2011; 33: 783–797 MEDICAL TEACHER

AMEE GUIDE

General overview of the theories used in assessment: AMEE Guide No. 57

LAMBERT W. T. SCHUWIRTH[1] & CEES P. M. VAN DER VLEUTEN[2]
[1]Flinders University, Australia, [2]Maastricht University, The Netherlands

Abstract

There are no scientific theories that are uniquely related to assessment in medical education. There are many theories in adjacent fields, however, that can be informative for assessment in medical education, and in the recent decades they have proven their value. In this AMEE Guide we discuss theories on expertise development and psychometric theories, and the relatively young and emerging framework of assessment for learning. Expertise theories highlight the multistage processes involved. The transition from novice to expert is characterised by an increase in the aggregation of concepts from isolated facts, through semantic networks to illness scripts and instance scripts. The latter two stages enable the expert to recognise the problem quickly and form a quick and accurate representation of the problem in his/her working memory. Striking differences between experts and novices is not *per se* the possession of more explicit knowledge but the superior organisation of knowledge in his/her brain and pairing it with multiple real experiences, enabling not only better problem solving but also more efficient problem solving. Psychometric theories focus on the validity of the assessment – does it measure what it purports to measure and reliability – are the outcomes of the assessment reproducible. Validity is currently seen as building a train of arguments of how best observations of behaviour (answering a multiple-choice question is also a behaviour) can be translated into scores and how these can be used at the end to make inferences about the construct of interest. Reliability theories can be categorised into classical test theory, generalisability theory and item response theory. All three approaches have specific advantages and disadvantages and different areas of application. Finally in the Guide, we discuss the phenomenon of assessment *for* learning as opposed to assessment *of* learning and its implications for current and future development and research.

Introduction

It is our observation that when the subject of assessment in medical education is raised, it is often the start of extensive discussions. Apparently, assessment is high on everyone's agenda. This is not surprising because assessment is seen as an important part of education in the sense that it not only defines the quality of our students and our educational processes, but it is also seen as a major factor in steering the learning and behaviour of our students and faculty.

Arguments and debates on assessment, however, are often strongly based on tradition and intuition. It is not necessarily a bad thing to heed tradition. George Santayana already stated (quoting Burk) that *Those who do not learn from history are doomed to repeat it.*1 So, we think that an important lesson is also to learn from previous mistakes and avoid repeating them.

Intuition is also not something to put aside capriciously, it is often found to be a strong driving force in the behaviour of people. But again, intuition is not always in concordance with research outcomes. Some research outcomes in assessment are somewhat counter intuitive or at least unexpected. Many researchers may not have exclaimed *Eureka* but *Hey, that is odd* instead.

This leaves us, as assessment researchers, with two very important tasks. First, we need to critically study which common and tradition-based practices still have value and

Practice points

- Neither good quality development of assessment in medical education, nor any scientific study related to assessment, can do without a sound knowledge of the theories underlying it.
- Validation is building a series of arguments to defend the principle that assessment results really represent the intended construct and without which validation is never complete.
- An assessment instrument is never valid *per se*, it is only valid for a specific goal or specific goals.
- The validity of an assessment instrument is generally not determined by its format but by its content.
- Reliability is the extent to which test results are reproducible and can be seen as one of the important components of the validity argument.
- When applying one of the theories on reliability, the user should be acquainted with the possibilities, limitations and underlying assumptions to avoid over- or underestimations of the reproducibility.
- In addition to calculating the reliability of an instrument, it is insightful to calculate the SEM as well and compare this to the original test data.
- When building an assessment programme, it is imperative to clearly define the goals of the assessment programme.

Correspondence: L. W. T. Flinders Innovations in Clinical Education, Health Professions Education, Flinders University, GPO Box 2100, Adelaide 5001, South Australia, Australia; email: lambert.schuwirth@flinders.edu.au

ISSN 0142–159X print/ISSN 1466–187X online/11/100783–15 © 2011 Informa UK Ltd.
DOI: 10.3109/0142159X.2011.611022

consequently which are the mistakes that should not be repeated. Second, it is our task to translate research findings to methods and approaches in such a way that they can easily help changing incorrect intuitions of policy makers, teachers and students into correct ones. Both goals cannot be attained without a good theoretical framework in which to read, understand and interpret research outcomes. The purpose of this AMEE Guide is to provide an overview of some of the most important and most widely used theories pertaining to assessment. Further Guides in assessment theories will give more detail on the more specific theories pertaining to assessment.

Unfortunately, like many other scientific disciplines, medical assessment does not have one overarching or unifying theory. Instead, it draws on various theories from adjacent scientific fields, such as general education, cognitive psychology, decision-making and judgement theories in psychology and psychometric theories. In addition, there are some theoretical frameworks evolving which are more directly relevant to health professions assessment, the most important of which (in our view) is the notion of 'assessment of learning' *versus* 'assessment for learning' (Shepard 2009).

In this AMEE Guide we will present the theories that have featured most prominently in the medical education literature in the recent four decades. Of course, this AMEE Guide can never be exhaustive; the number of relevant theoretical domains is simply too large, nor can we discuss all theories to their full extent. Not only would this make this AMEE Guide too long, but also this would be beyond its scope, namely to provide a concise overview. Therefore, we will discuss only the theories on the development of medical expertise and psychometric theories, and then end by highlighting the differences between the assessment of learning and assessment for learning. As a final caveat, we must say here that this AMEE Guide is not a guide to methods of assessment. We assume that the reader has some prior knowledge about this or we would like to refer to specific articles or to text books (e.g. Dent & Harden 2009).

Theories on the development of (medical) expertise

What distinguishes someone as an expert in the health sciences field? What do experts do differently compared to novices when solving medical problems? These are questions that are inextricably tied to assessment, because if you do not know *what* you are assessing it also becomes very difficult to know *how* you can best assess.

I may be obvious that someone can only become an expert through learning and gaining experience.

One of the first to study the development of expertise was by de Groot (1978), who wanted to explore why chess grandmasters became grandmasters and what made them differ from good amateur chess players. His first intuition was that grandmasters were grandmasters because they were able to think more moves ahead than amateurs. He was surprised, however, to find that this was not the case; players of both expertise groups did not think further ahead than roughly seven moves. What he found, instead, was that grandmasters were better able to remember positions on the board. He and his successors (Chase & Simon 1973) found that grandmasters were able to reproduce positions on the board more correctly, even after very short viewing times. Even after having seen a position for only a few seconds, they were able to reproduce it with much greater accuracy than amateurs.

One would think then that they probably had superior memory skills, but this is not the case. The human working memory has a capacity of roughly seven units (plus or minus two) and this cannot be improved by learning (Van Merrienboer & Sweller 2005, 2010).

The most salient difference between amateurs and grandmasters was not the number of units they could store in their working memory, but the richness of the information in each of these units.

To illustrate this, imagine having to copy a text in your own language, then a text in a foreign Western European language and then one in a language that uses a different character set (e.g. Cyrillic). It is clear that copying a text in your own language is easiest and copying a text in a foreign character set is the most difficult. While copying you have to read the text, store it in your memory and then reproduce it on the paper. When you store the text in your native language, all the words (and some fixed expressions) can be stored as one unit, because they relate directly to memories already present in your long-term memory. You can spend all your cognitive resources on memorising the text. In the foreign character set you will also have to spend part of your cognitive resources on memorising the characters, for which you have no prior memories (schemas) in your long-term memory. A medical student who has just started his/her study will have to memorise all the signs and symptoms when consulting a patient with heart failure, whereas an expert can almost store it as one unit (and perhaps only has to store the findings that do not fit to the classical picture or mental model of heart failure). This increasing ability to store information as more information-rich units is called chunking and it is a central element in expertise and its development. Box 1 provides an illustration of the role of chunking.

So, why were the grandmasters better than good amateurs? Well, mainly because they possessed much more stored information about chess positions than amateurs did, or in other words, they had acquired so much more knowledge than the amateurs had.

If there is one lesson to be drawn from these early chess studies – which have been replicated in such a plethora of other expertise domains that it is more than reasonable to

Box 1. The role of chunking in storing and retrieving information.

Through chunking, a person is able to store more information and, as long as the information is more meaningful, with even greater ease.
Suppose you were asked to memorise the following 20 characters:
Aomcameinaetaiodbtai
You will probably find it a difficult task (but doable)
Suppose we now increase the number of characters and ask you to memorise them again:
Assessment of medical competence and medical expertise is not an easy task, and is often dominated by tradition and intuition.
Now, the message contains 126 characters (including spaces and the full stop), but is much easier to memorise.

assume that these findings are generic – it is that a rich and well-organised knowledge base is essential for successful problem solving (Chi et al. 1982; Polsen & Jeffries 1982).

