Best Practice in the Assessment of Competence: A Literature Review

School of Medical Education
Newcastle University

September 2018

Bryan Burford
Charlotte Rothwell
Gillian Vance
School of Medical Education

Fiona Beyer
Louise Tanner
Evidence Synthesis Group, Institute for Health and Society

Acknowledgements

The research team would like to thank all the members of the Project Advisory Group for their thoughtful advice and expert insight to the research areas. The members of the group are listed in the appendix, but we give particular thanks to those many members who shared detail on their local practice and gave valuable critique on the project findings.

We would also like to thank colleagues at the General Medical Council, especially Steve Loasby and Ben Griffith, for their support throughout the project timeline, and to Ms Dawn Craig, Head of the Evidence Synthesis group, Newcastle University, for guidance and review of the research processes.

Executive Summary

Introduction and background

Doctors undergo assessment throughout their careers, from progress examinations during medical school, to final qualifying exams, progression through postgraduate training and membership of professional bodies. Doctors may also be assessed when moving between regulatory jurisdictions, or when returning to practice. Some of these assessments aim to support learning ('formative assessment'), while others provide evaluation of learning or performance at a key stage of progression ('summative assessment'). Assessments encompass scientific and clinical knowledge, clinical and practical skills, and elements of practice termed 'professional skills' such as 'professionalism', ethical judgement and awareness of patient safety. This report is concerned with the final group, and considers two research questions:

1. What evidence is there for good practice in the use, or potential use of summative assessments around professionalism, ethics and competence in relation to patient safety?
2. What evidence is there for the use of simulation or other technologically-mediated methods in summative assessments in medical and non-medical contexts?

The first of these is concerned with areas of clinical practice for which assessments may be less well established than for applied clinical knowledge or skills. The second reflects how changing technology may enable different approaches to the assessment of all aspects of clinical practice.

The study primarily consisted of a systematic literature review, supplemented by stakeholder consultation through a Project Advisory Group and identification of examples of good practice. The research aimed to inform assessments throughout medical careers, including the planned medical licensing assessment for practice in the United Kingdom.

Method

Medical and non-medical databases were systematically searched in order to identify papers describing assessments of professionalism, ethics and patient safety, or using novel approaches to assessment. Searches returned over 9900 papers, which were screened against inclusion criteria and prioritised in terms of their relevance to the research questions. This led to a final set of 140 highly relevant papers described in the main synthesis, and a further 108 considered in the introduction, discussion and appendices. We evaluated 'good practice' through considering the evidence provided of assessments' validity.

Key findings

Professionalism

We found many assessments of professional behaviour in simulated practice contexts, using role-player actors as simulated patients. Professionalism is operationalised as both global judgements of interpersonal conduct, and as judgement of specific constituent communication behaviour where the manner of communication is important (including 'complex' communication scenarios, perceived empathy, and interprofessional team practice). We noted that assessments of empathy may not be usefully distinct from assessments of general communication skills, but do include an important patient perspective. We also found assessments which took a substantially different approach, not using observation of behaviour. Paper-based situational judgment tests captured candidates' understanding of issues, reflected by their ability to select appropriate choices.

Good practice is apparent in examples of all types of assessment, but the key decision for future implementation is around what construct is to be assessed – global judgements, specifically defined behaviours, or the ability to demonstrate a professional response. Authenticity of responses in artificial assessment settings is a concern, but there is no indication that this would be more of a risk for professional skills than for practical skills.

There is not a clear construct which is unambiguously defined as 'professionalism'; rather there is a set of constructs and behaviours which are associated with 'professionalism'. The broad concept of professionalism may therefore be more usefully partitioned into more tractable and clearly defined elements such as communication and situated judgements. Which of these are of relevance for assessment is perhaps a matter for more 'top down' policy decisions (albeit evidence-based), following which evidence around specific assessments can then guide how those specific constructs and behaviours may best be assessed.

Ethics

We defined assessments of ethical judgement as those which were concerned explicitly with the application of ethical principles and knowledge. Examples of good practice were demonstrated in concordance tests, where candidates' decision making is compared to experts'.

Ethics assessments may need to be calibrated to candidate populations. Ethical challenges in medicine are universal, but expected standards of performance may vary with level of training and intended level of practice. There is some evidence that ethical practice may be rooted in culture. It is important that assessments in this area are not over-sensitive to cultural differences which may not be vital for patient care, while being able to appropriately identify doctors whose standards of practice are below those expected or required in the UK.

Patient safety

We focused on assessments of patient safety which considered candidates' understanding of safety as a process, rather than a consequence of technical competence. There were few of these, but we found good examples of candidates responding to, and so demonstrating understanding of, error in both practical procedures, and explaining the cause and consequences of errors to simulated patients.

Future developments of script concordance or situational judgement tests may provide useful assessments in this area, but we found no such examples.

The use of technology in assessment

We considered simulation, virtual reality and remote and mobile technology as potential tools for the development or enhancement of novel means of assessment. The use of simulation and virtual reality for technical skills is well-established, although evidence of good practice remains mixed. We found some evidence for the use of 'virtual patients' to present cases online, but the technology reported is somewhat outdated, and current technology may provide a more flexible and authentic platform. 'Virtual worlds' and paradigms adopted from computer games may have potential, but the literature we found indicated little more than proof of concept at this stage. Similarly, there was surprisingly little application of mobile technology given its ubiquity, but this may also betray a lag between development, evaluation and publication.

Sequential testing

Sequential testing is an approach to the structure of examinations whereby only borderline candidates complete a full assessment. Those whose performance is clearly adequate need only complete a reduced assessment. This increases the throughput of candidates and so the efficiency of the assessment. We found only two studies of sequential testing, whereby only borderline candidates complete a full assessment. The evidence from these studies however suggests this does provide a robust and cost-effective approach to assessment, with clear cost savings from reducing the total number of OSCE stations completed by candidates.

General issues

Questions of the content and process of assessment need to be considered for any future development. The necessary content for assessments of professional skills is not immediately apparent, and close consideration of the definition of

assessed skills and performance is necessary. Involvement of stakeholders, including patients, may help to define important criteria of performance. Where scenarios are used, the use of rating scales or checklists may have a profound effect, not just on the statistical properties of measurement, but also the nature of what is being assessed. Written examinations may be more precise and less confounded by behaviour for some constructs.

All assessments have costs, and studies tended not to describe these in detail. Development of assessments should consider one-off and running costs, as well as potential savings in reduced assessment workload, when considering feasibility.

Equality and diversity considerations are largely not addressed in the literature, although there is evidence of differential attainment between home and international graduates when taking licensing examinations in different countries. Risks of bias in assessment design should be considered against the justifiable and appropriate differentiation between levels of performance that may be culturally determined.

Finally, details of standard setting are not addressed in much of the literature. While these decisions are based in technical considerations of measurement, they should acknowledge the conceptual differences between different approaches to definition and measurement.

Contents

List of Appendices

(Appendices are available as a separate document)

Glossary

ABIM	American Board of Internal Medicine, which is one of 24 medical specialty boards that make up the American Board of Medical Specialties. The boards have a role in assessing and certifying doctors who met specific educational, training and professional requirements.
ACGME	Accreditation Council for Graduate Medical Education. The ACGME sets accreditation standards for graduate medical education in the USA.
Angoff method	A method that uses the judgement of subject-matter experts to assess the difficulty of each item in an exam and set cut-off score.
Borderline method	A standard setting method that sets the exam cut-off at the intersect of actual marks with the borderline 'global scores' for all candidates in an OSCE station.
CanMEDS	This is an educational framework describing the competencies of physicians in Canada that relate to medical expert, communicator, collaborator, leader, health advocate, scholar and professional.
COMLEX	The Comprehensive Osteopathic Medical Examination in the USA - a 3-stage national licensing examination for osteopathic medicine.
Content validity	The extent to which the items in an examination are relevant and representative of the construct they are being used to measure.
Construct validity	The ability of a test to measure what it is purported to measure.
Criterion validity	The extent to which a test score varies with a related external variable. This may be concurrent, or predictive.
Cronbach's alpha	A measure of internal consistency – namely, how closely a set of items are as a group.
Decision-study (D-study)	A model that estimates how consistency may be improved with increasing number of stations – ie, determines conditions where measurements would be most reliable.
Generalisability study (G-study)	Model that allows the estimation of multiple sources of error in a measurement process (eg, subject, rater, item)
FSMB	Federation of State Medical Boards – represents medical and osteopathic state regulatory boards in the USA.
Inter-rater reliability	Describes the strength of agreement in the scores between different raters.
ICC	Intra-class correlation coefficient – a measure of reliability.
IMG	International Medical Graduate. Generally, a doctor whose primary medical qualification is from a different regulatory jurisdiction. In the UK it refers to doctors who have qualified outside of European Union, and so are not subject to the EU Recognition of Professional Qualifications Directive (2005/36/EC).
Jefferson Scale of Physician Empathy	A validated questionnaire that captures perceptions of empathy – with separate versions for students, physicians and patients.
Mini-CEX	Workplace assessment that evaluate performance in a clinical encounter.
OSCE	Objective Structured Clinical Examination
OSLER	Objective Structured Long Examination Record
Rasch model	Model that specifies the probability of a right or wrong answer as a function of the respondent characteristics (candidate proficiency) and item parameters (item difficulty)
MCCEE MCCQE RSPSC	The Medical Council of Canada Evaluating Examination (MCCEE) is a prerequisite for eligibility to the MCC Qualifying Examinations (MCCQE). The Royal College of Physicians and Surgeons of Canada provide certifying exams for medical specialists.
Test-retest reliability	Describes temporal stability of scores – how well the tool/assessment produces similar result when administered for a second time.
Sequential testing	An approach to the structure of assessments whereby borderline candidates complete more stations/items than those who are clear passes. This increases reliability of the assessment for the most borderline candidates, while improving efficiency.

1 Introduction and background

Doctors undergo assessment throughout their careers, from progress examinations during medical school, to final qualifying exams, progression through postgraduate training and membership of professional bodies. Doctors may also be assessed when moving between regulatory jurisdictions, or when returning to practice. Some assessments aim to support learning ('formative assessment'), while others provide evaluation of learning or performance at a key stage of progression ('summative assessment').

Assessments encompass scientific and clinical knowledge, clinical and practical skills, and elements of practice termed 'professional skills', including 'professionalism', ethical judgement and awareness of patient safety. This report is concerned with the final group, and considers two research questions:

1. What evidence is there for good practice in the use, or potential use, of summative assessments of professionalism, ethics and competence in relation to patient safety?
2. What evidence is there for the use of simulation or other technologically-mediated methods in summative assessments in medical and non-medical contexts?

The study primarily consisted of a systematic literature review, supplemented by stakeholder consultation through a Project Advisory Group (PAG) and identification of examples of good practice. The research aimed to inform assessments throughout medical careers, including the planned medical licensing assessment for practice in the United Kingdom.

1.1 Contents of this report

This report presents the key findings of the project, with supplementary detail in the appendices. In this chapter we introduce some of the context for the study: the background of a planned licensing assessment in the UK, how 'good practice' in assessment may be determined, definitions of the key content domains in which we were interested, and the types of assessment that may be considered to be novel.

The second chapter summarises the methods used for the literature review, the composition and terms of reference of the expert PAG, and the use of examples of current UK practice.

The results chapters summarise and synthesise relevant literature around the content and type of assessments. More detailed descriptions of practice examples derived from the literature and from PAG members are given in boxes within the findings. Details of areas which were tangential to the research questions are provided in appendices.

Finally, a general discussion and synthesis draws together conclusions from these findings and commentary from the PAG. The implications of the findings for the specification of good practice are discussed, and gaps in knowledge to inform future work are set out.

1.2 A medical licensing examination for the UK

The initial impetus for the project was consideration by the General Medical Council (GMC) of a licensing examination for UK practice. However, our findings have more general relevance for assessments throughout medical careers.

Doctors working in the UK must be registered with the GMC and have a licence to practise. At present, UK graduates are eligible for a licence on graduation from medical school, while the majority of international medical graduates (IMGs) intending to work in the UK must complete the Professional and Linguistic Assessments Board (PLAB) test as a requirement for GMC registration and licensing (https://www.gmc-uk.org/registration-and-licensing/employers-medical-schools-and-colleges/the-plab-test). This consists of a knowledge test (PLAB Part 1) and a practical clinical skills test (PLAB Part 2).

In 2014 the GMC agreed to support the principle of establishing a UK national licensing exam for doctors wishing to practise in the UK. The Medical Licensing Assessment (MLA) will offer a means of ensuring that these doctors have met a common threshold of safe practice. It will be taken by UK medical students before they complete their degree programme, and IMGs who wish to practise in the UK.

The proposed structure of the MLA, agreed by the GMC in 2017 (Hart 2017) follows that of other assessments in medicine, including PLAB, in consisting of tests of knowledge and of practical skills. A common, online applied knowledge test (AKT) will be introduced from 2022, and completed by potential licensees in the UK and abroad. The AKT will be delivered within individual medical schools' assessment programmes, and replace PLAB Part 1. For practical skills, medical schools' own assessments will be assured against newly specified requirements to be set by the GMC, while a new clinical and professional skills assessment (CPSA) will replace the current PLAB Part 2. However, the introduction of a universal CPSA in the longer term has not been ruled out.

Such an assessment must be consistent and equitable, while reflecting that doctors are entering different levels of practice. While many candidates will be entering practice for the first time, international medical graduates may be taking up specialty training or sub-consultant level 'Specialty and Applied Specialist' (SAS), or 'staff grade' posts.

1.3 Good practice in assessment

All assessments should be demonstrably robust and fair, but this is even more important for summative assessment with high stakes outcomes, where passing or failing has consequences for students' or doctors' progression and careers. Such assessments should therefore be based on clear evidence of their rigour and fairness.

Norcini et al (2011) identified three categories of assessments in terms of their supporting evidence base: those where assessment practice is clearly based on an evidence base, those where practice does not yet reflect an existing evidence base and those where there is not yet an evidence base. These may be interpreted as reflecting the extent to which the functioning of an assessment is understood. In this review we have a particular interest in those assessments for which the evidence base is limited.

Norcini et al (2011) also defined seven criteria for good assessments: their validity or coherence, reproducibility or consistency, equivalence across different contexts, feasibility, having an educational effect on individuals, providing a catalytic effect on process, and their acceptability to stakeholders. The relative importance of these criteria varies with the type and purpose of assessment. We suggest these criteria have much in common with the sources of assessment validity described by Downing (2003, following multi-agency policy), and will return to this in the Methods section of the report.

The quality of assessment has also been linked to the level of performance which is assessed. A seminal typology known as 'Miller's pyramid' (Miller 1990), defines competence at four levels: 'knows', 'knows how', 'shows how', and 'does'. These levels represent a shift in assessment from recalled knowledge to situated practice. High stakes summative assessment should be as close to the top of this model as possible, but there are pragmatic limitations on the extent to which real practice can be assessed.

1.4 The content of assessment

Assessment in medicine has historically been focused on ensuring that doctors have the basic scientific and clinical knowledge to underpin practice, and the core competencies to perform practical procedures safely. However, recognition of the importance of other knowledge and skills to delivery of patient-centred care, and a need to assess them, has grown in recent years. The GMC identified three broad domains – 'professionalism', 'ethics' and 'patient safety' as areas relevant to practice. None of these have simple definitions, and here we briefly introduce the background to each domain.

1.4.1 Professionalism

Professionalism, and associated terms such as 'professional skills' and 'professional practice', are specified as assessable outcomes in regulatory guidance in different countries (eg the UK [GMC 2017], USA [FSMB/NBME 2014] and Canada [Frank et al 2015]). However, professionalism is not a simple construct.

Hodges et al (2011) identified three ways in which it has been considered in the literature: as individual traits, interpersonal interactions, and a function of societal–institutional structures or processes. Others have described it as a 'complex system' (Hafferty & Castellini 2010), a sociological phenomenon (Martimiakis et al 2009, Sullivan 2000), and as an aspect of professional identity development (Cruess et al 2016). West & Shanafelt (2007) describe it as an emergent property of personal and environmental influences. Wilkinson et al (2009) distinguished between assessments looking at adherence to ethical practice principles, interactions with patients and colleagues, reliability (in the sense of personal responsibility) and commitment to improvement (encompassing reflectiveness, leadership and advocacy, amongst other dimensions). These reflect fundamentally different underlying concepts, and any consideration of assessment should be clear what is, and can be, assessed.

Reviews of assessment methods for professionalism (Veloski et al 2005, Wilkinson et al 2009, Rodriguez et al 2012, Li et al 2017) have identified many ways in which professionalism may be assessed, but concluded that suitable approaches for high stakes usage are limited. Assessments of attitudes are conceptually problematic for summative use because they at best imply what doctors *may* do based on those attitudes, rather than assessing actual behaviour in practice. On the other hand, assessments of behaviour that are based on performance in real workplace contexts present a problem for doctors or medical students who are not currently working.

A consensus statement from emergency medicine professionals (Rodriguez et al 2012) did not identify any existing non-workplace-based summative approaches. Wilkinson et al's (2009) review identified 33 assessments of professionalism, but only six were neither workplace-based, nor reliant on self-administered or self-report scales. These included OSCEs (Objective Structured Clinical Examinations), simulations and paper-based tests. A series of papers by van Mook et al (2009a, 2009b, 2010) also emphasise the prevalence of workplace-based assessments, although they do note the role of OSCEs and simulated patients.

Our review will take a pragmatic approach to the operationalisation of professionalism, encompassing discrete skills such as communication and empathy, as well as more holistic and unarticulated definitions.

1.4.2 Ethics

Consideration of ethics in medicine has presented a dichotomy between ethical practice as a 'virtue', in the sense of traits or qualities possessed by a doctor, or as a set of skills (Eckles et al 2005). Assessment of the former has used theoretical models of moral development (eg the defining issues test, Murrell 2014), which may be problematic for summative assessment. Like definitions of professionalism based on attitudes, such abstractions may not reliably predict behaviour, and so may not be robust and fair. Our working definition therefore excludes this trait-based definition.

Ethical behaviour in practice has been closely associated with professionalism, semantically, conceptually and practically (eg Boon & Turner 2004). The GMC includes 'maintaining professionalism' within its ethical guidance (https://www.gmc-uk.org/ethical-guidance), which also contains other elements linked to professionalism, such as confidentiality and leadership. Assessments of these areas also illustrate this overlap: a literature review by Lynch et al (2004) identified 88 distinct assessments of professionalism, of which 49 were identified as relating to ethics.

Our definition of assessments relating to ethics will focus on distinct skills and judgements relating to patient-focused ethical practice, explicitly reflecting doctors' ability to identify, evaluate and address ethical issues in a patient's care. More global definitions will be identified with professionalism.

1.4.3 Patient safety

Superficially, 'patient safety' is an integral part of the holistic practice of medicine – all practice should be safe. However, rather than just the appropriate and safe application of medical knowledge and skills, patient safety has emerged as a distinct discipline concerned with awareness of the situational and interactional elements of care which may generate risk to patients (eg Cooper et al 2000). As such it is linked to healthcare initiatives such as quality improvement (Batelden & Davidoff 2007), and to the wider study of human error and the broader field of human factors (eg Reason 1990).

Our working definition of patient safety will therefore encompass a knowledge-based understanding of principles of safe practice, and the performance of behaviours reflecting those principles.

1.5 Types of assessment

The review will also consider different modes or types of assessment as they arise in the literature. Assessments may vary in the object of assessment (eg knowledge or behaviour), the process of assessment (eg through paper- or computer-based tests or simulation-based practice) and in the method by which performance is captured (direct or remote observation; checklists or global rating scales).

In this review we will prioritise novel processes and methods of assessment, for which literature may, by definition, be scarce. This may include simulation and other technology-based methods. Novel approaches to the organisation of assessments, such as sequential testing, where potentially failing students are examined more than those who are clearly passing, are also of interest.

1.6 Summary

This chapter has introduced key concepts and definitions, which will be referred to throughout this report. It has illustrated that understanding of terms is not necessarily definitive, and there remain debates around them, which will be returned to in this report.

In a changing political, clinical and educational landscape, this review sets out to examine systematically the evidence for assessment of a set of core professional skills and behaviours. We will identify and summarise key approaches in the literature and discuss these relative to indicators of good practice with a view to informing processes of development and implementation in GMC assessment strategy.

2 Methods

2.1 Literature review

Medical and non-medical databases were systematically searched in order to identify papers describing assessments of professionalism, ethics and patient safety, and using novel approaches to assessment. This was supplemented by further searches to capture grey literature and unindexed journals. Screening and prioritisation, guided by criteria agreed in discussion within the research team and in consultation with the GMC, reduced the initial set of more than 9900 papers to 248 with some relevance to the research questions. Of these, 140 are considered in detail in the results chapters, the remainder in the appendices and discussion. This final set is very large for this type of review, reflecting the breadth of the search (for example Archer et al [2015, 2016] considered 73 papers, focusing on 23 in detail; Havyer et al [2016] considered 70 studies, and Li et al [2017] considered 80).

Our synthesis adopted an 'evidence mapping' approach (Althuis & Weed 2013). This is suitable for heterogeneous topics and methodologies as found here. The approach involves description of the evidence relating to a broad question, summarising common features and identifying gaps and opportunities for further review or research.

Full details of the search strategy and screening process are provided in Appendix A.

2.2 Project Advisory Group

Alongside the systematic literature review, a Project Advisory Group (PAG) was convened in order to gain advice and opinion from relevant experts in assessment and domain areas. Specific functions agreed were to:

1. Critically review the approach to, and findings from, the systematic literature review.
2. Identify relevant assessment approaches that have not been published.
3. Facilitate access to current examples of good practice.
4. Contribute to interpretation of findings and their practical application for assessing competence, drawing on members' experience.

Ten organisational stakeholders were approached to select or nominate individuals with relevant expertise, and 22 domain and/or methodological experts agreed to participate, with representation across the UK. Two meetings were held by videoconference, with 14 and nine attendees respectively, with email commentary and discussion between and subsequent to these meetings.

The first meeting considered how participants defined elements of 'good assessment', and this discussion informed the approach to prioritisation and data extraction. The second meeting considered initial findings and potential specification of assessment practice. Comments from these meetings have been reflected in the discussion chapter. PAG members also provided comments on drafts of this report.

In addition, a number of PAG members with differing roles, geographical locations and subject expertise contributed to exemplars and discussion of local assessment practice, outlining details of current and developing assessment approaches. These exemplars are incorporated into the results chapter. PAG members are listed in Appendix B.

3 Findings of literature review

This chapter summarises the literature relating directly to our research questions. It includes papers which met two main eligibility criteria through our screening and review process:

1. they describe assessments that are used summatively, or were judged to have potential for summative use, in domains of professionalism, ethics and patient safety, *or*
2. they use novel technologies or present other novel processes of assessment.

Despite the deliberate sensitivity of our search, we found no eligible studies outside clinical professions, and few outside medicine.

Each section begins with a summary of key observations, with validity evidence described in each paper presented in tables. These tables include only those papers which met our inclusion and prioritisation criteria (see Appendix A for further details). Other references identified in the review are referred to, but are not included in tables. Examples of good practice from the literature are presented in green boxes, and from our PAG contributors in blue boxes.

We provide an overall judgement of the amount of evidence presented, analogous to the GRADE system (http://www.gradeworkinggroup.org/) which is used in systematic reviews of healthcare interventions. For each group we assigned a level of confidence as described in table 1. Due to our prioritisation of assessments for which there is less evidence, for most groups this judgement is less than 'high'.

Table 1. Global judgements of evidence based on GRADE system

High	Further research is very unlikely to change our confidence in our conclusions
Moderate	Further research is likely to have an important impact on our confidence in our conclusions and may change them
Low	Further research is very likely to have an important impact on our confidence in our conclusions and is likely to change them
Very low	Our conclusions are very uncertain, eg we found one small pilot study with no formal evaluation

We considered the evidence provided for each assessment against indicators of assessment validity described by Downing (2003, drawing on other sources). This describes five 'sources' of validity which we have interpreted as described in table 2. The conclusions for practice implied by each type of evidence are also indicated.

