A contemporary approach to validity arguments: a practical guide to Kane's framework

David A Cook,[1,2] Ryan Brydges,[3,4] Shiphra Ginsburg[3,4] & Rose Hatala[5]

CONTEXT Assessment is central to medical education and the validation of assessments is vital to their use. Earlier validity frameworks suffer from a multiplicity of types of validity or failure to prioritise among sources of validity evidence. Kane's framework addresses both concerns by emphasising key inferences as the assessment progresses from a single observation to a final decision. Evidence evaluating these inferences is planned and presented as a validity argument.

OBJECTIVES We aim to offer a practical introduction to the key concepts of Kane's framework that educators will find accessible and applicable to a wide range of assessment tools and activities.

RESULTS All assessments are ultimately intended to facilitate a defensible decision about the person being assessed. Validation is the process of collecting and interpreting evidence to support that decision. Rigorous validation involves articulating the claims and assumptions associated with the proposed decision (the interpretation/use argument), empirically testing these assumptions, and organising evidence into a coherent validity argument. Kane identifies four inferences in the validity argument: *Scoring* (translating an observation into one or more scores); *Generalisation* (using the score[s] as a reflection of performance in a test setting); *Extrapolation* (using the score[s] as a reflection of real-world performance), and *Implications* (applying the score[s] to inform a decision or action). Evidence should be collected to support each of these inferences and should focus on the most questionable assumptions in the chain of inference. Key assumptions (and needed evidence) vary depending on the assessment's intended use or associated decision. Kane's framework applies to quantitative and qualitative assessments, and to individual tests and programmes of assessment.

CONCLUSIONS Validation focuses on evaluating the key claims, assumptions and inferences that link assessment scores with their intended interpretations and uses. The *Implications* and associated decisions are the most important inferences in the validity argument.

Medical Education 2015: 49: 560–575
doi: 10.1111/medu.12678

Discuss ideas arising from the article at www.meduedu.com 'discuss'

[1]Mayo Clinic Online Learning, Mayo Clinic College of Medicine, Rochester, Minnesota, USA
[2]Division of General Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA
[3]Department of Medicine, University of Toronto, Toronto, Ontario, Canada
[4]Wilson Centre, University Health Network, Toronto, Ontario, Canada
[5]Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence: David A Cook, Division of General Internal Medicine, Mayo Clinic College of Medicine, Mayo 17-W, 200 First Street SW, Rochester, Minnesota 55905, USA. Tel: 00 1 507 266 4156; E-mail: cook.david33@mayo.edu

© 2015 John Wiley & Sons Ltd.

INTRODUCTION

Assessment is a central component of our medical education endeavours. We continually make judgements and decisions about learners based on various types of assessment including examinations, rating scales and clinical gestalt. As we move globally into an era of competency-based medical education,[1] with increased reliance on assessments of mastery,[2] milestones,[3,4] and readiness to perform key tasks,[5] the decision-making processes associated with how we judge learners' competencies have become increasingly relevant. In order to make sound judgements, we need to carefully understand the strengths and limitations of the assessment tools and processes upon which these decisions are based. Stated differently, we require evidence to support the validity of our decisions. The process of collecting and interpreting validity evidence is called 'validation'.

Messick[6] and Kane[7] have offered detailed reviews of how validation has evolved over the past 100 years. To summarise very briefly (see Fig. 1), educators initially recognised two types of validity: content validity (which relates to the creation of the assessment items), and criterion validity (which refers to how well scores correlate with a reference-standard measure of the same phenomenon). However, content validity nearly always supported the test, and investigators quickly recognised that identifying and validating a reference standard is very difficult, especially for intangible attributes (e.g. professionalism). As an alternative for contexts in which no definitive criterion existed, theorists proposed construct validity,[8] in which intangible attributes (constructs) are linked with observable attributes based on a conception or theory of the construct. Validity can then be tested by measuring observable attributes and evaluating the theorised relationships. Experts soon realised that all these different 'types' of validity, together with reliability metrics, ultimately had the common pathway of supporting (or refuting) the construct-related relationships. This led researchers (as detailed by Messick[6]) to abandon the different 'types' of validity in favour of a unified framework in which construct validity (the only type) is supported by evidence derived from multiple sources. However, although Messick's framework has subsequently been widely embraced,[9,10] it does not prioritise among the different evidence sources or indicate how priority might vary for different assessments.[11] For example, the key assumptions and weaknesses underlying a high-stakes multiple-choice examination might be very different from those of a low-stakes procedural assessment or an observation of clinical performance intended to guide formative feedback. The most recent evolution in validation theory, articulated by Kane,[7,12] addresses the issue of prioritisation by highlighting key phases or inferences in planning and evaluating the validity argument.

The beauty of Kane's framework is that it applies equally to an individual quantitative assessment tool, a qualitative assessment tool, or a programme of assessment. Such versatility will be required as we move into a 'post-psychometric era' of assessment in which qualitative and subjective data are increasingly valued[13] and multiple assessment data points of varying rigour are routinely integrated.[14] Schuwirth and van der Vleuten[15] have provided an eloquent review of how Kane's framework can be applied to programmatic assessment. In the present article, we aim to offer a practical introduction to the key concepts of Kane's framework that educators will find accessible and applicable to a wide range of assessment tools and activities.

A FOCUS ON DECISIONS AND CONSEQUENCES

When we assess a learner, we usually generate a number (although qualitative and portfolio-based assessments are increasingly used), but the number itself is of relatively little value. What we want –

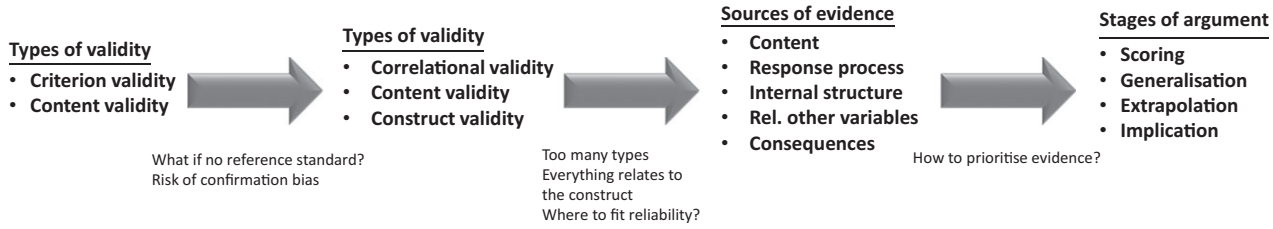

Figure 1 Evolution of concepts of validity. A narrative explanation is given in the text. Rel other variables = relationships with other variables

indeed, the reason we perform the assessment and obtain the number in the first place – is a decision about that learner. Did he pass? Does he need remediation, or should he be considered for student-of-the-year? What is he doing well and where does he need to improve? Each of these questions represents a decision that may have important consequences in the life of that learner and the patients and peers with whom he will work in future years, as well as the systems involved in that work. Ultimately, validation is all about collecting evidence to support the defensibility of that decision.

