Determining the Quality of Assessment Items in Collaborations: Aspects to Discuss to Reach Agreement

Developed by the Australian Medical Assessment Collaboration

A project funded by the Office for Learning and Teaching, Australian Government Department of Education

The Australian Medical Assessment Collaboration (AMAC) involves sixteen medical schools in Australia and New Zealand and the Australian Council for Educational Research. This collaboration has been funded by the Office for Learning and Teaching, part of the Australian Government's Department of Education.

The writing and compilation of this document was led by Lambert Schuwirth (Flinders University) and Jacob Pearce (ACER). Contributions to the document were made by members of AMAC across the medical schools involved. Further detail about AMAC can be found at www.acer.edu.au/amac.

AMAC partners are:

Macquarie University
Monash University
Australian Council for Educational Research
Flinders University
The University of Queensland
The University of Notre Dame Australia, Sydney
The University of Notre Dame Australia, Fremantle
The University of Wollongong
The University of New England/
University of Newcastle (Joint Medical Program)
The University of New South Wales
Griffith University
Deakin University
The Australian National University
Bond University
The University of Sydney
The University of Adelaide
The University of Otago

Support for the production of this report has been provided by the Australian Government Office for Learning and Teaching. The views expressed in this report do not necessarily reflect the views of the Australian Government Office for Learning and Teaching.

ISBN: 9781742862545

CONTENTS

1 BACKGROUND

The Australian Medical Assessment Collaboration (AMAC) project, funded by the Office of Learning and Teaching, seeks to provide an infrastructure and a road map to support collaboration between Australian medical schools in matters of assessment. This may not seem very new perhaps, because there are already several collaborations taking place in Australia, and, typically, they relate to joint item banks, (such as the IDEAL consortium), or joint test administration, (such as the International Foundation of Medicine tests). The AMAC project seeks to build on these existing collaborations in two ways: first, by tying these initiatives together and thus bundling the combined expertise and experiences in road maps, draft agreements and suggestions for governance structures; and, second, by combining joint examination item production and test administration into one. This should enable continuous meaningful quality comparisons between medical schools, with a view on continuous quality improvement.

One contentious issue in similar collaborations concerns differences in perceptions of the quality of test material. Often there are diverse views on what makes a test item high quality or not. This disagreement in views is a serious breakdown risk for collaborations when it cannot be reconciled (Schuwirth, Bosman, Henning, Rinkel & Wenink, 2010).

Unfortunately, the determination of 'quality' is an inexact science, and the medical education literature does not provide clear-cut answers to questions concerning quality. The role of this document is therefore to provide a framework for quality to help participants make perceptions more explicit and by this, support assessment collaborations.

2 WHAT IS QUALITY?

Trying to define the concept of 'quality' is not easy; such concepts (like 'health') are difficult to pin down. Yet it is extremely important to have a shared view on what constitutes quality of a test item in the case of collaborative item production and examination administration.

For test items we have chosen to use the extent to which an item is an optimal indicator for presence or absence of the requisite ability or knowledge. In other words, the item must be a sort of little diagnostic test for 'knowledge or competence'. As such, a high-quality item should have minimal false-positive and false-negative response. The former means that candidates can answer the item correctly without having the necessary knowledge or competence and the latter means that they answer the item incorrectly despite having sufficient relevant knowledge or competence.

A high-quality item is more than an item that just does not have any violations against agreed-upon item-construction rules; the item must also be creative, relevant for the discipline and appropriately difficult. It is clear that these are judgements and therefore require communication and agreement between partners.

In this part of the document we will discuss the following elements of quality:

2.1 indicators for knowledge/ability

2.2 creativity

2.3 relevance

2.4 format versus content

2.5 difficulty.

2.1 Indicators for knowledge/ability

Assessment can have different purposes, such as to determine whether candidates possesses sufficient knowledge or competence, to give feedback to students, to inform the school or faculty about the quality of the graduates, to ensure the quality of the graduates more broadly to meet the expectations of of governments and wider society, and so on. All purposes however, are based on the assumption that the test is valid and therefore each item is an optimal indicator for presence or absence of knowledge. Any situation in which a student answers a question correctly without having the knowledge (false-positive response) or answers a question incorrectly despite having sufficient knowledge (false-negative response) invalidates the assessment. One important aspect to define quality of a question therefore is the improbability of such false-positive and false-negative response. The literature provides ample guidelines to for item review to minimise false responses (Case & Swanson, 1996; Downing & Haladyna, 1997). Some of the most important guidelines (with examples) are discussed further in this paper.

2.1.1 Parts of a multiple-choice question

Ideally, a multiple-choice question consists of a stem or vignette in which the context of the question is described. This is the context in which the actual question is based. This actual question is often called the lead-in. The options consist of the correct option (the answer or key) and the incorrect ones (the distractors).

Example item:

Stem: *Mr Durmond is 35 years old. He has a bacterial bronchopneumonia. He has not been admitted to a hospital recently. Also, there are no factors that would compromise his immune system.*

Lead-in: *The most probable bacterial cause is:*

Options: *A Haemophilus influenza*
B Klebsiella pneumonia
C Pneumocystis carinii
D Staphylococcus aureus
E Streptococcus pneumoniae

Key: *E*

Though often all multiple-choice questions in a test have the same number of options (mostly four or five) this is not really necessary. Our own study demonstrated that there is no psychometric reason to stick to a certain number of options (Schuwirth, 1998). There is a plausible argument, on the other hand, to vary the number of options with the number of realistic options the author is able to produce and not to include nonsense options (so-called fillers). If a distractor option is simply filling space (i.e., no students are selecting it), then it should be removed.

When writing a multiple-choice question it is good to approach the question as a short-answer open-ended question first. This forces you to think very carefully about what you want to ask and focus the question on one aspect only.

2.1.2 Tips for constructing multiple-choice questions to test knowledge/ability

We will now describe some tips for the production of multiple-choice questions. We will briefly explain why each tip is necessary and give an example. In many of these tips the example of the question is flawed and meant to illustrate the specific item-construction error.

(i) Ensure that all options address the same aspect

It is important to avoid asking students to compare apples to pears. When having to write many items, authors sometimes unintentionally add distractors that could be correct from a different viewpoint.

The example below illustrates an item where this might be the case:

Sydney is:

A a large city.
B situated at the Pacific Ocean.
C the capital of Australia.

Although option C is obviously wrong (but a common misconception) one could debate whether A or B is true, or whether Sydney is more located at the Pacific Ocean than it is a large city.

The best way to prevent this is to use the so-called cover-up test. If one covers all the options the question should be phrased such that it can still be theoretically answered.

If you were to cover up the options the question would read: 'Sydney is:', which is an unanswerable question because you don't have a clue as to what the item writer would want you to know. This is not a trivial construction rule (that is why this is the first one we describe); actually research into strategies students use when answering multiple-choice questions shows that a fair number of students read the question, try to come up with the answer and only then look at the options. This is a kind of forward reasoning that one might want to stimulate in students, but this cannot take place if the questions are phrased in a way making forward reasoning impossible.

