Programmatic assessment and Kane's validity perspective

Lambert W T Schuwirth[1,2] & Cees P M van der Vleuten[2]

CONTEXT Programmatic assessment is a notion that implies that the strength of the assessment process results from a careful combination of various assessment instruments. Accordingly, no single instrument is superior to another, but each has its own strengths, weaknesses and purpose in a programme. Yet, in terms of psychometric methods, a one-size-fits-all approach is often used. Kane's views on validity as represented by a series of arguments provide a useful framework from which to highlight the value of different widely used approaches to improve the quality and validity of assessment procedures.

METHODS In this paper we discuss four inferences which form part of Kane's validity theory: from observations to scores; from scores to universe scores; from universe scores to target domain, and from target domain to construct. For each of these inferences, we provide examples and descriptions of approaches and arguments that may help to support the validity inference.

CONCLUSIONS As well as standard psychometric methods, a programme of assessment makes use of various other arguments, such as: item review and quality control, structuring and examiner training; probabilistic methods, saturation approaches and judgement processes, and epidemiological methods, collation, triangulation and member-checking procedures. In an assessment programme each of these can be used.

Medical Education 2012: ***46:*** *38–48*
doi:10.1111/j.1365-2923.2011.04098.x

[1]Flinders Innovation in Clinical Education, Flinders University, South Australia, Australia
[2]Department of Educational Development and Research, University of Maastricht, Maastricht, the Netherlands

Correspondence: Professor Lambert W T Schuwirth, Flinders University HPE, Flinders Innovation in Clinical Education, PO Box 2100, Adelaide 5001, SA, Australia. Tel: 00 61 8 8204 7174; Fax: 00 61 8 8204 5675; E-mail: Lambert.schuwirth@flinders.edu.au

INTRODUCTION

The medical education assessment literature has long been dominated by studies that try to demonstrate the intrinsic superiority of one assessment instrument over all others on the assumption that such a 'holy grail' for each of the separate constructs that make up medical competence will exist. Typical examples of this discourse include the many studies that have attempted to prove the innate superiority of open-ended questions over multiple-choice questions in the assessment of medical problem-solving.[1–4] Increasingly, however, it has become clear that the *content* of an assessment plays a far more important role than its *format*.[2–5] More importantly, there is increasing awareness that it is highly improbable that such a holy grail exists and even less likely that it will be applicable across different contexts. Instead, the notion that the utility of each assessment method is always a compromise between various aspects of quality has gained ground. Van der Vleuten suggested five criteria on which such a compromise could be made: reliability; validity; educational impact; cost efficiency, and acceptability.[6] However, many others can be discerned.[7]

A further step was taken when it became more generally accepted that the quality of assessment should be evaluated at a higher level. Thus, rather than evaluating an assessment at the level of the individual assessment method, the quality of the assessment should be determined across methods.[8,9] Two outcomes of this view are important. Firstly, it makes us realise that in any situation a single instrument may not be perfect (in reality almost all instruments are less than perfect). Secondly, it implies that strength derives from a more flexible and tailor-made approach to building a programme. A combination of (near-) perfect instruments may result in a weaker programme than a carefully combined set of perhaps less perfect components. In other words, it is not only the quality of the building blocks that is relevant, but also the ways in which they are combined. For example, in the context of the development of competency domains (such as those defined by the US Accreditation Council for Graduate Medical Education [ACGME] or the Royal College of Physicians and Surgeons of Canada's CanMEDS domains[10,11]), the traditional approach would dictate an assessment programme in which one superior instrument would require to be developed for each of the competency domains. Such a programme would follow a one-instrument-to-one-competency-domain design. In a programmatic approach one instrument can inform both students and teachers on various competency domains and a competency domain is assessed using information from various sources. Thus, rather than a 1 : 1 relationship, a so-called $n : n$ relationship is obtained.[12]

If there is no single perfect instrument and each instrument is considered to have its advantages and disadvantages (or indications, side-effects and contra-indications), this leads to a necessary reappraisal of methods that had been dismissed because of lack of reliability or construct validity, such as the *viva*, the long case, the oral examination, and so forth. Whereas the value of an instrument was traditionally judged in a more or less dichotomous manner (as good versus bad), it is now reappraised in terms of its strengths and weaknesses or its added value as a building block in an assessment programme.

Thus far, this is a logical sequence of views and it has led to a much more flexible approach to assessment methods. However, we think that the next step to be made requires the adoption of an equally programmatic approach to methods to determine the measurement quality of the assessment programme and its parts.

A good starting point for this is Kane's validity framework.[13,14] In this framework, validity is treated as a series of inferences for each of which sufficient data, rationales and arguments must be provided. Each of these must contribute to the validity of conclusions drawn about a candidate on the basis of assessment results.