An analogy with clinical medicine may help to illustrate this point. Is the prostate-specific antigen (PSA) test useful in screening for prostate cancer? Evidence suggests that values are quite reproducible on retesting and from year-to-year,[16,17] that high values correlate strongly with cancer,[18] and that cancer is diagnosed at a less aggressive stage when annual PSA tests are performed.[19,20] Yet despite this favourable evidence, professional organisations recommend against screening for most men.[21–23] This incongruity arises because of the unintended adverse consequences of further evaluation[24] and, more importantly, because large randomised trials have arrived at conflicting conclusions regarding the benefits of testing and subsequent treatment.[20,25]

From this clinical example we learn several important lessons relevant to the educational assessment of health professionals. Firstly, not all assessments are beneficial. In fact, an assessment with very careful development, reproducible results and correlation with important variables might ultimately cause more harm than good, especially at the level of the individual (if, for instance, a low score prompted unnecessary remediation activities). Secondly, people may rightly arrive at different conclusions when interpreting the same evidence, as might be reflected in differing perspectives in a residency promotion committee. Thirdly, an assessment might be useful in some contexts but not in others (e.g. PSA test properties vary by age; an education checklist may prove adequate for assessing procedural skills in a simulation-based context, but fail to capture important nuances of clinical practice). Fourthly, the usefulness of a test may vary for different purposes (e.g. the PSA test is generally considered useful in monitoring for cancer recurrence; the mini-clinical evaluation exercise [mini-CEX] seems appropriate as a tool for formative feedback, but may be less defensible when used for summative purposes or programme evaluation[26]). Fifthly, the act of assessment is in fact an intervention, as witnessed by research on test-enhanced learning,[27] and like all interventions can be evaluated for its impact (e.g. one can conduct a randomised trial of PSA testing versus no PSA,[19,20] or educational assessment versus no assessment[28,29]).

The principle of focusing on decisions and consequences is not limited to multiple-choice tests or objective structured clinical examinations. It applies equally well to portfolio-based assessment,[30] qualitative assessments,[31] and programmes of assessment,[15] to both formative (during instruction, to monitor learning and provide feedback) and summative (at the end of instruction, often judged in comparison with a standard) assessments, and to tests intended to directly enhance learning.[32] Sometimes test developers cannot envision all of the potential uses of the test, or existing data are repurposed to a new use. Regardless of the nature of the data or the format of the assessment activity, at some point a judgement will be made (e.g. 'meets standards', 'these areas need improvement…' or 'possesses qualities we admire') and a decision will result (e.g. 'pass', 'you could improve on …' or 'accept for admission').

The purpose of validation is to collect evidence that evaluates whether or not a decision and its attendant consequences are useful.

THE VALIDITY ARGUMENT

The validity argument guides the collection and interpretation of validity evidence. As an analogy, consider an attorney planning, collecting, organising and then presenting evidence in a legal argument before a court.[33] The intent of such an argument is to convince the judge or jury of the proposed interpretation of the evidence, namely that the defendant is guilty (or innocent), which will, in turn, lead to a decision (i.e. conviction or acquittal). That decision will depend on the strength, completeness and relevance of the evidence, the organisation and persuasion of the attorney's reasoning, and the personal feelings of those rendering judgement. Rarely is a single piece of evidence so incontrovertible that it single-handedly 'makes the case'. Rather, the argument usually consists of multiple pieces of evidence, individually incomplete but collectively sufficient to convince the jury.

Continuing the analogy of a legal argument, the amount of evidence required varies depending on

 © 2015 John Wiley & Sons Ltd.

the gravity of the pending decision. In a criminal case the threshold for conviction is that the evidence is convincing 'beyond reasonable doubt', yet the evidence required in a felony case would typically be much greater than that required in a minor driving infraction. By contrast, in a civil case, a judgement is rendered based on the 'preponderance of the evidence'.

Turning now to assessment validation, the same principles apply. First, an educator must consider the decision at hand (e.g. 'Who should pass the cardiology clerkship, and who needs remediation?') and a proposed interpretation that would support that decision (e.g. 'high scores reflect good cardiac examination skills; low scores indicate poor examination skills'). Next, with the desired decision in mind, the educator identifies the key claims, assumptions and inferences associated with this interpretation and use; Kane[12] labels this the 'interpretation/use argument'. The educator then develops a plan to test these assumptions and inferences. Finally, guided by this plan, he or she collects empiric evidence from multiple sources and organises this evidence into a validity argument (Fig. 2).

Brennan[34] observed: 'There may be devilish details to be considered, but the basic approach is straightforward.' Kane[12] declared: 'First, state the claims that are being made in a proposed interpretation or use (the IUA [interpretation/use argument]), and second, evaluate these claims (the validity argument).' This two-step approach – stating and then evaluating claims – is analogous to the routine research practice of stating and then testing a hypothesis. As with other forms of hypothesis-driven research, the hypothesis in validity research[8] (the interpretation/use argument) contributes greatly to the relevance, rigour, completeness and simple elegance of study results and interpretations.

KANE'S FRAMEWORK

However, identifying the weakest links and assumptions in the hypothesis, and planning the tests that will evaluate those assumptions, is rarely obvious (the 'devilish details' referred to by Brennan[34]). Fortunately, Kane has described a framework for thinking about the validity argument that helps educators identify the most important pieces of evidence when planning the evaluation (to prioritise the collection of evidence) and when interpreting the argument (to identify evidence gaps). Essentially, Kane traces an assessment from the *Scoring* of a single observation (e.g. multiple-choice examination question, skill station, clinical observation or portfolio item), to using the observation score(s) to generate an overall test score representing performance in the test setting (*Generalisation*), to drawing an inference regarding what the test score might imply for real-life performance (*Extrapolation*), and then to interpreting this information and making a decision (*Implications*) (Fig. 3). Each phase in this process represents an inference laden with several assumptions. Kane's validity framework specifies evidence that can be collected to support (or refute) each of these assumptions, thus strengthening (or weakening) the associated inferences and ultimately the overall validity argument; these sources of evidence