Another tip to keep in mind is always to try to formulate the lead-in as a complete question (so ending with a question mark), because this requires a more specific formulation of the actual question the students need to answer. Of course, it is still important to avoid non-informative lead-ins, such as:

Which of the following is correct …?
Which is true for …?
Which is NOT true for …?

(ii) Preferably include options of equal length

Ideally every item is a perfect predictor of the possession of knowledge or understanding; so those students who know should be able to give the correct answer and those who don't should not be able to give the correct answer. Let's call the situation in which a student without sufficient knowledge or understanding still produces the correct answer 'false-positive response' and the reverse (a student with sufficient knowledge or understanding produces an incorrect answer) 'false-negative response'. It may be clear that both error sources decrease the validity – the extent to which the test actually assesses what it purports to assess – of the test, especially if you would want to compare the test with a diagnostic for 'presence of competence'. Apart from false-positive responses due to random guessing there is also the factor of test-taking strategies. Students will know a number of tricks to increase the probability of a correct answer even if they don't know it. One of the simplest of these tricks is to select the longest option. The longest option is more likely to be the correct one, simply because you usually need more words to make an option defensibly correct than to make it incorrect.

What is the best treatment for pneumonia?

A antibiotics
B Aciclovir
C antimycotics
D This must be determined based on the specific cause of the disease.

Of course, this is a bit of an absurd example, but it is an item construction error that is frequently made and it leads to false-positive response.

(iii) Ensure all options are equally subtle

Often there is a difference in subtlety of options. This is logical as well; real life is often much more nuanced that what can be written down on paper. This is why it is logical that the most subtle option is more likely to be the correct one. The incorrect options don't need this level of subtlety and can be easily over-simplified.

What is the most indicated treatment in chronic benign low back pain?

A prescribe Tramadol HCL
B physiotherapy
C perform a surgical disc prosthesis
D multidisciplinary management

Option D may not be the longest one but it is certainly the most subtle one and it is pretty clear to the test-wise students that this is what the item writer intended as the correct option. In this case it would also lead to false-positive response.

(iv) Ensure that all options are in the same 'direction' (all positive or all negative)

It is very confusing if some options are worded affirmatively and others negatively.

A patient presents with complaints of headaches. The headaches have been present for more than two weeks. They come in attacks and typically start late in the afternoon and last for roughly one to three hours. The patient describes them as a sharp continuous pain on the right side of the head, above the eye and in the temporal region. Paracetamol brings some relief. The patient has had similar headaches last year but they were less severe and lasted for only two weeks.

Which is the correct deliberation concerning treatment?

A Tramadol is most likely not to have an effect on the pain.
B Pure oxygen is known to have an effect on the pain.
C Relaxation therapy is effective in more than 50 per cent of patients.

Such combinations of positively and negatively worded options are likely to produce reading errors and lead to false-negative responses.

(v) Test only one aspect per option

Two-in-one options not only make the item less clear; they also make the item vulnerable to the so-called conversion strategy. An example of two-in-one options is given below.

Not only repetitive nerve stimulation (RNS) and single-fibre electromyography (SFEMG) but also high titres of antibodies against acetylcholine receptors (AbAchR) can be used to diagnose myasthenia gravis.

Which of the following options is correct concerning the sensitivity of these tests?

A AbAchR higher than SFEMG and higher than RNS
B AbAchR lower than SFEMG and lower than RNS
C AbAchR lower than SFEMG and higher than RNS
D AbAchR equal to SFEMG and higher than RNS

In the first comparison the 'lower' is used twice and in the second 'higher' is used twice. Option C contains the combination of both most frequent comparisons and is therefore more likely to be the correct one. This is again logical, because typically the author starts with the correct option and then varies on it. Another, less conspicuous example is the following:

What is the normal value for the aspartate aminotransferase in a healthy adult?

A < 4.8 U/l
B < 48 U/l
C < 60 U/l
D < 480 U/l

Here again the conversion strategy works: the options with the 4 and 8 are all variations on 48 so they form one cluster and the 48 and 60 are variations of 48 in the same magnitude, so option B (being a member of both groups) is the most likely correct answer.

Of course, one cannot always avoid having two-in-one options but if it is necessary make sure that all combinations are covered. In the first example, the problem would be solved by changing the fourth option into:

AbAchR higher to SFEMG and lower than RNS

A specific case of the two-in-one problem is the use of qualifying statements in the options. For example:

> *A group of researchers want to compare the effectiveness of a new e-learning module on pharmacodynamics to the traditional approach of lectures and practicals. They employ a typical causal comparative research design with a pre-test to establish baseline knowledge and differences between the intervention and control group and a post-test to determine the differential effects of the interventions. The number of participants in each intervention arm is 50.*
>
> *Which of the following is the most appropriate statistical analysis in this case?*
>
> *A separate Mann-Whitney tests, because the scores are not normally distributed*
> *B a two-way ANOVA because the number of participants in each group high enough to use parametric statistics*
> *C separate chi-squares because it is necessary to establish the association between the intervention and the outcome*
> *D a Kruskall-Wallis test because with assessment results normality of the variable can be assumed*

It is often thought that adding explanations will ensure that students have to think harder and have to understand better the reasons why an option is correct or incorrect, but this is not the case. Students will often quickly rule out options simply because either the choice or the explanation is incorrect. So although there is more information contained in each option it actually makes the item easier.

(vi) Use clear, unambiguous wordings; in particular, be clear in wording the stem

The items on the test are aimed at testing whether a student possesses sufficient relevant knowledge or understanding of the subject matter. Other factors can be confounders in the 'measurement' of this knowledge and/or understanding. A text that is difficult to read can therefore be a confounding factor. This is not to say that it is unreasonable to expect an academic to be able to read complicated texts, but it may be better to use different instruments for this in the assessment program. So, try to be clear in the stem, place the sentences in a logical order and avoid unnecessarily complicated sentences. The item below is an exaggerated example of an attempt to make the question more difficult by using a complicated construction. However, once you have managed to decipher the sentence the question is really very easy.

> *It cannot be excluded that certain findings/symptoms are not present in a patient with purulent meningitis if this patient is not a member of the normal adult population.*
>
> *Such a finding or symptom is:*
>
> *A leucocytes in the spinal tap fluid*
> *B nuchal rigidity*
> *C inflammation of the meninges*

(vii) Ensure that options encompass the whole gamut, where possible

It is a pity if not all the possible realistic options are incorporated in the item. A somewhat exaggerated example of this is:

> *The sensitivity of a standard chest X-ray for the detection of lung cancer is:*
>
> *A greater than 95 per cent*
> *B smaller than 90 per cent*

The option of between 90 and 95 per cent is not included, so students who would have considered this option knows that their initial thoughts were wrong. This is a sort of cueing one would like to avoid. The reverse is also problematic; a subset of the options already covering the whole gamut rendering the rest of the options superfluous.

Administration of propranolol leads in the majority of cases to:

A a decrease in mean blood pressure.
B an increase in mean blood pressure.
C no measurable changes in blood pressure.
D a delayed response in blood pressure change.
E few side effects.