The next question then would be: *What does 'well-organised' mean?* Basically, it comes down to organisation that will enable the person to store new information rapidly and with good retention and to be able to retrieve relevant information when needed. Although the computer is often used as a metaphor for the human brain (much like the clock was used as a metaphor in the nineteenth century), it is clear that information storage on a hard disk is very much different from human information storage. Humans do not use a File Allocation Table to index where the information can be found, but have to embed information in existing (semantic) networks (Schmidt et al. 1990). The implication of this is that it is very difficult to store new information if there is no existing prior information to which it can be linked. Of course, the development of these knowledge networks is quite individualised, and based on the individual learning pathways and experiences. For example, we – the authors of this AMEE Guide – live in Maastricht, so our views, connotations and association with 'Maastricht' differ entirely from those of most of the readers of the AMEE Guides, although we may share the knowledge that it is a city (and perhaps that it is in the Netherlands) and that there is a university with a medical school, the rest of the knowledge is much more individualised.

Knowledge generally is quite domain specific (Elstein et al. 1978; Eva et al. 1998); a person can be very knowledgeable on one topic and a lay person on another, and because expertise is based on a well-organised knowledge base, expertise is domain specific as well. For assessment, this means that the performance of a candidate on one case or item of a test is a poor predictor for his or her performance on any other given item or case in the test. Therefore, one can never rely on limited assessment information, i.e. high-stakes decisions made on the basis of a single case (e.g. a high-stakes final VIVA) are necessarily unreliable.

A second important and robust finding in the expertise literature – more specifically the diagnostic expertise literature – is that problem-solving ability is idiosyncratic (cf. e.g. the overview paper by Swanson et al. 1987). Domain specificity, which we discussed above, means that the performance of the *same* person varies considerably across *various* cases, idiosyncrasy here means that the way different experts solve the *same* case varies substantially between *different* experts. This is also logical, keeping in mind that the way the knowledge is organised is highly individual. The assessment implication from this is that when trying to capture, for example, the diagnostic expertise of candidates, the process may be less informative than the outcome, as the process is idiosyncratic (and fortunately the outcome of the reasoning process is much less).

A third and probably most important issue is the matter of transfer (Norman 1988; Regehr & Norman 1996; Eva 2004). This is closely related to the previous issue of domain specificity and idiosyncrasy. Transfer pertains to the extent to which a person is able to apply a given problem-solving approach to different situations. It requires that the candidate understands the similarities between two different problem situations and recognises that the same problem-solving principle can be applied. Box 2 provides an illustration (drawn from a personal communication with Norman).

Most often, the first problem is not recognised as being essentially the same as the second and that the problem-solving principle is also the same. Both solutions lie in the splitting up of the total load into various parts. In problem 1, the 1000 W laser beam is replaced by 10 rays of 100 W each, but converging right on the spot where the filament was broken. In the second problem the solution is more obvious: build five bridges and then let your men run onto the island. If the problem were represented as: you want to irradiate a tumour but you want to do minimal harm to the skin above it, it would probably be recognised even more readily by physicians. The specific presentation of these problems is labelled as the surface features of the problem and the underlying principle is referred to as the deep structure of the problem. Transfer exists by the virtue of the expert to be able to identify the deep structure and not to be blinded by the surface features.

One of the most widely used theories on the development of medical expertise is the one suggested by Schmidt, Norman and Boshuizen (Schmidt 1993; Schmidt & Boshuizen 1993). Generally put, this theory postulates that the development of medical expertise starts with the collection of isolated facts which further on in the process are combined to form meaningful (semantic) networks. These networks are then aggregated into more concise or dense illness scripts (for example pyelonephritis). As a result of many years of experience, these are then further enriched into instance scripts, which enable the experienced clinician to recognise a certain diagnosis instantaneously. The most salient difference between illness scripts (that are a sort of congealed patterns of a certain diagnosis) and instance scripts is that in the latter contextual, and for the lay person sometimes seemingly irrelevant, features are also included in the recognition.

Box 2. The role of transfer in problem solving.

Problem 1: You are in possession of a unique and irreplaceable light bulb. Unfortunately, the filament is broken; so you cannot light the bulb anymore. There is no way of removing the glass without breaking the light bulb and to repair; you have to weld the filament with a laser beam. For this, you will need an energy output of 1000 W. Unfortunately, the glass will break if a laser beam with an intensity of more than 100 W runs through it.
How can you weld the filament?
Problem 2: You are an evil medieval knight. You want to conquer a tower from your enemy. The tower is located on a small piece of land, an island completely surrounded by a moat. To successfully conquer the tower, you must bring 500 men simultaneously onto the island. Unfortunately, any bridge you can build will only hold 100 men.
How do you bring 500 men on the island simultaneously?

Typically, these include the demeanour of the patient or his/her appearance, sometimes even an odour, etc.

These theories then provide important lessons for assessment:

(1) Do not rely on short tests. The domain specificity problem informs us that high-stakes decisions based on short tests or tests with a low number of different cases are inherently flawed with respect to their reliability (and therefore also validity). Keep in mind that unreliability is a two-way process, it does not only imply that someone who failed the test could still have been satisfactorily competent, but also that someone who passed the test could be incompetent. The former candidate will remain in the system and be given a re-sit opportunity, and this way the incorrect pass–fail decision can be remediated, but the latter will escape further observation and assessment, and the incorrect decision cannot be remediated again.

(2) For high-stakes decisions, asking for the process is less predictive of the overall competence than focussing on the outcome of the process. This is counterintuitive, but it is a clear finding that the way someone solves a given problem is not a good indicator for the way in which she/he will solve a similar problem with different surface features; she/he may not even recognise the transfer. Focussing on multiple outcomes or some essential intermediate outcomes – such as with extended-matching questions, key-feature approach assessment or the script concordance test – is probably better than in-depth questioning the problem-solving process (Bordage 1987; Case & Swanson 1993; Page & Bordage 1995; Charlin et al. 2000).

(3) Assessment aimed only at reproduction will not help to foster the emergence of transfer in the students. This is not to say that there is no place for reproduction-orientated tests in an assessment programme, but they should be chosen very carefully. When learning arithmetic, for example, it is okay to focus the part of the assessment pertaining to the tables of multiplication on reproduction, but with long multiplications, focussing on transfer (in this case, the algorithmic transfer) is much more worthwhile.

(4) When new knowledge has to be built into existing semantic networks, learning needs to be contextual. The same applies to assessment. If the assessment approach is to be aligned with the educational approach, it should be contextualised as well. So whenever possible, set assessment items, questions or assignments in a realistic context.

Psychometric theories

Whatever purpose an assessment may pursue in an assessment programme, it always entails a more or less systematic collection of observations or data to arrive at certain conclusions about the candidate. The process must be both reliable and valid. Especially, for these two aspects (reliability and validity) psychometric theories have been developed. In this chapter, we will discuss these theories.

Validity

Simply put, validity pertains to the extent to which the test actually measures what it purports to measure. In the recent century, the central notions of validity have changed substantially several times. The first theories on validity were largely based on the notion of criterion or predictive validity. This is not illogical as the intuitive notion of validity is one of whether the test predicts an outcome well. The question that many medical teachers ask when a new assessment or instructional method is suggested is: *But does this produce better doctors?*. This question – however logical – is unanswerable in a simple criterion-validity design as long as there is no good single measureable criterion for good 'doctorship'. This demonstrates exactly the problem with trying to define validity exclusively in such terms. There is an inherent need to validate the criterion as well. Suppose a researcher was to suggest a measure to measure 'doctorship' and to use it as the criterion for a certain assessment, then she/he would have to validate the measure for 'doctorship' as well. If this again were only possible through criterion validity, it would require the research to validate the criterion for the criterion as well – etcetera ad infinitum.

A second intuitive approach would be to simply observe and judge the performance. If one, for example, wishes to assess flute-playing skills, the assessment is quite straightforward. One could collect a panel of flute experts and ask them to provide judgements for each candidate playing the flute. Of course, some sort of blueprinting would then be needed to ensure that the performances of each candidate would entail music in various ranges. For orchestral applicants, it would have to ensure that all classical music styles of the orchestra's repertoire would be touched upon. Such forms of content validity (or direct validity) have played an important role and still do in validation procedures.

However, most aspects of students we want to assess are still not clearly visible and need to be inferred from observations. Not only are characteristics such as intelligence or neuroticism invisible (so-called latent) traits, but also are elements such as knowledge, problem-solving ability, professionalism, etc. They cannot be observed directly and can only be assessed as assumptions based on observed behaviour.

In an important paper, Cronbach and Meehl (1955) elaborated on the then still young notion of construct validity. In their view, construct validation should be seen as analogous to the inductive empirical process; first the researcher has to define, make explicit or postulate clear theories and conceptions about the construct the test purports to measure. Then, she/he must design and carry through a critical evaluation of the test data to see whether they support the theoretical notions of the construct. An example of this is provided in Box 3.