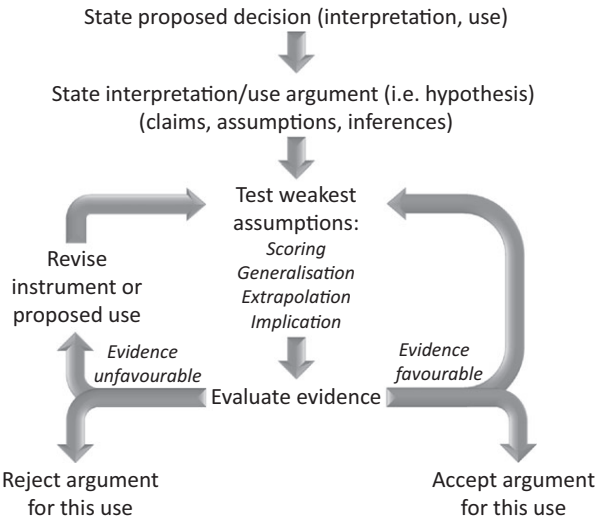

Figure 2 Evaluating the validity argument

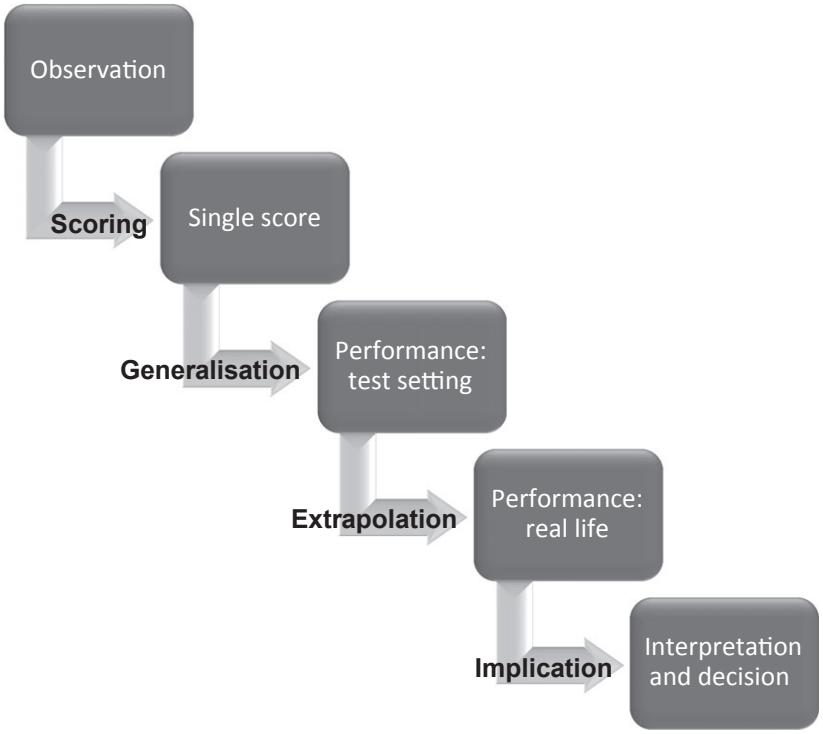

Figure 3 Key elements in the validity argument: inferences from observation to decision

are discussed below and in Table 1. Educators will be familiar with many of the types of validity evidence under each category. The key insight provided by Kane refers to the prioritisation, selection and organisation of this evidence into a comprehensive and coherent argument.

The validity argument should contain multiple sources of evidence that span several (if not all) inferences. It is also important to focus on the weakest links (most questionable assumptions). One advantage of Kane's framework is that it does not rely heavily on psychometric data, and thus the concepts apply readily to non-quantitative assessments (such as learning portfolios and narrative performance reviews). Kane's framework also applies to programmes of assessment (i.e. the use of multiple individual assessment activities to make judgements). Schuwirth and van der Vleuten[15] have explored this at length and we will not discuss it further.

Define the proposed use

Validation begins with a clear statement of the proposed use of the assessment scores (i.e. interpretations and decisions). Tests intended to result in pass/fail decisions may require different prioritisation of validity evidence than those reporting raw scores. The validation of an assessment of communication skills might vary substantially according to whether it is intended for a second-year medical student, first-year resident or practising physician, or for a psychiatrist versus an orthopaedic surgeon. Interpretations to guide formative feedback or to establish a minimal level of competence would require evidence that differs from that for interpretations suggesting competence to practise medicine independently.

***Scoring* inference**

Each assessment begins with an observation of some performance, such as a multiple-choice test question, a skill station, a clinical encounter or a self-reflection narrative in a learning portfolio. The intent is to use that observation to generate a fair, accurate, reproducible quantitative score (or an accurate and insightful narrative comment). The *Scoring* inference is greatly influenced by the construction of specific items, such as the wording of a written test question, training of a standardised patient, or procedural task specification. *Scoring* also depends on the selection of response options (e.g. dichotomous checklist versus global rating scale,[35] the number[36] and weighting of response options,[37] and the choice of specific scale anchors),

 © 2015 John Wiley & Sons Ltd.

Table 1 Specific evidence to support the validity argument

	Quantitative assessments		**Qualitative assessments**	
	Procedures to define, establish or select . . .	**Empirical evaluation of** . . .	**Procedures to define, establish or select** . . .	**Empirical evaluation of** . . .
Scoring	• Items and response options • Observation format (e.g. live versus video; written versus computer) • Standardisation, equating across forms or occasions • Scoring rubric/criteria and implementation procedures; pass/fail standard • Rater selection and training • Rules for combining related test elements from different sources (triangulation) or separating unrelated elements from one source (e.g. different skill domains on single OSCE) • Data security, quality control	• Item and response option performance (item difficulty, point biserial, response option analyses) • Observation format (e.g. empiric comparison of different formats, such as live versus video-based, or blinded versus unblinded scoring) • Standardisation, equating • Scoring rubric/criteria (e.g. empiric comparison of different procedures, think-aloud study) • Rater selection and training; rater accuracy and reliability • Data security, quality control	• Observation opportunities sufficient to inform meaningful narratives • Questions and prompts likely to stimulate a rich narrative response • Observers with credibility (e.g. appropriate experience and training) to provide the requested insights	• Observations actually conducted • The richness, accuracy, authenticity and fairness of qualitative data (e.g. individual narratives, other documents)
Generalisation	• Sampling strategy (items, raters, tasks, occasions), e.g. test blueprint; random versus purposive sampling • Sample size	• Reliability or generalisability (items, raters, tasks, occasions) • Item response theory	• Sampling strategy (prompts, observers, occasions, performance domains, complementary data sources and types), e.g. purposive sampling, sampling to saturation, triangulation • Interpretive process that is defensible and transparent (auditable) • Data interpreters with credibility to perform interpretation and synthesis • Response to conflicting data	• Sampling and triangulation; the variety of perspectives reflected in data being analysed (different observers, performance domains, time points, data types) • Defensibility, reflexivity, transparency, and responsiveness of the interpretive process • Thematic saturation and coherence of final interpretations • Consistency and reflexivity of interpretations formed by different interpreters