Table 2. Sources of validity evidence and examples of interpretation

Source	Examples of included evidence	Conclusions for practice
Content	Evidence that the content of the assessment – question areas, scenarios, rating scale anchors, etc – is authentic, and based on evidence rather than arbitrary judgements. This may include evidence of mapping to learning outcomes, involvement of experts/patients in development or validation, or empirical demonstration of practice relevance.	Assessment is authentic
Response process	Evidence that processes of assessment data collection are robust and consistent. This may include evidence of rater training, or of consistency in responses between raters ('inter-rater reliability').	Assessment is consistent and fair
Internal structure	Evidence relating to the statistical structure of assessment measures. This may include internal consistency metrics (Cronbach's alpha), and generalisability studies.	Assessment is reliable
Relationship to other variables	Evidence derived from correlations with other assessment variables indicating convergent or divergent validity (ie measuring the same or separate constructs) or from the presence or absence of hypothesised subgroup comparisons (eg male/female, trainee/consultant).	Assessment is fair Assessment is authentic
Consequences	Evidence is provided of data relating to immediate or distal outcomes of the assessment for examinees and organisations. This will include pass/fail outcomes and standard setting, and any associations between assessments and future outcomes. For organisations, evidence of feasibility, resource requirements, acceptability and scalability/sustainability fall within this category.	Assessment can discriminate Assessment is predictive of performance Assessment is sustainable

Sources of validity are derived from Downing (2003). Examples are interpreted to reflect the evidence available in the studies we have considered.

3.1 Content of assessment: Professionalism

We noted in the introduction that professionalism is a complex construct, identified with a number of elements. Our analysis of the literature identified that some assessments refer to specific contexts or types of communication – communication in sensitive or complex interactions, empathy and interprofessional collaborative practice or team working – behaviour, and these are described in distinct sections that follow. There were a number of assessments which considered professionalism as a more global or holistic construct, and we consider these here first.

3.1.1 Summary of evidence

We identified 20 studies (12 with postgraduate, eight with undergraduate participants) which assessed professionalism in global terms. Most of these were based on the observation of behaviour within scenarios, while two studies considered how candidates addressed written scenarios in paper-based tests. These are summarised in Table 3.

There is a lack of explicit definition of professionalism in these assessments, and so whether they constitute a homogeneous group is arguable, but to the extent that professionalism can be meaningfully defined at this level, the evidence is moderate to high.

There remain concerns though about a lack of specificity and clarity in the use of the term. If a global professionalism measurement is used, it must be clear what it relates to, rather than any vernacular meaning being assumed. This is perhaps demonstrated by more apparently reliable results when professionalism is broken down into more specific constructs (eg Berman et al 2009), although as with most observations contrary indications have also been found (Roberts et al 2011).

To the extent consensus exists around unprofessionalism being a 'behaviour that is inappropriate from a doctor', there are also concerns that such behaviour can be hidden or masked in one-off assessment settings, and that longitudinal, real-world assessments are necessary in order to capture low-frequency, but significant lapses. Our expert advisory group felt this strongly. This is not generally considered in the literature, although Berman et al (2009) found that ratings of professionalism given by training program directors (who have longitudinal knowledge of candidates) before an assessment were higher than, but correlated with, professionalism scores in an OSCE.

Jefferies et al (2007): assessment of professionalism using a competency-based framework

This paper illustrates that professionalism may be assessed in an OSCE by considering simultaneously multiple physician competencies. These were derived from the CanMEDS framework of: medical expert, communicator, collaborator, manager, health advocate, scholar and professional.

The pilot OSCE was in a neonatal-perinatal training programme and included 3-5 competencies in each of 10 stations. For example, a station that involved discussion with a mother about her infant's discharge related to expert, communicator and manager, as well as health advocate. Individual roles were rated on a 5-point, behaviourally anchored, scale.

All stations could assess expert and communicator roles, but consideration of the 'scholar' required particular planning and creativity.

There was moderate-high inter-station reliability for each of the CanMEDS roles, except 'scholar' - most likely due to the differences between relevant stations (one involved teaching of a technical skill and the other teaching of disease pathophysiology). Scores of 2nd year trainees were higher than 1st year trainees for each of the competencies, supporting construct validity.

The paper supports valid and feasible assessment of multiple professionalism competencies in a one-off approach.

There remains a fundamental debate around how professionalism can and should be defined that extends back from assessment to curricula and guidelines. Nonetheless, there is good validity evidence for both observational measures (Berman et al 2009, Roberts et al 2011), and a written paper relating to professional dilemmas (Tiffin et al 2011).

Table 3. Validity evidence for assessment of professionalism

Reference	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Abu Dabrh AM et al. (2016)	USA	PG	56	The instrument was developed, reviewed and pilot-tested, and revised by the study investigators, taking into consideration the ACGME definition of competencies and existing tools used for other OSCE scenarios and competencies evaluation	Live and video rating. SP and faculty training. Good IRR within faculty, less - but fair - between faculty and SP	None given	None given	> 60% outstanding across domains, SP and faculty
Berman JR, et al. (2009)	USA	PG	Not specified	Mapped to core competencies, and developed in practice	Good correlation between faculty and SP ratings	None given	Correlation between faculty and patients, program directors, but differences in absolute scores. Improvement with time identified.	Faculty more positive than trainees
Dwyer T, et al. (2014)	Canada	PG	25	Developed and blueprinted by specialists, based on CanMEDS roles	SP/SHP training	Interstation reliability > 0.8. Alpha > 0.8	Increase with year of training. Correlation with program directors rating. Some correlation with in-training assessment for previous year.	Residents felt scenario authentic, but low agreement good assessment
Hofmeister M, et al (2009)	Canada	PG (IMGs)	71	12-station Multiple Mini Interview. Determined by informal critical incident technique and job analyses	Interviewers attended a 2-hour training session, with demonstration of example interviews and practice using rating scale.	G-coefficient=0.7	Positive correlation with verbal communication scores of selection OSCE and MCCQE II. No correlation with overall OSCE score, MCCEE or MCCQE I.	None given
Jefferies A, et al. (2007)	Canada	PG	24	Scenarios and ratings written by faculty, mapping to CanMEDS roles. Ratings informed by literature	High correlation between examiner and SP/SHP	Alpha on means > 0.8. Alpha for roles variable <0.1 to 0.9	Higher scores in second year than first year	Candidates and examiners felt realistic and fair assessment
Kassam A, et al. (2016)	Canada	PG	63	Mapped to CanMEDS	None given	Checklist alpha all > 0.7	Moderate-high correlation between stations and global measures. Senior and non-IM residents scored higher on professionalism	None given
Kaul P, et al (2012)	USA	UG	289	Developed by clerkship directors and reviewed	SP training and detailed response guide	None given	None given	99% correct on prof'ism
Kaul P, et al. (2014)	USA	PG	47	Developed by program director with expert advisors. Mapped to ACGME competencies	SPs trained and check by independent observer (not stated how). When SP deemed reliable no external rating of OSCE	None given	None given	None given
Moniz T, et al. (2015)	Canada	UG	107	Reflection focused on CanMEDS intrinsic roles	REFLECT rubric based on literature. IRR alpha > 0.7	Low reliability across samples (ICC<0.3), but no difference in means	Divergent from MCQ, but not convergent with OSCE	None given
Neira VM, et al. (2013)	Canada	PG	24 pilot; 50 validation	GIOSAT based on literature and Delphi study, followed by piloting	IRR variable by item but adequate overall (>0.6)	Single measure ICCs for individual intrinsic items (0.36-0.69)	Moderate correlations between totals and PGY (0.36-0.42)	None given
Ponton-Carss A, et al (2011)	Canada	PG	14	None given for scenario, but drawing on CanMEDS. Checklists adapted from literature	Examiner orientation	Professionalism alpha=zero Communication alpha=0.69-0.92	No effect of stage of training on comms or prof'ism. Comms high correlations between checklist and GRS. Low for prof'ism. Mixed correlations for surgical and comm/prof scores	Station mean scores high for prof'ism (69 & 79%), but poorer for comms (53-65%)

Reference	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Ponton-Carss A, et al (2016)	Canada	PG	120	Derived from CanMEDS and literature	SP and SN trained and rehearsed	Checklist alpha variable for non-tech.	Convergent validity for non-tech roles, divergent from technical	None given
Roberts WL, et al (2011)	USA	UG	227	COMLEX exam	SP training and assessment	Strong convergent and discriminant validity	Data gathering and patient note scores associated with humanistic skills	None given
Sim JH, et al. (2015)	Malaysia	UG	185	Blueprinting, mapping to course objectives	Stations reviewed and field tests. Training of SPs and examiners	Overall alpha=0.68	Differences between measures (but no correlations reported)	Mean scores satisfactory
Tiffin PA, et al. (2011)	UK	UG	194	None given	Rasch model suggests professionalism items easier than others.	Rasch modelling	Divergence from anatomy and skills. No convergence with Conscientiousness Index.	Prof'ism poor at discriminating between candidates
Yang YY, et al. (2013)	Taiwan	PG	189	None given	IRR concept > 0.8 IRR behaviour > 0.49	Retest and internal consistency alpha > 0.7	No difference with gender on OSCE	None given
Zanetti M, et al. (2010)	USA	UG	20	Derived from ABIM	Rater variance components (and interactions) > 30%	Generalisability < 0.8	SP raters are less reliable than expert OR lay	None given
Zhang X & Roberts WL (2013)	USA	UG	4,564	COMLEX exam	SP training and assessment	Rasch analysis indicates reliability	None given	None given
Schubert S, et al. (2008)	Germany	UG	NA	Based on legal requirements and literature. Detailed account of correct item selection.	None given.	NA	NA	NA
Pugh D, et al. (2015)	Canada	PG	35	Blueprinted from college requirements, cases written by experts. Checklists by iterative agreement between experts.	Faculty and SP training and calibration. Pre-exam orientation for candidates and examiners. Distribution of examiner ratings narrower after training. G-study showed difference between stations by 'track'.	Item-total correlations low-moderate. G = 0.76 for NTS. 7 Stations required for NTS by D-study.	No divergence between tech and NTS (r=0.76 overall). Senior scored higher. No association between NTS and non-procedural OSCE.	Examinees felt more valid than model alone

Key:
BBN- Breaking Bad News; EOL= End of Life
UG= undergraduate; PG=postgraduate
SP=Standardised or simulated Patient
SN=Standardised Nurse
SHP=Standardised Health Professional
MCCEE=Medical Council of Canada Evaluating Examination
MCCQE I=Medical Council of Canada Qualifying Examination part I
MCCQE II=Medical Council of Canada Qualifying Examination part II
ICC= Intraclass Correlation Coefficient
IRR=Inter-rater reliability

3.1.2 Details of evidence

Professionalism as a holistic construct

In these examples, professionalism was assessed as a global or holistic construct, albeit in conjunction with other areas, including communication and practical skills. The approaches of Kaul et al (2012, 2014), Sim et al (2015) and Berman (2009) were similar, with professionalism and communication assessed as part of summative OSCE stations alongside other domains, such as physical examination and history taking. The content of these was established by blueprinting, mapping to learning objectives or expert review/Delphi study. Raters assessed performance on checklist items which may have a simple binary response, where the behaviour is observed or not (eg Sim et al 2015), or a scaled response where the extent of quality of the observed behaviour is rated (eg Kaul et al 2012, Kaul et al 2014). Other assessments used 'global rating scales' where higher level constructs are evaluated on a Likert-type scale (eg Berman et al 2009, Roberts et al 2011). These are less linked to questions of whether specific behaviours are present, but on the evaluation of a higher level, more abstract descriptor. While there can be some overlap, checklists reflect directly observable behaviours, while global rating scales reflect constructs that are not directly observable.

Berman et al (2009) initially used a single item to measure professionalism, but found this did not exhibit expected relationships with other variables – specifically correlation between OSCE raters and programme directors who observed candidates in the workplace. Berman et al revised their measure as a 12-item scale within four domains (respect, responsiveness, clarity in communication, and competency in patient interactions). With the modified measure, they found good correlations. This suggests that a more specified construct may be beneficial for consistent measurement. There were still differences in means between programme directors, faculty and SP raters, suggesting interpretation and calibration of measures is variable, but the correlation suggests similar aspects of performance are being assessed. However, the revised scale did not use the term 'professionalism', raising the question of how far a measure can move, semantically, from its purported construct before it ceases to be a measurement of that construct.

However, by contrast, Yang et al (2013) described a formative OSCE that used separate 'behaviour-based' and 'concept-based' checklists. The former contained detailed behavioural examples within three domains, while the latter were described as being based on 'attitude' and 'perception' of professionalism, and used higher level descriptors (eg 'Follows through on task', 'Is patient') more akin to a global rating scale. Both used a 3-point scale (pass/borderline/fail). Internal consistency was higher with the two scales combined, but inter-rater reliability was *lower* for the behaviour-based checklist, suggesting that broader (concept) categories may be interpreted more consistently.

Roberts et al (2011) and Zhang & Roberts (2013) provided detailed and technical statistical considerations of the internal structure of a 'Global Patient Assessment' tool, which combines ratings of professionalism, communication and interpersonal skills at an abstracted level. The technical details of their analyses are beyond the scope of this review, but they concluded that the tool provides a reliable measurement of a singular underlying construct, though with weak discrimination.

Abu Dabrh et al (2016) illustrated the multi-faceted nature of professionalism with elements that included communication, teamwork and patient safety. A scenario, developed with consideration of the Accreditation Council for Graduate Medical Education (ACGME) competencies and existing OSCE scenarios, and feedback from SPs and residents in a pilot, was based around a role-player playing a standardised nurse with a needlestick injury, who has been given an incorrect medicine for HIV prophylaxis. The resident must respond to this error. Six domains were assessed: the context of discussion, communication and detection of error, management of the error, empathy, use of electronic resources, and a global rating. Trainees were assessed by the role-player and remotely by faculty using a 3-point checklist (outstanding; satisfactory; unsatisfactory). No statistics were reported on relationships with other variables.

A different approach to assessment of professionalism as a global multifactorial construct was described by Hofmeister et al (2009). This was an example actually used in selection for postgraduate residency, so while not strictly summative it was high-stakes. The professionalism of international medical graduates applying to a family medicine residency programme was assessed using a Multiple Mini-Interview (MMI) format. Ten stations involved situational questions

based on common patient scenarios, such as a case conference called by the family of a patient with dementia who had wandered off and been missing for some hours. The applicant is asked to explain 'using the whole team, how would you manage the meeting to resolve the situation?'. Assessors, who included faculty, family doctors, family medicine residents and community members rated candidates on a 10-point scale on five global items: 'ability to understand and address the objectives'; 'interpersonal skills'; 'ability to function effectively in family medicine residency'; 'suitability to family practice' and 'overall performance'. Validation analyses included a G-study (G-coefficient=0.7) and D-study that indicated optimal reliability with a single interviewer and 12 stations. There was no bias due to applicant age, gender, years since medical school completion or language of medical school (even though the IMG candidates had a large number of different first languages). Criterion validity was suggested by positive correlations with the communication stations in the selection OSCE and scores in the Medical Council of Canada Qualifying Examination part II (MCCQE II).

The CanMEDS framework

The CanMEDS framework is a document used in Canada to specify the skills and qualities required of doctors (Frank et al 2015). As well as specific competencies, it includes seven 'intrinsic' competencies or roles, high level descriptors of what a doctor should be: Medical Expert, Communicator, Collaborator, Manager, Health Advocate, Scholar and Professional.

Several studies have described assessment against these high-level roles, in different ways. Jefferies et al (2007) used a single global item for most of these, except for 'communicator', which differentiated between communication with patients and other healthcare professionals, and 'professional', which differentiated between attitudes and ethics, while Neira et al (2013) and Dwyer et al (2014) both used more specific items within each role using scales of different lengths. Kassam et al (2016) took a different approach with stations designed to elicit aspects of two roles at a time (one primary and one secondary), and scoring on station-specific checklists. While validity evidence (in terms of internal structure and response process) was reported by all these studies, the details were variable – illustrating that psychometric results are specific to particular measures, even when they draw on similar content.

Pugh et al (2015) included the professional, communicator, collaborator and medical expert roles in all stations of a procedural skills OSCE. While focusing on procedural competence, they recognised that such competence is situated in a complex environment. Each of professional, communicator and collaborator were assessed on two global scales, with discrete behavioural anchors. However, these scores were not considered in isolation, but as components of a total OSCE score.

The Objective Structured Performance Related Examination (OSPRE) described by Ponton-Carrs et al (2011, 2016) also used scenarios designed to capture the possible tension between practical and professional skills in practice. Scenarios were based around communication with patients and colleagues while performing procedures that required professionalism. In the 2011 study professionalism was specifically operationalised in dealing with potential interprofessional conflict: a 'territorial and stressed' anaesthesiologist and a 'distracting' nurse. In the 2016 study however the role of 'professional' was not defined as part of these scenarios, but 'collaborator' was – illustrating perhaps the semantic difficulties around professionalism. Assessments were by checklists and rating scales. The professionalism scale was revised between the studies to remedy low internal consistency. The 2016 study demonstrated low inter-station reliability, indicating clear context specificity for all professional and technical skills.

Type of assessor

Most assessments were carried out by clinical faculty, or by the standardised patients involved in scenarios. Zanetti et al (2010) compared ratings provided by experts, SPs, and lay raters who observed the scenario without playing any part in it. The assessment drew on the American Board of Internal Medicine (ABIM) core set of professionalism attributes and included 21 items relating to interpersonal elements of practice under headings communication, trust, mannerisms, grooming, demeanour and professional manner. Using a generalisability analysis, Zanetti et al found that differences between raters made a substantial contribution to variance, indicating a lack of inter-rater reliability. This effect was least for the lay raters compared to SPs and experts. Overall, SP ratings were less reliable than other groups. They attributed this to either a partial attention to the construct by SP raters, or an ill-defined construct. Abu Dabrh et al (2016) also found that inter-rater reliability was good for faculty raters, but less so for SPs. These findings indicate that the underlying constructs being measured in these assessments may not be consistent between assessor populations.

'Paper-based' assessments of professionalism

While the role-player-scenario approach is dominant among assessments which explicitly define professionalism as one of their targets, we found three papers which used different approaches.

Moniz *et al* (2015) considered the reliability and validity of reflective writing as a meaningful assessment strategy for third year medical students. They used a published instrument for measuring 'reflective

Standardised patients in assessment

Standardised patients [SPs] are commonplace in undergraduate OSCE exams, but the degree to which they actively contribute to assessment content or process is more variable.

PAG members from **Manchester, Leicester** and **Belfast** gave detail about how SPs are involved in their assessments. In these centres SPs may contribute to scenario design, both in earlier course years (**Manchester**) and Final examinations (**Leicester**). They also provide formative feedback to students in **Manchester**, while in **Leicester** and **Belfast** SPs contribute to students' OSCE station score.

Patients may be either volunteer or professional actors. In **Leicester** they are selected in a rigorous process and are recruited from as broad a range of ages and ethnicities as possible. Once employed, they are required to attend training sessions ahead of participating in a circuit. In **Leicester** children or adolescents are not currently involved as SPs, but may be with chaperones in **Manchester**.

Cost of SPs is estimated in one centre at around £1500 for 1 OSCE station for 1 day (6 circuits), including training.

Content: SPs participate in a wide variety of content areas – in complex practice situations and when patient/family/carer emotions may be running high. For example, **Leicester**, run a 'diversity' scenario, where a lifelong smoker, who has just lost her partner to lung cancer, attends for routine asthma check-up, and is both tearful and wary of a 'lecture'.

In **Manchester** medical specialist trainees (ST) are trained to portray a standardised professional. For example, students are assessed on handover skills where the ST acts as the Foundation doctor who is keen to head home as soon as possible.

Process: In **Belfast**, SPs score students' overall performance on a global rating scale, while in **Leicester** the SP gives a global rating (5-point scale –fail to excellent), as well as rating individual checklist items (involving communication skills, confidence and trust in the doctor), and providing written feedback. In **Leicester**, the SP scores have improved reliability. Many of the SPs have worked with the university for years and across all student year groups, which may support reliable judgements.

Feedback: In **Leicester**, students receive SPs' scores and comments, as well those from examiners. In **Manchester**, feedback has been enhanced by the use of iPads, which afford opportunity for real-time detailed comments on a range of professional issues encompassed by the students' performance at the stations.

Good practice - authenticity

Real-world practice involves clinical encounters that trigger emotional reactions which challenge – and test - a doctor's professional behaviours. ***'It's very easy to be nice to a patient who is being nice to you!' (PAG member, RW).*** SPs can be trained to present such emotional responses in a consistent and standardised manner. Moreover, they are also well placed to assess how they experience the doctor's performance in that situation. Training is essential, but can be delivered feasibly and robustly with input from experienced faculty.

capacity' based on samples of writing. Four trained raters scored four samples, and while inter-rater reliability was high, indicating a valid response process, inter-sample reliability was low, indicating a volatile or inconsistent construct. Analysis indicated 14 writing samples would be required to achieve reasonable reliability, which limits feasibility.

Two papers described the use of situational judgement tests for the assessment of professionalism, where answers are selected based on contextual information provided in a written scenario. The example described by Schubert et al (2008) does not provide any assessment data, just the development of the approach, but does provide evidence of validity in the development of content, including two approaches to identifying best answers through rating or ranking by experts. Tiffin et al (2011) considered in detail the internal structure of a professionalism examination compared with questions on anatomy and applied skills. They found that the professionalism items were less difficult, and less discriminatory than those on anatomy and skills. In addition, there was no correlation between professionalism scores and another workplace-based tool – the longitudinal 'conscientiousness index' (McLachlan et al 2009) – indicating a lack of convergent validity and so a difference in concepts being assessed.

3.2 Complex Communication

We identified several aspects of communication as reflecting professionalism. Communication is the core of professional practice, and communication in complex, challenging and unpredictable situations is when it may be most put to the test. The lessons from the literature on complex communication are therefore relevant to all forms of professional communication. Details of assessments of more 'basic' communication (those that are routine and often protocol-driven), where professionalism may be less salient, are given in Appendix D.

3.2.1 Summary of evidence

Overall, we rate the weight of evidence in this area as moderate to high. There is a relatively large amount of evidence, demonstrating consensus in the overall approach to assessment being simulated scenarios, but clear conclusions are not apparent in the details of approaches.

Authentic assessment of communication requires assessment of skills at the 'shows how' or 'does' levels of Miller's pyramid – ie behavioural measures. While knowledge of communication protocols could be assessed, the quality of communication is best assessed from performance – indicating simulated or real patients. The use of simulated patient-based encounters in all 23 papers found in this area seems justified. Six of these described undergraduate, and 15 postgraduate assessments. Two involved both groups, with one also including senior clinicians.

Mortsiefer et al (2014): a dedicated complex communication OSCE

This paper illustrates an approach to assessment of complex communication issues in a dedicated summative undergraduate OSCE.

Issues included managing the guilt and shame of a patient attending A&E with multiple bruises due to domestic violence. Other stations considered breaking bad news, communicating with an aggressive patient and shared decision-making about steps to reduce cardiovascular risk in primary care (drawing on a computer-based decision tool).

Assessments used a 4-item Global Rating Scale, rated as having 'good' usability, and with advantages over a commonly used validated communication checklist that was regarded as time-consuming and detracting from observation of the student. Assessors all had 2 hours training, include rating and discussion of 2 sample videos portraying good and bad student performance.

The approach was validated comprehensively across 3 student cohorts, including details of psychometrics, standard setting and time for delivery (330 hours per cohort; administration, 390 hours; preparation, 462 hours). Assessments showed high (within) station and low (between station) item reliability, suggesting effects of case specificity. In particular, it highlights that selection, training, and re-training of assessors are key, especially when using a global rating method.

The key decision for effective and valid assessment here is in the content of scenarios, and specifically, ensuring that appropriate levels of complexity for different points of practice are authentically represented (eg Stroud et al 2009). This authenticity can be determined by expert consensus, but blueprinting can ensure clarity of focus. Decisions on content should consider evidence that communication competence may be highly context-specific (eg Balzora et al 2015), and so test cases should be designed to present different and complementary challenges. Decisions of content also extend to what behaviours are being assessed. There is some evidence that non-verbal behaviour is important for effective communication, and so should be included (eg Mortsiefer et al 2014, see also Collins et al 2011).