In this paper we will first describe and briefly explain Kane's validity approach; we will then map various quality procedures that may be used in an assessment programme to the different types of inferences in validity. We will confine ourselves to the validity of conclusions only. Predecessors of Kane, most importantly Messick,[7] have also stressed the importance of the consequences of the assessment and have seen them as essential to the validity of the assessment, but we will not discuss this inference here. We do not think it is irrelevant; on the contrary, we think this issue is so important that it deserves a separate discussion.

KANE ON VALIDITY

In essence, validity pertains to the question of whether the assessment in question actually captures

the aspect of competence or performance it purports to assess. Thus, in the case of medical education, the assessment programme aims to capture the entity 'medical competence'. Such an entity is a construct in that it is assumed to exist yet it cannot be observed directly. Therefore, it must be inferred from observed behaviour. Logically, inferences can only be made on the basis of theoretical assumptions about the nature of the construct we want to assess. The construct 'intelligence', for example, is assumed to be a construct associated with processing capacities in the human mind that include memory functions, flexibility in thinking, and so forth. It is also assumed to be relatively stable. Blood pressure (BP), by contrast, is assumed to vary during the day. If we were to measure someone's intelligence several times during a day we would expect to find roughly the same result, whereas if we were to find exactly the same BP readings on several occasions during a day we would probably doubt the validity of the measurement.

Although these examples are (too) simple, inferring validity from observation is far from easy. Kane states that inferring the validity of an assessment for a certain construct is an ongoing process of building and verifying (and trying to falsify) arguments.[13–15] His approach requires that inferences are made from observation to score, from score to universe score, from universe score to target domain, and from target domain to construct.

If, for example, we were to apply Kane's validity framework to an assessment of problem-solving ability as a measure of medical competence, we would involve the following inferences.

From observation to score

A typical approach in the assessment of problem-solving skills is to observe how students perform on various medical cases. A score must be derived from the 'raw' answers students give to questions. We could, of course, just count the number of relevant questions asked in history taking, the number of relevant physical examinations performed and the number of pertinent laboratory tests ordered, and add them all up to give a total score. However, theories on medical problem solving and expertise state that experts do not necessarily collect *more* information before they come to a conclusion, but that they collect information more *efficiently*.[16–18] In addition, there are individual differences between experts with respect to which information they collect (idiosyncrasy).[17–19] Therefore, scoring in the manner described above would not serve to properly translate observations to scores. In fact, the scoring of the 'patient management problem' (PMP), a type of patient simulation exercise, was based on how much relevant information a candidate collected, but it proved to represent an invalid inference from observation to score under the theoretical assumptions around medical expertise and problem solving.[19–22]

From observed score to universe score

It is generally known that one or two cases never provide evidence sufficient to support the drawing of general conclusions about a candidate's problem-solving expertise. Research in cognitive psychology has repeatedly demonstrated the phenomenon of the domain specificity of problem-solving expertise.[23,24] Therefore, small samples of long cases do not support inferences on general problem-solving ability. In assessment methods such as PMPs, in which each case takes a long time to complete, the number of cases completed per hour is simply too low to yield sufficiently generalisable scores, and thus any inference from observed scores to universe scores is not sufficiently defensible.

As a result, methods based on larger numbers of shorter cases, such as in the key-feature approach or in extended-matching items focused on medical decision making, have been designed.[25–27]

From universe score to target domain

Key-feature approaches with simple scoring schemes and good (broad but more superficial) sampling approaches lead to a good inference from observed score to universe score, but do they capture medical decision-making ability? A series of studies may be used to demonstrate that scores on such tests behave according to expectations. For example, studies have demonstrated that experienced experts outperform novices or intermediates (which they do NOT do in PMPs) and that assessment results on key-feature approach cases provide information about the candidate's ability which cannot be obtained otherwise. However, there is also the need for more judgemental evidence that the cases are authentic and that the questions ask for essential decisions and not for rote factual knowledge.[28,29] Thus, in order to support an inference from universe score to target domain (medical decision-making skills), information in support of the assumption that the decisions for which the questions ask are really essential or represent key-feature decisions must be collected.[28,30]

From target domain to construct

Finally, medical problem solving entails much more than simply making the right decisions on paper-based or computerised cases. In real life many other factors may play a role, such as ability to elicit information from a patient (e.g. communication ability), and ability to sift through information and distinguish relevant from non-relevant information. The extent to which the results on, for example, a key-feature examination contribute to the multifaceted construct of medical problem solving in real practice is an example of the last inference in Kane's validity argument. In other words, it is important to determine what the assessment of problem-solving ability, using key-feature approaches, adds to the construct of medical competence. What are the strengths or weaknesses of the method? How can the weaknesses be addressed or compensated for by using additional methods? What, for example, is the synergy between key-feature approach scores and scores on mini-clinical examinations (mini-CEXs), where the former is based on a high quantity of low-fidelity assessment and the latter on fewer samples of high-fidelity testing?