Table 1 (Continued)

	Quantitative assessments		Qualitative assessments	
	Procedures to define, establish or select ...	**Empirical evaluation of** ...	**Procedures to define, establish or select** ...	**Empirical evaluation of** ...
Extrapolation	• Scope of test (e.g. domain specification, construct definition) • Authenticity of assessment context (e.g. clinical setting, simulation) • Authenticity of item/scenario (e.g. real patient, task alignment)	• Needs analysis to define scope/objectives • Process-construct match (e.g. think-aloud study) • Relevance and authenticity (e.g. ratings by experts) • Correlation with another measure having an expected relationship (criterion-referenced or convergent; concurrent or predictive) • Discrimination (known groups comparison) • Responsiveness (sensitivity to change following intervention) • Construct profile (e.g. factor analysis, MTMM) • Differential item functioning	• Data sources that reflect key aspects of performance	• The relevance of data sources to performance • Agreement of relevant stakeholders (e.g. observers, learners, programme directors) with final interpretation (member check) • Agreement of stakeholders that interpretations will apply to new contexts in training or practice (transferability) • Relationship between qualitative interpretations and other measures of similar traits (e.g. quantitative data, independent decisions about remediation or honours)
Implications	• Pass/fail standard (e.g. Angoff method) • Planned actions based on assessment results (e.g. remediation)	• Pass/fail standard (e.g. ROC curve) • Effectiveness of actions based on assessment results • Intended or unintended consequences of testing (long-term follow-up; qualitative studies; consider impact on learners, raters and others) • Differential item functioning (if implications for consequences)	• Planned actions based on assessment results (e.g. remediation)	• Agreement of other experts with final judgement and decision • Effectiveness of actions based on assessment results • Intended or unintended consequences of testing (consider impact on learners, observers, interpreters and others)

This table lists many elements of evidence that could be used to test each inference, but is not inclusive of all possible elements. There is no expectation that all of these elements should be used in a given validation. Rather, an investigator should select those elements most salient to the intended use/decision (see text for details)
MTMM = multitrait, multimethod matrix;[80] OSCE = objective structured clinical examination; ROC = receiver operating characteristic

scoring rubric and procedures, and item analysis. Fairness requires consideration of whether everyone was given a similar test, the resolution of which may include standardisation and approaches to enhance test security. Standard-setting procedures can also influence dichotomous responses (e.g. definitions

 © 2015 John Wiley & Sons Ltd.

of pass/fail or done/not done) generated from a single observation.

In qualitative assessments, the questions that prompt a narrative response, the richness of the raw data, the credibility of the observers, and the use of 'thick description' (e.g. actual quotes or images) in the final report would support *Scoring*.

Generalisation inference

To understand *Generalisation* we need to distinguish performance in the 'test world' (formally the 'universe of assessment') from that in the 'real world'. *Generalisation* deals with test-world performance.

In the universe of assessment, there are in theory a limitless number of items that we could create or select to assess the performance domain under study. We could select 16 or 60 multiple-choice questions for a test of knowledge of cardiology, and in the process of question selection could give greater or lesser emphasis to valvular disease versus arrhythmias. For a skill examination, we might use four or 12 skill stations and could tweak the specifics of each station scenario or rating form in countless ways. The test items we ultimately select represent a *sample* of the items from this universe of possibilities. However, we are not really interested in this sample *per se*; rather, we ideally want to generalise from this sample to the entire assessment universe.

Thus, *Generalisation* seeks to answer the question: how well do the selected test items (the questions, cases, stations, raters, observations, survey response options, portfolio entries, self-reflection essays, etc.) in our sample represent all of the theoretically possible items in the relevant assessment universe? Evidence to answer this question comes from two primary sources: methods taken to ensure adequate and appropriate sampling within the test domain, and empiric studies to determine the likelihood of obtaining similar scores if we use an entirely new sample of items (reproducibility or reliability).

Methods to ensure appropriate sampling might include a test blueprint (across domains) or random sampling (within a domain) to assist in systematically selecting items. The required number of observations depends on both the scope of the domain (i.e. the size of the universe; more comprehensive tests will require more observations) and the reproducibility of individual observations. The qualitative research concept of saturation may be useful, especially for non-numeric data or if the universe is highly heterogeneous: does the new observation add important information beyond the information already collected?

The reproducibility of numeric scores can be empirically determined using reliability metrics. According to classical test theory, an observed score reflects the true score imperfectly because of measurement error. Scores are more reproducible, and presumably closer to the true score, as measurement error is reduced. Error can arise at each step or facet of the measurement activity – such as the individual item, station, rater or occasion – and error decreases as the number of replications increases (e.g. more items, more stations, more raters or more occasions). Generalisability theory allows a researcher to investigate the magnitude of error arising from each facet simultaneously, then to identify the sources of greatest error, and finally to optimise overall reproducibility by exploring the impact of varying numbers of replications for each facet.

In some assessments (e.g. surveys and checklists in which each item measures a unique point of interest), individual items are each intended to reflect a different domain. In such instances, aggregating or averaging responses is inappropriate, as is estimating reproducibility across items (inter-item internal consistency), and the *Generalisation* inference relies heavily on the sampling of each relevant domain and on other facets of reproducibility (raters and stations or cases).

For qualitative assessments, the synthesis of individual pieces of qualitative data to form an insightful, accurate and defensible interpretation is analogous to quantitative generalisation. Whereas we treat inter-rater variability as error for most numeric scores, in qualitative assessments we view observer variability as representing potentially valuable insights into performance (i.e. different perspectives[38,39]). The method for selecting and synthesising data from different sources (triangulation) and deciding when to stop (saturation) will inform the *Generalisation* inference for qualitative data.

Extrapolation inference

Test-world performance is important, but what we really want to measure or at least anticipate is real-world performance. This leap of faith from test performance to real-life performance requires *Extrapolation*. In other words, *Generalisation* takes us from a sample of observations to the test-world universe;

Extrapolation takes us from the test-world universe to the real world.

Evidence to support *Extrapolation* comes primarily from two sources: methods taken to ensure that the test domain reflects the key aspects of real performance, and empiric analyses evaluating the relationship between the test performance and real-world performance. To establish the relevant test domain, test developers might interview or poll experts, observe the actual task as performed by practitioners of varying levels of skill, ask experts to think aloud as they perform the task, and review past literature including relevant guidelines and authoritative texts. Once a complete picture of the desired clinical domain has been specified, principles of domain sampling can be applied to establish an appropriate test domain (as described for *Generalisation*; in addition, purposive sampling might deliberately over-represent areas of high importance).