The way in which performance is captured, and by whom, shapes the assessment. Tools used to capture performance fall mainly into classes of itemised checklists of discrete behaviours, or global ratings, while assessors are drawn from participating standardised patients or other role-players, and observing expert clinicians or faculty. There is not clear evidence here whether any of these approaches constitute better practice than others, with the process of development, including the training and calibration of assessors (eg Wouda and van der Wiel 2012), being most important. The description of a communication OSCE by Mortsiefer et al (2014) provides a very good example of good practice in this regard.

Table 4. Validity evidence for assessments of Complex Communication

Reference	Country	Group	Scenarios or detail	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Balzora S, et al. (2015)	USA	PG	BBN, disclosure, others	11	Developed from existing cases'	Validated OSCE scales SP training	None given	None given	None given
Chander B, et al. (2009)	USA	PG	BBN, disclosure, others	9	Developed from existing cases and reviewed by local experts	Previously validated	Alpha>0.65	None given	None given
Chipman JG, et al. (2007)	USA	PG	EOL, disclosure	8	Literature and ACGME outcomes, experts	Training for SPs. IRR mostly low within and between groups.	Alpha>0.77	No difference between year groups	None given
Chipman JG, et al. (2011)	USA	PG	EOL, disclosure	61	Cases based on patient encounters	Site and rater training	Alpha > 0.86; G-stat <0.9	No differences between year groups or rater groups	None given
Gorniewicz J, et al. (2017)	USA	UG and PG	BBN	66	Reference to literature and training programme development	SP training and selection process	None given	Pre-post change with intervention	None given
Gude T, et al. (2015)	Norway	PG	Patient fears	62	Tool based on literature	SPs trained for scenario, not for rating. ICC for experts 0.7	Alpha for expert tool =0.91	SP satisfaction was ~70% predictive of expert rating group	None given
Ju M, et al. (2014)	USA	PG	BBN	11	Developed by study team with SP programme	SP training and feedback. Rating based on Kalamazoo. IRR within group not given.	None given.	No difference in faculty and SP scores	None given
Lupi C, et al. (2016)	USA	UG	Pregnancy counselling	46	Designed against guidelines and reviewed.	Piloting. Rater training and practice. IRR high for most items.	Alpha=0.71	Correlations with OSCE and clerkship data mostly low.	None given
Matos FM & Raemer (2013)	USA	PG	Disclosure, Handling grief response	42	Developed by study team (details not given). Two-part scenario with manikin and SP	Two rating tools. Rater training and practice with 6 video scenarios. Cohen k overall 0.7	Correlation between tool 'element' and most constituent 'dimensions'	None given	None given
Mema B, et al (2016)	Canada	PG	BBN (among technical skills)	17	Test blueprint based on specialty and CanMEDS. Expert review. Feedback post-test.	Rater training. Tools from literature or clinical practice and piloted. Inter-rater ICC.	G-study; D-study	Divergent validity	None given
Mortsiefer A, et al. (2014)	Germany	UG	BBN, SDM, aggression, domestic violence	456	Expert development	SP training. ICC 0.38-0.74.	Station alpha>0.8. Overall alpha = 0.6. Projection 5 more stations for alpha=0.8	Convergent validity. Gender difference	Borderline groups method
Parikh PP, et al. (2015)	USA	UG	EOL	389	Scenarios adapted from EPERC website. Checklist and Kalamazoo scale.	None given	None given	Low correlations with trust and Kalamazoo scores. No gender difference	None given
Posner G & Nakajima A. (2011)	Canada	PG	Disclosure	14	None given for scenario. Scoring based on national guidelines	Performance assessed jointly and agreed.	None given	Improvement after training	None given
Raper SE, et al. (2014)	USA	PG	Disclosure	12	None given	Experienced raters. Tool adapted from previous report. No IRR, but faculty scored higher than SPs	None given	None given	Feedback indicated perceived authentic
Reed S, et al. (2015)	USA	PG	BBN	29	Based on GRIEV_ING protocol, literature	SP training. Expert review of process/scales.	ICC within rater < 0.5. ICC across raters > 0.7	Increase post training	30% sample scored <70%

Reference	Country	Group	Scenarios or detail	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Schildmann J, et al. (2012)	Germany	UG	BBN	37	Scale based on literature. Checklist adapted to GRS.	Rater training. IRR for checklist high (ICC>0.8), low for GRS (Friedman test difference between independent raters)	None given	Increase post training. Correlations between checklist and GRS 0/62-0.93. Low correlations between independent and SP raters	None given
Stroud L, et al. (2009)	Canada	PG	Disclosure	42	Scenario piloted. Checklist from study with patient involvement. Participant feedback.	SP training. IRR between SP and independent rater 0.5-0.8 on components. 0.7 overall	Alpha=0.91	No effect of sex, prior experience or training on disclosure	None given
Szmuilowicz E, et al. (2010)	USA	PG	BBN, EOL	49	Not given for scenarios. Tools derived from literature.	IRR good (> 0.5)	None given	Little effect of training	None given
Wong BM, et al. (2017)	USA	PG	Disclosure	49	None given for scenario. Scale referred to literature	SP training	None given	Difference between specialties on some dimensions. Difference between cohorts.	None given
Wong ML, et al. (2007)	USA	UG	BBN and others	213 (pilot), 233 (live)	Blueprinting. Expert review.	Examiner training. Good IRR between experts and non-experts.	Generalisability high if station treated as fixed, moderate if treated as random. Little examiner contribution to variance.	None given	None given
Wouda JC & van de Wiel HB. (2012)	Netherlands	UG, PG, consultants	BBN	110	None given for scenario. Scale based on model from literature. ICC >0.7	Rater training and manual. Scenarios observed twice to set and adjust ratings.	None given	ICC low compared with SP ratings. Difference between novice and consultants. Some other differences with grade.	None given
Wouda JC & van de Wiel HB. (2013)	Netherlands	PG	BBN, demanding patient, tissue donation, treatment restriction	50 (drawn from archive)	None given	Videos observed at least twice.	Inconsistency varies with similarity of scenarios.	Inconsistency varies with training	None given
Bloom-Feshbach K et al (2015)	USA	UG	Comm with patients with low health literacy	57	None given	Piloting, instructions and SP training	None given	Student who completed workshop scored higher	None given

Key:
BBN- Breaking Bad News; EOL= End of Life; SDM=Shared Decision-Making
UG= undergraduate; PG=postgraduate
ICC= Intraclass Correlation Coefficient; IRR=Inter-rater reliability
GRS=Global Rating Scale
SP=Standardised or simulated Patient
SN=Standardised Nurse
SHP=Standardised Health Professional

3.2.2 Details of evidence

This group of papers constitutes one of the largest in this report. Many of the assessments and papers are similar, and so details of each are not described. Rather the similarities and differences are described briefly, with additional detail only where relevant.

All scenarios explicitly describing complex communication assessment involved simulated scenarios with role-players playing standardised patients (SPs), or in some cases, standardised relatives. In some papers content was derived from blueprinting and guidelines, and in two there was reference to patient views (Stroud et al 2009 and Wong et al 2017 both cited work by Chan 2005, which fell outside the date range of our review), but all had prima facie validity in reflecting the types of scenarios doctors may face. These included breaking bad news around diagnosis (Mortsiefer et

al 2014, Chander et al 2009, Gorniewicz et al 2017, Szmuilowicz et al 2010, Mema et al 2016) and addressing chronic or life-changing trauma or illness (Wong et al 2007, Parikh et al 2015, Wouda & van de Wiel 2012). Discussions of end of life care and do not resuscitate (DNR) decisions have been used (Parikh et al 2015, Chipman et al 2007, 2011, Szmuilowicz et al 2010), and breaking the news of the death of a child to simulated parents (Reed et al 2015).

Other scenarios included disclosing a medical error (Mortsiefer et al 2014, Chander et al 2009, Stroud et al 2009, Wong et al 2017, Raper et al 2014) or complications (Chipman et al 2007, 2011, Ju et al 2014, Posner and Nakajima 2011) to a patient. Others included dealing with an angry patient or relative (Wong et al 2007, Mortsiefer et al 2014, Matos & Raemer 2013) or 'obnoxious' colleague (Chander et al 2009), or dealing with issues of sensitivity to the patient such as domestic violence (Mortsiefer et al 2014). Some of these studies also included assessment of basic communication. Some elements of communication may be less obviously challenging but require similar sensitivity, such as pregnancy counselling (Lupi et al 2016), addressing a patient's fear of cancer (Lupi et al 2016) and communicating with patients with low health literacy (Bloom-Feshbach et al 2016). Balzora et al (2015) added an explicit element of cultural competency to scenarios, where the attitudes or responses of SPs were designed to reflect their socioeconomic or cultural background.

Assessment of Complex Communication Skills

Summative assessment of communication skills and clinical reasoning in difficult or challenging circumstances is common in UK medical schools – reflecting the expectations of doctors' skills in real-life practice.

Many schools include at least one scenario in final clinical examinations to assess complex communication skills, but the School of Clinical Medicine, University of Cambridge, has extended this approach to a whole day, 10-station circuit, of structured clinical encounters

This number of stations facilitates assessment of a wide range of relevant outcomes. For example, handling errors, disclosing a missed fracture to an SP, formulating a multi-professional care plan or managing inter-professional conflict in addressing a nurse's concerns that a patient has died shortly after being connected to an intravenous morphine infusion. Routine communication is included, but presented to reflect workplace reality – so, for example, a station where the student has to negotiate a bed for their patient on the Intensive Care Unit over the phone.

Good practice - authenticity

The approach supports authenticity with a range of content that draws on GMC and curricular outcomes to blueprint scenario design.

A number of assessments of complex communication considered different phases of the interaction separately. Reed et al (2015) distinguished three parts to breaking bad news, reflecting pre-amble, breaking the news, and follow up, while Gorniewicz et al (2017) and Schildman et al (2012) described five stages, with further distinction in the middle phase. The tool devised by Matos & Raemer (2013) comprised four elements for error disclosure and six for handling grief. Each element had multiple dimensions, for example, 'listens actively and patiently' was a dimension of the 'posture towards patient' element of the 'handling grief' tool. While many approaches to communication considered similar progression through an interaction, not all assessed each phase separately.

Chipman et al (2007, 2011) described a simulated 'family conference' with relatives in two scenarios – end of life and disclosure of complications. In their initial pilot they included a novel temporal element, in which the end of life scenario unfolded across four separate conversations, separated by breaks. While they concluded that the necessary behaviours could be observed in a single session, the simulation of passing time, and so changing family needs was an interesting aspect of authenticity.

As in the previous section, the response process in these assessments varied between checklists and rating scales. While producing ostensibly similar numerical scores, these reflect different constructs in the type of judgement assessors are being asked to make. Checklists may include a simple binary judgement of whether specific behaviours are observed or not (eg Szmuilowicz et al 2010), or a judgement of how complete or well performed a behaviour is, thus allowing greater differentiation of performance (eg Stroud 2009, Wong et al 2007, Wong et al 2017, Chander et al 2009).

Checklists relate to observable behaviours, 'global rating scales' (GRS) generally reflect higher level judgements of performance. For example, the summative OSCE described by Mortsiefer et al (2014) used four global rating scales: 'response to the patient's feelings and needs', 'degree of coherence in the interview', 'verbal expression' and 'non-verbal expression'. Matos & Raemer (2013) used a seven-point scaled checklist for each of the items (element and constituent dimensions) in the disclosure and grief instruments, while some assessments used a combination of measures, for example, Schildman et al (2012) included a 22-item checklist scored with a scale response, and a global scale for each of five domains. Wouda & van de Wiel (2012) used a scaled checklist for expert rater responses, with behaviours mapped to dimensions of 'Control, Explaining, Listening and Influencing', but a GRS for SP ratings.

Overall, 14 studies described checklists, two global rating scales and five both. The choice of a checklist or global rating scale may change the nature of the construct being assessed. Whether an observable behaviour or non-observable construct is being assessed may have different cognitive overheads for assessors in terms of the judgement required. Both are open to threats to response process validity if raters' understanding of items and calibration of judgements is not consistent. The training of raters is therefore an important element to ensure the response process is consistent. Several studies referred to such training, but few gave any detail (although Mortsiefer et al [2014] and Wouda and van de Wiel [2012] provide particularly detailed examples). Questions of the relationship between checklists and rating scales are not limited to this domain and will be returned to in the overall Discussion.

Inter-rater reliability was reported for most tools, but varied between assessments, with some low and others high. This may be a function of internal structure, or of response process. This question is further complicated when raters are drawn from different populations. Some studies found inconsistent patterns of difference between SP and clinician ratings (eg Ju et al 2014, Schildman et al 2012), while some noted that clinical raters gave higher scores than SPs (eg Raper et al 2014). On the other hand, Stroud et al (2009) found the agreement between SP and expert rater groups to be sufficiently high that Wong et al (2017) included only SP ratings in their study using the same measure.

Both Matos & Raemer (2013) and Wouda & van de Wiel (2012) reported acceptable inter-rater reliability between *researchers* using a checklist tool. However, Wouda & van de Wiel (2012) found low agreement between researchers' scores and ratings given on a different, global, scale by SPs. Conversely Gude et al (2015) found that SPs' *satisfaction* with residents' performance was predictive of expert raters' classification of them as acceptable or unacceptable (as derived from a 14-item checklist). Note that apparently contradictory findings between papers may result from differences in constructs and statistical methods, rather than in differences in performance. The question of differences between lay and clinician raters will also be returned to in the Discussion.

Differences between clinician raters have also been studied. Wong et al (2007) found good agreement between clinician raters from the specialty portrayed in a scenario, and those from a different specialty, meaning same-specialty raters did not have to be found for all assessments. Mema et al (2016) reported good inter-rater reliability for rubric (checklist) based scores from medical and non-medical raters (drawn from nurses, respiratory physiotherapists and social workers).

There was some consensus that communication skills are case-specific, which leads to lower inter-station reliability (eg Mortsiefer et al 2014, Chipman et al 2011). There may be some transferable elements, but communication skills are not a single transferable set, and so assessments must account for the different requirements for these skills in different clinical cases (Wouda and van de Wiel 2013 provide a detailed account of this).

Some studies statistically projected the number of stations necessary for reliable measures, with implications for feasibility. For example, while many OSCEs used 6-10 stations, Mema et al (2016), based on data from an eight-station OSCE, showed that 17 stations would be needed to achieve good generalisability, with potential implications for feasibility.

Criterion validity was tested in comparisons between groups, with an expected association between training stage and performance. Chipman et al (2011, see also Gorniewicz et al 2017) found no such discrimination between medical students and residents. However, they concluded this was not an absence of validity, but that any assumptions of difference in skills may be unfounded. Wouda and van de Wiel (2012) found a significant difference only between novices and consultants, rather than intermediate stages. They concluded that this reflected an effective plateau in

performance even with a great deal of experience, with measures stalling in the mid-range of the scale. The lack of such discrimination may be problematic for summative assessments, but this would vary with details of distributions, content and standard-setting approaches of particular assessments rather than being necessarily a generalisable finding.

A comparison of scores for male and female candidates was reported by some studies. Mortsiefer et al (2014) inferred criterion validity from a gender difference, based on established findings in the literature. Stroud et al (2009) and Parikh et al (2012) on the other hand found no such difference. While concordance might be indicative of consistency with that literature, the absence of such an effect is not necessarily problematic.

Some studies reported correlations between their communication assessments and other scores. Mortsiefer et al (2014) found high correlation with another communication scale, indicating convergent validity. Mema et al (2016) found a strong association between their OSCE score and workplace assessments. Parikh et al (2015) found moderate, albeit significant, correlation between communication scales and overall OSCE scores.

Finally, validity derived from the consequences of assessments (in the sense of Downing 2003) is inferred from details of passing scores for individuals, and acceptability and cost from an organisational perspective. Details of passing scores can indicate the extent to which an assessment can discriminate good and bad performance. Mortsiefer et al (2014) reported a consistent cut score across three years, indicating consistency in the construct. Although not a summative assessment, Reed et al (2015) noted in their study that many students did not reach 70%, which may be indicative of a highly discriminatory tool for identifying only the best performers. Only Mema et al (2016) reported costs of implementing their assessment, which were minimal in the hire of two SP actors. However, costs associated with multi-professional faculty assessors were not calculated.

3.3 Content of assessment: Empathy

A number of papers considered the assessment of empathy as a discrete element of communication. While we considered complex communication as constituting the clinical scenarios in which professionalism may be demonstrated, empathy potentially provides a more precise construct on which to judge doctors' interpersonal performance. We define it as a doctor's ability to understand, and demonstrate understanding, of patients' feelings and perspective (Macnaughton 2009).

3.3.1 Summary of evidence

Overall, we rate the weight of evidence in this area as moderate. As with complex communication, there is a relatively large amount of evidence, but no clear examples of good practice. All nine assessments found in this area (all in undergraduate contexts) focused on the perception of empathy by patients (SPs), and imply empathy needs to be demonstrated behaviourally in simulated practice. Papers are summarised in table 5.

O'Connor et al (2014): assessment of empathy

This paper illustrates that simulated patients (SP) may make a valid assessment of empathy.

In a summative OSCE at the end of an undergraduate psychiatry module, empathy was assessed by both SPs and consultant psychiatrists in each of 4 stations using a single item 5-point Global Rating of Empathy (GRE) scale. Students self-assessed using the validated 'Jefferson Scale of Physician Empathy—Student Version'. SPs gave higher scores than examiners, and scored female students higher than males. But, their scores more closely correlated with students' self-assessment than those of the clinical examiners.

The findings suggest that SPs may be more valid in their assessments than the third person clinicians who are observing, rather than participating in the interaction, but a tendency to more lenient marking should be factored into training processes.

However, the first question is whether empathy needs to be assessed as a distinct construct. The scenarios in the previous section may all be expected to elicit empathic communication, but the assessments used were not precisely focused. If blueprinted behaviours or global evaluations indicate communication is of an acceptable standard at a functional level, is the isolation of empathy as a construct important? We found some evidence that empathy is not clearly distinct from generic consultation skills (Ogle et al 2013, McTighe et al 2016). This lack of divergent validity suggests that while empathy is important, its isolation as a discrete element of communication is functionally difficult.

In order to minimise this redundancy, assessments of empathy need to be distinct, and divergent, from other assessments of communication. In this, the contribution of patients to defining content, and which behaviours are assessed, is perhaps more essential than in other aspects of communication. Empathy is essentially how the doctor makes the patient feel, and so patients' definitions would seem to be most pertinent. We found examples of such patient involvement (eg Chen et al 2015), but drawing on earlier work rather than developing content directly with patients.

Even with authentic content, the selection of an assessment measure, and the training of raters is essential. The choice of checklist or global scale is again an important distinction. Assessments of empathy by global judgements were found (Chen et al 2010, O'Connor et al 2014, Wright et al 2014). High level domains may be more transferable between scenarios and contexts than behavioural checklists, regardless of their psychometric properties. For example, 'Did the student understand your concerns?' is applicable to any scenario, 'Did the student maintain eye contact?' may not be relevant to all physical examinations. While non-verbal behaviour was found to be important by some studies, the meaning or acceptability of some behaviours (such as eye contact and arm-touching) may be susceptible to cultural or individual differences. Isolating these effects to ensure fairness, while retaining the conceptual integrity of a subjective construct, will be a challenge for high stakes assessment.

While there is not a definitive example of good practice, examples with the most comprehensive validity evidence are Sennekamp et al 2012, O'Connor et al 2014 and Chen et al 2015. These examples do not provide a template for such assessments, but do indicate good process for assessment development and evaluation.

Table 5. Validity evidence for assessments of Empathy

Reference	Type of assessment	Type of response	Rated by	Country	UG/PG	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Berg K, et al. (2011)	Role-play scenario	JSPPPE GRS	SP	USA	UG	248	Scenarios - faculty committee. JSPPPE - from literature. GRS - none given	Rater training	None given	Women > Men White > Non-white	None given
Chen DC, et al. (2010)	Role-play scenario	GRS	SP	USA	UG	325	None given	SP training and experience	None given	None given	None given
Chen JY, et al. (2015)	Role-play scenario	JSPPPE GRS	SP	Hong Kong	UG	158	None given for scenario. Tool based in literature	SP training and experience	Alpha>0.9. Factor analysis	Convergent high, divergent moderate	None given
Deladisma AM, et al. (2007)	Role-play scenario	Scaled checklist GRS	Expert	USA	UG	84	None given	None given	Alpha > 0.6	Difference between SP and VP groups. Empathy associated with non-verbal behaviour	None given
McTighe AJ, et al. (2016)	Role-play scenario	JSPPPE GRS	SP	USA	UG	717	Scales from literature	None given	Alpha=0.76	Increase in scores from first to second year. No change to third year.	None given
O'Connor K, et al. (2014)	Role-play scenario	GRS	Expert SP	Ireland	UG	163	None given for scenario. Global rating based on literature.	Information provided to SPs. IRR high on correlation, but SPs rated higher.	None given	Correlation between SP and examiner ratings > 0.7 Woman > Men on SP rating, not on examiner rating. Effect of rotation order on SP rating. Some concurrent correlations.	None given
Ogle J, et al. (2013)	Role-play scenario	GRS	Expert	Australia	UG	57	Scale derived from literature	None given	None given	High observed empathy associated with higher rated clinical competency	None given
Sennekamp M, et al. (2012)	Role-play scenario	Binary checklist	Expert	Germany	UG	371	Tool: Expert group and piloting.	SP and examiner training, including video and manual. IRR mostly high.	Test-retest mostly high.	Prepared' scores > 'unprepared' scores from pilot	Overall test is easy (min is 50%)
Wright B, et al. (2014)	Role-play scenario	GRS	Expert SP	UK	UG	133	Derived from literature	IRR variable by station.	Alpha=0.74	Divergent – no correlation with skills-based OSCE Convergent – correlation with communication-based OSCE. Low correlation with OSLER.	None given

Key:
UG= undergraduate; PG=postgraduate
ICC= Intraclass Correlation Coefficient; IRR=Inter-rater reliability
SP=Standardised or simulated Patient; GRS=Global Rating Scale

3.3.2 Details of evidence

Empathy may be assessed as the holistic judgement of those receiving care, or as the verbal or non-verbal communication which may elicit that judgement. In these papers, scenarios may not be explicitly designed to elicit empathy, but measurement approaches included focused content. In these examples we again found use of both global rating scales (GRS) and checklists of observable behaviour.

Few papers reported empirical sources of content validity, but Chen et al (2015) drew on the literature around patients' perceptions of the important elements of empathic practice, such as 'letting the patient tell their story', 'really listening' and 'being interested in the patient as a whole' to develop an assessment scale. This 10-item GRS showed good internal structure, convergent and divergent validity. However, it was used for only one SP rating of each student in a formative context, and the authors cited earlier work that suggested a need for 50 patient raters in higher stakes assessment (Mercer 2005), which implies limitations for practical use.

Chen et al (2010) used a single-item rating of empathy. This demonstrated criterion validity, as shown by a difference between year groups, but the tool was felt to lack content validity by not indicating which behaviours contributed to scores. This raises a question considered in the professionalism assessment of Berman et al (2009) – at what point does decomposing a semantic label lose the specific sense of that label?

Two studies also considered the behaviours which constitute empathic communication in more detail. Sennekamp et al (2012) assessed seven verbal and non-verbal elements, albeit of varying precision, on a binary checklist: eye contact, mimic, body language, appropriate distance, respectful, atmosphere and understanding for patient. Here content validity was established from expert review and pilot testing. The paper reported good inter-rater and test-retest reliability, while criterion validity was inferred from higher scores observed among those who had had training in a communication course.

Deladisma et al (2007) similarly included non-verbal behaviours among dimensions identified by experienced clinician raters. Specific behaviours were eye contact, body lean, head nod, along with and more general items rating immersion level, anxiety, attitudes, empathetic comments and question clarity. These were rated on a four-point checklist, domain and overall global rating scales. Internal structure was indicated by moderate to good internal consistency across all elements, and convergent validity by positive correlations between the non-verbal elements and the global empathy measure.