There are many other examples of assessment that can serve to illustrate Kane's approach. Clauser *et al.*[31] provide a useful example with respect to the assessment of professionalism.

Although Kane's framework was developed in the context of educational assessment, it could be applied to all kinds of measurements of things that cannot be observed directly. We describe an example of the four inferences on the construct of BP.

A medical example of Kane's validity perspective

In medicine, BP is a good example of a construct that cannot be observed directly. Blood pressure is normally taken to aid in the evaluation of a patient's health. This evaluation requires that several inferences are made.

From observation to score

When taking a patient's BP, the doctor must convert acoustic (Korotkow sounds) signals and a visual reading of the sphygmomanometer to a numerical value. The inferences are based on the assumption that the doctor knows when to take the reading, does not let the sphygmomanometer run down too quickly or too slowly, and uses the right cuff, and so forth. Only when every aspect of the procedure is performed correctly can a valid inference from observation to score be made.

From observed score to universe score

The next inference refers to whether the observations are sufficiently representative of all possible observations. In our example, this refers to whether one measurement provides sufficient data on which to base a diagnosis. The Dutch guideline, for example, stipulates that hypertension can only be diagnosed if BP is taken twice during one consultation and is repeated during a second consultation.[32]

From universe score to target domain

Now the results of the BP measurements are used to draw conclusions about the cardiovascular status of the patient. This requires heart auscultation, pulse palpation and other results to be incorporated and the results triangulated in order for the conclusions to be valid.

From target domain to construct

The patient's cardiovascular status can now be used to establish his or her health status, but further information must be obtained from other sources and triangulated to support a more general conclusion.

MAKING INFERENCES

In Kane's view, inferences are based on arguments. These may be quantitative or qualitative, but they must always be theory-based and interpretive and thus cannot serve as arguments in isolation. Of course, not just any argument will do. Arguments in the validation process must be clear, specific, coherent, complete, plausible, verifiable and falsifiable.[14]

Arguments are required to be clear in order to ensure that every stakeholder or researcher is able to follow their logic. Therefore, the argument must include sufficient specific details. Coherence requires that the network of related inferences is such that the final conclusions and decisions follow plausibly from the observed performance. This requires the argument to be complete. The plausibility of the argument may often be self-evident, but some arguments will rely on empirical underpinning (preferably by not only verification, but also by multiple failed attempts at falsification) and others will rely on careful documentation and scrutiny of procedures. This may

involve the employment of not only deductive reasoning or inductive inferences, but also of other forms of defeasible reasoning, such as probabilistic reasoning. Defeasible arguments are arguments that contain a presupposition but accept that this may be overthrown if counterarguments are strong. Probability-based arguments are defeasible, whereas those based on sheer deductive logic are not. Although this may give the impression that the assessment developer or researcher has great latitude to use whatever arguments he or she needs, this is not the case: every argument must be carefully chosen in a strategic and programmatic way to ensure that it provides the optimal evidence for validity.

PROGRAMMATIC ASSESSMENT AND INFERENCES

A programme of assessment will use various assessment components (instruments). We believe that the quality of each of these cannot be determined using the same (psychometric) approaches. Instead, we think that a variety of methods and procedures should be used in assessment depending on the specific component of the programme. The choice of each of these must be based on a clear notion of the nature of the construct the component and the assessment programme are trying to capture. In this section, we will discuss most of the currently used quality approaches and measures using Kane's validity theory as an overarching structure. The methods described here do not represent different methods by which to determine the validity of a programme as a whole, but, rather, different methods by which to determine the validity of its various building blocks. We also do not pretend to include an exhaustive list of methods or to present all the procedures ever employed to ensure the validity of assessment. Rather, we will take some of the most widely used methods and put them in perspective. We do not wish to illustrate that any of these methods are either good or bad in themselves. On the contrary, the value and usefulness of a particular method can only be derived from the support it lends to an inference and thus to the validity of the assessment *for a certain construct.*

Inference 1. From observation to score

The most widely used methods to support the inference from observation to score are item construction rules, structuring, scoring rules and 'granularity', item analyses, relevancy evaluations, information selection procedures, reporting methods and feasibility issues.