Empiric analyses to support the *Extrapolation* inference evaluate the association between test scores and a comparable metric related to the real task. One common approach evaluates the ability of scores to discriminate among groups of learners who differ in a specific characteristic such as training status (the 'known-group' or 'expert–novice' comparison). However, known-group comparisons offer relatively weak validity evidence because association does not imply causation.[40] Stronger *Extrapolation* evidence can be collected by correlating test scores with scores from a conceptually related real-world assessment. In the absence of real-world scores, scores from another test measure can be used (i.e. making the argument that strong correlation between two independent measures of the same task supports the validity of both scores), although the inference is naturally weaker. The hypothesised correlation need not always be strong or positive; for example, a strongly positive correlation between two scores would undermine the inference if the constructs were conceptually independent. To avoid the pitfall of 'sheer exploratory empiricism [in which] any correlation of the test score with another variable is welcomed',[41] researchers should specify all hypothesised relationships prior to empiric evaluation. Cronbach labelled this practice the 'strong programme' of validation.[41] For qualitative assessment, *Extrapolation* might be further supported by evidence suggesting that stakeholders agree with the interpretations and anticipate that they will apply to new contexts in training or practice.

Unfortunately, *Generalisation* and *Extrapolation* are often at odds with one another. Kane[7] notes: 'We can strengthen extrapolation at the expense of generalisation by making the assessment tasks as representative of the target domain as possible, or we can strengthen generalisation at the expense of extrapolation by employing larger numbers of highly standardised tasks.'

Implications inference

The final inference moves from the target domain score to some interpretation about that score, and from that interpretation to a specific use, decision or action. This requires inference about the *Implications* of the assessment results. As Kane[7] states: 'It is generally inappropriate to assume that evidence supporting a particular *interpretation* of test scores automatically justifies a proposed *use* of the scores.' He also notes: 'A decision procedure that does not achieve its goals, or does so at too high a cost, is likely to be abandoned even if it is based on perfectly accurate information.'[7] In other words, even if we measure the attribute correctly, it doesn't necessarily mean this information will be useful (or used well). Thus, the final phase in the validity argument evaluates the consequences or impact of the assessment on the learner, other stakeholders and society at large.[42]

The most straightforward way to collect data regarding the consequences of assessment would be to offer the assessment (and the ensuing judgements and interventions [e.g. promotion or remediation]) to some learners but not to others, and to compare relevant outcomes including intended and unintended consequences (e.g. quantity and quality of feedback received, length and cost of training, drop-out rates, stress levels, scores on other short- and long-term performance measures, impact on raters, and effects on patient care). This approach would be similar to that used in comparative studies evaluating PSA testing in comparison with no testing, and looking for consequences both intended (improved cancer-free morbidity and mortality) and unintended (increased biopsies or surgeries, increased anxiety). However, such studies are difficult to conduct and exceed the reach of most investigators. More achievable studies evaluating the *Implications* inference include standard-setting studies (discussed under *Scoring*), non-comparative studies exploring intended and unintended consequences (e.g. what happens to learners who fail a key examination), and evaluations of differences in test performance among subgroups for which performance should be similar, such as men and women (differential item functioning). Likewise, in qualitative assessments, evaluating the agree-

 © 2015 John Wiley & Sons Ltd.

ment of experts with final interpretations and the impact of decisions on learners and raters would support the *Implications* inference.

Implications evidence of any variety (e.g. what happened to those learners who failed the test and those who passed? Did remediation result in improved performance on follow-up assessment?) is very rarely published,[43] but we suspect that relevant raw data are often available locally yet not rigorously analysed and disseminated. The absence of evidence of consequences represents an important gap in the literature.[42]

PUTTING THE ARGUMENT TOGETHER

Planning and presenting a coherent argument

Although Kane does not specify the order in which validity evidence should be collected and evaluated, there seems to be a natural progression that aligns the phases of the argument (from left to right in Fig. 3) with the priority and sequence of collecting empiric evidence. It seems natural to solidify evidence regarding the scoring rubric before analysing the generalisability of those scores, to evaluate generalisability before extrapolating to real life, and to confirm relationships with real-life performance before attempting to confirm the impact of assessment on meaningful outcomes. Of course, the issues related to domain specification and sampling will need to be addressed early in the process.

Although all of the inferences in the validity argument merit *some* attention, they are not all of equal importance. For a given interpretation and use, some assumptions are *a priori* more plausible, and some assumptions more vital, than others. *Generalisation* may be less important when the emphasis is on formative feedback, and the *Extrapolation* inference may be less important for assessments (both qualitative and quantitative) that rely on direct observation of real clinical performance as the underlying assumptions are relatively plausible. This underscores the need to clearly state the hypothesis (the interpretation/use argument) before collecting evidence! The prioritisation of specific inferences for specific test interpretations and decisions (e.g. admissions, promotions, licensure) is an area of active development, and some investigators have proposed validity arguments for specific assessment topics.[44,45]

Empiric findings often disagree within an inference (e.g. conflicting evidence for *Generalisation*), between inferences (e.g. favourable *Generalisation* but unfavourable *Extrapolation*), and across different contexts or research studies. A pre-specified interpretation/use argument and evaluation plan helps to integrate such findings. Further, several iterations through the 'revise instrument or proposed use' branch of Fig. 2 may be necessary, especially in early stages of development and validation.

Flaws in building the validity argument

Educators commonly make the mistake of assuming that a test validated for one purpose or context is valid for another. In reality, all assessments must be validated for each new proposed interpretation and use. Kane[7] identified a number of other flaws in building the validity argument. Firstly, educators often conclude that interpretations and decisions are valid after evaluating limited evidence. Secondly, critics, naïve investigators or inappropriate regulatory requirements might propose an argument that is more ambitious than required for a given purpose. Thirdly, investigators often collect easy-to-measure evidence for assumptions that are already plausible; this typically occurs at the expense of addressing other more questionable assumptions, and can be misleading if the sheer quantity of evidence obscures important omissions.

PRACTICAL APPLICATION OF KANE'S FRAMEWORK

We conclude by showing how Kane's framework can apply to three commonly used instruments: a clinical laboratory test (the PSA test); an assessment of procedural skills (the objective structured assessment of technical skills [OSATS]), and a qualitative assessment (narrative comments from in-training clinical assessments).