Several studies reported use of standardised patients as assessors. O'Connor et al (2014) examined SPs' rating of empathy in an undergraduate psychiatry OSCE using a 5-point GRS. No evidence was given for content validity, but a reliable response process was indicated by high inter-rater reliability. Validity from relationships with other variables was mixed: a high correlation between expert examiners' and SPs' assessments suggested a shared construct, but SPs scored students higher than experts, suggesting different calibration of the measure. Female students were scored higher than male counterparts, inferred as reflecting criterion validity. O'Connor et al also noted that experts' assessments of empathy in the OSCE stations correlated with all of the other summative assessments (overall OSCE score, continuous assessment, MCQ, and reflective essay) while SPs' scores correlated only with the overall OSCE and reflective essay scores. This suggested a lack of divergent validity of the empathy measure for clinician raters. O'Connor et al suggested that clinical examiners may be more influenced by their perceptions of students' knowledge of psychiatry, and so deriving scores from an overall appraisal.

Wright et al (2014) found no differences between SP and expert raters in a specific empathy score in an OSCE (a five-point global scale, with behavioural descriptors, across four stations). However, inter-rater reliability varied by station, suggesting possible case-specificity.

A risk of bias in even ostensibly objective measures of empathy is problematic for its use in assessments. A study by Berg et al (2011) reported a study considering the effects of student gender and ethnicity on SP assessment of empathy using a validated instrument (the Jefferson Scale of Patient Perceptions of Physician Empathy) and a five-point single-item global rating of empathy. They found that SPs assessed empathy significantly higher in female compared to male students, which could be interpreted as indicating criterion validity, as by O'Connor et al (2014). However, that this

may be a problematic assumption is indicated by a difference in ethnicity, with non-white students scoring lower than white, possibly because of cultural mannerisms (including accent) leading to more negative assessment of communication skills (echoing a finding of an earlier study that same-ethnicity raters rated IMGs more highly [Van Zantan et al, 2004]).

This has implications for the potential fairness of judgements, and so usefulness as an assessment. Subjective ratings of a construct such as empathy may be confounded by conscious or unconscious bias, and a fair assessment must be able to avoid this influence, while accurately reflecting an inherently subjective construct.

In relation to other variables, Wright et al (2014) found an association of empathy score with overall OSCE and OSLER (Objective Structured Long Examination Record) scores. Ogle et al (2013) compared empathy with clinical skills performance, and found positive correlations with both generic 'process' elements referring to their management of the interaction, and station-specific 'content' scores. Similarly, McTighe et al (2016) noted an association between empathy and communication skills, indicating that empathy may be a core element of communication behaviour without requiring specifically targeted assessment tools.

3.4 Content of assessment: Interprofessional collaboration and team-working

3.4.1 Summary of evidence

The assessments of communication considered in earlier sections were concerned with dyadic doctor-patient interactions. However, communication with colleagues is also an important element of professional practice, encapsulated by the concepts of teamwork and interprofessional collaboration.

The type and quality of evidence in this section is similar to other elements of communication, and overall we rate the weight of evidence in this area as moderate. Questions relating to scenario content, measurement content and response process of measurement are also relevant to this area, and there is a similar lack of clarity on best practice. We found 12 papers in this group, including two systematic reviews. Of the primary studies, all but two were in solely undergraduate contexts.

As with other areas, the central question to ensuring authentic assessment is in defining what is assessed. There is a focus on dyadic interactions between the candidate and another professional, rather than behaviour of an individual within a multi-professional team. While there is evidence that assessment of individuals within teams is possible (Lie et al 2015, Wright et al 2013), this has additional resource implications in requiring additional trained role-players, and also carries the risk that a scenario is less controlled, even with training of actors.

Zabar et al (2016): interprofessional practice skills are distinct

The findings of this formative OSCE for internal medicine residents illustrate assessment of interprofessional collaboration skills.

The multi-station OSCE included a scenario that assessed interprofessional collaborative practice (IPCP) skills that mapped to domains of: values and ethics, roles and responsibilities, interprofessional communication and teams and teamwork. For example, one scenario required the resident to work collaboratively with a standardised nurse (SN) over the telephone to agree a treatment plan for a patient with diabetes and hyperglycaemia.

The SN assessed the resident using a 32-44 item checklist of behaviourally anchored items, which included both generic and case-specific items, each scored on a 3-point (not done, partly done, or well done) scale.

A key finding was that IPCP performance did not correlate with core clinical skills, including patient communication and patient-centredness, suggesting that these skills are a distinct domain of competence.

The details of what can be assessed vary with this high-level focus on dyadic or team interactions. For dyadic interprofessional communication, structured communication protocols provide a framework for scenarios (Adams et al 2013, Foronda et al 2015, Zabar et al 2016). While some team-based activities may also be structured, the authenticity of team working, particularly in ward situations, may be lost if the scenarios are over-prescriptive. Measures of performance will therefore need to be set at an appropriate level in order to capture authentic practice.

For the assessment of interprofessional communication, the example described by Zabar et al (2016) provides detailed validity evidence. For assessment of performance within a team, Wright et al (2013) indicate a potentially robust assessment, but questions of logistics will remain a primary hurdle to implementation.

Table 6. Validity evidence for assessment of interprofessional collaboration and team-working

Reference	Type of assessment	Type of response	Rated by	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Havyer RD, et al. (2016)	Systematic review										
Havyer RD, et al. (2014)	Systematic review										
Adams J, et al. (2013)	Role-play scenario	Scaled checklist	SHCP	USA	UG	168	Uses SBAR and CUS protocols	None given	Alpha=0.7	Divergent from other OSCE scores	35% 'well done'
Dow AW, et al. (2016)	Online collaborative case	Knowledge test	NA	USA	UG	522	Designed by researchers to be authentic	Designed by researchers to be authentic	NA	Individual knowledge score correlated with online activity	None given
Farnan JM, et al. (2010)	SHCP scenario	Scaled checklist	SHCP	USA	UG	31	Scenario developed by faculty and piloted with student. Implicitly mapped to course. Tool based on mini-CEX	Resident-raters trained with scenario materials	None given	None given	None given
Foronda CL, et al. (2015)	Simulation	Binary checklist	Experts	USA and China	UG nurses	229	Item-content validity index calculated based on expert survey.	Rater training and pre-test IRR. Online introduction before session. IRR= 0.79	Alpha > 0.7 (USA) Alpha < 0.6 (China)	Difference between Chinese and US students	Acceptable to educators
Lie D, et al. (2015)	Role-play scenario	GRS	Expert	USA	Faculty	16	Derived from literature	Detailed instructions. G-study found systematic variation between raters	G-study	Faculty variable in identifying high and low performing individuals and teams	Faculty exhibited lenience – errors favoured lower performing
Oza SK, et al. (2015)	Role-play scenario	Binary checklist	SHCP	USA	UG	464	Case developed by authors. Measure based on core competencies.	SPs trained	Alpha > 0.9	Association with self-efficacy. No association with IP experience. Association with patient-centred comms (low divergent validity)	None given
Reising DL, et al. (2015)	Role-play scenario	GRS	Expert	USA	UG medicine and nurse	295	Scenario shaped by learning objectives. Tool developed iteratively from initial observation through expert review and theoretical grounding.	Justification for 5-point scale given. IRR high	Alpha > 0.8	Difference between senior and junior nursing students (medical students not examined)	None given
Saylor J, et al. (2016)	Role-play scenario	Scaled checklist	Expert	USA	UG and PG medicine and nurse	104	None given	Raters trained – video and practice scoring	Values reported from literature on TOSCE tool: internal consistency 0.73-0.87 for 2 raters	None given	None given

Reference	Type of assessment	Type of response	Rated by	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Wright MC, et al. (2013)	Role-play scenario	Scaled checklist	Expert	USA	UG medicine and nurse	38	Based on literature, confirmed by feedback from participants, SPs, raters	Detailed instructions. G-study showed little variance due to raters	G-study showed little influence of internal variables. Reliability moderate	Improvement after training	D-study indicates 12 scenarios required for G > 0.8
Zabar S, et al. (2016)	Role-play scenario	Scaled checklist	SHCP	USA	PG	178	Cases developed by medical and nursing educators based on commonly seen scenarios. Tool based on established competencies.	SN training.	Alpha > 0.77	Divergent validity - no association with other OSCE scores	None given

Key:
UG= undergraduate; PG=postgraduate
IRR=Inter-rater reliability; G-study=Generalisability study
SP= Standardised or Simulated Patient; SN=Standardised Nurse; SHCP=Standardised Health Care Professional
TOSCE=Team Objective Structured Clinical Examination tool

3.4.2 Details of evidence

Havyer et al (2014, 2016) conducted two systematic reviews relating to assessment of teamwork, one in postgraduate internal medicine (Havyer et al 2014) and the other in undergraduate medical education (Havyer et al 2016). Both reviews reported numerous assessment tools, but only some were of assessment of individuals within teams (30 of 73 tools in the 2014 (postgraduate) review; 17 of 64 tools in the 2016 (undergraduate) review). In undergraduate medical education assessments of attitudes to teamworking predominated. The validity evidence reported varied, with most of those cited by Havyer et al (2014) reporting content (54 tools; 74 %) and internal structure (51; 70 %), and fewer response process (12; 16 %), and relationships to other variables (25; 34 %), but there was robust validity evidence for only several tools, and these were setting-specific.

These assessments are typically based on role-player scenarios. Where standardised patients are the basis of doctor-patient communication scenarios, standardised healthcare professionals (SHCPs) are often used in this context.

Oza et al (2015) described a summative OSCE of medical students where a standardised nurse (SN), played by an actor, was present at the beginning and end of a scenario with a standardised patient. While interprofessional communication was not a central part of the scenario, the SN assessed interprofessional communication, collaboration and professionalism on a six-item binary checklist, based on competencies outlined by the Interprofessional Education Collaborative (IPEC 2011). While no other validity evidence was reported, feasibility was suggested by the simplicity and speed of use of the tool.

Zabar et al (2016) also drew on the IPEC model as a source of content validity. They described three scenarios where interprofessional collaboration was integral to the scenario and reflected everyday interactions in practice. Two scenarios required collaborative clinical care, while the third required a resident to challenge a standardised nurse's error. Another example from the same group involved elements of both collaboration and conflict (Adams et al 2013). Authenticity was supported by the need to use existing protocols: 'SBAR' (Situation, Background, Assessment, Recommendation) to present the case, or the 'CUS' format (Concern about situation; Uncomfortable with situation; Safety of patient at risk) for communication of error.

In these studies, a trained SN rated participants on a behaviourally anchored scaled checklist, derived from and mapped to the IPEC framework. Checklist scores did not correlate with other clinical skills assessed, including patient-centredness, suggesting that interprofessional communication was a distinct domain of competence. Internal consistency was adequate, and whilst the studies did not report a pass-fail score, they were able to identify weakness in specific areas.

Foronda et al (2015) also drew on an existing clinical protocol to assess structured communication between nurses and doctors. In this study the SBAR tool was adapted to include role 'Identification' on the part of the nurse (ie they identify themselves before describing the situation). This ISBAR rubric comprised 15 items rated on a four-point scale. Student nurses were assessed in a scenario requiring telephone communication with a physician about a deteriorating patient. Foronda et al found differences between participants from the USA and China, with Chinese students scoring higher, but the measure demonstrating lower internal consistency. This raises questions about possible cultural influences, not just on performance or the judgement of performance, but also on the underlying construct in the two countries. There were no differences in scores between two levels of nursing student in the USA.

Reising et al (2015) described the use of different scales for assessment of individual and team performance in groups consisting of one medical student and two nursing students. Evidence of content validity came from an extensive, theoretically informed development process, which defined distinct items for individual performance (body language, interpersonal skills, encouraging feedback and discussion, resource use, problem-solving, scenario management and patient communication), and team performance (role assignments, closed loop communication and clear language, use of team input, clinical impression management, patient education and reassessment). Internal consistency and inter-rater reliability were high, and the tool was sensitive to improvements in communications skills from year 1 to 2. However, the raters here were researchers rather than authentic clinician assessors.

Wright et al (2013) presented detailed validation evidence for an assessment of individual team-working skills, which involved six scenarios with standardised professionals presenting challenging behaviours. Authenticity of workplace practice was reflected in content that included high workload, unclear roles and responsibilities, multiple distractions, conflict and hierarchies. Four scenario-specific questions per scenario, mapped to team-working constructs, were answered on a three-point scale. Ratings were given by SPs and external raters. Less than 2% of variance was due to rater type – indicating high agreement between these groups. A generalisability study found moderate reliability, with a decision study indicating 12 scenarios would be necessary to reach an acceptable level. Wright et al also indicated feasibility, with the main cost/resource being the training of actors.

While strong evidence of validity suggests the value of this approach, notably there was a mixed response to this assessment from participants. While most were positive, some were seen not to be taking the scenarios seriously. A post-study survey found that some participants felt that realism was limited – for instance, through the SP playing the professional having little clinical knowledge or 'over-acting'. Perceived inauthenticity may thus be a threat to the content validity of a simulation if it does not 'feel' right to the candidate.

Threats to validity in the response process, and consequent implication for assessor training, were raised by several studies. Saylor et al (2016) described difference between medical and nurse raters of pairs of students, and recently qualified clinicians in a simulated palliative care scenario. Assessors scored their respective profession using the Team Objective Structured Clinical Examination tool. Physicians scored higher than nurses overall, and in each of the six competencies (communication, collaboration, roles and responsibilities, collaborative patient-family centred approach, conflict management, team functioning). However, the study did not clarify whether this arose from a difference in candidate behaviour as is inferred, or a variability in assessor performance. The risk of the latter is however mitigated by validity from consistent response process, ensured by extensive SP training of more than 30 hours, including observing oncology patients and viewing videotaped interviews.

A pilot study by Lie et al (2015) did not present actual assessment data, but rather validation of an assessment method. Simulated teams were trained to perform at a certain level of competence. Assessors rated the team as a whole, and four team members individually, at different points in a scenario. Ratings were on six global rating scales: communication, collaboration, roles and responsibilities, collaborative patient-centred approach, conflict management and team functioning. Results indicated calibration of team and individual judgements differed, with 'below expected' individual performance identified only 46% of the time, compared to 100% for team performance. A generalisability study identified high proportions of variance arising from raters, suggesting overall reliability was low, although having two raters improved this. Assessors also demonstrated an overall 'leniency error', indicating a need for specific training to ensure accuracy of high stakes decisions.

Other approaches

Whilst assessments in this domain were generally scenario-based, some novel approaches were identified in the review.

Dow et al (2016) reported an online approach, which exploited the reality of non-acute teams, which may interact asynchronously and often work without a clear hierarchy. Participants from different professions were involved in a longitudinal web-based scenario that followed a patient in simulated time across a range of settings. Participants were required to perform their professional role virtually, including collaborative information-sharing through a message board and entries to an electronic health record.

The primary assessment outcome was a knowledge test based on the case. However, teamwork behaviours logged by the online systems were also available and correlated with both individual and team knowledge scores. As presented, this longitudinal training approach may not be practical for summative assessment, but it illustrates how technology may be used to capture authentic team-based behaviours and online skills, in itself a requirement of modern healthcare practice. However, this tool came at a cost, as the bespoke software had development costs of US$200,000.

A prima facie element of team-based communication is handover. This is essentially a 'basic' communication skill, being protocol driven and routine, but we describe an innovative example here by way of illustration. Farnan et al (2010) reported a pilot formative assessment in which students gave handover to a standardised resident, played in this case by an actual resident rather than an actor. The station creatively incorporated multi-media, and also assessed written as well as verbal handover skills. Candidates reviewed a written patient history, but also had to identify cues for ‘to do’ items from a short video (for example, a need to chase radiology or laboratory results). Written handover skills were assessed for correct information relating to the specific scenario in categories ‘identification of information’, ‘problem list’, ‘medication list’, ‘anticipatory guidance’ (statements of the form ‘if…then’) and ‘to do tasks’.

3.5 Content of assessment: Ethics

3.5.1 Summary of evidence

Ethical practice of doctors can be defined in terms of individual moral virtue, and the ability to reason and resolve medical ethical dilemmas. While the former traits can be measured, teaching and assessment necessarily focuses on the latter skills (Eckles 2005). Within this, ethical behaviour may be linked to challenges present in complex communication scenarios, while ethical judgement is a cognitive skill.

Overall, we rate the weight of evidence in this area as moderate. While there is evidence of validation for many assessments, there is not a clear indication of what constitutes best practice in the content or form of these assessments. All but one of eight studies were in undergraduate settings, and most were tests of knowledge and reasoning skills reflecting the competencies needed to recognise ethical issues and reach practical and ethical solutions. Some scenario-based assessments of ethical behaviour were found.

For assessment of ethical judgement as a distinct area of applied knowledge and reasoning, written exams (paper- or computer-based) may be appropriate (Tsai et al 2012, Foucault et al 2015), with performed scenarios better used to demonstrate competence in ethical communication (Jameel et al 2015). Written exams are also more inherently scalable than scenario-based assessments.

Ethical judgements are sensitive to cultural norms and the context of learning and practice. While such effects may be apparent in all assessments, written assessments may allow the ethical imperatives candidates are invoking to be identified. They may also allow appropriate difficulty, and complexity of ethical cases to be more precisely presented, whilst being open to the range of ways an ethical dilemma can be interpreted.

Irrespective of specific approach, the selection of appropriate and authentic content is important. The types of ethical dilemmas and practice which are addressed need to be considered, and at the appropriate level for different candidate groups. The ethical decision-making required of new graduates will differ to those in more senior positions. The ethical knowledge and reasoning required may be similar, but the framing of the problem must be authentic to their level.

Table 7. Validity evidence for assessments of ethics

Reference	Country	Population	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Carlin et al. (2011)	USA	UG	327	Developed by faculty; used health professional literature.	Raters (x6) reached consensus on sample. IRR 0.9.	Intra-rater reliability=0.85	None given	None given
Favia et al. (2013)	USA	UG	137	Developed by authors. Some reference to literature and theory. Reviewed by experts.	ICC low, but absolute difference small. IRR with outside raters mixed.	None given	None given	22% 'high competence'
Foucault A, et al. (2015)	Canada	UG	79	Vignettes derived from literature and reviewed by students and residents.	None given	Alpha 0.4 with all items. Review of items identified reduced set with alpha > 0.6	None given	Post-test survey
Jameel A, et al. (2015)	Pakistan	PG	136	Scenarios derived from 'Project Professionalism' (ABIM 1995).	SP training. Instructions translated into Urdu. Test-retest moderate.	Alpha=0.61 across stations, 0.31 across scenarios. Generalisability=0.65. Item-test correlations moderate	Correlation between OSCE and written components moderate. Improvement following teaching	Evaluation score high
Lohfeld L, et al. (2012)	Canada/ UK	UG	62	Based on literature. Reviewed by experts against Conventional Validity Index. Scoring derived from literature. Four scores derived from each short answer	Details of scoring. IRR and test-retest overall low.	Generalisability between cases low. No improvement indicated by D-study	No association with MCCQE, overall or ethics	None given
Reinert A, et al. (2014)	USA	UG	262	Developed by surgical fellow with faculty consensus	Rater training and practice.	Alpha=0.67	No difference with rotation/year. Moderate correlations with most other exams/evaluations. Predicted by Step 2 CK and clerkship evaluations	Positive feedback from students
Tsai TC, et al. (2009)	Canada/ Taiwan	UG	49	Scripts developed from CMA document and reviewed by experts and students.	Think aloud restricted to one hour. 'Decision score' based on CMA and reviewed by experts. Reasoning inventory selected by experts.	Alphas high. Rasch analysis	Decision score: Difference between countries. No difference between experts, residents and students. Inventory: difference between countries and levels of expertise in Taiwan only	None given
Tsai TC, et al. (2012)	Taiwan	UG	22	Steps of SCT and scoring keys derived from ethics experts.	None given	Alpha = 0.81	Scores higher for experts than students or laypersons	None given

Key:
UG= undergraduate; PG=postgraduate
ICC=Intraclass Correlation Coefficient; IRR=Inter-rater reliability; D-study=Decision study
SP=Standardised or simulated Patient; SCT=Script Concordance Test; CMA=Canadian Medical Association; Step 2 CK=Step 2 Clinical Knowledge exam of USMLE

3.5.2 Details of evidence

We noted in the introduction that while ethical practice or behaviour may be synonymous with more global concepts of professionalism, there is also a specific competency that relates to the application of principles of ethical practice. In contrast to previous sections, this is something that has been operationalised more through applied knowledge tests, rather than behavioural assessment.

Such tests adopt a number of formats. Lohfeld et al (2012) used a 'single best answer' format, with best answers established by experts. Students also had to provide a short answer to justify their responses to a series of vignettes. This assessment showed little evidence of validity from internal structure or relationships with other variables, while a

decision-study showed that increasing the number of raters and cases made little improvement in reliability. The content may be authentic, and the method have some appeal on grounds of scalability, but this fails in terms of performance as an effective assessment.

A short answer approach was also reported by Carlin et al (2011), who designed a written exercise consisting of four open ended questions with responses scored 'insufficient', 'acceptable' or 'proficient' on the basis of relevant criteria. The case was derived from the student's own clinical experiences, ensuring content validity and authenticity. Acceptable inter-rater and test-retest reliability was reported, but no evidence of discrimination was provided. The assessment was feasible in terms of the time take to complete and assess (completed in 15 minutes by the student; five minutes by the assessor) implied feasibility. It was suggested as one component of a wider ethics assessment strategy.

A computer-based written exam for students in surgical clerkships was described by Reinert et al (2014). Alongside surgical knowledge and clinical reasoning, this included a 'professionalism' section, which concerned ethical principles related to the provision of, or withholding of clinical care. Content was developed by an individual fellow with faculty consensus. For the test overall, evidence of test-retest reliability and construct validity from relationships with other variables was reported, but details of each sub-section were not reported.

Tsai et al (2012) developed an ethics Script Concordance Test (eSCT). SCTs are a method of assessing clinical reasoning in the context of uncertainty, by assessing how candidates respond to changing information within a scenario. Tsai et al used ethical vignettes designed by experts (no details are given of content). In each case candidates had to make an initial decision, and then consider changing that decision as further information was supplied. Scoring benchmarks of the 43 items were derived from answers of ethics experts. Internal consistency was high. While the paper did not report a pass mark, Tsai et al demonstrated criterion validity with expert scores being higher than those of medical students and laypersons.

Foucault et al (2015) also reported an online resource based on a Script Concordance Test approach. A number of written vignettes presenting ethical issues (albeit termed 'professionalism') with four responses were completed by an expert panel, who provided brief justification for their response. Medical students completed the test and could then view the experts' justification. While this was used for formative feedback and internal consistency was low for high stakes purposes, it demonstrates feasibility for adaptation to summative assessment. However, a post-test survey indicated that students differentiated what they felt was the 'correct' answer as identified by experts from what they would actually do in practice (illustrating the distinction between 'knows how' or 'shows how' and 'does' in Miller's pyramid). Foucault et al implied that knowing the 'correct' answer was a function of expertise, and so the test constituted learning for the students, rather than directly questioning the authenticity of the assessment.

Assessment of Ethics and Law

Practice in the UK mirrors that of the literature review with assessments of ethics and law taking both paper- and scenario-based processes.

One clinical station focusing on ethics and law is included in final undergraduate examinations at the Universities of Manchester and Leicester. An example of a typical scenario is that of a 'reluctant' patient who must be counselled by the 'doctor' about informing the DVLA, his employer and insurance company after having had a first grand-mal seizure. This scenario assesses the student's knowledge and skills in making a risk assessment, and balancing patient confidentiality with potential harm to others. It also assesses the student's ability to negotiate with the patient and 'hand back' responsibility to them for information sharing, whilst demonstrating an empathetic understanding of the patient's social context.

At Cambridge a short answer ethics and law written paper is included in Final examinations. The paper includes 6 clinical vignettes that 'unfold' and draw on student's knowledge of ethical and legal principles and skills in reasoning a moral argument. Topics include assisted suicide and teenage pregnancy, including prescription of the contraceptive pill without parents' consent and the responsibilities of a doctor who has conscientious objection to termination of pregnancy. Answers are double marked and model answers support reliability of assessment.

The written approach allows assessment of students' ethical reasoning skills across a range of subjects and how this informs decision-making in ethical dilemmas. There is opportunity to include contemporary and controversial topics, for example, gender reassignment or rationing of health resources, where the student must balance their own value laden perspective against the patient's and society's position.