Item construction rules

These are designed to optimise the probability that a student who has mastered the subject matter will answer the item correctly and those without sufficient mastery will answer incorrectly.[33,34] In other words, they serve to minimise the chance that a student will give a false negative or false positive response. If, for example, a student answers a multiple-choice question correctly because he or she has chosen the longest option, or answers an open-ended question correctly because he or she has successfully applied a so-called blunderbuss technique, the scores this student obtains are invalid as they are based on 'test-wiseness' and not on subject matter mastery. In a programme of assessment this is especially clear in the written or computer-based elements of the programme, but it also pertains to test-taking strategies in oral examinations (e.g. find out the hobby horses of the examiner and capitalise on them).

Structuring of the assessment

If the construct of interest is uniformity, such as in the assessment of advanced trauma life support (ATLS) procedural skills, in which all candidates should respond similarly to the tasks demanded by the assessment, structuring the assessment improves the conversion of observations to scores. However, if the quality of the interaction between the candidate and the problem at hand (e.g. as in workplace-based assessment) is an issue, structuring does not work well. It trivialises the assessment in the perception of the users and validity theory explains why.

This is exemplified by early objective structured clinical examinations (OSCEs), which were highly structured. Many examiners complained that adding up all the scores on the individual items did not really indicate ability in the competence the OSCE was intended to assess. Depending on the specific definition of the construct, structuring the assessment may strengthen the validity argument in some cases, but weaken it in others.

Scoring rules

Of course, the determination of scoring rules plays an important role. The extensive debates about whether or not to apply a penalty for guessing are a good example of this.[35] If we summarise the conclusions, we find they converge on the notion that the purpose of the assessment, the intended construct, is an essential decision. If one tries to capture the student's

knowledge as a construct in his or her head, willingness to guess is a source of error in the inference from observation to score. If, by contrast, one is interested in assessing which knowledge the student is willing to actually use, willingness to make an educated guess may well be seen as a source of construct-relevant variance. Further, more complicated scoring methods are not inherently better than simple 1-0 approaches[19] because although they generally do introduce more variance, this is seldom construct-relevant variance. This is an issue related to 'granularity'. Overly detailed scoring can increase the construct-irrelevant variance: a mark of 7.35 out of 10 for a thesis suggests an accuracy that is simply not there.

Item analyses

Item analyses can be used to improve the inference from observation to score because they can identify items that might have a negative influence on validity. However, this ability depends on the construct the assessment aims to test. If the construct is assumed to be homogeneous and stable, item analyses often lead to the elimination of items. This improves the measurement properties of the test by weeding out construct-irrelevant variance. If, however, the test is seen as a collection of intrinsically meaningful and relevant items (as in the case situations in ATLS training), item analysis results can only serve to flag up the need to carefully review an item and check whether it is actually as relevant, unambiguous and meaningful as it was thought to be on construction.

Relevancy evaluations

There is probably complete agreement that items or assessment parts need to be relevant, but how relevance is defined is again dependent on theoretical conceptions about the construct. Relevance can be defined as what most people know. In that case high p-values would constitute an argument for validity. If relevance is defined as what competent people *need* to know, low p-values in conjunction with high item–total correlations (R_{it}) would be a better argument for validity. By contrast, if relevance is defined as what *all* people *should* know, p-values and R_{it} are not useful parameters for relevancy. In this case, qualitative arguments for the relevancy of an item need to be made: for example, if a student doesn't understand the biofeedback mechanism of the thyroid gland and its hormones, he or she will not interpret laboratory results well.

Reporting and summarising

Reporting and summarising in oral assessments, portfolios and workplace-based assessments is one way of converting information to 'scores'. Whereas means, standard deviations and so forth represent standard ways of converting large amounts of data into scores in quantitative methods, in qualitative assessment an expert summary plays this role. In quantitative assessment methods the supporting evidence is based on the application of the correct (statistical) descriptive techniques and correct calculations. In the qualitative context it is based on examiner expertise and its development (teacher training).

Feasibility of the instrument

The user must be fully comfortable with using the assessment instrument. How else can he or she correctly translate his or her observations into scores? If the user is unsure about how to score an observation using the instrument or where to score certain observations, the strength of the inference from observation to score is seriously limited. Another such situation exists if the instrument is so complicated to use (e.g. a 60-item OSCE form) that the observer's 'cognitive load' is occupied by finding out how to manage the instrument rather than by observing and judging the performance. A valid inference from observation to score can therefore only be made if the instrument is sufficiently user-friendly or the examiner has been carefully familiarised with the instrument through training.

In summary, for all inferences from observation to score, validity arguments are based on the quality procedures used to construct the measurement instrument, the expertise of the user and the interplay between both factors. We cannot stress enough that the strength of each argument is determined by the extent to which it supports the theoretical notions of the construct.