For a screening test for a pre-symptomatic disease (e.g. the PSA) to support the *Scoring* inference, we would expect to have well-defined procedures for standardisation and for combining scores with other clinical information such as physical examination findings and other test results. To support *Generalisation*, we would expect low test–retest variability (high test–retest reliability) and, if relevant, high inter-rater reliability, and to support *Extrapolation* we would expect that different assays correlate well, that scores discriminate among patients with and without the target disease (high sensitivity and specificity), and that the test normalises after the disease is treated. Finally, to support the proposed

Table 2 Applying Kane's framework to three assessments

	PSA test	OSATS	In-training assessment narratives
Proposed use (decision)	Screen patients for prostate cancer (Does patient need further testing?)	Determine procedural competence of surgery residents (Can resident operate on real patients?)	Determine clinical competence of internal medicine residents (Can resident advance to next training year?)
Scoring	• Calibration procedures are well defined[16,46] • Cystoscopy and biopsy produce transient rise (i.e. informs conditions or standardisation of testing)[46] • Cut-point thresholds vary by age[47] • Measurement of free PSA adds value to total PSA[48] • Measurement of rate of change over time adds value to single PSA measurements[17]	• Description of checklist and GRS item development and selection[55] • Adding checklist to GRS does not improve discrimination[56] • Higher inter-rater reliability with surgeon raters than with family practice raters[57] • Live scores consistently higher than videotaped scores[58]	• Most in-training assessments contain narrative comments[64] • Comments map to discrete and overlapping competencies, and to non-competency characteristics[64] • Observers consider multiple themes when forming opinions about residents[65] • Observers often suppress negative comments*,[66,67] • Richness of comments improves following faculty development†,[68] • Observers with different roles (physician, nurse, administrator) emphasise different aspects of performance in their comments‡,[69] • Observer engagement enhances perceived authenticity/credibility[70]
Generalisation	• Test–retest variation on the same sample is 1–5%[16,18]	• Inter-rater reliability high for both checklist and GRS, but typically higher for GRS[55] • Inter-station correlations typically higher for GRS than checklist[59,60]	• Each resident can receive narrative comments from multiple observers (14 per resident in one study[71]) • Different groups of interpreters form relatively consistent judgements about trainees[72]
Extrapolation	• Results from different assays correlate relatively well[16] • Average scores discriminate among patients with normal prostate, benign prostatic hypertrophy, and prostate cancer, but there is substantial overlap[18,46] • PSA levels drop substantially following total or subtotal prostatectomy[18,46]	• Detects expert–novice differences across postgraduate year[59]	• Residents requiring remediation had longer comments and more negative comments than those in good standing[73] • Qualitative classifications correlate well with rotation scores,[71] course grades,†,[74] and peer evaluations‡,[75] • Written comments are better than numeric ratings in identifying deficiencies§,[76]
Implications	• Test performance varies for different test cut-points*,[49] • Alternate-year screening may be preferable[23,50] • The clinical benefit of screening + treatment is small and controversial*,[19,20,51] • Prostate cancer mortality varies by baseline PSA level[52]	• None identified[63]	• Remediation, when recommended based on qualitative or quantitative assessments, is resource-intensive but can be highly successful[77]

 © 2015 John Wiley & Sons Ltd.

Table 2 (Continued)

	PSA test	OSATS	In-training assessment narratives
	• Short-term unintended consequences of testing include pain, fever, haematuria after biopsy*,[24] • Long-term unintended consequences include overdiagnosis and overtreatment*,[53]		

There may be other acceptable proposed uses/decisions for each test; we have selected one specific use for each assessment. The evidence presented for each test use reflects what has been published rather than an ideal collection of evidence; hence there are likely to be important gaps. In addition, some evidence is unfavourable to the validity of interpretations and decisions
GRS = global rating scale; OSATS = objective structured assessment of technical skills; PSA = prostate-specific antigen.
* Unsupportive evidence (i.e. suggests invalidity of interpretations/decisions)
† Study involves ratings of medical students (not residents)
‡ Study involves ratings of physicians in practice (not residents)
§ Study involves ratings of surgery (not internal medicine) residents

Implications we would want to know that screening for a disease and then treating it yields better long-term clinical outcomes than waiting for the disease to become clinically apparent, and that adverse effects of the treatment do not outweigh the benefits. As Table 2 shows,[16–20,23,24,46–53] abundant evidence supports the first three elements of the proposed argument for the PSA test. However, the clinical benefits have been called into question, which, in turn, fundamentally challenges the utility of the test. It is for this reason that the US Preventive Services Task Force now advises *against* using the PSA test as a screen for prostate cancer.[21]

We next consider a rater-based assessment of procedural skills that is used to inform decisions about whether a resident is ready to operate on real patients under supervision. To support the *Scoring* inference, we would expect to see an evidence-based scoring rubric, to know that raters have been appropriately trained and that the observation format (e.g. video-based review) provides sufficient information. To support *Generalisation*, we would expect broad sampling across different tasks and levels of difficulty (simple cases and more complex cases), and high reproducibility (considering raters and stations/cases as sources of measurement error). To support *Extrapolation*, we would expect experts to agree on the key task elements and the simulator to appropriately represent these key elements ('functional task alignment'[54]), that scores correlate strongly with an independent rating of performance on the same task (measuring skill or performance in clinical practice), and that scores improve following training. Finally, to support the proposed *Implications*, we would want to know that decisions to delay operating privileges improve patient care, that remediation leads to objective improvement, that residents perceive a benefit, and that the delay does not impose an excessive burden on residents or training programmes. As Table 2 shows, substantial evidence supports the first three inferences for the OSATS,[55–60] although some evidence is less favourable.[61–63] However, virtually no evidence has been reported to support the *Implications* inference.[63]

Finally, we consider the use of narrative comments (qualitative data) from supervisors assessing residents' clinical performance to make decisions about promotion to the next training year. To support the *Scoring* inference we would expect to see that questions prompt a variety of relevant narrative data, that assessors have actually observed the behaviours they are asked to assess, and that narrative comments provide a rich, detailed description of observed behaviours. To support *Generalisation*, we would expect to see that narratives have been

solicited from people representing a variety of clinical roles, that the narratives collectively form a coherent picture of the resident, and that those conducting the interpretive analysis have appropriate training or experience. To support *Extrapolation*, we would anticipate that those providing raw narratives agree with the synthesised 'picture' and that the qualitative narrative agrees with other data (qualitative or quantitative) measuring similar traits. Finally, to support the proposed *Implications*, we would want to know that both those providing narratives and the residents themselves agree with the decision based on these narratives, and that actions based on these decisions have the desired effect. We found evidence to support many, but not all, of these propositions (Table 2).[64–77]

ALTERNATIVE PERSPECTIVES

Not everyone fully agrees with Kane. Most notably, the 2014 *Standards for Educational and Psychological Testing*[9] wholly endorse the argument-based approach to validity but do not embrace Kane's focus on four key inferences, choosing instead to emphasise the five sources of evidence proposed by Messick.[6] The need for a separate interpretation/use argument is also a matter of debate.[78] Kane[79] recently responded to other critiques; we refer readers to this discourse for further elaboration.