The so-called 'intermediate effect', as described in the example (Box 3) (especially when it proves replicable) is an important falsification of the assumption of validity of the test.

Box 3. An example of a construct validation procedure.

Suppose a test developer wants to design a new test to measure clinical problem solving. He decides to follow real life as closely as possible and to design a set of authentic patient simulations. In such a test, the candidates are given the initial complaint and they then have to work their way through the simulation, asking relevant history questions, 'performing' physical examinations, ordering additional diagnostics, etc. In order to determine the total score, all decisions are scored. Every relevant history taking-question, relevant physical examination or additional diagnostic is score with a mark. The total mark determines the total score.

It is clear from the theoretical perspective of problem solving that this is not a valid test. Current theories highlight the emergence of scripts and schemata, enabling the expert to come to the right conclusion with less information than the novice. In short, experts in general are more efficient in their data gathering and not necessarily more proficient. The marking system rewards thoroughness and not efficiency. So, there is good reason to doubt the construct validity of the method, as the translation from observation to scoring is not in accordance to the theory behind the construct of interest.

Empirical data have confirmed this. The method described, the PMP (Berner et al. 1974), showed that intermediates outperformed experts, mainly because the expert efficiency was penalised rather than rewarded (Schmidt et al. 1988).

We have used this example deliberately, and there are important lessons that can be drawn from it. First, it demonstrates that the presence of such an intermediate effect in this case is a powerful falsification of the assumption of validity. This is highly relevant, as currently it is generally held that a validation procedure must contain 'experiments' or observations which are designed to optimise the probability of falsifying the assumption of validity (much like Popper's falsification principle2). Evidence supporting the validity must therefore always arise from critical 'observations'. There is a good analogy to medicine or epidemiology. If one wants to confirm the presence of a certain disease with the maximum likelihood, one must use the test with the maximum chance of being negative when disease is absent (the maximum sensitivity). Confirming evidence from 'weak' experiments therefore does not contribute to the validity assumption.

Second, it demonstrates that authenticity is not the same as validity, which is a popular misconception. There are good reasons in assessment programmes to include authentic tests or to strive for high authenticity, but the added value is often more prominent in their formative than in their summative function. An example may illustrate this: Suppose we want to assess the quality of the day-to-day performance of a practising physician and we had the choice between observing him/her in many real-life consultations or extensively reviewing charts (records and notes), ordering laboratory tests and referral data. The second option is clearly less authentic than the first one but it is fair to argue that the latter is a more valid assessment of the day-to-day practice than the former. The observer effect, for example, in the first approach may influence the behaviour of the physician and thus draw a biased picture of the actual day-to-day performance, which is clearly not the case in the charts, laboratory tests and referral data review.

Third, it clearly demonstrates that validity is not an entity of the assessment *per se*; it is always the extent to which the test assesses the desired characteristic. If the PMPs in the example in Box 3 were aimed at measuring thoroughness of data gathering – i.e. to see whether students are able to distinguish all the relevant data from non-relevant data – they would have been valid, but if they are aimed at measuring expertise they failed to incorporate efficiency of information gathering and use as an essential element of the construct.

Current views (Kane 2001, 2006) highlight the argument-based inferences that have to be made when establishing validity of an assessment procedure.

In short, inferences have to be made from observations to scores, from observed scores to universe scores (which is a generalisation issue), from universe scores to target domain and from target domain to construct.

To illustrate this, a simple medical example may be helpful: When taking a blood pressure as an assessment of someone's health, the same series of inferences must be made. When taking a blood pressure, the sounds heard through the stethoscope when deflating the cuff have to be translated into numbers by reading them from the sphygmomanometer. This is the translation from (acoustic and visual) observation to scores. Of course, one measurement is never enough (the patient may just have come running up the stairs) and it needs to be repeated, preferable under different circumstances (e.g. at home to prevent the 'white coat'-effect). This step is equivalent to the inference from observed scores to universe scores. Then, there is the inference from the blood pressure to the cardiovascular status of the patient (often in conjunction with other signs and symptoms and patient characteristics) which is equivalent to the inference from universe score to target domain. And, finally this has to be translated into the concept 'health', which is analogous to the translation of target domain to construct. There are important lessons to be learnt from this.

First, validation is building a case based on argumentation. The argumentation is preferably based on outcomes of validation studies but may also contain plausible and/or defeasible arguments.

Second, one cannot validate an assessment procedure without a clear definition or theory about the construct the assessment is intended to capture. So, an instrument is never valid *per se* but always only valid for capturing a certain construct.

Third, validation is never finished and often requires a plethora of observations, expectations and critical experiments.

Fourth, and finally, in order to be able to make all these inferences, generalisability is a necessary step.

Reliability

Reliability of a test indicates the extent to which the scores on a test are reproducible, in other words, whether the results a candidate obtains on a given test would be the same if she/he were presented with another test or all the possible tests of the domain. As such, reliability is one of the approaches to the generalisation step described in the previous

section on validity. But even if generalisation is'only' one of the necessary steps in the validation process, the way in which this generalisation is made is subject to theories in its own. To understand them, it may be helpful to distinguish three levels of generalisation.

First, however, we need to introduce the concept of the 'parallel test' because it is necessary to understand the approaches to reproducibility described below. A parallel test is a hypothetical test aimed at a similar content, of equal difficulty and with a similar blueprint, ideally administered to the same group of students immediately after the original test, under the assumption that the students would not be tired and that their exposure to the items of the original test would not influence their performance on the second.

Using this notion of the parallel test, three types of generalisations are made in reliability, namely if the same group of students were presented with the original and the parallel test:

(1) Whether the same students would pass and fail on both tests.
(2) Whether the rank ordering from best to most poorly performing student would be the same on both the original and the parallel tests.
(3) Whether all students would receive the same scores on the original and the parallel tests.

Three classes of theories are in use for this: classical test theory (CTT), generalisability theory (G-theory) and item response theory (IRT).

Classical test theory. CTT is the most widely used theory. It is the oldest and perhaps easiest to understand. It is based on the central assumption that the observed score is a combination of the so-called true score and an error score ($O = T + e$).[3] The true score is the hypothetical score a student would obtain based on his/her competence only. But, as every test will induce measurement error, the observed score will not necessarily be the same as the true score.

This in itself may be logical but it does not help us to estimate the true score. How would we ever know how reliable a test is if we cannot estimate the influence of the error term and the extent it makes the observed score deviate from the true score, or the extent to which the results on the test are replicable?

The first step in this is determining the correlation between the test and a parallel test (test–retest reliability). If, for example, one wanted to establish the reliability of a haemoglobin measurement one would simply compare the results of multiple measurements from the same homogenised blood sample, but in assessment this is not this easy. Even the 'parallel test' does not help here, because this is, in most cases, hypothetical as well.

The next step, as a proxy for the parallel test, is to randomly divide the test in two halves and treat them as two parallel tests. The correlation between those two halves (corrected for test length) is then a good estimate of the 'true' test–retest correlation. This approach, however, is also fallible, because it is not certain whether this specific correlation is a good exemplar; perhaps another subdivision in two halves would have yielded a completely different correlation (and thus a different estimate of the test–retest correlation). One approach is to repeat the subdivision as often as possible until all possibilities are exhausted and use the mean correlation as a measure of reliability. That is quite some work, so it is simpler and more effective to subdivide the test in as many subdivisions as there are possible (the items) and calculate the correlations between them. This approach is a measure of internal consistency and the basis for the famous Cronbach's alpha. It can be taken as the mean of all possible split half reliability estimates (cf. e.g. Crocker & Algina 1986).

Although Cronbach's alpha is widely used, it should be noted that it remains an estimate of the test–retest correlation, so it can only be used correctly if conclusions are drawn at the level of the whether the rank orderings between the original and the parallel tests are the same, i.e. a norm-referenced perspective. It does not take into account the difficulty of the items on the test, and because the difficulty of the items of a test influences the exact height of the score, using Cronbach's alpha in a criterion-referenced perspective overestimates the reliability of the test. This is explained in Box 4.

Although the notion of Cronbach's alpha is based on correlations, reliability estimates can range from 0 to 1. In rare cases, calculations could result in a value lower than zero, but this is then to be interpreted as being zero.

Although it is often helpful to have a measure of reliability that is normalised, in that for all data, it is always a number between 0 and 1, in some cases, it is also important to evaluate what the reliability means for the actual data. Is a test with a reliability of 0.90 always better than a test with a reliability of 0.75? Suppose we had the results of two tests and that both tests had the same cut-off score, for example 65%. The score distributions of both tests have a

Box 4. Difference between reliability from a norm- and criterion-referenced perspective.