It is clear that options D and E do not have to be considered, because A, B, and C have covered all realistic possibilities: the drug either increases or decreases the mean blood pressure or has no influence at all. There are no other options. So a student who has no idea about propranolol will still be able to increase the probability of a successful guess from 0.2 to 0.33.

(viii) Ensure that options are mutually exclusive

If there is any overlap between the options, students are given a powerful cue to strategically sort out what the correct answer would be.

The overall five-year survival rate of patients with a metastasised oat cell lung carcinoma lies:

A between 0 and 10 per cent.
B between 10 and 30 per cent.
C between 20 and 40 per cent.
D between 30 and 50 per cent.
E between 40 and 60 per cent.

Any percentage between 20 and 50 per cent would make two options correct; 20 to 30 per cent would make options B and C correct, 30 to 40 per cent would make C and D correct, and so on. So, only percentages between 0 and 20 and between 50 and 60 will have to be considered. Students who had originally thought of another percentage will now know that their thoughts were incorrect and will have to guess between only three options (A, B and E).

(ix) Ensure that one option is defensibly correct and the others are defensibly incorrect

This may sound like an open door but often it is not. Often the formulation of the question is such that the key is not fully correct or that other options can be correct as well. If your examination rules and scoring system allow for more than one answer being correct (or at least give partial credits) this may not be a big problem, but it is always better to avoid such situations.

A 40-year-old woman has been suffering from stomach aches, especially after eating. She is diagnosed with a duodenal ulcer. Which of the following drugs would be most indicated if it is decided to start with medication?

A an antacid
B a prokinetic drug
C a histamin-2-blocker
D a proton pump inhibitor

This item had to be withdrawn from the test because both the options C and D were considered defensible. In this case it was a content-related item-construction problem, but there are also formulation-based problems.

Ovulation occurs after the luteinising hormone (LH) peak. A certain period of time passes between the LH peak and the moment of ovulation. This period is:

A 18 hours.
B 36 hours.
C 54 hours.
D 72 hours.

The stem does not describe exactly enough which exact measurement moment of the LH peak was intended (some define it as the maximum LH level, others as the whole period of spiking) leading to more than one option being defensibly correct.

(x) Do not use collective options, such as ‘all of the above’ or ‘none of the above’

Perhaps not every aspect of item construction rules has been thoroughly studied, but research shows that the use of collective options has a negative impact on the purity of the measurement of knowledge/understanding of the test.

This is easiest to understand in the case of an ‘all-of-the-above’ option. Suppose there are five options, four with a unique content and one collective. In this case, every student who knows at least two of the other options to be true can automatically conclude that option E must be the correct one. This is not to say that you could not ask an item to which more than one option would be correct (the so-called multiple true-false items), but in a standard single-best option multiple choice it is preferable to avoid collective options.

In an item with a ‘none-of-the-above’ option, it is easy for a candidate to answer the item correctly based on incorrect information. For example:

In which part of Australia is Uluru located?

A New South Wales
B Western Australia
C South Australia
D None of the above

The candidates who think that Uluru lies in the Australian Capital Territory, Victoria or Queensland will also respond with option D and produce false-positive response.

(xi) Be aware of grammatical misalignment between lead-in and options

Students will use all information at their disposal to produce the correct answer. Grammatical misalignments are a simple cue for the strategic student.

Ipratropium is an anti-asthma drug. It is an:

A anti-cholinergic.
B beta-2-sympathomimetic.
C corticosteroid.
D xanthine derivative.

Simple item for the clever reader; only option A starts with a vowel and this aligns with the article ‘an’ from the lead-in. All the others are consonants and would have required the article ‘a’. The simple solution here is to put the correct articles in the options.

(xii) Do not provide logical cues

Apart from grammatical cues, there could also be logical cues.

A patient consults you because of a radiating pain from his lower back to his left gluteal region and his left leg. The pain increases when he coughs or sneezes. There are no complaints of loss of sensitivity or motor functions. All reflexes of the lower extremities are intact and symmetrical. The most indicated treatment is:

A non-steroidal anti-inflammatory drugs for pain relief.
B physiotherapy.
C surgery.
D electromyography.
E magnetic resonance imaging (MRI).

This is a bit of a caricature but certainly an easy item for most students: D and E are not realistic options as they are not treatment options. So, the students only have to consider A, B and C.

(xiii) Do not use too absolute or too open options

Options with ‘can’ and ‘is possible’ are so open that it is very hard to defend that they are incorrect (see tip ix). On the contrary, option with ‘always’, ‘never’, ‘excluded’, etc. are so absolute that they are most likely not correct.

Patients with diabetes mellitus:

A never have heart disease.
B are always adults.
C can have complaints of poorly healing wounds.
D have to be treated with insulin.

Apart from the problem of this item not passing the ‘cover-up test’ (tip i), it is also formulated in such a way that even for a student who has no knowledge whatsoever about diabetes mellitus it will be easy to produce the correct answer. Options A and B contain ‘never’ and ‘always’ and D suggests that this is ‘always’ the necessary treatment. In addition, C has the open ‘can’ in it and will therefore be the correct answer. This is not to say that ‘can’ and ‘never’ cannot be used, but in such cases it is best to introduce this in the stem.

(xiv) Avoid semi-quantitative terminology

Although our text books are replete with vague semi-quantitative terms (‘often’, ‘seldom’, ‘frequently’, ‘usually’, and so on), they are best avoided when writing items. It is not clear how often ‘often’ is, how seldom ‘seldom’ is, etc. Research has shown the large variation in how people use these terms if you ask them to express the meaning in a percentage (Hakel, 1968).

Patients with diabetes mellitus:

A seldom have heart disease.
B are often adults.
C sometimes have complaints of poorly healing wounds.
D frequently have to be treated with insulin.

It is now impossible to answer the question; all options are defensible depending on how you define the semi-quantitative terms. Yet these could be phrases directly copied from a text book. Often such items are reformulated into percentage items. Though understandable as a remedy for the semi-quantitative terms items asking for percentages are often perceived as trivial both by staff and students. There is not standard solution for this and often close collaboration between author and item reviewer needs to take place to find a good alternative.

(xv) Check for ambiguities in formulation

It is always sensible after having produced questions to put them aside for a couple of days and then review them with a fresh pair of eyes. Often ambiguities become clear that were not apparent before. Option D from the previous example:

D frequently have to be treated with insulin

could mean two things. It could suggest that a patient with diabetes mellitus would normally not require continuous treatment but only intermittent treatment. Alternatively, it could mean that patients require continuous therapy with insulin. It is obviously the latter, but the sentence could be misconstrued by students. Incorrect interpretations lead to incorrect answers and thus to false-negative response.