Inference 2. From scores to universe scores

This second inference is often referred to as 'reliability'; this notion is the basis for the adage that unreliable tests can never be valid. However, the relationship is more nuanced than that. The inference from observed scores to universe scores is based on the argument that the observed set of scores is sufficiently representative of the universe of all possible scores. An idea of the nature of this

'universe' is therefore indispensible. For example, test–retest correlations are only valid inferences for universe representation under the assumption that the universe – the target domain or the construct – is internally consistent or homogeneous. If the 'universe' is assumed to be heterogeneous, it will be neither logical nor plausible to find high test–retest correlations. In this case test–retest reliability would indicate poor rather than good universe generalisation.[36] In a programme of assessment, various methods for generalisation may be used for the various parts.

Classical test theory

Procedures based on classical test theory (CTT), such as Cronbach's alpha and Kuder–Richardson formulas, refer to the notion of a test–retest correlation. In fact, they determine the internal consistency of the test results. Of course a test–retest correlation is only a useful approach to universe generalisation if we assume that the universe itself is so homogeneous that two independently taken samples can be expected to lead to the same results. A consequence of this assumption is that all variation between observations is generally treated as construct-irrelevant variance.

Another assumption is that the object of measurement does not change during the observations. In the examples we used before, of BP and intelligence, we expect the former to change from moment to moment and the latter to remain stable. If we were to take repeated measurements of both during the day and were to find perfect agreement within subjects and systematic differences between subjects, we would regard this as an argument in favour of the validity of the intelligence test and against that of the BP measurement.

In cases in which unidimensionality or homogeneity are not part of the theory about the construct or where qualitative data are collected, CTT does not work well. For multidimensional theories about the construct, a stratified alpha can be used, but in practice it rarely gives different results from a standard alpha.

Generalisability theory

Generalisability theory (GT) is much more flexible. It requires the user to define exactly which elements of variance are to be seen as construct-relevant and which as construct-irrelevant. It still, however, starts from the notion that there is one universe score and this has implications. If, for example, a generalisability analysis is performed on the total scores of the stations on an OSCE, the underlying assumption is that the trait 'skills' is such that it is defensible to combine the scores on a resuscitation station with those on an abdominal examination station, and that both are interchangeable. Another example is the mini-CEX, where a generalisability analysis must make the automatic assumption that history-taking skills are completely interchangeable with humanistic qualities. On this assumption, someone who asks stupid questions but does so in a skilled communicative manner is as competent as someone who asks the right questions in an unpleasant manner. Teachers often argue that both are equally important and one should not be allowed to compensate for the other, whereas assessment experts and psychometricians often argue that one should. There is no saying who is right, but it is clear that there are hugely different views on the nature of the construct 'skills'.

Probabilistic approaches

Another issue concerns whether every situation requires the same amount of sampling. Does the candidate who has performed very poorly or extremely well on seven mini-CEX observations really require an eighth? Certainly the candidate who has performed excellently in four situations and very poorly in the remaining four requires more observations in order to achieve a good generalisation to the universe score. Probabilistic approaches, which are equivalent to those underlying positive and negative predictive values in epidemiology, cater more specifically to such differences. Let us take factual knowledge as an example. One theory may start from the assumption that knowledge is a construct of a single trait which implies that there will be a uniform increase in the probability that a candidate will give a correct answer with increasing ability (e.g. if a student has good knowledge about left-sided heart failure, it will be safe to assume that he or she knows about heart failure, about Frank–Starling mechanisms and about heart physiology). In such a theoretical context, item response theory (IRT) models are useful means of generalisation. In a situation in which the possession of knowledge is seen as an unrelated set of items (e.g. if a student knows that surfactant is produced by type II pneumocytes, this does not automatically mean that he or she knows what the origo and insertion of the pronator teres muscle are), IRT is less useful and other models, such as binomial models, may be more applicable.[37]

Saturation of information

Saturation of information approaches originate from qualitative research methodologies. If we assume or theorise the construct to be heterogeneous and non-dimensional, internal consistency measures are not the best way to generalise. Saturation of information basically means that new observations do not add important new information to that already obtained. This is comparable with the diagnostic adage that if additional diagnostics do not change the diagnosis or the therapeutic actions, they should not be ordered, but it does not stipulate that only one diagnosis can be made. If one wants to design an assessment component aimed at bedside manners, especially from an assessment-for-learning approach, converting all observations to a score and calculating the generalisability coefficient would not really do justice to the assessment of such a complex phenomenon. Making assumptions about whether a new observation would add anything to the kaleidoscope of information about how a candidate is doing is much more useful and information-rich.