Suppose a test was administered to five students: A, B, C, D and E and their scores on the original test are the ones in the first column and those of the parallel test are in the second column:

A	64	85
B	59	79
C	56	62
D	53	61
E	47	48

The test–retest correlation is perfect; so one could assume that reliability is good. But the absolute scores on the original test are consistently lower than those in the parallel test. Especially, when, for example, the cut-off score is set to 60%, 4 out of 5 students will fail the original test and only one would fail the parallel test. There are therefore some differences in pass–fail decisions between both test, whereas Cronbach's alpha would indicate perfect reliability. This is not a flaw in Cronbach's alpha but only to illustrate than any measure used incorrectly will produces false results.

Table 1. Descriptive statistics of two hypothetical tests.

	Cut-off score (%)	SD (%)	Mean (%)	Minimum (%)	Maximum (%)	Reliability
Test 1	65	5	83	66	97	0.75
Test 2	65	5	68	53	81	0.90

standard deviation (SD) of 5%, but the mean, minimum and maximum scores differ, as shown in Table 1.

Based on these data, we can calculate a 95% confidence interval (95%-CI) around each score or the cut-off score. For this, we need the standard error of measurement (SEM). In the beginning of this section, we showed the basic formula in CTT (observed score = true score + error). In CTT, the SEM is the SD of the error term or, more precisely put, the square root of the error variance. It is calculated as follows:

$$\text{SEM} = \text{SD}\sqrt{1-\alpha}$$

If we use this formula, we find that in test 1, the SEM is 2.5% and in test 2, it is 1.58%. The 95% CIs are calculated by multiplying the SEM by 1.96. So, in test 1 the 95% CI is ±4.9% and in test 2 it is ±3.09%. In test 1 the 95% CI around the cut-off score ranges from 60.1% to 69.9% but only a small proportion of the score of students falls into this 95% CI.4 This means that for those students we are not able to conclude, with a $p \leq 0.05$, whether these students have passed or failed the test. In test 2, the 95% CI ranges from 61.9% to 68.1% but now many students fall into the 95% CI interval. We use this hypothetical – though not unrealistic – example to illustrate that a higher reliability is not automatically better. To illustrate this further, Figure 1 presents a graphical representation of both tests.

Generalisability theory. G-theory is not *per se* an extension to CTT but a theory on its own. It has different assumptions than CTT, some more nuanced, some more obvious. These are best explained using a concrete example. We will discuss G-theory here, using such an example.

When a group of 500 students sit a test, say a 200-item knowledge-based multiple-choice test, their total scores will differ. In other words, there will be variance between the scores. From a reliability perspective, the goal is to establish the extent to which these score differences are based on differences in ability of the students in comparison to other – unwanted – sources of variance. In this example, the variance that is due to differences in ability (in our example 'knowledge') can be seen as wanted or true score variance. Level of knowledge of students is what we want our test to pick up, the rest is noise – error – in the measurement. G-theory provides the tools to distinguish true or universe score variance from error variance, and to identify and estimate different sources of error variance. The mathematical approach to this is based on analysis of variance, which we will not discuss here. Rather, we want to provide a more intuitive insight into the approach and we will do this stepwise with some score matrices.

In Table 2, all students have obtained the same score (for reasons of simplicity, we have drawn a table of five test items

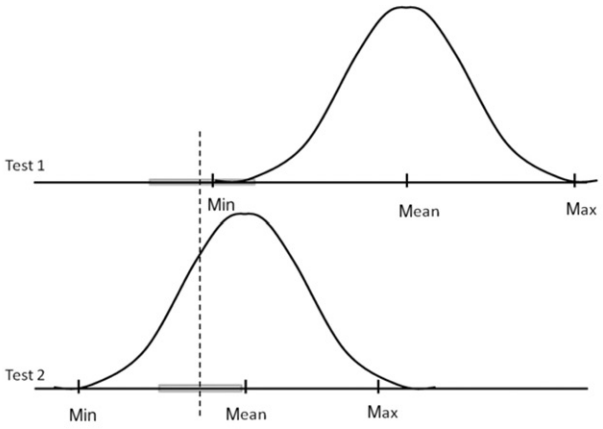

Figure 1. Two tests, in which the one with a lower reliability produces fewer incorrect pass–fail decisions, used to illustrate the value of calculating an SEM.

Table 2. Item variance (I-variance).

	Items					
Students	1	2	3	4	5	Total score
A	1	0.9	0.5	0.1	0	2.5
B	1	0.9	0.5	0.1	0	2.5
C	1	0.9	0.5	0.1	0	2.5
D	1	0.9	0.5	0.1	0	2.5
E	1	0.9	0.5	0.1	0	2.5
p-Value	1	0.9	0.5	0.1	0	

and five candidates). From the total scores and the *p*-values, it becomes clear that all the variance in this matrix is due to systematic differences in items. Students collectively 'indicate' that item 1 is easier than item 2, and item 2 is easier than item 3, etc. There is no variance associated with students. All students have the same total score and they have collected their points on the same items. In other words, all variance here is item variance (I-variance).

Table 3 draws exactly the opposite picture. Here, all variance stems from differences between students. Items agree maximally as to the ability of the students. All items give each student the same marks, but their marks differ for all students, so the items make a consistent, systematic distinction between students. In the score matrix, all items agree that student A is better than student B, who in turn is better than student C, etc. So, here, all variance is student-related variance (person variance or P-variance).

Table 4 draws a more dispersed picture. For students A, B and C, items 1 and 2 are easy and items 3–5 difficult, and the reverse is true for students D and E. There seems to be a clearly discernable interaction effect between items and students. Such a situation could occurs if, for example, items 1 and 2 are on cardiology and 3–5 on the locomotor system, and students A, B and C have just finished their clerkship in cardiology and the other students just finished their orthopaedic surgery placements.

Of course, real life is never this simple, so matrix 5 (Table 5) presents a more realistic scenario, some variance can be

Table 3. Student or person variance (P-variance).

	Items					
Students	1	2	3	4	5	Total score
A	1	1	1	1	1	5.0
B	0.9	0.9	0.9	0.9	0.9	4.5
C	0.5	0.5	0.5	0.5	0.5	2.5
D	0.1	0.1	0.1	0.1	0.1	0.5
E	0	0	0	0	0	0
p-Value	0.5	0.5	0.5	0.5	0.5	

Table 4. Systematic interaction between items and students (P × I variance).

	Items					
Students	1	2	3	4	5	Total score
A	1	1	0	0	0	2.0
B	1	1	0	0	0	2.0
C	1	1	0	0	0	2.0
D	0	0	1	1	1	3.0
E	0	0	1	1	1	3.0
p-Value	0.6	0.6	0.4	0.4	0.4	0

Table 5. Systematic and non-systematic effects.

	Items					
Students	1	2	3	4	5	Total score
A	1	1	0	0	0	2.0
B	1	1	0	0	0	2.0
C	1	1	0	0	0	2.0
D	0	0	1	1	1	3.0
E	0	0	1	1	1	3.0
p-Value	0.6	0.6	0.4	0.4	0.4	0

attributed to systematic differences in item difficulty (I-variance), some to differences in student ability (P-variance), some to the interaction effects (P × I-variance), which in this situation cannot be disentangled from general error (e.g. perhaps student D knew the answer to item 4 but was distracted or he/she misread the item).

Generalisability is then determined by the portion of the total variance that is explained by the wanted variance (in our example, the P-variance). In a generic formula:

$$g = \frac{\text{wanted variance}}{\text{wanted} + \text{error variance}}$$

Or in the case of our 200 multiple choice test example:5

$$g = \frac{\mathrm{P}}{\mathrm{P} + \mathrm{I/ni} + \mathrm{P} \times \mathrm{I}, e/\mathrm{ni}}$$

The example of the 200-item multiple-choice test is called a one-facet design. There is only one facet on which we wish to generalise, namely would the *same* students perform similarly if *another* set of items (another 'parallel' test) were administered. The researcher does not want to draw conclusions as to the extent to which *another* group of students would perform similarly on the *same* set of items. If the latter were the purpose, she/he would have to redefine what is wanted and what is error variance. In the remainder of this paragraph we will also use the term 'factor' to denote all the components of which the variance components are estimates (so, P is a factor but not a facet).

If we are being somewhat more precise, the second formula is not always a correct translation of the first. The first deliberately does not call the denominator 'total variance', but 'wanted' and 'error variance'. Apparently, the researcher has some freedom in deciding what to include in the error term and what not. This of course, is not a capricious choice; what is included in the error term defines what type of generalisations can be made.

If, for example, the researcher wants to generalise as to whether the rank ordering from best to most poorly performing student would be the same on another test, the I-variance does not need to be included in the error term (for a test–retest correlation, the *systematic* difficulty of the items or the test is irrelevant). For the example given here (which is a so-called P × I design), the generalisability coefficient without the $\mathrm{I}_{/\mathrm{ni}}$ term is equivalent to Cronbach's alpha.