(xvi) Place the options in a logical or alphabetical order

We don't know why, but option C is most often the correct answer to multiple-choice questions, especially to questions with four options. We assume this has to do with the way item writers work. Option A is unattractive for an item writer because of the feeling of 'giving the answer away'. So a distractor has to be sought for option A and a second one for option B. Often finding the third distractor is more difficult so the item writer fills in the correct answer as option C and then spends time finding the third distractor. Whatever the underlying explanation might be, students know that C is more often the correct option and will choose this one if they don't know the answer. The remedy is simple; either put the options in a logical order (increasing severity of disease, invasiveness of procedures, and so on) or just to put them in alphabetical order. Another way of remediating this problem is look at the distribution among the options of the correct answer. Often it is found that the first and final options are less likely to be the correct answer key, so this may prompt you to reorder the options in some of the items to produce a more equal distribution (but never a completely equal distribution, to avoid predictability).

(xvii) Avoid complicated formulations

For a while it was assumed that questions with complicated constructions would test understanding or insight better than straightforward multiple-choice questions. This turned out to be untrue. A more essential difference exists between items with a vignette, case or problem description combined with a question asking for decisions on the one hand and rote factual knowledge questions on the other. Both types can be highly relevant, but the thinking steps in the former items types are more at the level of weighing probabilities while in the latter item types are more at the level of yes/no deliberation (Schuwirth, Verheggen, van der Vleuten, Boshuizen & Dinant, 2001). Complicated formulations only lead to unnecessary complexity and not to a better measurement of knowledge or understanding. Another aspect is that they are more likely to provide cues as to the correct answer.

The most important symptoms associated with cardiovascular disorders:

1 are chest pain, dyspnoea and palpitations.
2 appear or increase at exertion.
3 are fatigue, dizziness and syncope.
4 appear in rest.

A (1), (2) and (3) are correct
B (1) and (3) are correct
C (2) and (4) are correct
D only (4) is correct
E all are correct

Now it is easy for students to start using their common sense. Either (1) or (3) is true; it is highly unlikely that both are true at the same time. The same applies to (2) and (4). All options that include (1) and (3) or (2) and (4) can be excluded. One could also argue that if A were correct B would automatically be correct as well and therefore A cannot be the answer key. So theoretically only option D would remain (though we find it hard to believe that this would be the correct one). Regardless of whether this line of reasoning is correct, it will lead to either false-positive or false-negative response. Or, to put it in other words, the item induces all kinds of reasoning that has nothing to do with the knowledge or understanding the question seeks to assess, and therefore most likely introduces so-called construct-irrelevant variance or noise. This decreases the validity of the item. To be honest, we don't know what the item writer's intentions were here and which cardiovascular disorder he or she had in mind (varicose veins?). So the question in its current form does not convey the item writer's intention very well and would require a constructive conversation between item writer and item reviewer.

2.2 Creativity

Because the literature seems to focus almost entirely on the restrictions surrounding item writing – all the dos and don'ts of the previous section – it is easy to neglect that item writing is also a creative effort. Indeed, it is possible to follow all of the rules above and still produce a poor item. Opening a book and taking a random fact to be asked rarely leads to a good question, even if all the item construction rules have been heeded. Often we seek to test more than mere rote factual knowledge. In this section we want to provide some tips in this area.

2.2.1 Contextualise items

There is shared opinion that having good and well-organised knowledge is a necessary requirement for (medical) problem solving, but this does not mean that it is also sufficient for successful problem solving. The assessment of higher-order skills, application of knowledge or even clinical problem-solving ability has been high on the agenda for a long time. It is fair to say that asking questions in a relevant context, for example by presenting students with a problem and then asking them for a solution, generally leads to questions which are perceived to be more interesting. In addition, such questions elicit thinking steps which are substantially different from questions without a vignette asking for factual knowledge (Schuwirth, et al., 2001). Typical examples of these can be found in the form of Extended-Matching Items (Case & Swanson, 1993) or in key-feature approach items (Bordage, 1987).

For case-based items, the following tips and rules apply (Schuwirth, et al., 1999).

(i) Use, whenever possible, cases that are derived from real life

These can be basic sciences problems, clinical problems, public health problems, and so on. Real-life cases ensure a better authenticity and better relevance, and they provide a relatively easy source of items.

(ii) Ensure that the description of the information is as clear as possible

Avoid vague terminology and shorthand. Remember that reading is a skill that most of us manage quite well and that reading some extra words does not take long. Having to think about what the item writer intended takes much more time. Also, bear in mind that each discipline has its own jargon and that this may not be evenly well known across disciplines or even in the same discipline in another centre.

(iii) Provide sufficient realistic 'clinical' information

When writing the case think of all the information the candidate needs to answer the question. Is all the information present that is needed to make one option defensibly the correct one and all the others defensibly incorrect? After writing the question and the options, go back to the case and add or revise if needed. Of course 'clinical' here also stands for basic science or public health information, as relevant.

(iv) Provide sufficient realistic contextual information

Do not provide 'clinical' information only but also contextual information, for example: where do you see the patient, what is your role, what is the setting (remote rural, urban primary care, hospital)?

(v) Provide sufficient negative information

Sometimes it is also wise to describe findings that are NOT abnormal, for example, 'no rebound tenderness'. Sweeping statements such as 'otherwise normal' sometimes do not suffice. Especially in physical examination everybody has their own routine, so with sweeping statements the candidate might not know what procedures were performed and what not.

(vi) Provide information that is not pre-interpreted ('raw')

In patient charts it may be good to describe information in jargon and interpretation but for a case description for a test it might be better if the students were asked to do their own interpretation. When lab values are used, though, it might be good to include normal values (as they vary somewhat from centre to centre) and leave it to the candidate to interpret whether the lab values are markedly abnormal or still within reason.

(vii) Link the problems directly to the case

It is not useful to present a case and then ask a question that could also be answered without having read the case. Students will lose valuable time over it and it does not lead to different results or scores. Time is precious in assessment and it should not be wasted.

(viii) Focus on essential problems only

This is an essential element of case-based items. The question must focus on essential decisions (key features), and the diagnosis or even the treatment may not be essential. In a rural general-practice setting, for example, the decision whether to evacuate the patient may be more important than the exact diagnosis, or whether or not to perform a risky diagnostic procedure. Generally a key feature is asked when the problem is based on combining the different information parts of the case and when an incorrect decision automatically leads to an incorrect management of the case. It is good to consult colleagues and check whether they agree with your selection.

(ix) Phrase the questions as clearly as possible

This pertains to all the suggestions of the previous section.

(xi) Ensure that the answer is defensibly correct and the distractors defensibly false

This can also be ensured by the wording of the question. There is some room for creativity here. You might ask, for example: 'If you could only ask five questions during history taking, which of the following would then be most relevant?' This question is actually asking for the most sensitive or specific questions. It is also important to make sure that while defensibly false, the distractors are plausible options. To what extent, however, will depend on the purposes of the test.

2.2.2 Transformation of information (Ebel, 1972)

Rarely is it a good idea to randomly pick a piece of information from the literature and turn it into a question. Often necessary contextual information is lost or the topic is just not relevant or suitable for the specific test. It may be helpful to use the literature as the basis for an item, but often you cannot simply ask questions verbatim from the literature. Some useful suggestions are to:

- Restate the concept in different words or paraphrase what was said in the literature.
- Restate parts of what was described.
- Ask for the opposite.
- Ask for the exception.
- Ask for a relationship between the concept from the literature and other concepts.
- Ask for implications of the concept.
- Ask for a problem situation in which the concept needs to be applied.