Credibility

Although authority-based arguments are not in vogue at present, the issue of credibility does, of course, play a role in universe generalisation. Research in diagnostic expertise shows that experienced experts need less information to reach valid decisions about diagnosis and treatment. This can be easily translated to the assessment field. It is highly likely – and, in many contexts, normal – that an expert requires fewer observations than a novice assessor to make the inferences from observation to universe score. Studies into the nature of assessor expertise and the development of person and performance scripts support this notion.[38,39] Therefore, the observed-to-universe-score inference argument is stronger if the inference is made by an expert assessor than by a novice.

Sampling schemas

Sampling schemas – such as blueprinting – support the universe generalisation inference by arguments based on the domain to be sampled and the representativeness of the sample of observations (items, mini-CEXs, etc.) for the universe. They are crucial in the argument for all notions of the universe. Even if the universe is seen as homogeneous, sampling must be broad enough to average out all unwanted sources of variance. By contrast, if the universe is seen as heterogeneous, sampling must be such that all aspects of the universe are included in the sample.

In the validity argument, the issue does not concern the question of whether or not universe representation is needed, but, rather, that of how the most convincing argument of universe representation can be made. In some cases reproducibility and internal consistency represent a better argument; in others procedures that seek to add observations until no new information is acquired are more useful.

Inference 3. From universe score to target domain

At some point during the process the representative results must be combined in such a way that conclusions about the target domain can be drawn. In a programme of assessment this requires that the results of various instruments be combined. In our example about assessment of clinical reasoning, these might include the results of observations in practice, part of the results of an OSCE, results of key-feature approach tests, and so forth. This demands that decisions be made not only about what the standards are, but also on how to combine the results of various instruments (especially if they combine quantitative and qualitative information).

Standard setting

Standard setting is a heavily debated issue in assessment. This is logical because it concerns the optimal way to reduce much of the measurement information to arrive at a dichotomous yes/no decision about the target domain. Again, the type of inferences and the strength of the argument depend on the theoretical notion of the target domain. For modular target domains (e.g. ability to perform an examination of the knee), which should be mastered by a certain time-point, standard setting is typically used to define the minimally acceptable level of mastery. Longitudinal components (e.g. progress testing) assess characteristics that constantly improve during life. In these cases, relative or ipsative standards (relative to the phase of the training or a peer group or relative to the candidate's past performance) are more applicable.

Epidemiological or criterion-based approaches

In cases in which a numerical outcome can be defined as a criterion for the target domain, the arguments in this third inference can be based on positive and negative predictive values and odds ratios. In such cases, receiver operating characteristic

(ROC) curves can be used to support the inference argumentation. However, not only numerical approaches can benefit from epidemiological arguments. Some more theoretical epidemiological concepts are also useful. The idea of the positive predictive value gives us to understand that, logically, senior year classes should show dissimilar failure rates to more junior classes, simply because the ongoing selection process has decreased the *a priori* probability that an incompetent student will remain in the class. Thus, if a failure rate of 25% is considered acceptable in a first-year cohort, this should not mean that the same failure rate is acceptable in a final-year group. In designing an assessment programme for a curriculum, it would therefore be very defensible to focus on selection or on identifying unsuitable students during the more junior years of training and on detecting remediable problems and the optimal ways of remediating them during more senior years.

Of course, in the combination of many different (quantitative) elements, arguments can be based on statistical approaches such as multiple regression, correlation or multitrait-multimethod approaches.

Compensation, conjunction and collation

In order to arrive at a good inference about the target domain, separate assessment elements must be combined. However, randomly choosing a certain method of combining information does not provide a strong basis for argumentation.

Despite the robust finding that things generalise well across formats if the content is the same and vice versa,[2,3,5] we often combine elements because they are of the same format (e.g. OSCE stations on abdominal examination and knee examination) rather than because they have similar meaningful content. This is based on the implicit notion of skills as a unidimensional trait, rather than as a selected set of intrinsically relevant observed abilities. In the former, compensation is the best way of making an inference to the target domain; in the latter conjunction is. If information from various sources or assessment elements needs to be combined (e.g. an OSCE station on knee examination and the part of a written examination that focuses on knee anatomy), collation and triangulation are more suitable bases for argument. Here, human judgement and the expertise of the person doing and interpreting the triangulation form the basis for the completeness and plausibility of the inference (in much the same way as the expertise of the doctor is needed to make meaning of the combination of information on sodium level and a thirst complaint).[9,38,39]

Member checking

Member checking refers to all processes in an assessment programme that not only includes the views of various contributors to the assessment process (such as in a 360-degree approach), but also includes in-built steps designed to continually evaluate whether the intermediate and final conclusions with respect to the target domain accord with the views of these contributors and whether inferences made on the basis of these views are valid. As such, member checking supports the ownership of all actors of the final decisions and conclusion with respect to the target domain and thus to the plausibility of the inference.