Two studies described approaches combining written tests and other methods to assess knowledge and behaviour. Jameel et al (2015) described a written exam used in conjunction with an OSCE for assessment of postgraduate residents in Pakistan. The written exam covered a number of ethical issues derived from professional vignettes in the 'Project Professionalism' document from the USA (ABIM 1995). The OSCE scenarios included patient autonomy, confidentiality, shared decision making, do not resuscitate orders and probity. OSCE performance was rated on a checklist by SPs, but details were not given of how written papers were marked. Reliability of the OSCE was low, as were correlations with the written exam, indicating a lack of convergent validity and so casting doubt on the constructs being assessed. Decision-study results suggested that 13 written and OSCE scenarios were needed to achieve borderline-acceptable generalisability. Jameel et al (2015) also noted found that recruitment of female SPs from the local community was problematic due to cultural norms, and staff had to play female SPs.

Favia et al (2013) described a formative process which used a short written assignment and oral presentation based on ethical issues identified in simulated clinical encounters. Content validity was indicated in feedback from external ethics experts. The written assessment was graded on a standardised rubric with a scaled checklist. This showed a high degree of inter-rater agreement between faculty raters, but not a group of external raters who assessed a number of selected cases. However, it did discriminate between students. The poor reliability of external raters poses a challenge for wider use, but clearer articulation of the expectations around student competency, and review of the rating rubric to fully explain and communicate instructions, might support more consistent grading of the assignment.

A final study used a novel oral approach, albeit one with questionable feasibility. Tsai et al (2009) described an approach where respondents were scored through a 'think aloud' interview based on 15 ethical vignettes deriving from a resource from the Canadian Medical Association. 'Thinking aloud' required candidates to verbalise their thoughts about initial, and supplementary information until they reached an ethical decision. While this showed acceptable reliability, it did not discriminate between levels of experience, and logistically offers little advantage over an OSCE, at up to one hour per candidate. Tsai et al (2009) also found differences between Canadian and Taiwanese participants, with Taiwanese subjects (including experts) performing less well, suggesting that ethical judgements may be shaped by local context, and perhaps cultural differences. Nonetheless, it is a novel approach to assessment which may bear further consideration.

3.6 Content of assessment: Patient safety

3.6.1 Summary of evidence

While all practice, and so all assessment, should reflect safe clinical care, patient safety *per se* is defined as a discrete knowledge base and associated skill set to recognise and respond to challenges to patient safety in the form of medical errors.

Evidence from the review was weakest in this area, both in the number of studies and the validity evidence provided. We found just eight studies in this domain, six of which were in undergraduate settings, with disparate approaches. Some addressed error identification or recovery directly, but most operationalised safety as a specific element of practical or knowledge-based tasks. It is somewhat surprising there are so few given current levels of interest in patient safety, error and human factors. It may be that the focus of activity to date has been on in-practice training and continuing professional development, rather than assessment per se. Examples here draw on some theory and empirical precedent (Daud-Gallotti et al 2011, Sternbach et al 2017), but deeper consideration of theoretical literature on the causes of error (eg Reason 2000) could provide more detailed grounding for the development of appropriate content.

The assessment of the necessary skills to directly recognise and reduce error, like those of ethical practice, may lend themselves to 'knows how' methods such as script concordance or situational judgement tests that incorporate clinical uncertainty, rather than purely behavioural measures. With such limited evidence it is hard to firmly conclude what constitutes good practice, but the assessment described by Sternbach et al (2017) provides an encouraging example.

Table 8. Validity evidence for assessments of patient safety

Reference	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Chowriappa AJ, et al. (2013)	USA	PG	27	Delphi exercise	Automated data. Weighting agreed by experts.	None given for measurements. Correlations between tasks low-moderate.	Experts performed better on all tasks	None given
Daud-Gallotti RM, et al. (2011)	Brazil	UG	95	None given for scenarios. Checklist developed by experts.	SPs trained and scenarios piloted.	Subscales correlated 0.4-0.6	None given	Highly rated by students
Ginsburg LR et al (2015)	Canada	UG	18	Developed by experts from Safety Competency Framework	1hr training for assessor pairs. Guidance sheet provided. Calibration video. Good IRR.	Alpha > 0.75	Nurse students scored less than medics on 3 stations	39% scored borderline on at least one station.
Morris MC, et al. (2014)	Ireland	UG	37	Tools derived from literature.	Written information provided to SPs and examiners in advance. Examiners met to agree criteria. SPs standardisation meeting. IRR high.	None given	None given	Fail more likely than with long case. Examiner and participant responses good.
Sternbach JM, et al. (2017)	USA	PG	15	Embedded errors selected from most common errors identified in a previous study.	IRR acceptable. Videos reviewed for consistent set-up.	Range of difficulty and discrimination across items. Overall alpha=0.61. Intra-rater correlation high	Differences between interns and PGY3	None given
Tweed M and Wilkinson T. (2009)	New Zealand	UG	210	Items selected from question bank. 'Safety' of distractor responses rated by experts.	Student randomised to 1 of 4 sets of instructions around unsafe responses & 'guessing'.	Alpha (for control group) =0.7	Year 5 students performed better than year 4	'don't know' response more common if mark reduction for 'unsafe' response.
Tweed M, et al. (2013)	New Zealand	UG	372	Items selected from question bank. Safety of distractors rated by experts.	Students provided with scoring grid.	None given	Improvement of performance with year	None given
Varkey and Natt (2007)	USA	UG	42	None given	None given	None given	Low correlation with other station Hx scores, negative correlation with interpersonal skills	4/42 students did not meet cut point.

Key:
UG= undergraduate; PG=postgraduate
IRR=Inter-rater reliability
SP=Standardised or simulated Patient

3.6.2 Details of evidence

Safe practice is intrinsic to clinical competence, but safe practice alone does not indicate an awareness of patient safety. We found relatively few studies that placed an understanding of safety or human error at the centre of assessment. This is somewhat surprising given the profile of patient safety and human factors awareness in practice. It may be that such training is seen as part of continuing professional development, rather than an issue for summative assessment. Nonetheless, whilst limited, the evidence offers ways to deliver valid assessments in this domain.

Daud-Gallotti et al (2011) assessed performance in error-focused scenarios using a scaled checklist, on which eight of 21 items related to medical error (the remainder to patient-physician relationship and humanism). This end-of-clerkship formative OSCE followed a theory-driven course on medical error and contained scenarios that focused on effective interpersonal communication, invasive procedures, and resuscitation. The approach extended assessment to

the principles of patient safety, as the checklist items reflected the students' understanding of errors, rather than just their disclosure of them (eg, 'did the student explain to you what type of error occurred and how it will impact your health?').

A practical assessment of error identification and recovery was described by Sternbach et al (2017) in an assessment of surgical residents' error identification and recovery skills using a thoracoscopic lobectomy simulator. In each of five stations, the resident was asked to take over a procedure from another surgeon who was feeling unwell. The resident was expected to identify and correct any errors that had been made by the previous surgeon, and complete the procedure. These errors represented common mistakes in this procedure, as identified in an earlier study (Meyerson et al 2012), which supported content validity. Video recordings of each station were scored by four raters using a scaled checklist for each step of each task. Inter-rater reliability was high and all stations adequately discriminated between high- and low-performing residents.

Two studies described OSCEs considering elements of patient safety. Ginsburg et al (2015) described a four-station OSCE including a 'near miss', a complex discharge, challenging authority and medication error disclosure. Rating was on global rating scales based on a safety competence framework, encompassing awareness of patient safety culture, managing risks, communicating and responding to risk. Good inter-rater reliability and internal consistency was found. Varkey and Natt (2007) described a single OSCE station looking at medication error, comprising tasks of conducting a root cause analysis, communication with an SP, and completing patient notes. History taking and patient notes were scored on a checklist, and a global score given for overall performance. Standard setting using the modified Angoff method was reported, with pass rates comparable to other stations in the OSCE.

While technical competence alone is not necessarily indicative of understanding of safety, Chowriappa et al (2013) derived a patient safety score from expert-weighted evaluation of technical skills required in robotic surgery. The resultant metric indicated not just completion of a task as a simple checklist would do, but *safe* completion. While indirect, this adds an element of understanding to simple competence. Construct validity of this score was established with experts scoring significantly higher than novices. However, no pass-fail standard was specified and data on reliability of the tool were not reported.

Tweed (Tweed & Wilkinson 2009, Tweed et al 2013) considered the issue of 'safety' by exploiting tests of clinical knowledge to examine the students' response when distractor 'unsafe practice' items were included in the questions. In a pilot randomised controlled trial (Tweed & Wilkinson 2009) they noted that 4th year students were more likely to give a response 'I do not know and would seek advice' if the examination had negative marking for incorrect, unsafe responses. In the second study (Tweed et al 2013), multiple choice distractor responses were rated by experts on a scale of safety, and weighted by students' self-reported confidence in their responses. Tweed et al inferred 'insight' from the association between increasing confidence and more correct, and fewer incorrect answers. Conversely, a lack of 'foresight' (or hazardous ignorance) was inferred in responses that were unsafe to any degree, but held with high confidence. Criterion validity was indicated by improving insight and foresight with year of study.

These studies imply that paper-based assessment can discern differing responses of candidates to 'safeness', including the fail-safe option of deferring to a senior colleague. However, the relevance of these response choices for behaviour in practice is not known, though confidence-based scoring might further help identify those with unsafe gaps in knowledge.

Another approach based on negative marking, and focusing on the identification of *unsafe* practice, was described by Morris et al (2014), in a pilot of a wide-ranging assessment of eight domains of professional surgical practice. In this, safety was operationalised as the absence of an egregious error within four patient encounters. *Unsafe* practice was identified, rather than requiring *safe* practice to be explicitly defined and scored. Similarly, professionalism was defined as the avoidance of specified *unprofessional* behaviours. Students were assessed on four SP scenarios by two examiners, and inter-rater reliability was good. Concurrent validity was low based on comparison with a long case examination, with students more likely to fail the new assessment, suggesting either a different construct, or a need for improved calibration of assessment. Examiners' judgements of feasibility and workload indicated an acceptable assessment.

3.7 The use of technology in assessment

Our discussion of assessment content has so far described a number of approaches to assessment methods. While most have involved role-player enacted scenarios, paper- and computer-based examinations have also been described.

In this section we consider other types of assessment which utilise different technological approaches. Most of these differ in the 'process' of assessment, how it is enabled and delivered, though some also differ in the 'outcome' of the assessment or how performance is captured and encoded.

Some technological approaches are well-established and commonplace, and so were treated as 'low priority' in our screening of the literature. Nonetheless, here we acknowledge the many applications of simulation to the assessment of technical skills, and give a brief summary of some of those approaches, before focusing on more novel approaches.

3.7.1 Types of simulation

A systematic review by Cook et al (2013) found 350 studies of simulated assessment of doctors, most of which focused on surgical or procedural skills. They concluded evidence of validity was sparse and the overall quality of the literature was poor. Our review similarly found a large number of pilot studies where simulation-based assessment tools were predominantly used formatively, and sources of validation were minimal (eg, Knudson et al 2008).

A number of types of simulation were described in the literature, including the use of role-player actors which we have discussed in some detail already. Other common forms of simulation for procedural and technical skills included part-task trainers and manikins. These modalities have typically been used across specialties and professions in the training and assessment of technical skills (eg, the MISTELS simulator used for certification of laparoscopic surgery skills in the USA, Peters et al 2004), though manikins can feature as an element of more immersive environments where cognitive and behavioural elements may be examined (Banerjee 2015; Bensfield et al 2012). Till et al (2015) described validation of an assessment for final year medical students in Scotland that used a simulation ward, where students had to manage six patients over 20 minutes. This created an authentic context for assessment of skills that reflected workplace practice, such as 'safe medical practice' and 'response to interruptions'.

The outcome of such assessments may be similar to the OSCEs already described, in ratings and checklists of performance by trained raters, but other approaches are possible, including automated data capture of hand movements by the simulator itself (eg Saleh et al 2008, Howells et al 2008). Machine learning approaches have also been tested for the automated assessment of images of suturing, with algorithmic analysis identifying differences between novices and experts (Frischknecht et al 2013), and written exams (Latifi et al 2016). Currently these approaches may be limited, but future applications of machine learning and artificial intelligence may have more flexibility.

It is worth noting that simple manikins have been used in the training and certification of Basic (BLS) and Advanced Life Skills (ALS) since their inception for both clinicians and lay participants (Boet et al 2017). However, Boet et al (2017) also flagged that assessment on a part-task trainer may not transfer to manikin-based (high fidelity) scenario, as, in this study, only one of 20 lay participants, each of whom had passed the BLS course, subsequently passed a simulated cardiac arrest scenario.

Hybrid simulation can involve different modalities within a single scenario. For example, Black et al (2010) described assessments which combined simulators with role-players to achieve a balance of authenticity for clinical and communication skills. Both reported findings that this approach was appropriate and authentic.

Benedict et al (2017) described a blended simulation approach for assessment of practice readiness of pharmacy students. This process involved a 5-station OSCE that followed the course of a single diabetic patient and included review of a simulated Electronic Health Record, single best answer responses to online case scenarios and scenarios with a standardised colleague and standardised patient. This approach demonstrated assessment of different levels of Miller's pyramid, from 'knows' and 'knows how' to 'shows how'.

The USMLE step 3 exam incorporates computer-based simulation cases (Dillon & Clauser 2009). In this, candidates can request additional history and details of examination, make free text entries to order investigations and 'move' the patient's location (eg, emergency department, ward, discharge home). The patient's condition changes in a realistic manner as time progresses and in response to treatments given. Here, the case content, including clinical presentation and details of appropriate and inappropriate actions, have been developed with physician experts. Automated scoring takes a regression-based approach and has been shown repeatedly to closely approximate the scores of experts. Hence the computerised format has offered cost effective and efficient assessment of complex clinical scenarios (two million scenarios scored, estimated as avoiding 100,000 hours of expert assessors' time). Multimedia items (sound or video) can add authenticity to tests, but were noted to introduce a threat to validity in the Comprehensive Osteopathic Medical Licensing Examination [COMLEX] (Shen et al 2010). In this exam item difficulty varied with the inclusion of multimedia content, affecting item discrimination. Also, examinees needed a significantly longer time to respond to the multimedia items. However, benefits of multimedia have yet to be fully explored, for example, to support fairness of assessment when English is the examinee's second language.

Technology in Final examinations: PAG experience

Among PAG member' organisations, simulation technology is predominantly used as a tool for teaching and training. It is used in Final examinations (example, **Leicester** and **Edinburgh**) to assess competencies in acute care management (deteriorating patient), where resources are 'low tech' (eg, manikin, airway adjuncts, prescription sheets, monitoring devices), but may involve a standardised professional in the scenario.

Key challenges to greater use include cost and scalability of the approach, especially for new technologies (such as virtual reality simulation). More especially, realism is seen to be less a feature of the technical environment, but the nature of scenario design. For example, in **Belfast**, assessment of students' ability to make a telephone referral to another specialty team member represents a common requirement of real-life practice, but does not have the cost or practical limitations of high-fidelity simulation.

Setting a standard of what constitutes 'safe practice' needs careful definition, though may be easier in emergency scenarios that are based on a standardised approach.

Other technology-driven approaches to assessment include online programmes that range from e-scenarios that require the user to select certain actions in response to cues for patient deterioration (FIRST2ACTWEB™ , as described by Bogossian et al 2015) and imaging studies (CT and MRI) for assessment of radiological interpretation skills (Gondim Teixeira et al 2017). The remainder of this chapter will focus on more novel technological approaches.

3.8 Types of assessment: Virtual Reality

3.8.1 Summary of evidence

While simulation is traditionally largely based on the physical representation of a clinical event or workplace, developments in technology provide new ways of presenting simulated environments. Virtual reality (VR) refers to computer-generated virtual representations of the physical world with which users interact. While VR technology can provide fully immersive environments, we found no instances of such systems being used in assessment. We found 18 papers (and one systematic review) describing some form of non-immersive or partially-immersive VR, of which ten referred to postgraduate doctors, six to medical students, one to undergraduate and postgraduate medicine, and one to nursing students. Two papers (Chowriappa et al 2013 and Deladisma et al 2007) have already been mentioned with reference to the content of their assessments.

We found uses of VR that fitted into three main categories, with the evidence for each varying. The dominant use of VR currently is for the presentation of intra-corporeal procedures. This is well established, and the evidence here we judge to be good. While improvements in technologies will allow enhancements to content through improved displays and haptic (sensation) feedback, the basics of practical authenticity and response processes are robust.

The second use is in the presentation of virtual patients. These interactive computer simulations of patients are generally based in non-immersive online systems, and the underlying technology of many appear primitive. However, the paradigm of assessment they illustrate is not tied to the technology, and these papers give some indication of how virtual patients may provide a standardised medium for assessment of different elements of practice. For the most part, these create interactive wrappers for script concordance-type assessments,

Virtual patients and virtual worlds

Virtual reality technology has a spectrum of sophistication as illustrated by 2 papers.

Waldmann et al (2008) created a 'Virtual General Practice' that was used to formatively assess 147 final year medical students. Students had to deal with 3 virtual patients, who had common primary care presentations, and were required to select appropriate history items, examination modes (inspection, palpation etc) and tests from an online 'menu'. Responses and results were displayed as text, and included multimedia elements (pictures, audio, video etc)

Assessment was automated, and scores took account of right and necessary steps taken, as well as avoidance of wrong, unnecessary or harmful actions.

In this validation study no technical issues were encountered, but students' lack of familiarity with the format led to incomplete data capture. Content was mapped to the primary care curriculum, but views on authenticity were not captured. Rather, that some students left the assessment early or took silly online actions might suggest that authenticity was lacking.

Nonetheless, the potential to deliver mass, standardised assessment is illustrated.

By contrast, **Heinrichs et al (2008)** described development of virtual worlds – 3-dimensional environments where participants can role-play, communicate and interact in real-time.

One scenario is of a radioactive ('dirty') bomb blast. The virtual emergency department has photorealistic driveway, entrance, waiting area, treatment area, hospital staff and patient avatars. The vital signs of the virtual victims reflect the severity of the injuries, as well as the patient's age, sex and comorbid conditions. These are responsive to fluid, blood and drug therapy and appear in real-time allowing participants to make clinical decisions. Text responses to queries, sketches of results and reports from diagnostic procedures are presented on a pop-up interface.

Participants take on an avatar and select a leader, who then assigns roles. Teams move their avatars to the appropriate areas and begin management. Each member uses a headset and microphone to speak to and hear others online. Once the patient is stabilised, they can request admission or transfer to surgery.

Pilot data suggested that participants found the approach acceptable and realistic. Further, the approach could be used for assessment, whether in rating of individual and team performance using validated team-working tools, or in capture of outcome metrics, for eg, the number of patients treated appropriately, the time taken to stabilise casualties and the number who died.

but one (Deladisma et al 2007) using a more immersive animated interface, shows that interpersonal skills may also be assessed. That said, the validity evidence for these systems is low, and details of content, response and authenticity will need to be addressed.

The final use is in the representation of virtual worlds, that is partially or fully immersive environments in which users can move and interact freely. The evidence for these is low, with just one study found (Heinrichs et al 2008), but there is potential here for integration with both virtual patients and virtual procedures to provide an integrated, controlled and standardised assessment environment.

However, this potential should be considered against cost, functional as well as monetary. Virtual representations of patients and environments may provide opportunities in terms of scalability and standardisation, perhaps in conjunction with remote assessment, but this should be offset against the risk of a loss of flexibility and authenticity – virtual systems are ultimately limited by their design and coding.

Table 9. Validity evidence for assessments involving virtual reality

Reference	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Thijssen AS & Schijven MP (2010)	Systematic review							
Botezatu M, et al (2010)	Sweden	UG	216	VP and SP cases matched	ICC > 0.95	Alpha > 0.75	Web-SP higher scores than control. Web-SP exam results higher than normal exam for both	None given
Chowriappa AJ, et al. (2013)	USA	PG	27	Delphi exercise	Automated data. Weighting agreed by experts.	None given for measurements. Correlations between tasks low-moderate.	Experts performed better on all tasks	None given
Courteille O, et al. (2008)	Sweden	UG	110	Scenarios redesigned and simplified for web VP by surgeons	Assistant interactions observed by video	None given	Female > male	Students reported limitations in interaction, but overall felt engaging and realistic
Deladisma AM, et al. (2007)	USA	UG	84	None given	None given	Alpha > 0.6	Difference between SP and VP groups. Empathy associated with non-verbal behaviour	None given
Forsberg E, et al. (2015)	Sweden	PG nursing students	19	Scenarios reviewed by senior paediatrician. Correct responses agreed by experts	Scoring rubric piloted. Semi-automated assessment.	None given	Increase with progression	None given
Heinrichs WL, et al. (2008)	USA	UG and PG	30	Cases adapted from simulator cases	None given	None given	Increase after training	None given
Jacobsen ME, et al. (2015)	Denmark	PG novice / expert orthopaedic surgeons	26	Test piloted by expert.	Tasks translated and checked. Participants familiarised with simulator. Automated data.	ICC for metrics and cases high	Differences between novice and experienced surgeons	None given
Konge L, et al. (2013)	Denmark and Netherlands	PG respiratory physicians – differing experience and training	22	None given	Automated data. Differentiating metrics combined into quality score.	G-study=0.67. D-study, 0.8 achievable with 2 more procedures.	No differences between novices with and without training.	Pass derived from contrasting groups. Just one untrained novice passed test.
McGrath et al (2015)	USA	PG	35	Scenario identical to traditional exam, scored with standard tools	None given	None given	No significant different between virtual and traditional	Virtual less intimidating, and preferred by most candidates
Noureldin YA, et al. (2016)	Canada	PG	26	None given	Automated data. Participants given 3 min orientation.	None given	Differences by age and sim experience, but not procedural experience. Global score correlated with number of attempts	Cut-score based on attending performance. Half sample passed/failed.
Oliven A, et al. (2011)	Israel	UG	262	None given	Automated data	Alpha > 0.8	Alpha larger than for standard OSCE. Good correlations between modes	None given
Pedersen P, et al. (2014)	Denmark	PG novice / expert orthopaedic surgeons	20	None given	Automated data	Inter-case reliability > 0.8 with three procedures	Higher scores for consultants	Cut-score derived by contrasting groups

Reference	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes
Raison N, et al. (2017)	21 European countries	Novice-expert surgeons	223	Exercises selected by experts from those in course.	Automated data	None given	Only complex tasks discriminated between novice and intermediate	Cut-score based on expert performance.
Vassiliou MC, et al. (2014)	USA	PG - surgeons and physicians with endoscopy role; residents	111	Skills defined by expert group	Automated data. Retest ICC 0.85	Alpha>0.7	Score correlated with previous experience	Cut-score established by ROC curve. Expert group achieved 92% pass rate
Waldmann UM, et al. (2008)	Germany	UG	147	Cases reviewed by experts. Correct answers based on consensus and guidelines	Automated data	Inter-case correlation > 0.5	Correlation with written exam < 0.4	Pass rate derived from written exam
Willaert WI, et al. (2012)	UK	PG - surgery, cardiology, radiology trainees	20	Content based on expert consensus. Case derived from real case	Participants trained with 10 generic sim cases. Automated data. Rating scales derived from literature. IRR for scales moderate - high	None given	Improvement with experience in study	Participants perceived assessment to be authentic
Williams K, et al. (2011)	Sweden	PG - psychiatry residents	10	Cases reviewed by expert	Automated data	None given	None given	All participants felt VP was realistic, useful and accurate
Yang RL, et al. (2013)	USA	UG	27	Cases developed by students, residents and faculty to reflect learning objectives, reviewed by experts	None given	None given	Increase at end of rotation	Participants positive about VP

Key:
UG= undergraduate; PG=postgraduate
IRR=Inter-rater reliability; G-study=Generalisability study; D-study=Decision study
SP= Standardised or Simulated Patient; VP=Virtual Patient
ROC curve=Receiver Operating Characteristic curve

3.8.2 Detail of evidence

Assessment of technical skills

The use of virtual reality (VR) is best established for the teaching and assessment of technical skills, in particular for invasive procedures where a camera may be used in real practice. For example, a literature review by Thijssen & Schijven (2010) identified 42 examples of VR simulators being used in training and assessment of laparoscopic skills. In such procedures, the doctor in real practice will be looking at a screen, so the virtual representation should not inherently detract from face validity or authenticity.