2.2.3 Six Steps Approach (Miller, 1976)

Writing an item is not always easy and for most of us something we do not do on a day-to-day basis. More often we only do it once a year. Therefore it is important to keep in mind that it is a complex design task and breaking it down in smaller steps often is more efficient. Suggested steps are outlined on the following page.

1. Select the information to be tested.
2. Condense the information.
3. Select the task on how the information is to be used.
4. Write the item stem.
5. Write the answer.

Or alternatively:

1. Define the area.
2. Define the subject.
3. Define the topic.
4. Define the problem.
5. Write the question in the easiest format.
6. Write the question in the desired format.

2.2.4 Notebook method

The most difficult aspect of writing items, especially if larger numbers are needed, is to come up with relevant topics for the items. This is a pity because during our normal teaching or patient care we often encounter situations that would be perfectly suited to turn into a relevant question. When carrying a (paper) notebook, a smart phone or tablet it is easy to quickly voice record or type in these topics for later use. Ideally however, you would write the question as soon as possible and store it for future use. Typical triggering events are:

- misconceptions of students
- main points of lectures
- points from practice
- own inspirations
- results of our own additional study
- patient encounters
- discussions with family and friends (for example, as a trigger for items concerning professional behaviour, ethics, health economics, and so on).

2.2.5 Communities of practice approach

Working together is often the fastest method for producing creative and relevant items. Such meetings typically work best if you already have your topics. Typically group members will be critical, asking you to explain the relevance of the items you propose and help you in making them more creative, relevant and challenging. In such group meetings it is important to deploy the following activities:

- brainstorm
- critique others' questions
- question the relevance of items
- use literature
- work together
- make notes of various solutions to item-writing problems and develop standard strategies for recurring problems.

2.2.6 Item modelling

There may be situations in which item writers have a good conception of a question but are stuck on finding suitable distractors. What may help is to think of other steps in the clinical journey from the stem to identify new types of questions – such as involving investigations, levels of acuity, epdiemiology – that were not considered by the item writer originally but which may lead to better distractors. Sometimes they may lead the item writer to discard the original question, or to use the stem but as a different question with better distractors (for example, to change from a question on the diagnosis to one on diagnostic management).

2.3 Relevance

Relevance is difficult to define. Often it is described as a global judgement by a panel of experts, as, for example, in the course of the Ebel standard-setting process (cf. Livingston & Zieky, 1982). But this is of limited usefulness in writing items. Firstly, because it involves a decision after the item has been produced, and therefore unhelpful in writing items. Secondly, because it is unreliable, and therefore unreasonably large panels would be needed to ensure reproducible judgements (the decision of relevance has to be made for each item individually and not on the total of the test). Because relevance is a subjective process based on human judgement it is more helpful to provide arguments according to reasoning lines to discuss and decide on the relevance of items.

An example of such an instrument is outlined in the table below.

	NOT RELEVANT	SOMEWHAT RELEVANT	VERY RELEVANT
Medical knowledge	Knowledge is an element that is not necessarily specific to a doctor; the baker on the corner knows the answer.	Knowledge is specific to medicine but also known to the interested layperson.	Knowledge is specifically for medicine and requires a proper study and understanding of the subject.
Ready knowledge	The knowledge is not easily recalled but is easy to find. Even specialists in practice cannot remember it.	The knowledge is easy to find, but should be typically recalled when confronted with it in practice.	Any medical doctor has this knowledge at the ready at any time of day. It is a prerequisite for functioning in a practical situation.
Incidence in practice	There is no medical situation (not necessarily clinical) in which this knowledge is important.	While there are medical situations in which this knowledge is important, these situations are not frequent.	This knowledge is important for many practical situations.
Prevalence or high-risk	The knowledge is usually only found in highly specialised centres, is low risk or is rarely found.	The knowledge is found in high-prevalence or high-risk situations in practice, but is not essential for successfully handling the situation.	The knowledge is found in high-prevalence or high-risk situations in practice, and is essential for successfully handling the situation.
Knowledge foundations in the medical curriculum	The knowledge is a fact or an isolated event and is not required for building other concepts in the curriculum.	The knowledge is needed to further understand concepts but the specific knowledge may itself be forgotten (e.g., the Bohr/Haldane effect for understanding why haemoglobin releases oxygen into the tissues in the lung).	The knowledge forms the basis for one or more other concepts in the curriculum and it should remain known as explicit knowledge (e.g., the Frank-Starling mechanism as a basis for congestive heart failure).

2.4 Format versus content

There has always been a debate about question format and whether certain formats are suitable to test difficult or higher-order cognitive skills. When we summarise the literature on this, three recurrent issues are worth mentioning: the cueing effect; case-based questions; and question format.

2.4.1 Multiple-choice questions are subject to the cueing effect

The cueing effect was first documented in 1954 (Hurlburt, 1954) and basically states that in multiple-choice questions, recognition of the correct option suffices to give a correct answer, whereas in open-ended questions, spontaneous generation of the correct answer is needed. Often, this recognition is not seen as a higher-order cognitive skill and therefore multiple-choice questions are seen as unfit to test these skills. The literature, however, converges on the notion that even if the cueing effect occurs, it does not interfere with the type of skill the item tests (Norman, Swanson & Case, 1996; Norman, et al., 1987; Schuwirth, van der Vleuten & Donkers, 1996; Ward, 1982). The format of the item determines only to a very limited extent what the item tests, and the content is much more important. Please compare the following items:

> *Name the premiers of all Australian states and territories in 2013.*

and

> *Three students have dinner in a restaurant. Right before dessert arrives they all fall asleep. The dessert is brought: stuffed dates.*
>
> *Student #1 wakes up, eats what she thinks is her share and falls asleep again. Then, student #2 wakes up, eats what he thinks to be his share and falls asleep again.*
>
> *The same happens to student #3.*
>
> *Finally, all three wake up and they start a discussion about who ate how many dates. They eventually decide to distribute the remaining eight dates between students 2 and 3.*
>
> *How many dates were there originally?*
>
> *A 21*
> *B 24*
> *C 27*
> *D 30*

Regardless of which item is more difficult it is clear that for someone who has never solved the problem in the second example, deduction, reasoning and even some creativity in problem solving (you have to think of putting yourself in the position of the first two students to deduce their reasoning) are needed to produce the correct answer, whereas in the first example simple memorisation suffices. This has nothing to do with the question format and everything with the question content. This was experimentally convincingly demonstrated by William Ward in 1982 (Ward, 1982) but repeated many times in medical education afterwards. (Norman, et al., 1996; Norman, et al., 1987; Schuwirth, et al., 1996).