For this inference, both quantitative and qualitative methods are available. Whenever purely quantitative results need to be combined, issues such as compensation and conjunction, and predictive values, are more convincing. Whenever qualitative results are used (either in isolation or in conjunction with quantitative results), human judgement plays a role, and thus the expertise (teacher training) and credibility of the people making the judgements are necessary elements of the arguments.

Inference 4. From target domain to construct

Basically the same methods and procedures used in the previous two inferences are used to make inferences from target domain to construct. For the construct of medical competence, especially in the light of the currently popular view of this construct as a set of competency domains, it is important to have a theoretical and practical notion of how these competency domains make up the final construct. Health as a final analogy is defined by the World Health Organization[40] as: '...a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.'[40] This is a useful theoretical construct, but it is useless in medical practice because it will almost never be attained in any real patient. In practice, health is more often used in the sense that both the patient and doctor are satisfied about the outcome of the process and have decided that further actions are neither needed nor wanted.

Competency domains, such as those defined in the CanMEDS and ACGME competencies, are useful in theory, but, for assessment purposes, they form a

construct that currently creates more problems than it solves. What information should be mapped onto which competency? How should we deal with information that maps onto different competencies? How should we manage different sources of information that map onto one competency? Can competencies compensate for one another or should they be treated conjunctively? These are all questions that must be addressed before we can make a valid inference from each target domain to the complete construct of medical competence.

Another issue of discussion may be even more central. Is medical competence the ability to act in a manner that accords with protocol in every situation or is it the ability to be act in a manner that is sufficiently flexible to allow for the optimal adaptation of diagnostic, communicative and therapeutic decisions to each situation? In the former, more structured approaches to inferences are more plausible. In the latter, more interpretative arguments must be made. We cannot stress enough that the arguments of validation can only be made if the construct we want to assess is defined clearly enough and when all theoretical notions about it are sufficiently concrete.

CONCLUSIONS

This overview is far from complete, but we must emphasise that it was not our intent to be complete in this paper. We think that the content of each heading alone could support a fully fledged paper or a chapter in a book. What we wanted to demonstrate is that the plethora of practices in the assessment of medical competence are all of good value in some situations and of poor value in others. A one-size-fits-all approach does not work (Cronbach's alpha on portfolio results is at best a less than convincing argument for generalisation).

This paper represents a follow-up of our previous publications on programmatic assessment, which pose as central the notion that the quality of the programme is built on the quality of the combination of its building blocks and not on the superiority of any one of them.[8] We have argued here that the validity of the assessment of medical competence – especially if it is based on a programme of assessment – is based on a programme of inferences, each of which must be coherent, but which must also contribute maximally to the forming of one consistent and coherent argumentation series.

Contributors: the content of this paper is the result of the many discussions between its authors and of their exploration of the literature. LWTS drafted the first version of the paper. CPMvdV commented on this draft and suggested revisions.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Newble DI, Baxter A, Elmslie RG. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Med Educ* 1979;**13**:263–8.
2 Norman G, Swanson D, Case S. Conceptual and methodological issues in studies comparing assessment formats, issues in comparing item formats. *Teach Learn Med* 1996;**8**:208–16.
3 Norman G, Tugwell P, Feightner J, Muzzin L, Jacoby L. Knowledge and clinical problem-solving. *Med Educ* 1985;**19**:344–56.
4 Schuwirth LWT, van der Vleuten CPM, Donkers HHLM. Open-ended questions versus multiple-choice questions. In: Harden R, Hart IR, Mulholland H, eds. *Approaches to the Assessment of Clinical Competence. Proceedings of the Fifth Ottawa Conference*. Norwich: Page Brothers 1992;486–91.
5 Maatsch J, Huang R. An evaluation of the construct validity of four alternative theories of clinical competence. In: *Proceedings of the 25th Annual Research in Medical Education Conference*. Chicago, IL: Association of American Medical Colleges 1986;69–74.
6 van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996;**1**:41–67.
7 Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educ Res* 1994;**23**:13–23.
8 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39**:309–17.
9 Van der Vleuten CPM, Schuwirth LWT, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol* 2010;**24**:703–19.
10 Accreditation Council for Graduate Medical Education. Common Program Requirements Document. http://www.acgme.org/outcome/comp/compCPRL.asp. [Accessed 15 August 2011.]
11 Royal College of Physicians and Surgeons of Canada. CanMeds 2005 framework. http://rcpsc.medical.org/publications/index.php#canmeds. [Accessed 15 August 2011.]