The situation is different if the reliability of an exact score is to be determined. In that case, the systematic item difficulty *is* relevant and should be incorporated in the error term. This is the case in the second formula.

To distinguish between both approaches, the former (without the I-variance) is called 'generalisability coefficient' and the latter 'dependability coefficient'. This distinction further illustrates the versatility of G-theory, when the researcher has a good overview on the sources of variance that contribute to the total variance she/he can clearly distinguish and compare the wanted from the unwanted sources of variance.

The same versatility holds for the calculation of the SEM. As discussed in the section on CTT the SEM is the SD of the error term, so in a generalisability analysis it can be calculated as the square root of the error variance components, so either

$$\sqrt{\mathrm{I/ni} + \mathrm{P} \times \mathrm{I}, e/\mathrm{ni}} \text{ or } \sqrt{\mathrm{P} \times \mathrm{I}, e/\mathrm{ni}}$$

In this example the sources of variance are easy to understand, because there is in fact one facet, but more complicated situations can occur. In an OSCE with two examiners per station, things already become more complicated. First, there is a second facet (the universe of possible examiners) on top of the first (the universe of possible stations). Second, there is crossing and nesting. A crossed design is most intuitive to understand. The multiple-choice example is a completely crossed design (P × I, the '×' indicating the crossing), all items are seen by all students. Nesting occurs when certain 'items' of a factor are only seen by some 'items' of another factor. This is a cryptic description, but the illustration of the OSCE may help. The pairs of examiners are nested within each station. It is not the same two

examiners who judge all stations for all students, but examiners A and B are in station 1, C and D in station 2, etc. The examiners are crossed with students (assuming that they remain the same pairs throughout the whole OSCE), because they have judged all students, but they are not crossed with all stations as A and B have only examined in station 1, etc. In this case examiner pairs are nested within stations.

There is a second part to the analyses in a generalisability analysis, namely the decision study or D-study. You may have noticed in the second formula that both the I-variance and the interaction terms have a subscript/ni. This indicates that the variance component is divided by the number of elements in the factor (in our example the number of items in the I-variance) and that the terms in the formula are the mean variances per element in the factor (the mean item variance). From this, it is relatively straightforward to extrapolate what the generalisability or dependability would have been if the numbers would change (e.g. what is the dependability if the number of items on the test would be twice as high, or which is more efficient, using two examiners per OSCE station or having more station with only one examiner?), just by inserting another value in the subscript(s). Although it may seem very simple, one word of caution is needed: such extrapolations are only as good as the original variance component estimates. The higher the number of original observations, the better the extrapolation. In our example, we had 200 items on the test and 500 students taking it, but it is obvious that this leads to better estimates and thus better extrapolations than 50 students sitting a 20 item test.

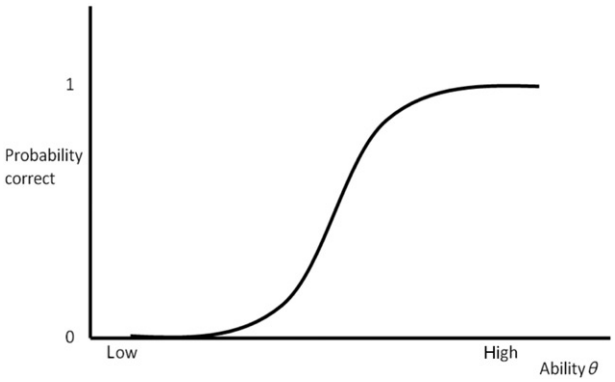

Figure 2. A generic example of an IRF ogive.

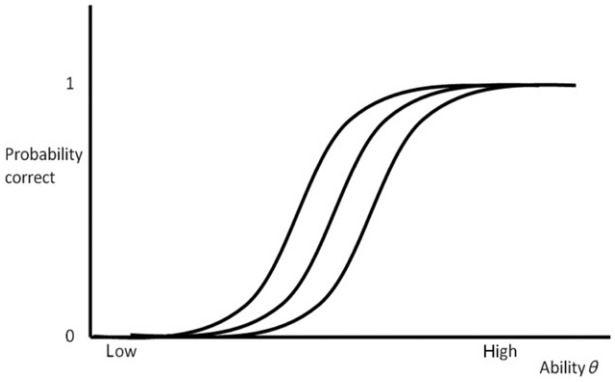

Figure 3. An example of a one-parameter model.

Item response theory. Both CTT and G-theory have a common disadvantage. Both theories do not have methods to disentangle test difficulty effects from candidate group effects. If a score on a set of items is low, this can be the result of a particularly difficult set of items or of a group of candidates who are of particularly low ability level. Item response theories try to overcome this problem by estimating item difficulty independent of student ability, and student ability independent of item difficulty.

Before we can explain this, we have to go back to CTT again. In CTT, item difficulty is indicated by the so-called *p*-value, the proportion of candidates who answered the item correctly, and discrimination indices such as point biserials, R_{it} (item-total correlation) or R_{ir} (item-rest correlation), all of which are measures to correlate the performance on an item to the performance on the total test or the rest of the items. If in these cases a different group of candidates (of different mean ability) would take the test, the *p*-values would be different, and if an item were re-used in a different test, all discrimination indices would be different. With IRT the response of the candidates are modelled, given their ability to each individual item on the test.

Such modelling cannot be done without making certain assumptions. The first assumption is that the ability of the candidates is uni-dimensional and the second is that all items on a test are locally independent except for the fact that they measure the same (uni-dimensional) ability. If, for example, a test would contain an item asking for the most probable diagnosis in a case and a second for the most appropriate therapy, these two items are not locally independent; if a candidate answers the first items incorrectly, she/he will most probably answer the second one incorrectly as well.

The third assumption is that modelling can be done through an item response function (IRF) indicating that for every position on the curve, the probability of a correct answer increases with a higher level of ability. The biggest advantage of IRT is that difficulty and ability are modelled on the same scale. IRFs are typically graphically represented as an ogive, as shown in Figure 2.

Modelling cannot be performed without data. Therefore pre-testing is necessary before modelling can be performed. The results on the pre-test are then used to estimate the IRF. For the purpose of this AMEE Guide, we will not go deeper into the underlying statistics but for the interested reader some references for further reading are included at the end.

Three levels of modelling can be applied, conveniently called one-, two- and three-parameter models. A one-parameter model distinguishes items only on the basis of their difficulty, or the horizontal position of the ogive. Figure 3 shows three items with three different positions of the ogive. The curve on the left depicts the easiest item of the three in this example; it has a higher probability of a correct answer with lower abilities of the candidate. The most right curve indicates the most difficult item. In this one-parameter modeling, the forms of all curves are the same, so their power to discriminate (equivalent to the discrimination indices of CTT) between students of high and low abilities are the same.

A two-parameter model includes this discriminatory power (on top of the difficulty). The curves for different items not only differ in their horizontal position but also in their steepness. Figure 4 shows three items with different discrimination (different steepness of the slopes). It should be noted that the curves do not only differ in their slopes but also in their positions, as they differ both in difficulty and in discrimination (if they would only differ in slopes, it would be a sort of one-parameter model again).

A three-parameter model includes the possibility that a candidate with extremely low ability (near-to-zero ability) still produces the correct answer, for example through random guessing. The third parameter determines the offset of the curve or more or less its vertical position. Figure 5 shows three items differing on all three parameters.

As said before, pre-testing is needed for parameter estimation and logically there is a relationship between the number of candidate responses needed for good estimates; the more parameters have to be estimated, the higher the number of responses needed. As a rule of thumb, 200–300 responses would be sufficient for one-parameter modelling, whereas a three-parameter model would require roughly 1000 responses. Typically, large testing bodies employ IRT mix items to be pre-tested with regular items, without the candidates knowing which item is which. But it is obvious that such requirements in combination with the complicated underlying statistics and strong assumptions limit the applicability of IRT in various situations. It will be difficult for a small-to-medium-sized faculty to produce enough pre-test data to yield acceptable estimates, and, in such cases, CTT and G-theory will have to do.

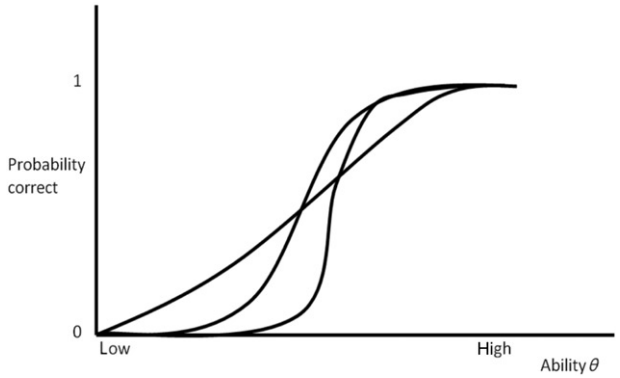

Figure 4. An example of a two-parameter model.