2.4.2 Case-based questions are more likely to test higher-order cognitive skills than simple questions

The difference is marked; case-based questions asking for decisions typically induce thinking steps which are more based on using personal experience and weighing possible options, whereas isolated factual knowledge questions are more a matter of knowing or not knowing (Schuwirth, et al., 2001). This is not an issue about which type is more difficult but about the purpose of the test. It is even not about relevance because both simple factual knowledge and application or problem solving can both be relevant. Research demonstrates that closed or open questions de facto test the same skill if the content is the same, and that the main difference lies in whether they are case-based or isolated factual knowledge questions.

2.4.3 There are no superior question formats

Contrary to a widely held belief, there are no superior formats. All formats have their strengths and weaknesses (van der Vleuten, 1996). For open ended questions, the spontaneity and creativity that can be required will be an advantage, but their resource intensiveness – especially for marking – and logistical complexity may be a downside. There may even be the argument of lower reliabilities (though this is disputed as well). For multiple-choice questions, the opposite may be the case. A good assessment program combines strengths and weaknesses of various test and assessment formats (van der Vleuten & Schuwirth, 2005). It may be that for a certain project a certain format has to be chosen (for example, multiple-choice for a national test or a test produced in collaboration between institutes), and it is always good to bear in mind the downsides of that choice but not helpful to completely discard the approach because of them.

2.5 Level of difficulty

Difficulty is still a largely ill-understood concept and we do not claim to have the definitive answer to it in this document. A detailed discussion would also be beyond the scope of this report. Still though, we want to discuss the aspects surrounding difficulty that are relevant for judging item quality.

2.5.1 Psychometrics and difficulty

Often it is assumed that the so-called p-value, the proportion of candidates answering the item correctly, is the equivalent of difficulty. But it is fair to say that this is not correct. Of course the probability that many students will answer a question correctly is associated with the difficulty of the item: 'How many arms does a normal human being have?' is intrinsically easier than 'Name the amino acid sequence of insulin', simply because more elements are being asked in the second question compared to the first. But, there are complex abilities with high p-value (being able to walk upright is a rather complex motor skill but it has a high p-value in medical students; most of them can do it). So p-values are not a perfect indicator for difficulty but for the probability that a candidate knows the answer, and therefore a reasonable proxy for difficulty. Regardless of this, there is always an interaction effect between the candidate and the item; what is an easy item for the one candidate is a hard one for the other and vice versa.

Another, more fashionable, way of calculating difficulty is with the Rasch model (cf. Smith, et al, 2004). With these calculations, though, it is important to recognise that the values obtained for an item are not objective, rather, they are relative to the cohort of students and the set of items which they undertook. There are more complex ways of linking items between cohorts and scaling items to produce a more nuanced difficulty metric for a particular item in a comparable context.

2.5.2 Difficulty and purpose of the test

Another important consideration with respect to producing good test items in this context is the purpose of the test. In discussions these purposes are often convoluted.

A first and most often used purpose is of selection. In this, the test is viewed as an instrument that clearly distinguishes between people, for example to determine who is admitted to a program or not. Typically, such tests need to have candidates who fail and candidates who pass. If a test were designed, for example, to determine who is admitted into medical school (with only a limited number of places) and all the candidates would pass the test then it does not serve its purpose very well. Tests such as these often have many difficult items to discriminate between good and very good candidates to a high degree of precision.

Another purpose is to assess (the development of) the extent to which the student is developing or has gained competence. This could be either competence using the total score of the test or to detect strengths and weaknesses in the road to competence. Such a test is designed to test whether all candidates have sufficient competency of the topic to progress to the next phase in their education (or to determine which further action is needed before a student can progress). This requires a different perspective on producing items; they now have to be constructed in such a way that they do test relevant and valid aspects of the competence but the test does not necessarily have to contribute to passing or failing students. The test is less focused on selection but more on being an integral part of the educational process.

These two functions are often convoluted. Many assessment programs, for example, consist of a series of selective tests only. This is often criticised by the saying that ‘no patient has ever been cured by taking their temperature’.

In an assessment program both types of test can – or even should – play a role but writing items for each type of tests is quite different. In a selection-orientated test you would include items that only the best of the best can answer and you would want to exclude items that everybody can answer (they do not contribute to the distinction between passing and failing students). In the more education-oriented test you do want to focus on optimising the test by including relevant items than every competent student should be able to answer. Theoretically, that test could contain items all with a p-value of 1.00, as long as they are valid, relevant items and constitute a valid test.

In general it is therefore important to also consider asking questions that test knowledge *you want your students to possess* and not focus on *knowledge that they most likely will not possess*. Of course it is not useful to include items that even non-medical people could answer (such as ‘How many arms does a normal person have?’) but it is also not sensible to include items that nobody could answer.

Finally, experience shows that it is quite difficult to predict the exact difficulty of an item at the level of the group of candidates. So monitoring afterwards – by psychometric analyses – and feedback to the item writer is important.

3 ORGANISATION OF QUALITY CONTROL

In the early 1980s a short debate took place in the literature about validity (Cronbach, 1983; Ebel, 1983). Where Cronbach (from the famous Alpha) contended that validity is purely a matter of how the test scores 'behave' (for example, do they increase with increasing levels of expertise of the candidates), Ebel stated that validity also has to be built into the test. An item asking whether students know how to treat a pneumococcal pneumonia is not only relevant because it adds to other items to produce a score on 'medical knowledge', but it is also relevant and valid in its own right. The typical quality control and quality assurances practices in producing test items reside in Ebel's (and later Mike Kane's) view on validity (Kane, 2006). Quality therefore has to be built into the test and has to be built into the organisation. In this part we will discuss this from five viewpoints:

3.1 pathways of items in the quality assurance process

3.2 review panels and composition

3.3 feedback to item writers

3.4 item analyses

3.5 organisation of joint or multicentre quality control.

3.1 Pathways of items in the quality assurance process

It is fair to say that items that are not reviewed before they are on put on the test generally suffer from more item-construction flaws than items that have been reviewed. It is a repeated finding that we all, as item authors, have our blind spots or lapses of attention and may produce items that are flawed and therefore contribute to false-positive or false-negative response (cf. part 2.1). Therefore many organisations have review panels that critically review draft items and provide the author with suggestions on how to improve the item.

These panel meetings can be positioned differently in the process of producing a test. The most common setup is one where the draft test items are collected some time before the test and then reviewed in the panel. This has the advantage that the purpose of the panel is clearly visible in the organisation, namely to help produce THIS particular test. Another option is to have a setup in which items are produced on a regular basis (when authors encounter situations or have inspiration) and the review panel meets regularly to review draft items. The test is then produced from the stack of items that have been reviewed and agreed upon. Both set-ups are shown in Figure 1 below.

Though an item bank will be helpful in the first set-up it is almost indispensable in the second setup. (These figures, by the way, show the main function of an item bank, namely to support and manage item quality assurance processes.)

These two schemes are based on a quality control process that uses only one cycle. A second cycle of quality control can be added by having a review process that incorporates information that is collected after the test administration. This information can come from (psychometric) item analyses and even from student comments. These setups are shown in Figure 2.

