE afp@racgp.org.au