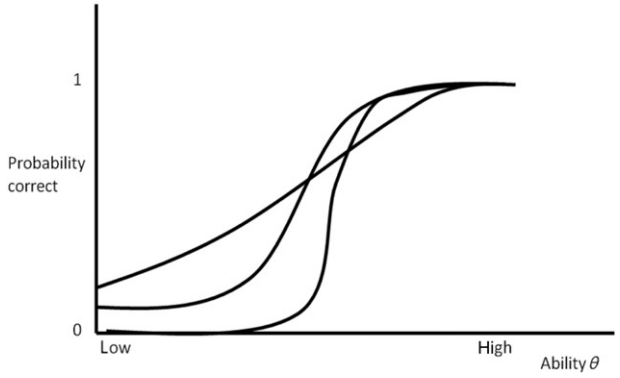

Figure 5. An example of a three-parameter model.

On the other hand, IRT must be seen as the strongest theory in reliability of testing, enabling possibilities that are impossible with CTT or G-theory. One of the 'eye-catchers' in this field is computer-adaptive testing (CAT). In this approach, each candidate is presented with an initial small set of items. Depending on the responses, his/her level of ability is estimated, and the next item is selected to provide the best additional information as to the candidate's ability and so on. In theory – and in practice – such an approach reduces the SEM for most if not all students. Several methods can be used to determine when to stop and end the test session for a candidate. One would be to administer a fixed number of items to all candidates. In this case, the SEM will vary between candidates but most probably be lower for most of the candidates then with an equal number of items with traditional approaches (CTT and G-theory). Another solution is to stop when a certain level of certainty (a certain SEM) is reached. In this case, the number of items will vary per candidate. But apart from CAT, IRT will mostly be used for test equating, in such situations where different groups of candidates have to be presented with equivalent tests.

Recommendations. The three theories – CTT, G-theory and IRT seem to co-exist. This is an indication that there is good use for each of them depending on the specific test, the purpose of the assessment and the context in which the assessment takes place. Some rules of thumb may be useful.

- CTT is helpful in straightforward assessment situations such as the standard open-ended or multiple choice test. In CTT, item parameters such as *p*-values and discrimination indices can be calculated quite simply with most standard statistical software packages. The interpretation of these item parameters is not difficult and can be taught easily. Reliability estimates, such as Cronbach's alpha, however, are based on the notion of test–ret7est correlation. Therefore, they are most suitable for reliability estimates from a norm-orientated perspective and not from a domain-orientated perspective. If they are used in the latter case, they will be an overestimation of the actual reproducibility.
- G-theory is more flexible in that it enables the researcher to include or exclude source of variance in the calculations. This presupposes that the researcher has a good understanding of the meaning of the various sources of variance and the way they interact with each other (nested *versus* crossed), but also how they represent the domain. The original software for these analyses is quite user unfriendly and requires at least some knowledge of older programming languages such as Fortran (e.g. UrGENOVA; http://www.education.uiowa.edu/casma/GenovaPrograms.htm, last access 17 December 2010). Variance component estimates can be done with SPSS, but the actual g-analysis would still have to be done by hand. Some years ago, two researchers at McMaster wrote a graphical shell around UrGenova to make it more user friendly (http://fhsperd.mcmaster.ca/g_string/download.html, accessed 17 December 2010). Using this shell prevents the user from knowing and employing a difficult syntax. Nevertheless, it still requires a

good understanding of the concept of G-theory. In all cases where there is more than one facet of generalisation (as in the example with the two examiners per station in an OSCE), G-theory has a clear advantage over CTT. In CTT multiple parameters should be used and somehow combined (in this OSCE Cronbach's alpha and Cohen's Kappa or an ICC for inter-observer agreement), in the generalisability analysis both facets are incorporated. If a one-facet situation exists (like the multiple choice examination) from a domain-orientated perspective (e.g. with an absolute pass–fail core), a dependability coefficient is a better estimate than CTT.

- IRT should only be used if people with sufficient understanding of the statistics and the underlying concepts are part of the team. Furthermore, considerably large item banks are needed and pre-testing on a sufficient number of candidates must be possible. This limits the routine applicability of IRT in all situations other than large testing bodies, large schools or collaboratives.

Emerging theories

Although we by no means possess a crystal ball, we see some new theories or extension to existing theories emerging. Most of these are related to the changing views from (exclusively) assessment *of* learning to more assessment *for* learning. Although this in itself is not a theory change but more a change of views on assessment, it does lead to the incorporation of new theories or extensions to existing ones.

First, however, it might be helpful to explain what assessment *for* learning entails. For decades, our thinking about assessment has been dominated by the view that assessment's main purpose is to determine whether a student has successfully completed a course or a study. This is epitomised in the summative end-of course examination. The consequences of such examinations were clear; if she/he passes, the student goes on and does not have to look back; if she/he fails, on the other hand, the test has to be repeated or (parts of) the course has to be repeated. Successful completion of a study was basically a string of passing individual tests. We draw – deliberately – somewhat of a caricature, but in many cases, this is the back bone of an assessment programme. Such an approach is not uncommon and is used at many educational institutes in the world, yet there is a growing dissatisfaction in the educational context. Some discrepancies and inconsistencies are felt to be increasingly incompatible with learning environments. These are probably best illustrated with an analogy. Purely selective tests are comparable in medicine to screening procedures (e.g. for breast cancer or cervical cancer). They are highly valuable in ensuring that candidates lacking the necessary competence do not graduate (yet), but they do not provide information as to how an incompetent candidate can become a competent one, or how each student can achieve to become the best possible doctor she/he could be. Just as screening does not make the patients better, but tailored diagnostic and therapeutic intervention do, assessment of learning does not help much in improving the learning but assessment for learning can.

We will mention the most striking discrepancies between assessment of and assessment for learning.

- A central purpose of the educational curriculum is to ensure that students study well and learn as much as they can; so, assessment should be better aligned with this purpose. Assessment programmes that focus almost exclusively on the selection between the sufficiently and insufficiently competent students do not reach their full potential in steering student learning behaviour.
- If the principle of assessment *of* learning is exclusively used, the question all test results need to answer is: *is John better than Jill?*, where the pass–fail score is more or less one of the possible 'Jills'. Typically CTT and G-theory cannot calculate test reliability if there are no differences between students. A test–retest correlation does not exist if there is no variance in scores, generalisability cannot be calculated if there is no person variance. The central question in the views of assessment *for* learning is therefore: *Is John today optimally better than he was yesterday, and is Jill today optimally better than she was yesterday*. This gives also more meaning to the desire to strive for excellence, because now excellence is defined individually rather than on the group level (if everybody in the group is excellent, 'excellent' becomes mediocre again). It goes without saying that in assessment *for* learning, the question whether John's and Jill's progress is good enough needs to be addressed as well.
- A difficult and more philosophical result of the previous point is that the idea of generalisation or prediction (how well will John perform in the future based on the test results of today) in an assessment *of* learning is mainly based on uniformity. It states that we can generalise and predict well enough if all students sit the same examinations under the same circumstances. In the assessment *for* learning, view prediction is still important but the choice of assessment is more diagnostic in that there should be room for sufficient flexibility to choose the assessment according to the specific characteristics of the student. This is analogous to the idea of (computer) adaptive testing or the diagnostic thinking of the clinician, tailoring the specific additional diagnostics to the specific patient.
- In the assessment *of* learning view, developments are focussed more on the development (or discovery) of the optimal instrument for each aspect of medical competence. The typical example of this is the OSCE for skills. In this view, an optimal assessment programme would incorporate only *the* best instrument for each aspect of medical competence. Typically, such a programme would look like this: multiple-choice tests for knowledge, OSCEs for skills, long simulations for problem-solving ability, etc. From an assessment *for* learning, view information needs to be extracted from various instruments and assessment moments to optimally answer the following three questions:

(1) Do I have enough information to draw the complete picture of this particular student or do I need specific additional information? (the 'diagnostic' question)

(2) Which educational intervention is most indicated for this student at this moment? (the 'therapeutic' question)
(3) Is this student on the right track to become a competent professional on time? (the 'prognostic' question).

- It follows logically from the previous point that this cannot be accomplished with one single assessment method or even with only a few. A programme of assessment is needed instead, incorporating a plethora of methods, each with its own strengths and weaknesses, much like the diagnostic armamentarium of a clinician. These can be qualitative or quantitative, more 'objective' or more 'subjective'. To draw the clinical analogy further: if a clinician orders an haemoglobin level of a patient she/he does not want the laboratory analyst's opinion but the mere 'objective' numerical value. If, on the other hand, she/he asks a pathologist, s/he does not expect a number but a narrative ('subjective') judgement. Similarly, such a programme of assessment will consist of both qualitative and quantitative elements.