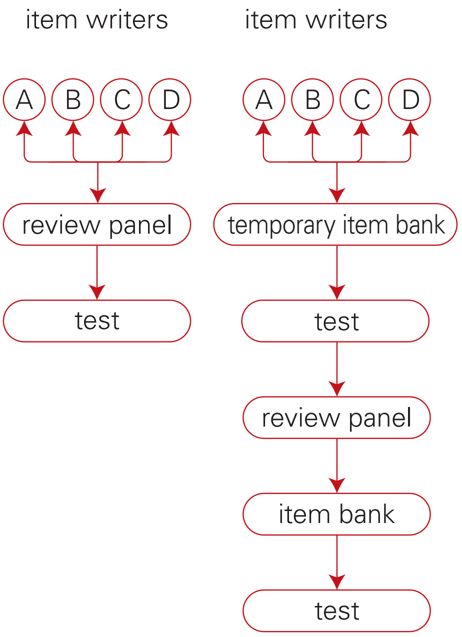

Figure 1 Single-cycle quality-assurance processes

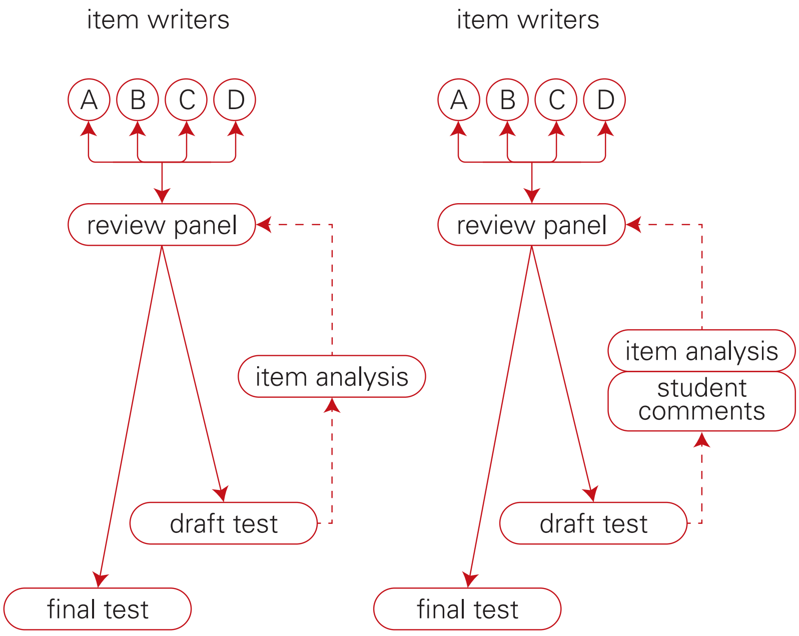

Figure 2 Dual-cycle quality-assurance processes

3.2 Review panels and composition

Review panels are best composed of critical people with sufficient knowledge of the matter to understand the questions and the answers, enabling them to critically question the content, phrasing and relevance of items. Super-specialists may be less well positioned to note content, phrasing or relevance issues with items, simply because they may overestimate the knowledge of the average candidate, the relevance of the item or because they are so well versed in the matter that they overlook obvious problems with the phrasing. Diversity in backgrounds is also helpful; for example, combining basic scientists with clinicians in panels. In review panels it often becomes quickly evident that all members have areas of deep knowledge and areas of relative ignorance. It is that combination that is most sensitive for picking up item-construction issues. Therefore, the best contribution a panel member can make is to ask: ‘I don’t understand why A is the correct answer; could anyone explain this to me?’, or to actively seek to misunderstand the phrasing. It is important for all members to understand that ‘not knowing’ something which is asked in an item is not a demonstration of lack of expertise, but that we all have things we know and things we don’t. A safe atmosphere and a culture in which open discussion about an item can take place are therefore prerequisite for an effective panel.

3.3 Feedback to item writers

When providing item writers with feedback it is important to acknowledge that writing items is always difficult and that making an error is not an indication of lack of expertise of the item writer; it is often just an oversight which happens to everybody.

Four elements should preferably be present in the feedback:

1. What is the incorrect element of the item? What is the content problem, the flaw in formulation or the issue with relevance that has caused the panel to flag the item? Here it is best to use more or less standard feedback as there are often standard item-construction flaws. You could decide to use or adapt the text of this document.
2. Why is this a problem? For example, how would this induce a false-positive response or a false-negative response or how would this produce random results?

3. How could the item best be rephrased or changed to eliminate the flaw or at least mitigate its influence? Often concrete suggestions or examples of a revision are most helpful.
4. Why is the suggestion for revision or the revised version better than the original?

Another opportune moment for feedback to item writers is after the test administration. Item writers are often experts in their field and therefore may find it difficult to gauge the appropriate level of difficulty of an item, especially when they are writing items for collaborative assessment (Muijtjens, Schuwirth, Cohen-Schotanus & van der Vleuten, 2007). Providing feedback on item performance (with an explanation of what it means) supports the item writers in better aligning the difficulty of their items to the level of the students.

Another way of using the feedback is to inform members of standard setting panels. Standard setting is a difficult issue and there are more than 35 different methods in the literature (Cusimano, 1996). Regardless of the method, however, they are all based on judgements of experts of what is reasonable to expect from the individual candidate or group of candidates. Methods such as Angoff and Ebel (cf. Livingston & Zieky, 1982) require panels of experts to judge the difficulty of each item (for the specific group of candidates), Hofstee requires judgements about acceptable pass/fail levels and acceptable pass/fail proportion (Hofstee, 1983) but even completely norm-referenced methods use assumptions and judgements about what can be expected of the candidates (cf. Cohen-Schotanus & van der Vleuten, 2010). Judgements become better if they are better informed and feedback on the performance of the students and of the items (item analyses) is therefore a unique way to ensure that the judgements in the standard-setting process are more accurate.

3.4 Item analyses

During quality control, item analyses can be calculated and used. Item analyses give an impression about how this group of students performed on the test. They do not provide a completely neutral picture of the qualities of the items but always in relation to the group of students who took the test. If for example an item is only answered correctly by 10 per cent of the students (so has a p-value of 0.10), that could mean that the item is difficult in itself, or it could mean that the students were on average not competent enough to master it (despite it being taught), or that the students weren't taught this at all. In the first case it could imply that nothing has to be done; in the second it would mean that the educational process has to be better aligned with the ability of the students; and in the third case it might mean that it is best to withdraw the item from the test. The best action to take based on the results of item analyses is therefore always a matter of judgement and often of further investigation to understand why the item performed poorly. It is therefore rarely a good idea to eliminate an item merely based on its item statistics. This is like eliminating a data point from your research simply because it does not fit your expectation or because you don't like it. There always has to be a good argument to remove an item.

There a number of often-used parameters.