CONCLUSIONS

In conclusion, we emphasise four points. Firstly, validation is not an endpoint but a process. Stating that a test has been 'validated' merely means that the process has been applied, but does not indicate the intended interpretation, the result of the validation process or the context in which this was done. Secondly, validation ideally begins with a clear statement of the proposed interpretation and use (decision), continues with a carefully planned interpretation/use argument that defines key claims and assumptions, and only then proceeds with the collection and organisation of logical and empirical evidence into a substantiated validity argument. Thirdly, educators should focus on the weakest links (most questionable assumptions) in the chain of inference. Fourthly, in all of the clinical and educational examples cited herein, the *Scoring*, *Generalisation* and *Extrapolation* evidence is fairly strong; only when we attempt to infer actionable *Implications*, moving from the real-world score to specific decisions, do important deficiencies come to light. For this reason, we believe that the *Implications* and associated decisions are ultimately the most important inferences in the validity argument.

Contributors: DAC drafted the initial manuscript. All authors contributed to the conception of the work, revised the manuscript for important intellectual content, and approved the final version for publication.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach* 2010;**32**:676–82.
2 Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Mastery learning for health professionals using technology-enhanced simulation: a systematic review and meta-analysis. *Acad Med* 2013;**88**:1178–86.
3 Caverzagie KJ, Iobst WF, Aagaard EM *et al*. The internal medicine reporting milestones and the next accreditation system. *Ann Intern Med* 2013;**158**:557–9.
4 Green ML, Aagaard EM, Caverzagie KJ, Chick DA, Holmboe E, Kane G, Smith CD, Iobst W. Charting the road to competence: developmental milestones for internal medicine residency training. *J Grad Med Educ* 2009;**1**:5–20.
5 ten Cate O. Trust, competence, and the supervisor's role in postgraduate training. *BMJ* 2006;**333**:748–51.
6 Messick S. Validity. In: Linn RL, ed. *Educational Measurement*, 3rd edn. New York, NY: American Council on Education and Macmillan 1989;13–103.
7 Kane MT. Validation. In: Brennan RL, ed. *Educational Measurement*, 4th edn. Westport, CT: Praeger 2006;17–64.
8 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.
9 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: AERA 2014.
10 Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003;**37**:830–7.
11 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;**19**:166.e7–16.
12 Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013;**50**:1–73.

 © 2015 John Wiley & Sons Ltd.

13 Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach* 2013;**35**:564–8.
14 Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ* 2006;**40**:296–300.
15 Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012;**46**:38–48.
16 Chan DW, Bruzek DJ, Oesterling JE, Rock RC, Walsh PC. Prostate-specific antigen as a marker for prostatic cancer: a monoclonal and a polyclonal immunoassay compared. *Clin Chem* 1987;**33**:1916–20.
17 Carter HB, Pearson JD, Metter EJ, Brant LJ, Chan DW, Andres R, Fozard JL, Walsh PC. Longitudinal evaluation of prostate-specific antigen levels in men with and without prostate disease. *JAMA* 1992;**267**:2215–20.
18 Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. *N Engl J Med* 1987;**317**:909–16.
19 Schroder FH, Hugosson J, Roobol MJ *et al.* Prostate-cancer mortality at 11 years of follow-up. *N Engl J Med* 2012;**366**:981–90.
20 Andriole GL, Crawford ED, Grubb RL III *et al.* Mortality results from a randomised prostate-cancer screening trial. *N Engl J Med* 2009;**360**:1310–9.
21 Moyer VA. Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *Ann Intern Med* 2012;**157**:120–34.
22 Qaseem A, Barry MJ, Denberg TD, Owens DK, Shekelle P. Screening for prostate cancer: a guidance statement from the Clinical Guidelines Committee of the American College of Physicians. *Ann Intern Med* 2013;**158**:761–9.
23 Carter HB, Albertsen PC, Barry MJ *et al.* Early detection of prostate cancer: AUA guideline. *J Urol* 2013;**190**:419–26.
24 Rosario DJ, Lane JA, Metcalfe C *et al.* Short term outcomes of prostate biopsy in men tested for cancer by prostate specific antigen: prospective evaluation within ProtecT study. *BMJ* 2012;**344**:d7894.
25 Schroder FH, Hugosson J, Roobol MJ *et al.* Screening and prostate-cancer mortality in a randomised European study. *N Engl J Med* 2009;**360**:1320–8.
26 Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VE. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomised, controlled trial. *J Gen Intern Med* 2009;**24**:74–9.
27 Roediger HL, Karpicke JD. The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci* 2006;**1**:181–210.
28 Cook DA, Thompson WG, Thomas KG, Thomas MR, Pankratz VS. Impact of self-assessment questions and learning styles in web-based learning: a randomised, controlled, crossover trial. *Acad Med* 2006;**81**:231–8.
29 Cook DA, Thompson WG, Thomas KG. Test-enhanced web-based learning: optimising the number of questions (a randomised crossover trial). *Acad Med* 2014;**89**:169–75.
30 Driessen EW, Overeem K, van Tartwijk J, van der Vleuten CPM, Muijtjens AMM. Validity of portfolio assessment: which qualities determine ratings? *Med Educ* 2006;**40**:862–6.
31 Kuper A, Reeves S, Albert M, Hodges BD. Assessment: do we need to broaden our methodological horizons? *Med Educ* 2007;**41**:1121–3.
32 Larsen DP, Butler AC, Roediger HL III. Test-enhanced learning in medical education. *Med Educ* 2008;**42**:959–66.
33 Cook DA. When I say… validity. *Med Educ* 2014;**48**:948–9.
34 Brennan RL. Commentary on 'Validating the interpretations and uses of test scores'. *J Educ Meas* 2013;**50**:74–83.
35 Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ* 2015;**49**:161–73.
36 Cook DA, Beckman TJ. Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Adv Health Sci Educ Theory Pract* 2009;**14**:655–64.
37 Sandilands DD, Gotzmann A, Roy M, Zumbo BD, De Champlain A. Weighting checklist items and station components on a large-scale OSCE: is it worth the effort? *Med Teach* 2014;**36**:585–90.
38 Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgements: rethinking the aetiology of rater errors. *Acad Med* 2011;**86** (10 Suppl):1–7.
39 Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly observed performance assessments. *Adv Health Sci Educ Theory Pract* 2013;**18**:325–41.
40 Cook DA. Much ado about differences: why expert–novice comparisons add little to the validity argument. *Adv Health Sci Educ Theory Pract* DOI: 10.1007/s10459-014-9551-3 [Epub ahead of print 2014 Sep 27].
41 Cronbach LJ. Five perspectives on validity argument. In: Wainer H, Braun HI, eds. *Test Validity.* Hillsdale, NJ: Routledge 1988;3–17.
42 Haertel E. Getting the help we need. *J Educ Meas* 2013;**50**:84–90.
43 Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med* 2013;**88**:872–83.
44 Clauser BE, Margolis MJ, Holtman MC, Katsufrakis PJ, Hawkins RE. Validity considerations in the assessment of professionalism. *Adv Health Sci Educ Theory Pract* 2012;**17**:165–81.