Much of the theory to support the approach of assessment *for* learning still needs to be developed. Parts can be adapted from theories in other fields; parts need to be developed within the field of health professions assessment research. We will briefly touch on some of these.

- What determines the quality of assessment programmes? It is one thing to state that in a good assessment programme the total is more than the sum of its constituent parts, but it is another to define how these parts have to be combined in order to achieve this. Emerging theories describe a basis for the definition of quality. Some adopt a more ideological approach (Baartman 2008) and some a more utilistic 'fitness-for-purpose' view (Dijkstra et al. 2009). In the former, quality is defined as the extent to which the programme is in line with an ideal (much like formerly quality of an educational programme was defined in terms of whether it was PBL or not); in the latter the quality is defined in terms of a clear definition of the goals of the programme and whether all parts of the programmes optimally contribute to the achievement of this goal. This approach is more flexible in that it would allow for an evaluation of the quality of assessment *of* learning programmes as well. At this moment, theories about the quality of assessment programmes are being developed and researched (Dijkstra et al. 2009, submitted 2011).
- How does assessment influence learning? Although there seems to be complete consensus about this – a complete shared opinion, much empirical research has not been performed in this area. For example, much of the intuitive ideas and uses of this notion are strongly behaviouristic in nature and do not incorporate motivational theories very well. The research, especially in the health professions education, is either focussed on the test format (Hakstian 1971; Newble et al. 1982; Frederiksen 1984) or on the opinions of students (Stalenhoef-Halling et al. 1990; Scouller 1998). Currently, new theories are emerging incorporating motivational theories and describing better which factors of an assessment programme influence learning behaviour, how they do that and what the possible consequences of these influences are (Cilliers et al. submitted 2010, 2010).
- The phenomenon of test-enhanced learning has been discussed recently (Larsen et al. 2008). From expertise theories it is logical to assume that from sitting a test, as a strong motivator to remember what was learned, the existing knowledge is not only more firmly stored in memory, but also reorganised from having to produce and apply it in a different context. This would logically lead to better storage, retention and more flexible retrieval. Yet we know little about how to use this effect in a programme of assessment especially with the goal of assessment for learning.
- What makes feedback work? There are indications that the provision of feedback in conjunction with a summative decision limits its value, but there is little known about which factors contribute to this. Currently, research not only focusses on the written combination of summative decisions and formative feedback, but also on the combination of a summative and formative role within one person. This research is greatly needed as in many assessment programmes it is neither always possible nor desirable to separate teacher and assessor role.
- In a programme of assessment the use of human judgement is indispensible. Not only in the judgement of more elusive aspects of medical competence, such as professionalism, reflection, etc., but also because there are many situations in which a prolonged one-on-one teacher-student relationship exists, as is for example the case in long integrated placements or clerkships. From psychology it is long known that human judgement is fallible if it is compared to actuarial methods (Dawes et al. 1989). There are many biases that influence the accuracy of the judgement. The most well-known are primacy, recency and halo effects (for a more complete overview, cf. Plous 1993). A primacy effect indicates that the first impression (e.g. in an oral examination) often dominates the final judgement unduly; a recency effect indicates the opposite, namely that the last impressions determine largely the judgement. There is good indication that the length of the period between the observation and the making of judgement determines whether the primacy or the recency effect is most prominent effect. The halo effect pertains to the inability of people to judge different aspects of someone's performance and demeanour fully independently during one observation, so they all influence each other. Other important sources of bias are cognitive dissonance, fundamental attribution error, ignoring base rates, confirmation bias. All have their specific influences on the quality of the judgement. As such, these theories shed a depressing light on the use of human judgement in (high-stakes) assessment. Yet, from these theories and the studies in this field, there are also good strategies to mitigate such biases. Another theoretical pathway which is useful is the one on naturalistic decision making (Klein 2008; Marewski et al. 2009). This line of research does not focus on why people are so poor judges when compared to clear-cut and number-based decisions, but why people still do such a

good job when faced with ill-defined problems with insufficient information and often under less than ideal situations. Storage of experiences, learning form experiences and the possession of situation-specific scripts seem to play a pivotal role here, enabling the human to employ a sort of expertise-type problem solving. Much is based on quick pattern recognition and matching. Both theoretical pathways have commonality in that they both describe human approaches that are based on a limited representation of the actual observation. When, as an example, a primacy effect occurs, the judge is in fact reducing information to be able to handle it better, but when the judge uses a script, she/he is also reducing the cognitive load by a simplified model of the observation. Current research increasingly shows parallels between what is known about medical expertise, clinical reasoning and diagnostic performance and the act of judging a student's performance in an assessment setting. The parallels are such that they most probably have important consequences for our practices of teacher training.

- An important underlying theory to explain the previous point is cognitive load theory (CLT) (Van Merrienboer & Sweller 2005, 2010). CLT starts from the notion that the human working memory is limited in that it can only hold a low number of elements (typically 7 ± 2) for a short-period of time. Much of this we already discussed in the paragraphs on expertise. CLT builds on this as it postulates that cognitive load consists of three parts: intrinsic, extraneous and germane load. Intrinsic load is generated by the innate complexity of the task. This has to do with the number of elements that need to be manipulated and the possible combinations (element interactivity). Extraneous load relates to all information that needs to be processed yet is not directly relevant for the task. If, for example, we would start the medical curriculum by placing the learners in an authentic health care setting and require them to learn from solving real patient problems, CLT states that this is not a good idea. The authenticity may seem helpful, but it distracts, the cognitive resources needed to deal with all the practical aspects would constitute a high extraneous load even to such an extent that it would minimise the resources left for learning (the germane load).
- Finally, new psychometric models are developed and old ones are being rediscovered at this present time. It is clear that, from a programme of assessment view, in incorporating many instruments in the programme not one single psychometric model will be useful for all elements of the programme. In the 1960s and 1970s, some work was done on domain-orientated reliability approaches (Popham and Husek 1969; Berk 1980). In the currently widely used method internal consistency (like Cronbach's alpha) is often used as the best proxy for reliability or universe generalisation, but one can wonder whether this is the best approach to all situations. Most standard psychometric approaches do not handle a changing object of measurement very well. By this we mean that the students – hopefully – change under the influence of the learning programme. In the situation of a longer placement for example, the results of repeatedly scored observations (for instance, repeated mini-CEX) will differ in their outcomes, with part of this variance being due to the learning of the student and part to measurement error (Prescott-Clements et al. submitted 2010). Current approaches do not provide easy strategies to distinguish between both effects. Where internal consistency is a good approach to reliability, then stability of the object of measurement and of the construct can be reasonably expected; it is problematic when this is not the case. The domain-orientated approaches therefore were not focussed primarily on the internal consistency but on the probability that a new observation would shed new and unique light on the situation, much like the clinical adage never to ask for additional diagnostics if the results are unlikely to change the diagnosis and/or the management of the disease. As said above, these methods are being rediscovered and new ones are being developed, not to replace the existing theories, but rather to complement them.

Epilogue

In this AMEE Guide, we have tried to describe currently used theories in assessment. We chose to spend the larger part of this Guide on expertise development and on psychometric theories. These are well established theories at the moment, their importance is clear and are of increasing relevance to health sciences education.

What we have tried to advocate also is that these theories are necessary but not sufficient, medical education is neither cognitive psychology nor only psychometrics. There is a need to build our own theories of assessment, to cater better for our specific educational needs and lacunae. It is with this in mind that we have included our views on emerging theories and fields in which new theories are needed. We do realise that this is our view and that it is highly individual. Therefore we hope that the future will not prove us wrong on our predictions. What we do hope, however, is that this Guide will be completely outdated in 5 years, because this would mean that the scientific discipline of medical education and assessment has evolved rapidly in a direction so desperately needed. It will also be an indication that our scientific discipline has started to build and test theories itself. For a relatively young and rapidly evolving scientific field, this is a sheer necessity. We truly hope that this AMEE Guide then has made a contribution to this effect.

***Declaration of interest*:** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

CEES VAN DER VLEUTEN, PhD, is a Psychologist and a Professor of Medical Education and Chairman of the Department of Educational Development and Research at Maastricht University in the Netherlands.

LAMBERT SCHUWIRTH, MD PhD, is strategic professor for medical education at the Flinders Innovation in Clinical Education, Flinders University.

Notes

1. From: George Santayana (1905) Reason in Common Sense, volume 1 of The Life of Reason.

2. Which he explained first in *Logik der Forschung*. Julius Springer Verlag, Vienna, 1935 and later in The Logic of Scientific Discovery. Hutchinson, London, 1959.