- p-values
  This is simply the proportion of students that answered the question correctly. If all students answered the item correctly the p-value is 1.00 and if nobody answered the question correctly the p-value is 0.00. For a competence-orientated test you may want to accept any p-value as long as the item is relevant for the topic and has been taught in the course. For a selection-orientated test you may want to have p-values that are not too close to either 1.00 or 0.00. As a rule of thumb, often ranges between 0.25 and 0.75 or 0.30 and 0.70 are used.
- a-values
  These are simply the proportion of students that selected this option from the options of a multiple choice. The a-value associated with the correct option is therefore the same as the p-value. Again, their interpretation is based on the purpose of the test. In a selection-orientated test you want distractors (false options) to be attractive for those who don't know because that way the item will contribute well to distinguishing between the passes and fails. For a more competence-orientated test it does not matter if a certain distractor is not chosen, as long as it indicates that the students know that this is not the right answer (and not for example because the distractor was so poorly worded that the student could guess it wasn't the right answer).

- q-values
  These are the opposite of the p-value, i.e., the proportion of students that answered the question incorrectly. In multiple-choice questions they are the sum of the a-values of the distractor and the interpretation is similar to that of the p-value and a-values.
- Rit or item-total correlation (or discrimination/point-biserial)
  The Rit is the correlation between the item and the total score on the test. In other words, whether the item was answered correctly mainly by those students who also had a high score on the test (and incorrectly by those with a low total test score) or the other way around (answered correctly mainly by those with low total scores and vice versa). It is therefore an indication of whether the item aligns well with the rest of the test. So if an item has a low p-value but a high Rit this probably means that the item was difficult and could only be answered by the bright students, whereas if the item has a low p-value and a low Rit it more probably means that the item was not very relevant for the test. Because the Rit is a correlation it can run between 1.00 and – 1.00. The former would mean that the item is a perfect indicator for the type of competence the test measures and the latter that it is the most imperfect indicator. In tests with low numbers of items, for example a 15-item short-answer test, the total score is for a large proportion influenced by the item – of which the Rit is calculated itself – (in this case 6 to 7 per cent) which creates the problem of an auto-correlation, and if a correlation with itself is included it spuriously increases the Rit. Therefore an alternative is the Rir, which is the correlation between the item and the total score on the rest of the test. This item-rest correlation is also called corrected item-total correlation. If an item has a high Rit we say that the item is highly discriminating. The point-biserial correlation is another way of describing this relationship. It is worth mentioning again that these values are highly dependent on the cohort of students and the other items in the test. Also, the values will be different depending on what metric is used (how the correlations are calculated). So be careful. Don't take values as gospel; always ask more questions about how they were calculated and then go back to the items and see what these statistics can tell you. They are there to inform your judgement.

4 CONCLUSION

As discussed in the introduction to this document, it is impossible to define item quality so clearly that there will be full agreement. This means that in joint item production and test administration there will be items about which differences of opinion exist. This document is not intended to be used as a cookbook recipe to decide whether an item is good enough or not. Instead we have tried to bundle the currently available knowledge on determinants of quality and procedures to achieve high quality to enable a well-informed discussion between institutes, should disagreement about the quality of items arise.

5 REFERENCES

Bordage, G. (1987). An alternative approach to PMPs: the ‘key-features’ concept. In I.R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59–75). Montreal: Can-Heal Publications Inc.

Case, S.M., & Swanson, D.B. (1993). Extended-matching items: a practical alternative to free response questions. *Teaching and Learning in Medicine*, 5(2), 107–115.

Case, S.M., & Swanson, D.B. (1996). Constructing written test questions for the basic and clinical sciences. Retrieved 29 April 2014 from <http://www.nbme.org/publications/item-writing-manual-download.html>.

Cohen-Schotanus, J., & van der Vleuten, C.P.M. (2010). A standard setting method with the best performing students as point of reference: practical and affordable. *Medical Teacher*, 32, 154–160.

Cronbach, L.J. (1983). What price simplicity? *Educational Measurement: Issues and Practice*, 2(2), 11–12.

Cusimano, M.D. (1996). Standard setting in medical education. *Acad Med*, 71(10 Suppl), S112–120.

Downing, S.M., & Haladyna, T.M. (1997). Test item development: validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10(1), 61–82.

Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R.L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice*, 2(2), 7–10.

Hakel, M.D. (1968). How often is often? *American Psychologist*, 23(7), 533–534.

Hofstee, W.K.B. (1983). The case for compromise in educational selection and grading. In S.B. Anderson & J.S. Helmick (Eds.), *On educational testing* (pp. 109–127). San-Francisco: Josey-Bass.

Hurlburt, D. (1954). The relative value of recall and recognition techniques for measuring precise knowledge of word meaning, nouns, verbs, adjectives. *Journal of Educational Research*, 47(8), 561–576.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement*, Fourth Edition. Westport, CT: American Council on Education and Praeger, pp. 17–64.

Livingston, S.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton NJ: Educational Testing Service.

Miller, G.E. (1976). Continuous assessment. *Medical Education*, 10, 81–86.

Muijtjens, A.M., Schuwirth, L.W.T., Cohen-Schotanus, J., & van der Vleuten, C.P.M. (2007). Origin bias of test items compromises the validity and fairness of curriculum comparisons. *Medical Education*, 41(12), 1217–1223.

Norman, G., Swanson, D., & Case, S. (1996). Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. *Teaching and Learning in Medicine*, 8(4), 208–216.

Norman, G.R., Smith, E.K.M., Powles, A.C., Rooney, P.J., Henry, N.L., & Dodd, P.E. (1987). Factors underlying performance on written tests of knowledge. *Medical Education*, 21, 297–304.

Schuwirth, L.W.T., Bosman, G., Henning, R.H., Rinkel, R., & Wenink, A.C. (2010). Collaboration on progress testing in medical schools in the Netherlands. *Medical Teacher*, 32(6), 476–479.

Schuwirth, L.W.T. (1998). *An approach to the assessment of medical problem solving: Computerised Case-based Testing.* Universiteit Maastricht, Maastricht.

Schuwirth, L.W.T., Blackmore, D.B., Mom, E., Van den Wildenberg, F., Stoffers, H., & van der Vleuten, C.P.M. (1999). How to write short cases for assessing problem-solving skills. *Medical Teacher*, 21(2), 144–150.

Schuwirth, L.W.T., van der Vleuten, C.P.M., & Donkers H.H.L.M. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education*, 30, 44–49.

Schuwirth, L.W.T., Verheggen, M.M., van der Vleuten, C.P.M., Boshuizen, H.P.A., & Dinant, G.J. (2001). Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education*, 35(4), 348–356.

Smith, E.V. Jr., Smith, R.M. (Eds.) (2004) *Introduction to Rasch measurement: theory, models, and applications*. Maple Grove, MN: JAM Press.

van der Vleuten, C.P.M. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Science Education*, 1(1), 41–67.

van der Vleuten, C.P.M., & Schuwirth, L.W.T. (2005). Assessing professional competence: from methods to programmes. *Medical Education*, 39(3), 309–317.

Ward, W. (1982) A Comparison of Free Response and Multiple-Choice Forms of Verbal Aptitude Tests, *GRE Board Professional Report GREB No. 79-8P*; ETS Research Report 81–28, January 1982.