45 Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-clinical evaluation exercise: a review of the research. *Acad Med* 2010;**85**:1453–61.
46 Oesterling JE. Prostate specific antigen: a critical assessment of the most useful tumour marker for adenocarcinoma of the prostate. *J Urol* 1991;**145**:907–23.
47 Oesterling JE, Jacobsen SJ, Chute CG, Guess HA, Girman CJ, Panser LA, Lieber MM. Serum prostate-specific antigen in a community-based population of healthy men. Establishment of age-specific reference ranges. *JAMA* 1993;**270**:860–4.
48 Vashi AR, Oesterling JE. Percent free prostate-specific antigen: entering a new era in the detection of prostate cancer. *Mayo Clin Proc* 1997;**72**:337–44.
49 Schroder FH, Roobol MJ. Defining the optimal prostate-specific antigen threshold for the diagnosis of prostate cancer. *Curr Opin Urol* 2009;**19**:227–31.
50 Ross KS, Carter HB, Pearson JD, Guess HA. Comparative efficiency of prostate-specific antigen screening strategies for prostate cancer detection. *JAMA* 2000;**284**:1399–405.
51 Barry MJ. Screening for prostate cancer – the controversy that refuses to die. *N Engl J Med* 2009;**360**:1351–4.
52 Wilt TJ, Brawer MK, Jones KM *et al*. Radical prostatectomy versus observation for localised prostate cancer. *N Engl J Med* 2012;**367**:203–13.
53 Hayes JH, Barry MJ. Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence. *JAMA* 2014;**311**:1143–9.
54 Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering fidelity in simulation-based training. *Acad Med* 2014;**89**:387–92.
55 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;**84**:273–8.
56 Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;**73**:993–7.
57 Friedlich M, MacRae H, Oandasan I, Tannenbaum D, Batty H, Reznick R, Regehr G. Structured assessment of minor surgical skills (SAMSS) for family medicine residents. *Acad Med* 2001;**76**:1241–6.
58 Hance J, Aggarwal R, Stanbridge R, Blauth C, Munz Y, Darzi A, Pepper J. Objective assessment of technical skills in cardiac surgery. *Eur J Cardiothorac Surg* 2005;**28**:157–62.
59 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative 'bench station' examination. *Am J Surg* 1997;**173**:226–30.
60 Goff BA, Lentz GM, Lee D, Fenner D, Morris J, Mandel LS. Development of a bench station objective structured assessment of technical skills. *Obstet Gynecol* 2001;**8**:412–6.
61 Datta V, Bann S, Beard J, Mandalia M, Darzi A. Comparison of bench test evaluations of surgical skill with live operating performance assessments. *J Am Coll Surg* 2004;**199**:603–6.
62 Bann S, Davis IM, Moorthy K, Munz Y, Hernandez J, Khan M, Datta V, Darzi A. The reliability of multiple objective measures of surgery and the role of human performance. *Am J Surg* 2005;**189**:747–52.
63 Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Adv Health Sci Educ Theory Pract.* [Epub ahead of print 2015 Feb 22].
64 Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H. Competencies 'plus': the nature of written comments on internal medicine residents' evaluation forms. *Acad Med* 2011;**86** (10 Suppl):30–4.
65 Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Acad Med* 2010;**85**:780–6.
66 Watling CJ, Kenyon CF, Schulz V, Goldszmidt MA, Zibrowski E, Lingard L. An exploration of faculty perspectives on the in-training evaluation of residents. *Acad Med* 2010;**85**:1157–62.
67 Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med* 2005;**80** (Suppl):84–7.
68 Dudek NL, Marks MB, Wood TJ, Dojeiji S, Bandiera G, Hatala R, Cooke L, Sadownik L. Quality evaluation reports: can a faculty development programme make a difference? *Med Teach* 2012;**34**:e725–31.
69 Vivekananda-Schmidt P, MacKillop L, Crossley J, Wade W. Do assessor comments on a multi-source feedback instrument provide learner-centred feedback? *Med Educ* 2013;**47**:1080–8.
70 Watling CJ, Kenyon CF, Zibrowski EM, Schulz V, Goldszmidt MA, Singh I, Maddocks HL, Lingard L. Rules of engagement: residents' perceptions of the in-training evaluation process. *Acad Med* 2008;**83** (Suppl):97–100.
71 Ginsburg S, Eva K, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med* 2013;**88**:1539–44.
72 Regehr G, Ginsburg S, Herold J, Hatala R, Eva K, Oulanova O. Using 'standardised narratives' to explore new ways to represent faculty opinions of resident performance. *Acad Med* 2012;**87**:419–27.
73 Guerrasio J, Cumbler E, Trosterman A, Wald H, Brandenburg S, Aagaard E. Determining need for remediation through postrotation evaluations. *J Grad Med Educ* 2012;**4**:47–51.
74 Cohen GS, Blumberg P, Ryan NC, Sullivan PL. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med* 1993;**5**:10–5.
75 Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text

 © 2015 John Wiley & Sons Ltd.

comments from the UK General Medical Council Colleague Questionnaires. *Med Educ* 2009;**43**:757–66.
76 Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? *Acad Med* 2004;**79**:453–7.
77 Guerrasio J, Garrity MJ, Aagaard EM. Learner deficits and academic outcomes of medical students, residents, fellows, and attending physicians referred to a remediation programme, 2006–2012. *Acad Med* 2014;**89**:352–8.
78 Newton PE. Two kinds of argument? *J Educ Meas* 2013;**50**:105–9.
79 Kane MT. Validation as a pragmatic, scientific activity. *J Educ Meas* 2013;**50**:115–22.
80 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959;**56**:81–105.

Received 3 October 2014; editorial comments to author 20 November 2014, accepted for publication 19 December 2014