In our search, we found examples relating to flexible endoscopy (Vassiliou et al 2014), fluoroscopy guided percutaneous renal access (Noureldin 2016) and hip fracture surgery (Pedersen et al 2014). These assessments required candidates to perform psychomotor tasks of accurately guiding the instrument in the virtual body and then identifying 'targets', 'popping balloons' (Noureldin 2016) or fixing a screw (Pedersen et al 2014) – analogous to tasks in practice.

Vassiliou et al (2014) reported that while raters and participants felt the interface presented the gastrointestinal tract with 'reasonably high fidelity', some aspects of authenticity, and so content validity, were lacking. Specifically, haptic feedback in response to some manoeuvres commonly used in practice was not possible, and experts noted the

absence of these typical 'body' responses. One dimension of the assessment – 'thoroughness of gut examination' – also had lower internal consistency, possibly related to this gap in fidelity. To enhance reality, Noureldin et al (2016) used VR in a hybrid format with a manikin to represent the physical patient and the physical aspects of the endoscopy.

Performance on these systems can be captured directly by the simulators, intrinsically providing high reliability. Automated data encompasses a wealth of parameters, not all of which will be informative. We found examples that identified which parameters were important through theoretical considerations and expert consensus (Raison et al, 2017; Chowriappa et al 2013, Willaert et al 2012), or empirically from the simulation data itself. For example, Konge et al (2013) established which variables indicated a difference between novice and experienced doctors performing endobronchial ultrasound, and used those variables to derive a pass/fail cut score based on overlap between the distributions (the contrasting groups method). This method for standard-setting was also used by Jacobsen et al (2015) in an assessment of knee arthroscopy simulations. Noureldin et al (2016), on the other hand, assessed performance independently by an observer using a rating scale, despite the availability of automated data.

Jacobsen et al (2015) identified a possible limitation of VR systems with a lack of discrimination between groups on the only therapeutic, rather than investigative, orthopaedic procedure included in the assessment (resection of a tear in the medial meniscus). While a detailed explanation was not offered, the implication for consequential validity is noted and it is possible a therapeutic simulation requires more interactivity than a relatively static investigation.

Scalability may also be an issue in VR, and with that associated costs. VR systems may be restricted by specific hardware and software needed for procedures (eg Chowriappa et al 2013), limiting what can be assessed without multiplying costs. On the other hand, costs for equipment cleaning are reduced (Vassiliou et al 2014).

Assessment of professional skills

Virtual patients constitute a different form of virtual reality, being interactive computer-based representations of patients rather than the physical environment. These can range from highly immersive to text only. At the upper end, Deladisma et al (2007) compared second year medical students' non-verbal communication during interactions with a 'real' SP or an interactive virtual patient (VP) presented as a life-sized computer animation projected onto a wall. The VP and SP had identical scripted responses to student questions, and to prompt an empathetic response during the interview the VP or SP stated "I am scared; can you help me?". Clinician raters rated non-verbal communication skills, empathy and overall performance higher in the encounters with the SP than the VP, and this was suggested to be the result of both student and assessor bias regarding the artificial nature of the VP interaction.

McGrath et al (2015) described a virtual patient encounter as an alternative to a mock oral examination where faculty describe a case and the candidate responds verbally. The virtual patient described was intended to add a degree of authenticity absent from the standard oral exam. The patient was controlled and voiced by a member of faculty. No difference in performance was found between virtual and oral exams, and participants found the virtual exam less intimidating.

The example described by Deladisma et al (2007) and McGrath et al (2015) are relatively sophisticated. We also found a number of more primitive, less immersive technologies which are described as virtual patients, but which rely on text and multimedia content in order to present an interactive case vignette for the assessment of clinical reasoning. A common platform is Web-SP ((http://websp.lime.ki.se/about), developed at the Karolinska Institute in Sweden. This, and similar technologies were used in studies looking at interactions in general practice (Waldmann et al 2008), surgery (Yang et al 2013), psychiatry (Williams et al 2011) and nursing (Forsberg et al 2015). These studies used technology which appears quite dated, although Web-SP is still a current system, but there is potential to update the basic paradigm with more sophisticated interface design and technology.

In terms of assessment mechanics, these approaches offer a form of script concordance test, with candidate questions, investigations and decisions being matched to expert-generated checklists of responses – in many cases through an automated process. Positive and negative marking systems can be applied based on the extent and type of decision which is taken (for example, deductions if too many tests are ordered).

However, the authenticity of an assessment may be increased if the scenarios are presented through an interface that gives that gives the appearance of a real patient. We found two studies comparing these systems with other forms. Botezatu et al (2010) found that students performed better on Web-SP exams than on paper-based scenarios, even if they had not used the Web-SP system before. While Botezatu et al recognised study limitations and confounds, it is possible that even a relatively static VP provides a more authentic experience than a paper case, and so more situated recall. Oliven et al (2011) compared performance in a web-based VP OSCE (not specified, but described similarly to Web-SP) with an SP OSCE, and found good correlation (>0.6) and no significant difference between means.

Finally, the study by Heinrichs et al (2008) described the use of a 'virtual world' for training and assessment in different emergency medicine scenarios. Virtual worlds are larger, animated and immersive virtual environments, derived from or inspired by computer games, and allow the user to move around and interact with people and objects. Virtual patients can be presented in virtual worlds (as in McGrath et al 2016), but the ability to explore the world is what makes the application described by Heinrichs et al distinct.

In their first study Heinrichs found no difference in performance on a leadership assessment scale, or attitudes to the simulation, between students who completed a matched six-scenario assessment in a low-fidelity virtual world or with a full-size simulator. In two other studies they piloted a higher-fidelity photo-realistic representation of a real emergency department to assess performance in two major incident scenarios. These did not gather assessment data, but found that trainees reported increased confidence after the scenarios, and had felt moderately immersed in the scenarios.

Practical implications

Virtual reality in all its forms offers ostensible advantages in flexibility, in cost savings of SP time and training, and greater standardisation. However, SPs provide more diversity and challenge, and assessment of a broader range of skills. In this way Deladisma et al (2007) suggested that the VR patient would not be suitable for high stakes exams with 'experienced' students and doctors.

Another advantage of VR, and indeed other simulators, is the easy and secure capture of a large amount of intrinsically reliable performance data. However, both Noureldin et al (2016) and Courteille et al (2008) noted the potential consequences of computer system failures leading to a loss of data. This illustrates the need for robust systems in high-stakes examinations.

Finally, VR may incur more capital outlay than other methods which may use existing space and resources, and cost effectiveness was not established in these studies. Virtual reality simulators are expensive, and may involve a dedicated assessment centre (as in Vassiliou et al, 2014). An in-depth cost-benefit analysis of implementation, including the costs of purchase and ongoing maintenance, is needed.

3.9 Types of assessment: Remote and mobile technology

3.9.1 Summary of evidence

Video technology to allow assessors to view simulated performance remotely is well established, but the growth of mobile technology means that this functionality is being constantly extended. This section includes just two assessments that used mobile technology, which is somewhat surprising. However, with the smartphone revolution barely a decade old, it may simply be that applications have not yet reached the literature. Four of the 13 studies involved medical students, four postgraduate trainees, one just faculty raters, and the remainder a mixture.

As such, we rate the evidence in this section as low. While there are validated assessments, developments in available technology are such that questions of authenticity in content and process are likely to be rendered moot by current, and future iterations.

Mobile technologies allow synchronous or asynchronous remote assessment to be carried out worldwide (Okrainec et al 2013, Everett et al 2013). This has potential to improve access to assessment for overseas doctors, and associated quality assurance of remote testing sites. Risks around security, and fairness, will need to be considered.

Chan et al (2014): Remote assessment of an undergraduate OSCE

OSCEs are usually carried out in a central location, which can preclude participation of experienced educators who are based at a distance. Chan et al (2014) demonstrated that technology may support remote assessment.

Forty 3rd year medical students taking a formative OSCE were assessed by examiner pairs, one based remotely and one present in the station. Remote examiners viewed the encounter via 1 or 2 internet protocol webcams installed in the station and had the facility to pan, tilt and zoom, as well as adjust the volume on the camera. Both remote and station examiners completed checklist and global rating scales. Password protected access to the webcams and checklists ensured security.

Scores of the station and remote examiner moderately-highly correlated on checklists, though local examiner scores were significantly higher. Correlation in global rating scores was more varied, perhaps reflecting differing observations of students' non-verbal interactions by remote examiners, but there was no significant difference in mean ratings given.

The approach was acceptable to the majority of students and examiners. There were no equipment failures, but there were some difficulties with audio quality. There was also a loss of data through the delayed start of some remote examiners in setting up the system. Investment in high quality cameras and technical support may surmount these issues.

The findings encourage further work to validate use in high stakes assessments.

There are no clear examples of good practice here, and in some respects the aim of such developments will be to integrate technology with such transparency that there is no measurable effect. To this end the studies by Chan et al (2014) and Ma et al (2015) which report comparison of the ratings of local and remote assessors may be important.

Table 10. Validity evidence for assessments involving remote or mobile technology

Reference	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes	Cost
Cendan JC, et al. (2017)	USA	Supervisors	26 faculty	Rubric defined by faculty group	Faculty training. Pre-post survey with quant and qual data	None given	None given	None given	1 hr training
Chan J, et al. (2014)	Canada	UG	40	None given	Rater training. High local-remote agreement on checklists, less on GRS	None given	Moderate agreement with checklist and GRS	Pass/fail by borderline groups method. Student and examiners largely felt it was valid assessment.	45-60 minute training. Purchase and installation of webcams.
Everett TC, et al. (2013)	Canada	Anaesthesia trainees	30	Scenarios and tools based on literature, reviewed and rehearsed to achieve consensus.	Rater training. ICC moderate to good - projected all good with 3 raters	Single-measure ICC moderate-good for checklists, lower for GRS	None given	Overall acceptable as assessment	3 x 3-hour training sessions
Ang WJJ, et al. (2014)	UK	Medical students-consultants	25	None given	Automated data	None given.	Students and junior doctors scored less than seniors on total movement but not average. No divergent validity for total from time alone. Hand dominance more apparent in juniors. Lower total and fewer fast movement associated with better outcome.	None given	None given
Jensen JT, et al. (2014)	Denmark	Novice and experienced endoscopy nurses	12	Developed through Delphi process	Rating scale justified. Participants given theoretical course. IRR moderate for scale, good for pass/fail. D-study predicts max G of 0.7	None given.	Experienced scored higher than nurses.	Global pass/fail judgement	None given
Kiehl C, et al. (2014)	Germany	UG	155	Scenarios written by surgeon	SP training. IRR >85%	Alpha > 0.5 overall. Low but acceptable for communication. Item difficulty and discrimination good	None given	None given	None given
Kneebone R, et al. (2007)	UK	PG trainees	> 200	Based on literature	None given	None given	None given	None given	None given
Lucas NC, et al. (2016)	New Zealand	UG	75	None given for scenario. Scale derived from literature.	IRR > 0.6	None given	No difference between year groups	20% passed. Pass mark with reference to literature.	None given
Ma IW, et al. (2015)	Canada	Not specified	18	None given	Assessor recall minimised by delay and distractors. Comments on recording quality. IRR high.	Alpha adequate for checklist and scale	No difference between direct and video rating	No detail given.	None given
Millington SJ, et al. (2009)	Canada	IM residents	30	Rating tool derived from literature.	IRR fair-moderate	None given	Performance improved after training	None given	None given
Nikouline A, et al. (2013)	Canada	Surgical team	28	Established assessment	ICC >0.9	None given	None given	Participants prefer Skype to Google Glass	None given
Okrainec A, et al. (2013)	Canada	Surgical trainees	20	Established assessment	ICC >0.9. Identification of barriers from video	None given	None given	Acceptable	None given

Reference	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes	Cost
Rutherford JS, et al (2015)	UK	Anaesthetic practitioner	48 raters	Video scenario scripts based on interview data.	Rater training and handbook. Rater feedback. IRR by element v variable (0.47-0.86)	Alpha > 0.7 for different dimensions. Biased by one video.	None given	Overall acceptable to raters	None given

Key:
UG= undergraduate; PG=postgraduate; IM=internal medicine
ICC= Intraclass Correlation Coefficient; IRR=Inter-rater reliability; GRS=Global Rating Scale; D-study=Decision study

3.9.2 Details of evidence

The technology enabling remote assessment by video link, whether real-time or deferred, is well established, and its use in assessment in medicine and other health professions is often routine, fairly trivial and accepted by students and faculty (Chan et al 2014). Many examples discussed in other sections have included video recording in the assessment process, though often without detailed commentary or analysis (eg Stroud et al 2009, Abu Dabrh 2016, Chipman 2011). Examples may be found in assessment of technical skills (eg Millington et al 2009, Jensen et al 2014), as well as basic communication skills (Rutherford et al 2015, Kiehl et al 2014, Lucas et al 2016; see Appendix C).

A more novel use was reported by Okrainec et al (2013) who used live video-conferencing to allow an examiner to monitor both a surgeon's behaviour and their performance inside a laparoscopic simulator remotely – a video feed of the surgeon and the simulator output were displayed simultaneously. They found very good inter-rater reliability between the remote and local raters. A conference paper by the same group (Nikouline et al 2013) extended this to use 'Google Glass' to provide the surgeon's first-person view, although they found no additional value of this.

Remote assessment by video may also allow assessors to be distributed geographically, including internationally. While this may have benefits, there is a risk that heterogeneity may limit reliability. Everett et al (2013) used raters both in Canada and the UK to assess paediatric anaesthetic trainees and found acceptable inter-rater reliability, and evidence that nationality, background, and practice culture of the assessor did not appear to affect scoring. However, this may not be true of more disparate cultures than Canada and the UK.

Two studies compared how local and remote assessors judge the same cases, with conflicting results. Chan et al (2014) compared local and remote real-time physician-examiners in an OSCE. Correlations between local and remote checklist scores were moderate to high, but were low to high between local and remote global rating scales, and varied by scenario. Mean checklist scores of local examiners were also found to be significantly higher than remote. In contrast, Ma et al (2015) compared assessment of video-recordings versus direct observation of central venous catheterization and found no significant differences in mean rating or checklist scores. Interrater reliabilities for both were high.

There were few examples where mobile technology was used to support assessment, perhaps because the ubiquity of powerful, versatile internet connected mobile devices is still relatively recent. However, there were suggestions in the literature that the 'always on' capability of modern mobile technology could be exploited to facilitate real-time assessment of professionalism.

An early example was a computer-based hybrid training and assessment tool described by Kneebone et al (2007). SPs and residents used handheld computers to input data in situ during a range of clinical scenarios (involving technical and non-technical skills, professionalism and empathic behaviour). Simulations also used multiple video viewpoints allowing remote assessment. The paper was published on the cusp of the smartphone explosion, and the model of assessment described could now be delivered even more flexibly.

More recently, Cendan et al (2017) described a mobile web-based platform and examined its feasibility and utility among a small group of faculty members. This application was designed for workplace settings, and allowed faculty to record examples of positive or negative behaviour mapped to a professionalism rubric with 25 behavioural elements in six domains. Faculty could therefore record behaviour in situ, and near contemporaneously. Faculty felt there was benefit, although they emphasised a need for a shared understanding of behaviours. While this was based on

workplace usage, a similar approach may have potential for simulated cases, capturing contemporaneous behaviour in interactions.

Finally, mobile and specifically wearable technology has potential for real-time automated data collection in real or simulated settings. This was indicated in a study by Ang et al (2014) who evaluated the use of motion data captured from a smartphone accelerometer to capture data on a surgeon's wrist movements, as a marker of precision and performance. This method was able to discriminate between training stages. While the clinical relevance of this data was questioned, and no performance correlates were reported, it illustrates potential applications for technologies now becoming widespread in 'smart watches', as well as smartphones.

3.10 Process of assessment: Sequential testing

3.10.1 Summary of evidence

Sequential testing refers to a way of improving the efficiency of testing by not requiring all students or candidates to undergo a full assessment. Put simply, those candidates who will safely pass all assessments do not need to demonstrate their competence as much as those who are borderline, or may fail. These results are combined with the initial screening assessment results, in order to provide a larger number of stations on which to assess the candidates. Statistically, this increases reliability and confidence in the internal structure of an assessment where it is needed – for those at risk of failure.

Our consideration of sequential testing is in contrast to the other areas we have considered, being concerned not with what is assessed, nor with how the assessed knowledge or skills behaviour are elicited or recorded, but rather with the *structure* of the assessment. The three examples we describe here all related to undergraduate OSCEs, but the principles of sequential testing are applicable to any content or format of assessment.

Due to a small number of studies, we rate the overall evidence here as low, although the validity evidence which is reported indicates the sequential process is robust and enables savings in time and resources.

Although we report three examples in the literature, just two described 'true' sequential testing. The third study features second marking only for those who do not clearly meet a passing standard. We suggest the underlying principle is the same, although the process is different. Both reduce the need for remediation or retesting based on a single examination, and may be economically appealing.

Judgements of content and response process validity are not inherently relevant here as they are functions of the assessments themselves rather than the sequential process. However, evidence of reliability across the phases of the sequential process is important. For any potential future application, the modelling approach described by Pell et al (2013), where the impact of sequential testing can be calculated from historical data, appears to be a useful means of assessing likely consequences.

The greater outstanding question is, perhaps, how this paradigm may be extended to other assessments, and also what its limitations are. Of all assessments it has the potential to most directly reduce costs without capital investment in equipment or estates, and so may be appealing when large cohorts are assessed in a short time period.

Table 11. Validity evidence for assessments involving sequential testing

Reference	Country	Group	Sample size	Content	Response process	Internal structure	Other variables	Outcomes	Cost
Cookson J, et al. (2011)	UK	UG	127	Not given	16% of variance from examiner x candidate	Generalisability 0.63 - 0.77. OSCE lower (0.38 - 0.55). Case-case variation high (49% of variance).	Moderate correlation between OSCE and OSLER.	Pass/fail on accumulation of penalty points. First stage - threshold to retain 1/3 of candidates. Final set on borderline groups	GBP55,608. Savings from sequential design approx GBP30,000
Mavis BE, et al. (2013)	USA	UG	132	Worksheet drafted from literature and reviewed by faculty	Rater training with review of videos from previous year. Rater agreement by item 63-98%	None given	Pass modelled as 1 or 2 'satisfactory' - with stricter criterion, closer match to SP ratings.	None given	None given
Pell G, et al. (2013)	UK	UG	228	None given	None given	Generalisability 0.67 with 12 stations, 0.8 with 24 stations	None given	Threshold placed at pass mark + 2SEM. Empirical results confirm model	Estimated at saving GBP29,000 on standard cost of GBP124,000

Key:
UG= undergraduate; PG=postgraduate
SP=Standardised or Simulated Patient

3.10.2 Details of evidence

Pell et al (2013) described the introduction of sequential testing in a UK medical school. After modelling the effects of introducing sequential testing using historical data, they reported findings on its introduction to a live assessment round. The majority of students passed the first round of 16 stations. Of 31 who progressed to the second round, seven failed, six of whom had scored lowest on the first round. Performance on the first round was therefore predictive of overall performance. Pell et al concluded that sequential testing produced robust results when compared to the use of a full OSCE sequence, and was more efficient for borderline cases than a resit of exams. Regarding practical impact, the authors estimated savings amounting to approximately £29,000 within a single year of assessment. They also stated that the elimination of false positives (identification of those who might inappropriately pass) in the screening test is a key quality feature that helped to gain external stakeholder acceptance of the shorter assessment.

A sequential process was also described by Cookson et al (2011; see also Wright et al 2014 discussed in the 'empathy' section) in a study involving retrospective analysis of results from medical finals that were based on Objective Structured Long Examination Record examinations (OSLERs, in which candidates see real patients), and OSCEs. The first stage included four OSLER cases and six OSCE stations, the second a further eight OSLER patients and six OSCE stations. Data from 127 participants were analysed. Performance on the first stage of the assessment was highly predictive of the results of the second stage of assessment (for candidates initially considered 'unsatisfactory'). However, there was only a weak correlation between the OSLER and the OSCE components of the exam, and the OSCE was less reliable across both stages, casting some doubt of construct validity. Cookson et al also reported savings from the design that amounted to approximately £30,000.

A different approach, but which has elements of sequential testing in avoiding the need for students to retake exams, was described by Mavis et al (2013). Here, the performance of students scored as borderline by SPs in an OSCE was reviewed by faculty. Recordings of those OSCEs were re-scored on a structured worksheet containing 26 items and an overall rating. Raters were blinded to the rating given by SPs in the OSCE. This effectively constituted a second mark for borderline cases, without requiring a second OSCE. The final decision on whether a student passed or failed the exam was based on the final overall assessment of both SPs and faculty. There was 85% agreement between rater pairs in the final assessment. Findings suggested that the review form/worksheet enabled faculty to validate pass/fail decisions based on SP ratings.

Sequential testing

The **University of Edinburgh** medical school introduced sequential testing to their Final examination OSCE format 4 years ago.

All students take the first 8 x10 minute OSCE station on a single site, with a variable number of students (usually 10-15%) attending for repeat testing in a second 8-station OSCE around 2 weeks later. Exempt students are those who achieve a score more than 2 Standard Errors of Measurement (SEM) above the cut-score. For those sitting both parts, pass-fail decisions are made on their performance averaged across both circuits.

Reliability of the whole exam is satisfactory (Cronbach's alpha=0.68) and advantages include fairness of approach, with less assessment burden for those students who have clearly passed, and a second chance for those who may have underperformed due to anxiety on the day. The model also allows examiner resource to be focused on the group of students where accurate decision-making is crucial.

Embedding the approach has required work to address perceptions that the second circuit is a 'resit exam'. Student anxiety has been managed through sharing formative feedback from the first circuit, which allows them to manage their preparation and direct practice to areas of noted weakness.

A key practicality is the administrative pressure to produce scores, psychometric analyses and feedback in order to facilitate set up of the second circuit within a timely interval. Use of electronic exam delivery systems, with marking on iPads, significantly reduces this administration time.

4 Discussion

In this review we set out to consider two broad research questions:

1. What evidence is there for good practice in the use, or potential use, of summative assessments around professionalism, ethics and competence in relation to patient safety?
2. What evidence is there for the use of simulation, or other technologically-mediated methods, for summative assessments in medical and non-medical contexts?

The initial search identified an extremely large set of potential papers even after screening, and so this was further constrained by prioritising those which best advanced the research questions. Nonetheless, the final review has included 248 papers from the search, a substantial number for a review of this kind (for comparison, systematic reviews in similar areas have considered 80 [Li et al 2017], 70 [Havyer 2016] and 73 [Archer et al 2015] papers).

We evaluated the included papers against a framework of validity (from Downing 2003), which also reflects consensus on criteria of good assessment (Norcini et al 2011). There was notable variability in the extent and quality of evidence in the literature, and very few examples where good practice was demonstrated comprehensively. We have highlighted some of these examples.

Here we recapitulate some of the key points from the review and consider the wider implications of these findings. We bring in other papers which did not meet inclusion criteria, but which inform interpretation, and the perspectives and experiences of our expert advisors.

4.1 Professionalism

The conceptual difficulties involved in defining professionalism mean that assessment in this area is similarly challenging. This notwithstanding, we found examples of well-evidenced assessments which address aspects of the broad concept of professionalism.

The term ‘professionalism’ is used in a number of ways, and we described in the introduction how it contains behavioural, attitudinal and social elements. Our review has identified that professional behaviour, based on performance in simulated scenarios, has been assessed in two main ways – either as a global judgement of interpersonal conduct (eg Berman et al 2009, Wright et al 2013, Zabar et al 2016), or as a judgement of specific communication performance where the manner of communication is important, such as in the expression of empathy (Sennekamp et al 2012), ‘complex’ communication scenarios (Mortsiefer et al 2014) and interprofessional team practice (eg Saylor et al 2016). We noted that assessments of empathy may not be statistically distinct from assessments of general communication skills, but do directly provide an important patient-centred perspective on those skills. We found just two paper-based tests of professionalism rather than observed behaviour, representing contrasting approaches (Moniz et al 2015, Tiffin et al 2011). We excluded assessments of attitudes or performance before medical school from detailed review, but there are examples of the assessment of related constructs as part of selection, such as a situational judgement test for integrity tested with applicants to medical school (de Leng et al 2018), and the use of multiple mini-interviews which may encompass professionalism (Reiter et al 2007, Hofmeister et al 2009).