3. Of course, this is not the only assumption that is needed for the application of CTT, another important assumption is that of local independence of individual observations, i.e. that all data points are independent of each other except for the construct the test aims to measure. An extensive discussion of the theoretical assumptions for each of these theories falls outside the scope of this AMEE Guide. Also understanding the assumptions mentioned in this AMEE Guide suffices for almost all normal everyday test situations.

4. It may seem a bit enigmatic how these conclusions are drawn but one has to bear in mind that the SDs are 5%. In a normal distribution, roughly 68% of the observations is located between the mean minus 1 SD and the mean plus 1 SD. From this, it is logical to infer that in test 1 more observations will fall into the 95-CI area than in test 2. This is an example based on a somewhat normally shaped symmetrical distribution, needless to say that if the distribution is more extremely skewed towards more high scores, the influence of the reliability on the reproducibility of pass–fails decisions is even less.

5. In fact, this formula does not describe a generalisability coefficient but a dependability coefficient. We have used this formula because it is more intuitive and therefore more helpful in understanding G-theory. We will explain the difference between a generalisability and dependability coefficient later on in this section.

Recommended reading ([a] 'must read')

General

Dent and Harden (2009)

On the value of theories

[a]Bordage G. 2009. Conceptual frameworks to illuminate and magnify. Med Educ 43: 312–319.

On learning and expertise theories

Ericsson KA, Charness N. 1994. Expert performance. Am Psychol 49: 725–747.

[a]Eva (2004)

[a]Regehr and Norman (1996)

[a]Schmidt and Boshuizen (1982)

On psychometric theories

[a]Berk (1980)

[a]Crocker L, Algina J. 1986. Introduction to classical and modern test theory. Fort Worth Holt: Rinehart and Winston, Inc.

Cronbach L, Shavelson RJ. 2004. My current thoughts on coefficient alpha and successor procedures. Educ Psychol Measur 64: 391–418.

[a]Kane (2006)

[a]Swanson et al. (1987)

On assessment for learning

[a]Shepard (2009)

References

Baartman LK. 2008. Assessing the assessment: Development and use of quality criteria for competence assessment programmes. Universiteit Utrecht, Utrecht.

Berk RA. 1980. A consumers' guide to criterion-referenced test reliability. J Educ Meas 17:323–349.

Berner ES, Hamilton LA, Best WR. 1974. A new approach to evaluating problem-solving in medical students. J Med Educ 49:666–672.

Bordage G. 1987.An alternative approach to PMP's: The 'key-features' concept. In: Hart IR, Harden R, editors. Further developments in assessing clinical competence, proceedings of the second Ottawa conference. Montreal: Can-Heal Publications Inc. pp 59–75.

Case SM, Swanson DB. 1993. Extended-matching items: A practical alternative to free response questions. Teach Learn Med 5:107–115.

Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. 2000. The script concordance test: A tool to assess the reflective clinician. Teach Learn Med 12:185–91.

Chase WG, Simon HA. 1973. Perception in chess. Cognit Psychol 4:55–81.

Chi MTH, Glaser R, Rees E. 1982. Expertise in problem solving. In: Sternberg RJ, editor. Advances in the psychology of human intelligence. Vol. 1. Hillsdale NJ: Lawrence Erlbaum Associates. pp 1–75.

Cilliers FJ, Schuwirth LWT, Adendorff HJ, Herman N, van der Vleuten CPM. 2010. The mechanisms of impact of summative assessment on medical students' learning. Adv Health Sci Educ 15:695–715.

Cilliers FJ, Schuwirth LWT, Herman N, Adendorff HJ, van der Vleuten CPM. (2011). A model of the sources, consequences and mechanism of impact of summative assessment on how students learn. DOI: 10.1007/s10459-011-9292-5.

Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. Psychol Bull 52:281–302.

Dawes RM, Faust D, Meehl PE. 1989. Clinical versus actuarial judgment. Science 243:1668–1674.

De Groot AD. 1978. Thought and choice in chess. The Hague, The Netherlands: Mouton publishers.

Dent JA, Harden RM. (eds.) 2009. A practical guide for medical teachers. Edinburgh: Churchill Livingstone Elsevier.

Dijkstra J, Galbraith R, Hodges B, Mcavoy P, Mccrorie P, Southgate L, van der Vleuten CPM, VDV, Wass V, Schuwirth LWT. 2011. Development and validation of guidelines for designing programmes of assessment: A modified Delphi-study.

Dijkstra J, van der Vleuten CPM, Schuwirth LWT. 2009. A new framework for designing programmes of assessment. Adv Health Sci Educ 15:379–393.

Elstein AS, Shulmann LS, Sprafka SA. 1978. Medical problem-solving: An analysis of clinical reasoning. Cambridge, MA: Harvard University Press.

Eva KW. 2004. What every teacher needs to know about clinical reasoning. Med Educ 39:98–106.

Eva KW, Neville AJ, Norman GR. 1998. Exploring the etiology of content specificity: Factors influencing analogic transfer and problem solving. Acad Med 73:S1–S5.

Frederiksen N. 1984. The real test bias: Influences of testing on teaching and learning. Am Psychol 39:193–202.

Hakstian RA. 1971. The effects of type of examination anticipated on test preparation and performance. J Educ Res 64:319–324.

Kane MT. 2001. Current concerns in validity theory. J Educ Meas 38:319–342.

Kane MT. 2006. Validation. In: Brennan RL, editor. Educational measurement. Vol. 4. Westport, CT: ACE/Praeger. pp 17–64.

Klein G. 2008. Naturalistic decision making. Hum Factors 50:456–460.

Larsen DP, Butler AC, Roediger HL. 2008. Test-enhanced learning in medical education. Med Educ 42:959–966.

Marewski JN, Gaissmaier W, Gigerenzer G. 2009. Good judgements do not require complex cognition. Cogn Process 11:103–121.

Newble D, Hoare J, Baxter A. 1982. Patient management problems, issues of validity. Med Educ 16:137–142.

Norman GR. 1988. Problem-solving skills, solving problems and problem-based learning. Med Educ 22:270–286.

Page G, Bordage G. 1995. The medical council of Canada's key features project: A more valid written examination of clinical decision-making skills. Acad Med 70:104–110.

Plous S. 1993. The psychology of judgment and decision making. Englewood Cliffs, NJ: McGraw-Hill Inc.

Polsen P, Jeffries R. 1982. Expertise in problem solving. In: Sternberg RJ, editor. Advances in the psychology of human intelligence. Hillsdale NJ: Lawrence Erlbaum Associates. pp 7–75.

Popham WJ, Husek TR. 1969. Implications of criterion-referenced measurement. J Educ Meas 6:1–9.

Prescott-Clements LE, van der Vleuten CPM, Schuwirth LWT, Rennie JS. (submitted 2010). Investigating the reliability of observed workplace-based assessment in vivo.

Regehr G, Norman GR. 1996. Issues in cognitive psychology: Implications for professional education. Acad Med 71:988–1001.

Schmidt HG. 1993. Foundations of problem-based learning: Some explanatory notes. Med Educ 27:422–432.

Schmidt HG, Boshuizen HP. 1993. On acquiring expertise in medicine. Special issue: European educational psychology. Educ Psychol Rev 5:205–221.

Schmidt HG, Boshuizen HPA, Hobus PPM. Transitory stages in the development of medical expertise: The 'intermediate effect' in clinical case representation studies. In: Proceedings of the 10th annual conference of the cognitive science society, 1988 August 17–19, Montreal, Canada: Lawrence Erlbaum Associates.

Schmidt HG, Norman GR, Boshuizen HPA. 1990. A cognitive perspective on medical expertise: Theory and implications. Acad Med 65:611–622.

Scouller K. 1998. The influence of assessment method on students' learning approaches; multiple choice question examination versus assignment essay. High Educ 35:452–472.

Shepard L. 2009. The role of assessment in a learning culture. Educ Res 29:4–14.

Stalenhoef-Halling BF, van der Vleuten CPM, Jaspers TAM, Fiolet JBFM. 1990. A new approach to assessing clinical problem-solving skills by written examination: Conceptual basis and initial pilot test results. In: Bender W, Hiemstra RJ, Scherpbier A, Zwierstra RJ, editors. Teaching and assessing clinical competence, proceedings of the fourth Ottawa conference. Groningen, The Netherlands: Boekwerk Publications. pp 552–557.

Swanson DB, Norcini JJ, Grosso LJ. 1987. Assessment of clinical competence: Written and computer-based simulations. Assess Eval High Educ 12:220–246.

Van Merrienboer J, Sweller J. 2005. Cognitive load theory and complex learning: Recent developments and future directions. Educ Psychol Rev 17:147–177.

Van Merrienboer JJ, Sweller J. 2010. Cognitive load theory in health professional education: Design principles and strategies. Med Educ 44:85–93.