12 Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach* 2011;**33**:478–85.
13 Kane MT. Current concerns in validity theory. *J Educ Measure* 2001;**38**:319–42.
14 Kane MT. Validation. In: Brennan RL, ed. *Educational Measurement*. Westport, CT: ACE/Praeger 2006;7–64.
15 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.
16 Polsen P, Jeffries R. Expertise in problem solving. In: Sternberg RJ, ed. *Advances in the Psychology of Human Intelligence*. Hillsdale, NJ: Lawrence Erlbaum Associates 1982;367–411.
17 Chi MTH, Glaser R, Rees E. Expertise in problem solving. In: Sternberg RJ, ed. *Advances in the Psychology of Human Intelligence*. Hillsdale, NJ: Lawrence Erlbaum Associates 1982;7–76.
18 Schmidt HG, Boshuizen HP. On acquiring expertise in medicine. *Educ Psychol Rev* 1993;**5**:205–21.
19 Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assess Eval High Educ* 1987;**12**:220–46.
20 Schmidt HG, Boshuizen HP. On the origin of intermediate effects in clinical case recall. *Mem Cognit* 1993;**21**:338–51.
21 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*. Montreal, QC: Lawrence Erlbaum Associates 1988;139–45.
22 Schuwirth LWT, Gorter SL, van der Heijde D, Rethans JJ, Brauer J, Houben H, Van der Linden SJ, Van der Vleuten CPM, Scherpbier AJJA. The role of a computerised case-based testing procedure in practice performance assessment. *Adv Health Sci Educ Theory Pract* 2005;**10**:145–55.
23 Elstein AS, Shulmann LS, Sprafka SA. *Medical Problem-Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press 1978.
24 Eva K. On the generality of specificity. *Med Educ* 2003;**37**:587–8.
25 Page G, Bordage G. The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Acad Med* 1995;**70**:104–10.
26 Bordage G. An alternative approach to PMPs: the 'key-features' concept. In: Hart IR, Harden R, eds. *Further Developments in Assessing Clinical Competence. Proceedings of the Second Ottawa Conference*. Montreal, QC: Can-Heal Publications 1987;59–75.
27 Case SM, Swanson DB. Extended-matching items: a practical alternative to free response questions. *Teach Learn Med* 1993;**5**:107–15.
28 Schuwirth LWT, Blackmore DB, Mom E, van de Wildenberg F, Stoffers H, van der Vleuten CPM. How to write short cases for assessing problem-solving skills. *Med Teach* 1999;**21**:144–50.
29 Schuwirth LWT, Verheggen MM, van der Vleuten CPM, Boshuizen HPA, Dinant GJ. Do short cases elicit different thinking processes than factual knowledge questions do? *Med Educ* 2001;**35**:348–56.
30 Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad Med* 1995;**70**:194–201.
31 Clauser BE, Margolis MJ, Swanson DB. Issues of validity and reliability for assessments in medical education. In: Holboe ES, Hawkins RE, eds. *Evaluation of Clinical Competence*. Philadelphia, PA: Mosby Elsevier 2008;10–23.
32 Nederlands Huisartsen Genootschap. Richtlijn cardiovasculair management. http://nhg.artsennet.nl/kenniscentrum/k_richtlijnen/k_nhgstandaarden/Samenvattingskaartje-NHGStandaard/M84_svk.htm#N65743. [Accessed 15 August 2011.]
33 Downing SM, Haladyna TM. Test item development: validity evidence from quality assurance procedures. *Appl Meas Educ* 1997;**10**:61–82.
34 Case SM, Swanson DB. Item Writing Manual. http://www.nbme.org/publications/item-writing-manual.html. [Accessed 15 August 2011.]
35 Muijtjens AMM, van Mameren H, Hoogenboom RJI, Evers JLH, van der Vleuten C. The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Med Educ* 1999;**33**:267–75.
36 Schuwirth LWT, van der Vleuten CPM. A plea for new psychometrical models in educational assessment. *Med Educ* 2006;**40**:296–300.
37 Rickets C. A plea for the proper use of criterion-referenced tests in medical assessment. *Med Educ* 2009;**53**:1141–6.
38 Govaerts MJB, Schuwirth LWT, van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ* 2011;**16** (2): 151–65.
39 Govaerts MJB, van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ Theory Pract* 2007;**12**:239–60.
40 World Health Organization. *Preamble to the Constitution of the World Health Organization as Adopted by the International Health Conference*. Geneva: Official Records of the WHO 1946.

Received 14 March 2011; editorial comments to authors 14 April 2011; accepted for publication 19 July 2011