There were also examples of negative scoring or the specification of unprofessional behaviour (eg de Leng et al 2018, Morris et al 2014). Such approaches may allow more precise definitions by focusing on what is clearly unacceptable, and may avoid the circularity of professionalism being defined essentially as behaviour that should be expected of a professional.

There were concerns expressed by assessment experts in our PAG that unprofessional behaviour may not be demonstrated in artificial assessment contexts. These include risks that when being visibly or consciously assessed candidates may be more guarded or cautious in how they express themselves. However, the question of whether assessment can elicit authentic behaviours is also true of practical skills and indeed knowledge tests. While workplace-

based or longitudinal assessments may represent the ideal case for capturing authentic behaviours, some artificiality and compromise is necessary for standardised and scalable assessment. Where performance is context specific (eg Balzora et al 2015, Baig et al 2009), careful construction of content may ensure assessment is triangulated. As evidence to mitigate the risk of inauthentic behaviour in assessments, Berman et al (2009) found ratings of professionalism given by programme directors in the workplace correlated with professionalism ratings in an OSCE. This is not a generalisable observation, but is encouraging.

The conceptualisation and understanding of professionalism in medicine has a long, and some may say protracted, history which appears to be no closer to clear resolution. It is a term with shared understanding to a large extent, but not in all details. Green et al (2009) found general agreement in the 'signs of physician professionalism' identified by patients, doctors and nurses, although some of these were rather broad (eg 'Practices in an ethical manner'). There were differences though, for example with 'keep patient and/or family up to date' seen as important by patients and nurses, but less than 75% of doctors. Conversely 'is personable and polite' was important to nurses, but not patients or doctors. There are therefore differences in the interpretation and definition of behavioural cues, depending on viewpoint.

'Professionalism' may not in fact be a useful term to consider in the context of assessment. A balance in the debate may be achieved by firstly identifying what elements are a priority for assessment, agreeing how these are defined, and then to establish how and where best to assess them. Where possible, assessed behaviours and skills will need to push beyond 'professionalism' to more precise language. Assessment of an undifferentiated holistic construct may then be considered on its own terms without conflation with those more defined areas (cf Burford et al 2014).

4.2 Content of assessment: Ethics

We defined ethics assessments as those concerned explicitly with the application of ethical principles and knowledge, rather than generic 'ethical behaviour', without clear definition of what makes such behaviour ethical. Defined in this way, good practice is indicated in concordance test approaches (Tsai et al 2012, Foucault et al 2015). These compare candidates' ethical reasoning with normative judgements and provide relative precision compared to behavioural assessments which may be confounded by practical constraints and communication abilities. Written exams also have the advantage of scalability across a large cohort at minimal marginal cost.

For ethical scenarios it is perhaps more important that content is appropriate to candidates' level of practice than it is for professionalism or communication scenarios. The ethical judgements and reasoning required of newly qualified doctors will differ from those entering more senior levels. For example, the ethical challenge of withdrawing treatment from a patient will not be a junior doctor's decision. Their ability to recognise and reason about such a case would still be expected, but they would not be expected to reach a decision in practice. For all doctors, good ethical practice may also vary over time, as the clinical, technical and legal context changes. This will require recognition and revision of assessment content so that candidates' up-to-date knowledge and application to practice is being assessed.

Standards of ethical practice may not be universal, and cultural norms may frame what constitutes good clinical practice in these domains (see Tsai et al 2009, Foronda et al 2015). It is important that assessments in this area are able to identify doctors who are not applying standards of practice expected or required in the UK, but are not over-sensitive to cultural differences which may not be vital for patient care.

4.3 Content of assessment: Patient safety

We focused on assessments of patient safety which considered candidates' understanding of safety as a process, rather than a consequence of technical competence. While the latter were well-represented, there was very little evidence falling into the former group despite a great deal of relevance and apparent interest across healthcare. We surmised this may be because it is seen as an element of in-practice continuing professional development, rather than something suitable for summative assessment.

As with ethics, the understanding of patient safety would seem to lend itself to assessments of ability to apply knowledge, such as script concordance or situational judgement tests which would isolate such ability from technical performance. We found no evidence of these being used, but they may be a fruitful avenue of further development.

Performance-based assessments relating to patient safety (eg Sternbach et al 2017) may involve candidates responding to artificially introduced error, but as with professionalism, this may test 'shows how' rather than 'does' levels of behaviour. We can speculate that candidates may be more attuned to error in such cases, unless completely blinded to the purpose of an assessment, although we did not find studies examining this. A good example approaching error laterally was provided by Daud-Gallotti et al (2011). This assessed candidates' understanding of error through the way in which they explained it to patients. While other 'complex communication' scenarios addressed disclosure of error, Daud-Gallotti et al's example explicitly marked candidates on their ability to explain *why* an error had occurred.

4.4 The use of technology in assessment

We considered simulation, virtual reality and remote and mobile technology as potential tools for the development or enhancement of novel assessment.

There is a wealth of literature on the use of simulation for the assessment of procedural and technical skills, although previous reviews have identified the evidence for these as being lacking (eg Cook et al 2013). Simulation is used extensively in basic and advanced life support training and assessment, but even in such a relatively constrained and protocol-driven activity, the question of the authenticity of behaviour is present, with limited transfer of learning from lower to higher fidelity simulators (Boet et al 2017).

Virtual reality is more novel, and its use for some procedural skills which rely on imaging that can be virtually represented (eg laparoscopic procedures) appears to be fairly robust, albeit for a narrow range of scenarios (Thijssen & Schijven 2010). Examples of assessments using interactive virtual patients – computer-based representations of simulated patients – were primitive, but rapid improvements in cost-effective animation and virtual reality may provide more scalable and cheaper ways of presenting standardised patient encounters than manikins or role-players. The example of Deladisma et al (2007), where a virtual patient was projected life-size onto a wall, illustrates what these could look like, but given the pace of technological development, it is hard to base judgements of good practice on examples from even a few years ago. There is potential with virtual technology to increase standardisation regardless of physical location. Augmented reality, where a computer-generated layer is superimposed real-time on real images, is increasingly available in consumer technology, and has the potential to provide new hybrid approaches such as adding detail to manikins, or displays to equipment.

We found extremely limited applications of mobile technology, but the ubiquity of highly sophisticated technology achieved in recent years may allow remote assessment in a wider variety of environments and locations, while also allowing assessors a closer but less intrusive view of candidate performance. One of our examples of practice from PAG members found the use of mobile technology elicited more detail in comments on workplace-based assessments, and it may be that the integration of technology into assessment practice would similarly allow more robust and detailed assessment of behaviours.

There are also potential trade-offs with the use of technology. While acceptance of remote assessment has been found to be high (eg Langenau et al 2014), the authenticity of assessments may be lower than in controlled physical environments, and reported discrepancies between local and remote assessors (eg Chan et al 2014) may need further examination. The literature and practice of telemedicine may have some bearing on how assessments may be developed (Europe Economics 2018).

However, the practical benefits of allowing IMGs, for example, to complete an assessment overseas while being assessed from the UK may outweigh marginal loss of authenticity. Remote technology may allow raters with specific expertise to assess higher-level competencies elsewhere in the world, or where confidentiality and blinding are challenges to summative assessment, for example in small specialties with a limited number of candidates and assessors.

These technologies would all require some capital investment in hardware and software, as well as in case development, against which any economies from increased throughput of candidates would need to be weighed. The technological infrastructure required to support real-time delivery and security of test material in mass assessments will not be trivial.

4.5 Sequential testing

Finally, we looked for evidence for sequential testing as a process of assessment. We found just two examples where 'true' sequential testing was used, both of which reported similar cost savings in the delivery of OSCEs, and robust pass rates (Cookson et al 2011; Pell et al 2014). Such savings would need to be modelled and confirmed with any new assessment, but the potential reduction of several OSCE stations for a proportion of any given cohort greatly improves capacity and scalability (just 30% of Cookson et al's sample and 14% of Pell et al's were recalled for the second stage of testing).

More speculatively, there is potential for sequential testing to be combined with technological approaches to assessment – for example, remote assessment for a first stage, followed by local assessment only for those below a threshold.

4.6 General issues

There are a number of issues which arose across all areas of the review.

4.6.1 Content of assessments

Specification of the content of assessments is not straightforward for the domains of 'professional skills' we have considered. The definition of concepts needs to be clear. In the literature, content validity is often provided by reference to consensus, and while this is legitimate, the scale and uniformity of consensus may not always be straightforward. For terms such as professionalism, consensus may mask semantic nuance, and agreement should be around specific rather than broad labels (see section 4.1). Assessments which purport to assess a particular element of practice may in fact be assessing something else (for example, an analysis of 280 communication checklists reported by Setyonugroho et al [2016] found that 34% of items were not actually considered by experts to reflect communication skills).

The potential of patients to contribute to assessments is possibly underutilised. While falling outside our core domains, a study by Hoffman et al (2015) described student and patient involvement in developing the content of an OSCE assessment of 'patient-centred care', identifying 25 discrete patient-centred behaviours. Patients also informed the format of the assessment, highlighting the importance of using visual aids.

Furman et al (2010) advocated a 'committee approach' to case development as this encouraged clarity and detail, as well as consideration of logistical needs. An extension to this, broadly supported by our PAG, may be a cooperative approach by national stakeholders to share, refine and agree standards of content. Lay or patient involvement may ensure that consensus regarding professional skills is not driven by professional norms or dominated by clinical judgements. Hodges & McNaughton (2009) suggested that role-players may be susceptible to the emotions elicited by playing the role and that their judgement of a candidate's competence might be influenced by their affective response to the encounter. However, for some domains such as empathy, that affective response *is* the judgement of competence, and eliminating it may negate the authenticity of an assessment.

A general note, regardless of the conceptual focus of assessment content, is that scenarios should be matched to the appropriate level of practice candidates are entering, and reflect the background of students and doctors who will be taking the assessment. While the same construct or performance may be being assessed, the appropriateness of the challenge and expected performance may drive the authenticity and plausibility from the candidates' point of view.

4.6.2 Process of assessment

How assessment is performed

Most assessments of professional skills use standardised patient based simulated scenarios, using checklists or rating scales. The overall approach is well-established in robust high-stakes usage. Despite this, there is conflicting evidence regarding the type of measurement, and the types of assessor, to be used. The statistical properties of different scales or checklists may vary with different populations, different scenarios, and different raters (Brannick et al 2011). A 'good', reliable assessment in one context may not be so in another, and development of assessments should consider how different uses may influence reliability. While concrete checklists may place less cognitive demand on assessors, if they require less interpretation of what is being observed, the literature suggests more abstract rating scales are often preferable (Adler et al 2011, Walzak et al 2015, Wass et al 2001, Ma et al 2012).

Response process also has a direct effect on what can be elicited. Kim et al (2009) found that multiple choice questions did not allow students' thought processes in clinical communication to be elicited, in comparison with free text. This is akin to what in research questionnaires are called the 'demand characteristics' of a tool – the details of a response are constrained or afforded by its modality.

Turner et al (2016) looked at the effect of the timing of SP feedback – whether given during a scenario, rather than just at the end as is usual. SP participants gave feedback on verbal and non-verbal performance of a standardised doctor in a videotaped scenario, either twice during as well as at the end of the scenario, or just at the end. The raters giving periodic feedback identified variation in performance over time, with more verbal cues in the middle section, and more non-verbal cues across the scenario. This difference in sensitivity may have relevance for summative applications. There may of course be damage to authenticity if a live scenario is interrupted, but recorded scenarios could be 'chunked' in this way.

Who performs the assessment

A second key element of process is *who* performs the assessment. Some variability can arise from different perspectives of assessors. Mazor et al (2007) found assessors varied in the behaviours they attended to, and how behaviours were evaluated. Different behaviours were considered when giving specific or global evaluations of professionalism. Chahine et al (2016) explored the rationale of examiners' decision-making in an OSCE for international medical graduates in Canada, and identified that examiners prioritised competency in 'Investigation & Management' over other domains. Lurie et al (2008) reported that for a given case, a student may receive high and low scores from different raters, while some assigned every student the same score, and others used the entire range of the scale. Conversely, Shirazi et al (2014) found that non-clinicians who had been extensively trained to use the Calgary-Cambridge checklist showed high inter-rater and test-retest reliability, and good correlation with expert raters. Variability can therefore be reduced, but such rater effects are a vulnerability.

Across the literature, the reliability of role-players' assessments is variable. Some studies reported good reliability measures and positive correlations with expert raters in assessment of communication skills (eg, Shirazi et al 2014; Bergus et al 2009) and clinical reasoning (Berger, 2012). Others reported only moderate reliability (Mema et al, 2016) and limited agreement with faculty (Whelan et al, 2009), or even substantial variability in scores. Mortsiefer et al (2017) found agreement between assessors was significantly higher in pairs of the same gender, and there were non-significant trends for greater consistency when both were practitioners, and, counterintuitively, when neither had received training. A national selection assessment for dentistry in the UK found that trained SPs showed moderate agreement with clinicians in scores for communication performance (Wiskin et al 2013).

The calibration of assessors' performance may also vary within an assessment. Hope & Cameron (2015) demonstrated that examiner stringency increased over a two-day summative OSCE for third year students, perhaps reflecting increasing experience of the performance of successful candidates.

The use of 'real' patients in assessments – ie those who are currently receiving treatment – is quite common in practice, although few examples were found in this review suggesting they may be used for clinical rather than

professional skills. Real patients are not standardised, which can be seen to limit the robustness of an assessment, but could be expected to elicit more authentic behaviours than SPs. However, Jabeen (2013) found that student performance in basic communication did not differ between SPs and real patients.

Our PAG voiced reservations about use of lay assessors in summative assessments, though there was some experience of simulated patients co-assessing with faculty. However, the implication though is that raters, whether lay or expert, require training. The extent of training required may depend on the outcome of the assessment: a checklist, where a behaviour needs to be recognised, may be simpler to train than a global rating scale that requires some subjective evaluation. Comprehensive training and calibration of rater response is time-consuming (eg Wouda and van der Wiel 2012), and this should be, but it seems rarely is, factored into assessment design. Training may also be desirable to offset assessor bias arising from personality traits. Finn et al (2014) noted that examiner stringency correlated with certain personality traits – negatively with openness to experience and positively with neuroticism.

The implication is that an understanding of the judgements made by raters is necessary to inform a valid and fair assessment process. Authentic assessment should take account of how examiners approach the assessment process and conceptualise the relevant competency. Regardless of approach, good practice should involve consideration of psychometric properties of an assessment format from the outset (Furman et al, 2010).

4.6.3 Feasibility of assessments

Few studies explicitly offered details of the feasibility of assessments in any detail. There is a complex balance of costs and benefits which is not addressed in the literature. Capital investment in infrastructure and technology, recurrent costs of hiring and training role-players, and hidden costs of faculty time, are all factors in the sustainability and scalability of an assessment.

At a basic level, statistical projections of the required numbers of stations or tests provide some evidence, but these are linked to the statistical properties of specific assessments and are not readily transferable to future assessments. The threshold of feasibility in this regard may also vary with other parameters, such as rater availability, and potential technological solutions, such as remote assessment. However, there is likely to be a ceiling on the number of practicable stations, and developing authentic content for a sufficiently large number of stations may preclude use of a particular assessment tool even where statistical models indicate reliability. The psychometric benefits of an expensive multi-station OSCE may not justify the additional resources required over an alternative approach if those benefits are marginal. For example, Lievens and Patterson (2011) found that simulation-based assessment was a marginally better predictor of real-world performance than a situational judgement test (SJT), but the SJT was considerably cheaper and more feasible for mass assessment.

4.6.4 Equality and diversity in assessment

For some assessments there is evidence of systematic differences between men and women, which appear to be assumed to reflect underlying differences in communication skills with gender. While this is often seen as unproblematic, and indeed an indication of criterion validity, the risks of biased design or measurement should not be discounted.

Such bias is more immediately problematic where it may lead to differences between ethnic or national groups, and is particularly relevant when assessments may be undertaken by home and international graduates. Studies have found lower pass rates among international medical graduates (eg Schenarts et al 2008, MacLellan et al 2010, Guttormsen et al 2013 – see Appendix D), but the root cause of this variation is not necessarily consistent across assessments.

There is a problem here to tease out – exams should be fair, but fairness is not necessarily demonstrated by equal performance. Differential attainment within an educationally homogenous group (for example ethnic minorities within a cohort of UK graduates) is a concern, but differences in communication performance, for example, may be pertinent if they reflect different values or norms of communication. Some, such as withholding information from patients or

telling families of diagnoses before patients, are normal and acceptable in some cultures, but would not be acceptable in UK practice. Other differences, such as non-verbal expressions of empathy, may not be directly pertinent to practice.

Identifying differences that are relevant is not a statistical question, but requires detailed consideration of meaning. Differences which are not relevant to successful and safe practice should not present obstacles to candidates, but those that have adverse consequences for patient care or experience must be addressed. Further work establishing the impact on practice and patient perceptions of culturally-different communication norms may be helpful.

The ways in which doctors identify and address some ethical issues (including end of life care, harassment, patient autonomy, gender issues and conflict of interest) may also be affected by local cultural norms (Jameel et al 2015, Tsai et al 2009, Sobani et al 2013). Norms around family-social structures (accepted roles of children and seniors), views on the value of life and use of health care resources may affect the way in which these issues are considered. They may have greater practical relevance for preparation of overseas doctors wishing to practise in the UK, but provide important contextual information for an assessment strategy affecting doctors from differing cultural backgrounds.

This balance of fairness, pragmatism and safety is a delicate one. Any regulatory obstacle to practice will need to be secure and evidenced, to instil confidence in examinees, and the public.

4.6.5 Approaches to standard setting

Standard setting for assessments was rarely presented in the literature we have considered, and where reported at all lacks transparency and justification. We found very few instances where standard setting, or the calculation of cut-scores was clearly referred to. This reflects the few examples of 'live' assessments found, which may reflect that such assessments tend to be developed and evaluated in-house rather than disseminated. Even where such evidence was provided, little justification was given, and did not present consideration of what standard setting method may be appropriate (Norcini 2003, Academy of Medical Royal Colleges 2015). One example considered different approaches to standard setting in an OSCE for physicians' assistants, and differences between the borderline groups and Angoff methods (Carlson et al 2010). The borderline groups method resulted in a higher cut-score (76% compared to 62%), and was more reliable, but harder to implement.

There are also considerations of how scores are combined in multi-faceted assessments. Park et al (2016) found that differential weighting of components (clinical skills and patient notes), as indicated by faculty, affected the reliability of the examination, but not the pass/fail outcomes, for USMLE candidates. Detailed consideration of these technical approaches to scoring and standard setting fell outside the scope of our review, but the weight given to different elements and the derivation of pass scores may have consequences for the functioning of an assessment.

4.6.6 Implications for professional skills assessments

We have set out some broad implications for those involved in assessments, which may help to define new areas of development or research. In this section we describe elements that may inform the specification of future GMC assessments, including the MLA Clinical and Professional Skills Assessment.

Table 12 sets these out. The broad 'Outcomes' were identified by our PAG as important high-level requirements for any high-stakes assessment such as the MLA. The 'specification' combines our observations from the evidence, and PAG members' comments, to suggest how these outcomes may be assured. We stress that this is an illustrative suggestion rather than a recommendation.

Table 12. Suggested outcomes and specifications for high stakes assessment

Outcome	Specification
Authentic content Assessment should demonstrate that content has been developed to reflect the context in which doctors will be practising.	'Patient' involvement in scenario design
	Domains (professionalism, ethics, patient safety) reflect contemporary issues in real-world practice, eg, rationalising NHS resources
	Domains reflect issues of cultural sensitivity, eg, asylum seekers
	Reflect multi-professional workforce
	A cooperative approach among stakeholders to allow sharing and refinement of content.
Authentic assessment Format of assessment should reflect the holistic reality of practice.	'Patient' involvement in design of assessment rubrics
	Involvement of lay assessors (in addition to expert assessors), with suitable training, to ensure standardised and reliable outcomes.
	Involvement of non-medical healthcare professional assessors
	Assessment setting should recreate the workplace in a) use of technology in authentic practice, b) representation of inter-professional workforce (eg through use of one or more 'standardised professionals')
	Assessments should reflect appropriately independent decision-making rather than elicit default 'safe' behaviour (ie, simply 'calling senior' may not necessarily be a positive outcome)
Fairness / equity Assessments should not unfairly penalise doctors from different educational or cultural backgrounds, while ensuring that standards of competence are appropriate for UK practice. (This is primarily relevant to IMGs, but there may be other issues relevant to UK graduates).	Training of assessors (lay and professional) to minimise risk of unconscious bias. A cooperative approach among stakeholders to allow sharing and standardisation of training.
	Assessment criteria should remove cultural differences and clearly distinguish inadequate cultural awareness.
	Assessments should avoid biasing of previous simulation experience (eg by providing orientation trials)
Feasibility Assessments should demonstrate sustainable capacity for throughput of students / applicants.	Assessments should specify time and resources (infrastructure, equipment, staff, training).
	Wherever possible, assessment environments should use technological resources in assessment design efficiently and cost-effectively, while retaining robustness and authenticity.
Standard setting	Process of standard setting and establishing cut-scores should not only be statistically robust, but should reflect practice relevance.

5 Conclusion

The central message of this research is that while the practice of assessment overall is mature and well-evidenced, there are gaps in the approach to assessment of professional skills. Assessments of these areas of practice may not be simply adapted from assessments of clinical or technical skills, but need consideration of the theoretical underpinnings of *what* is being assessed, and *how* this should be done.

The evidence is too grounded in the context and assumptions of specific use to indicate appropriate 'off the shelf' assessments. However, with some 'top down' specification of what an assessment should be considering, there are examples which suggest good practice.

New technologies may allow different paradigms of assessment, but these have not as yet reached published literature. Ubiquitous mobile technology may open up new avenues for remote assessment, using virtual and augmented reality, and this remains to be explored for its potential impact on scalability and accessibility of assessments. However, such potential will need to be weighed up in terms of genuine gains, and set against risks of authenticity, cost, flexibility and security.

5.1 Future work

Finally, while this project has identified a great deal of empirical literature – more than was anticipated – there remain gaps in understanding. Much of the literature is small scale, or locally focused. There are practical and political questions to be addressed in future work, both in terms of primary research and policy development. Here we suggest some of these avenues of work. A programmatic body of work to specify, develop and evaluate assessments is indicated.

Specification of content

Firstly, we have noted that there is still conceptual confusion around many of the domains we have considered in this review. Blueprinting is essential for good assessment, and it is important that the blueprint has sufficient detail to allow content to be specified appropriately. Theoretically informed research may resolve some of the conceptual and semantic problems necessary to achieve meaningful consensus on relevant and authentic content for assessments, particularly around the nature and limitations of 'professionalism' as a concept. It may be that such over-arching terms have to be left to one side when it comes to the detailed specification of assessments.

Specification and quality assurance of assessments

Above, we have suggested some broad outcomes that may guide the development of future assessments. These offer a starting point to inform development of standards for a national MLA, which can be delivered by medical schools, and potentially other organisations, and can be quality assured by the GMC. However, concurrent sense-checking and evaluation of assessment content and process as these standards develop will be appropriate.

Developing scalable approaches

With increasing numbers of UK graduates, and unknown numbers of IMGs wishing to take the MLA in future, development of measures to ensure the scalability of the CPSA, and medical schools' parallel versions, will be important. Cooperative approaches to assessment – whether through sharing of resources and scenarios in consortia, or even in developing assessment centres – may be one way. Technology, specifically remote, mobile and virtual reality may, in future, also support such scalability. Feasibility studies and piloting will be necessary to establish the potential and limitations of such approaches.

References

ABIM. (1995) Project Professionalism. Philadelphia, PA: American Board of Internal Medicine

Abu Dabrh AM, Murad MH, Newcomb RD, et al. (2016) Proficiency in identifying, managing and communicating medical errors: feasibility and validity study assessing two core competencies. BMC Medical Education, 16: 233

Academy of Medical Royal Colleges. (2015) Guidance for standard setting: A framework for high stakes postgraduate competency-based examinations. London: AoMRC [https://www.aomrc.org.uk/wp-content/uploads/2016/05/Standard_setting_framework_postgrad_exams_1015.pdf accessed 7 July 2018]

Adams J, Triola M, Djukic M, et al. (2013) Patient Safety and Interprofessional Collaboration Assessment: A Distinct Skills Set for Medical Students. Journal of General Internal Medicine, 28: S136-S137

Adler MD, Vozenilek JA, Trainor JL, et al. (2011) Comparison of checklist and anchored global rating instruments for performance rating of simulated pediatric emergencies. Simulation in Healthcare, 6: 18-24

Althuis MD, Weed DL. (2013) Evidence mapping: methodologic foundations and application to intervention and observational research on sugar-sweetened beverages and health outcomes. American Journal of Clinical Nutrition, 98: 755-768

Ang WJJ, Hopkins ME, Partridge R, et al. (2014) Validating the use of smartphone-based accelerometers for performance assessment in a simulated neurosurgical task. Neurosurgery, 10 Suppl 1: 57-64; discussion 64-55

Archer J, Lynn N, Coombes L, et al. (2016) The impact of large scale licensing examinations in highly developed countries: a systematic review. BMC Medical Education, 16: 212

Archer J, Lynn N, Roberts M, et al. (2015) A Systematic Review on the impact of licensing examinations for doctors in countries comparable to the UK. London: General Medical Council

Baig LA, Violato C, Crutcher RA. (2009) Assessing clinical communication skills in physicians: are the skills context specific or generalizable. BMC Medical Education, 9: 22

Balzora S, Abiri B, Wang XJ, et al. (2015) Assessing cultural competency skills in gastroenterology fellowship training. World Journal of Gastroenterology, 21: 1887-1892

Banerjee A. (2015) Using simulation for primary certification. International Anesthesiology Clinics, 53: 42-59

Batelden PB, Davidoff F. (2007) What is “quality improvement” and how can it transform healthcare?. Quality & Safety in Healthcare, 16: 2-3

Benedict N, Smithburger P, Donihi AC, et al. (2017) Blended Simulation Progress Testing for Assessment of Practice Readiness. American Journal of Pharmaceutical Education, 81: 14

Bensfield LA, Olech MJ, Horsley TL. (2012) Simulation for high-stakes evaluation in nursing. Nurse Educator, 37: 71-74

Berg K, Majdan JF, Berg D, et al. (2011) Medical students' self-reported empathy and simulated patients' assessments of student empathy: an analysis by gender and ethnicity. Academic Medicine, 86: 984-988

Berger AJ, Gillespie CC, Tewksbury LR, et al. (2012) Assessment of medical student clinical reasoning by "lay" vs physician raters: inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination. American Journal of Surgery, 203: 81-86

Bergus GR, Woodhead JC, Kreiter CD. (2009) Trained lay observers can reliably assess medical students' communication skills. Medical Education, 43: 688-694

Berman JR, Lazaro D, Fields T, et al. (2009) The New York City Rheumatology Objective Structured Clinical Examination: five-year data demonstrates its validity, usefulness as a unique rating tool, objectivity, and sensitivity to change. Arthritis & Rheumatism, 61: 1686-1693

Black SA, Nestel DF, Kneebone RL, Wolfe JH. (2010) Assessment of surgical competence at carotid endarterectomy under local anaesthesia in a simulated operating theatre. British Journal of Surgery, 97: 511-516

Bloom-Feshbach K, Casey D, Schulson L, et al. (2016) Health Literacy in Transitions of Care: An Innovative Objective Structured Clinical Examination for Fourth-Year Medical Students in an Internship Preparation Course. Journal of General Internal Medicine, 31: 242-246

Boet S, Bould MD, Pigford AA, et al. (2017) Retention of Basic Life Support in Laypeople: Mastery Learning vs. Time-based Education. Prehospital Emergency Care, 21: 362-377

Bogossian FE, Cooper SJ, Cant R, et al. (2015) A trial of e-simulation of sudden patient deterioration. (FIRST2ACT WEB) on student learning. Nurse Education Today, 35: e36-42

Boon K, Turner J. (2004) Ethical and professional conduct of medical students: review of current assessment measures and controversies. Journal of Medical Ethics, 30: 221-226

Botezatu M, Hult H, Tessma MK, Fors UG. (2010) Virtual patient simulation for learning and assessment: Superior results in comparison with regular course exams. Medical Teacher, 32: 845-850

Brannick MT, Erol-Korkmaz HT, Prewett M. (2011) A systematic review of the reliability of objective structured clinical examination scores. Medical Education, 45: 1181-1189

Burford B, Morrow G, Rothwell C, et al. (2014) Professionalism education should reflect reality: findings from three health professions. Medical Education, 48: 361–374

Carlin et al. (2011) The Health Professional Ethics Rubric: Practical Assessment in Ethics Education for Health Professional Schools. Journal of Academic Ethics, 9: 277-290

Carlson J, Tomkowiak J, Knott P. (2010) Simulation-based examinations in physician assistant education: A comparison of two standard-setting methods. The Journal of Physician Assistant Education, 21: 41821

Cendan JC, Castiglioni A, Johnson TR, et al. (2017) Quantitative and Qualitative Analysis of the Impact of Adoption of a Mobile Application for the Assessment of Professionalism in Medical Trainees. Academic Medicine, 92: S33-S42

Chahine S, Holmes B, Kowalewski Z. (2016) In the minds of OSCE examiners: uncovering hidden assumptions. Advances in Health Sciences Education, 21: 609-625

Chan DK, Gallagher TH, Reznick R, et al. (2005) How surgeons disclose medical errors to patients: a study using standardized patients. Surgery, 138: 851-858

Chan J, Humphrey-Murto S, Pugh DM, et al. (2014) The objective structured clinical examination: can physician-examiners participate from a distance?. Medical Education, 48: 441-450

Chander B, Kule R, Baiocco P, et al. (2009) Teaching the competencies: using objective structured clinical encounters for gastroenterology fellows. Clinical Gastroenterology & Hepatology, 7: 509-514

Chen DC, Pahilan ME, Orlander JD. (2010) Comparing a self-administered measure of empathy with observed behavior among medical students. Journal of General Internal Medicine, 25: 200-202

Chen JY, Chin WY, Fung CS, et al. (2015) Assessing medical student empathy in a family medicine clinical test: validity of the CARE measure. Medical Education Online, 20: 27346

Chipman JG, Beilman GJ, Schmitz CC, Seatter SC. (2007) Development and pilot testing of an OSCE for difficult conversations in surgical intensive care. Journal of Surgical Education, 64: 79-87

Chipman JG, Webb TP, Shabahang M, et al. (2011) A multi-institutional study of the Family Conference Objective Structured Clinical Exam: a reliable assessment of professional communication. American Journal of Surgery, 201: 492-497

Chowriappa AJ, Shi Y, Raza SJ, et al. (2013) Development and validation of a composite scoring system for robot-assisted surgical training--the Robotic Skills Assessment Score. Journal of Surgical Research, 185: 561-569

Collins LG, Schrimmer A, Diamond J, Burke J. (2011) Evaluating verbal and non-verbal communication skills, in an ethnogeriatric OSCE. Patient Education & Counseling, 83: 158-162

Cook DA, Brydges R, Zendejas B, et al. (2013) Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. Academic Medicine, 88: 872-883

Cookson J, Crossley J, Fagan G, et al. (2011) A final clinical examination using a sequential design to improve cost-effectiveness. Medical Education, 45: 741-747

Cooper JB, Gaba DM, Liang B, et al. (2000) The National Patient Safety Foundation agenda for research and development in patient safety. Medscape General Medicine, 2: E38

Courteille O, Bergin R, Stockeld D, et al. (2008) The use of a virtual patient case in an OSCE-based exam--a pilot study. Medical Teacher, 30: e66-76

Cruess RL, Cruess SR, Steinert Y. (2016) Amending Miller's Pyramid to Include Professional Identity Formation. Academic Medicine, 91: 180-185

Daud-Gallotti RM, Morinaga CV, Arlindo-Rodrigues M, et al. (2011) A new method for the assessment of patient safety competencies during a medical school clerkship using an objective structured clinical examination. Clinics, 66: 1209-1215

de Leng WE, Stegers-Jager KM, Born MP, Themmen APN. (2018) Integrity situational judgement test for medical school selection: judging 'what to do' versus 'what not to do'. Medical Education, 52: 427-437

Deladisma AM, Cohen M, Stevens A, et al. (2007) Do medical students respond empathetically to a virtual patient?. American Journal of Surgery, 193: 756-760

Dillon GF, Clauser BE. (2009) Computer-delivered patient simulations in the United States Medical Licensing Examination. (USMLE). Simulation in Healthcare, 4: 30-34

Dow AW, Boling PA, Lockeman KS, et al. (2016) Training and Assessing Interprofessional Virtual Teams Using a Web-Based Case System. Academic Medicine, 91: 120-126

Downing SM. (2003) Validity: on meaningful interpretation of assessment data. Medical Education, 37: 830-837

Dwyer T, Glover Takahashi S, Kennedy Hynes M, et al. (2014) How to assess communication, professionalism, collaboration and the other intrinsic CanMEDS roles in orthopedic residents: use of an objective structured clinical examination. (OSCE). Canadian Journal of Surgery, 57: 230-236

Eckles RE, Meslin EM, Gaffney M, Helft PR. (2005) Medical ethics education: where are we? Where should we be going? A review. Academic Medicine, 80: 1143-1152

European Economics. (2018) Regulatory approaches to telemedicine. London: General Medical Council. [https://www.gmc-uk.org/-/media/documents/Regulatory_approaches_to_telemedicine.docx_73978543.docx accessed 7 July 2018]

Everett TC, Ng E, Power D, Marsh C, Tolchard S. (2013) The Managing Emergencies in Paediatric Anaesthesia global rating scale is a reliable tool for simulation-based assessment in pediatric anesthesia crisis management. Paediatric Anaesthetics, 23: 1117-1123

Farnan JM, Paro JA, Rodriguez RM, et al. (2010) Hand-off education and evaluation: piloting the observed simulated hand-off experience (OSHE). Journal of General Internal Medicine, 25: 129-134

Favia A, Frank L, Gligorov N, Birnbaum S et al. (2013) A model for the assessment of medical students' competency in medical ethics. AJOB Primary Research, 4: 68-83

Finn Y, Cantillon P, Flaherty G. (2014) Exploration of a possible relationship between examiner stringency and personality factors in clinical assessments: a pilot study. BMC Medical Education, 14: 1052

Foronda CL, Alhusen J, Budhathoki C, et al. (2015) A mixed-methods, international, multisite study to develop and validate a measure of nurse-to-physician communication in simulation. Nursing Education Perspectives, 36: 383-388

Forsberg E, Ziegert K, Hult H, Fors U. (2015) Evaluation of a novel scoring and grading model for VP-based exams in postgraduate nurse education. Nurse Education Today, 35: 1246-1251

Foucault A, Dube S, Fernandez N, et al. (2015) Learning medical professionalism with the online concordance-of-judgment learning tool (CJLT): A pilot study. Medical Teacher, 37: 955-960

Frank JR, Snell L, Sherbino J. (eds). (2015) CanMEDS 2015: Physician Competency Framework. Ottawa: Royal College of Physicians and Surgeons of Canada

Frischknecht AC, Kasten SJ, Hamstra SJ, et al. (2013) The objective assessment of experts' and novices' suturing skills using an image analysis program. Academic Medicine, 88: 260-264

FSMB/NBME. (2014) USMLE® Physician Tasks/Competencies. Federation of State Medical Boards of the United States, Inc. (FSMB), and the National Board of Medical Examiners

Furman GE, Smee S, Wilson C. (2010) Quality assurance best practices for simulation-based examinations. Simulation in Healthcare, 5: 226-231

Ginsburg LR, Tregunno D, Norton PG, et al. (2015) Development and testing of an objective structured clinical exam. (OSCE) to assess socio-cultural dimensions of patient safety competency. BMJ Quality and Safety, 24: 188-194

GMC. (2017) Generic professional capabilities framework. London: General Medical Council

Gondim Teixeira PA, Cendre R, Hossu G, et al. (2017) Radiology resident MR and CT image analysis skill assessment using an interactive volumetric simulation tool - the RadioLOG project. European Radiology, 27: 878-887

Gorniewicz J, Floyd M, Krishnan K, et al. (2017) Breaking bad news to patients with cancer: A randomized control trial of a brief communication skills training module incorporating the stories and preferences of actual patients. Patient Education & Counseling, 100: 655-666

Green M, Zick A, Makoul G. (2009) Defining professionalism from the perspective of patients, physicians, and nurses. Academic Medicine, 84: 566-573

Gude T, Grimstad H, Holen A, et al. (2015) Can we rely on simulated patients' satisfaction with their consultation for assessing medical students' communication skills? A cross-sectional study. BMC Medical Education, 15: 225

Guttormsen S, Beyeler C, Bonvin R, et al. (2013) The new licencing examination for human medicine: from concept to implementation. Swiss Medical Weekly, 143: w13897

Haffery FW, Castellini B. (2010) The increasing complexities of professionalism. Academic Medicine, 85: 288-301

Hart M. (2017) Medical Licensing Assessment – proposals for delivery. GMC Council paper, 12 December 2017. London: General Medical Council [https://www.gmc-uk.org/M05___MLA_proposals_for_delivery.pdf_72840551.pdf accessed 7 July 2018]

Havyer RD, Nelson DR, Wingo MT, et al. (2016) Addressing the interprofessional collaboration competencies of the association of american medical colleges: a systematic review of assessment instruments in undergraduate medical education. Academic Medicine, 91: 865-888

Havyer RD, Wingo MT, Comfere NI, et al. (2014) Teamwork assessment in internal medicine: a systematic review of validity evidence and outcomes. Journal of General Internal Medicine, 29: 894-910

Heinrichs WL, Youngblood P, Harter PM, Dev P. (2008) Simulation for team training and assessment: case studies of online training with virtual worlds. World Journal of Surgery, 32: 161-170

Hodges B. (2007) A socio-historical study of the birth and adoption of the Objective Structured Clinical Examination. (OSCE).

Hodges BD, Ginsburg S, Cruess R et al. (2011) Assessment of professionalism: recommendations from the Ottawa 2010 Conference. Medical Teacher, 33: 354-363

Hodges BN, McNaughton N. (2009) Who should be an OSCE examiner? Academic Psychiatry, 33: 282-284

Hoffman KG, Griggs M, Donaldson JF, et al. (2015) Through patient eyes: Can third-year medical students deliver the care patients expect? Medical Teacher, 37: 684-692

Hofmeister M, Lockyer J, Crutcher R. (2009) The multiple mini-interview for selection of international medical graduates into family medicine residency education. Medical Education, 43: 573-579

Hope D, Cameron H. (2015) Examiners are most lenient at the start of a two-day OSCE. Medical Teacher, 37: 81-85

Howells NR, Brinsden MD, Gill RS, et al. (2008) Motion analysis: a validated method for showing skill levels in arthroscopy. Arthroscopy, 24: 335-342

IPEC. (2011) Core Competencies for Interprofessional Collaborative Practice. Interprofessional Education Collaborative [https://www.umassmed.edu/globalassets/office-of-educational-affairs/ipeg/collaborativepractice.pdf.pdf accessed 7 July 2018]

Jacobsen ME, Andersen MJ, Hansen CO, Konge L. (2015) Testing basic competency in knee arthroscopy using a virtual reality simulator: exploring validity and reliability. Journal of Bone and Joint Surgery, 97: 775-781

Jameel A, Noor SM, Ayub S, et al. (2015) Feasibility, relevance and effectiveness of teaching and assessment of ethical status and communication skills as attributes of professionalism. Journal of the Pakistan Medical Association, 65: 721-726

Jefferies A, Simmons B, Tabak D, et al. (2007) Using an objective structured clinical examination. (OSCE) to assess multiple physician competencies in postgraduate training. Medical Teacher, 29: 183-191

Jensen JT, Konge L, Moller A, et al. (2014) Endoscopy nurse-administered propofol sedation performance. Development of an assessment tool and a reliability testing model. Scandinavian Journal of Gastroenterology, 49: 1014-1019

Ju M, Berman AT, Hwang WT, et al. (2014) Assessing interpersonal and communication skills in radiation oncology residents: a pilot standardized patient program. International Journal of Radiation Oncology Biology Physics, 88: 1129-1135

Kassam A, Cowan M, Donnon T. (2016) An objective structured clinical exam to measure intrinsic CanMEDS roles. Medical Education Online, 21: 31085

Kaul P, Barley G, Guiton G. (2012) Medical student performance on an adolescent medicine examination. Journal of Adolescent Health, 51: 299-301

Kaul P, Gong J, Guiton G, et al. (2014) Measuring pediatric resident competencies in adolescent medicine. Journal of Adolescent Health, 55: 301-303

Kiehl C, Simmenroth-Nayda A, Goerlich Y, et al. (2014) Standardized and quality-assured video-recorded examination in undergraduate education: informed consent prior to surgery. Journal of Surgical Research, 191: 64-73

Kim S, Spielberg F, Mauksch L, et al. (2009) Comparing narrative and multiple-choice formats in online communication skill assessment. Medical Education, 43: 533-541

Kneebone R, Bello F, Nestel D, et al. (2007) Training and assessment of procedural skills in context using an Integrated Procedural Performance Instrument. (IPPI). Studies in Health Technology & Informatics, 125: 229-231

Knudson MM, Khaw L, Bullard MK, et al. (2008) Trauma training in simulation: translating skills from SIM time to real time. Journal of Trauma, 64: 255-264

Konge L, Annema J, Clementsen P, et al. (2013) Using virtual-reality simulation to assess performance in endobronchial ultrasound. Respiration, 86: 59-65

Langenau E, Kachur E, Horber D. (2014) Web-based objective structured clinical examination with remote standardized patients and Skype: resident experience. Patient Education & Counseling, 96: 55-62

Latifi S, Gierl MJ, Boulais A-P, De Champlain AF. (2016) Using automated scoring to evaluate written responses in english and french on a high-stakes clinical competency examination. Evaluation & the health professions, 39: 100-113

Li H, Ding N, Zhang Y et al. (2017) Assessing medical professionalism: A systematic review of instruments and their measurement properties. PLoS ONE, 12: e0177321

Lie D, May W, Richter-Lagha R, et al. (2015) Adapting the McMaster-Ottawa scale and developing behavioral anchors for assessing performance in an interprofessional Team Observed Structured Clinical Encounter. Medical Education Online, 20: 26691

Lievens F, Patterson F. (2011) The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. Journal of Applied Psychology, 96: 927-940

Lohfeld L, Goldie J, Schwartz L, et al. (2012) Testing the validity of a scenario-based questionnaire to assess the ethical sensitivity of undergraduate medical students. Medical Teacher, 34: 635-642

Lucas NC, Walker N, Bullen C. (2016) Using a videotaped objective structured clinical examination to assess Knowledge In Smoking cessation amongst medical Students (the K.I.S.S. Study). Medical Teacher, 38: 1256-1261

Lupi C, Ward-Peterson M, Coxe S, et al. (2016) Furthering the validity of a tool to assess simulated pregnancy options counseling skills. Obstetrics and Gynecology, 128: 12s-16s

Lurie SJ, Mooney CJ, Nofziger AC, et al. (2008) Further challenges in measuring communication skills: accounting for actor effects in standardised patient assessments. Medical Education, 42: 662-668

Lynch DC, Surdyk PM, Eiser AR. (2004) Assessing professionalism: a review of the literature. Medical Teacher, 26: 366-373

Ma IW, Zalunardo N, Brindle ME, et al. (2015) Notes from the field: direct observation versus rating by videos for the assessment of central venous catheterization skills. Evaluation and the Health Professions, 38: 419-422

Ma IW, Zalunardo N, Pachev G, et al. (2012) Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. Advances in Health Sciences Education, 17: 457-470

MacLellan AM, Brailovsky C, Rainsberry P, et al. (2010) Examination outcomes for international medical graduates pursuing or completing family medicine residency training in Quebec. Canadian Family Physician, 56: 912-918

Macnaughton J. (2009) The dangerous practice of empathy. The Lancet, 373: 1940-1941

Martimianakis MA, Maniate JM, Hodges BD. (2009) Sociological interpretations of professionalism. Medical Education, 43: 829–837

Matos FM, Raemer DB. (2013) Mixed-realism simulation of adverse event disclosure: an educational methodology and assessment instrument. Simulation in Healthcare, 8: 84-90

Mavis BE, Wagner DP, Henry RC, et al. (2013) Documenting clinical performance problems among medical students: feedback for learner remediation and curriculum enhancement. Medical Education Online, 18: 20598

Mazor KM, Zanetti ML, Alper EJ, et al. (2007) Assessing professionalism in the context of an objective structured clinical examination: an in-depth study of the rating process. Medical Education, 41: 331-340

McGrath J, Kman N, Danforth D, et al. (2015) Virtual alternative to the oral examination for emergency medicine residents. West J Emerg Med, 16: 336-343

McLachlan JC, Finn G, Macnaughton et al. (2009) The conscientiousness index: a novel tool to explore students' professionalism. Academic Medicine, 84: 559-565

McTighe AJ, DiTomasso RA, Felgoise S, Hojat M. (2016) Correlation between standardize patients' perceptions of osteopathic medical students and students' self-rated empathy. J Am Osteopath Assoc, 116: 640-646

Mema B, Park YS, Kotsakis A. (2016) Validity and feasibility evidence of objective structured clinical examination to assess competencies of pediatric critical care trainees. Critical Care Medicine, 44: 948-953

Mercer SW, McConnachie A, Maxwell M et al. (2005) Relevance and practical use of the Consultation and Relational Empathy (CARE) measure in general practice. Family Practice, 22: 328-334

Meyerson SL, Tong BC, Balderson SS, et al. (2012) Needs assessment for an errors based curriculum on thoracoscopic lobectomy. Annals of Thoracic Surgery, 94: 368-373

Miller GE. (1990) The assessment of clinical skills/competence/performance. Academic Medicine, 65: S63–S67

Millington SJ, Wong RY, Kassen BO, et al. (2009) Improving internal medicine residents' performance, knowledge, and confidence in central venous catheterization using simulators. J Hosp Med, 4: 410-416

Moniz T, Arntfield S, Miller K, et al. (2015) Considerations in the use of reflective writing for student assessment: issues of reliability and validity. Medical Education, 49: 901-908

Morris MC, Gillis AE, Smoothey CO, et al. (2014) An alternative certification examination. ("ACE") in surgery. J Surg Educ, 71: 779-789

Mortsiefer A, Immecke J, Rotthoff T, et al. (2014) Summative assessment of undergraduates' communication competence in challenging doctor-patient encounters. Evaluation of the Dusseldorf CoMeD-OSCE. Patient Education & Counseling, 95: 348-355

Mortsiefer A, Karger A, Rotthoff T, et al. (2017) Examiner characteristics and interrater reliability in a communication OSCE. Patient Education & Counseling, 100: 1230-1234

Murrell VS. (2014) The failure of medical education to develop moral reasoning in medical students. Int J Med Educ, 5: 219-225

Neira VM, Bould MD, Nakajima A, et al. (2013) GIOSAT: a tool to assess CanMEDS competencies during simulated crises. Can J Anaesth, 60: 280-289

Nikouline A, Jimenez MC, Okrainec A. (2013) Feasibility of remote administration of the fundamentals of laparoscopic surgery (FLS) skills test using Google Glass. Surgical Endoscopy and Other Interventional Techniques, 1: S355

Norcini J. (2003) Setting standards on educational tests. Medical Education, 37: 464-469

Norcini J, Anderson B, Bollela V et al. (2011) Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Medical Teacher, 33: 206-214

Noureldin YA, Fahmy N, Anidjar M, Andonian S. (2016) Is there a place for virtual reality simulators in assessment of competency in percutaneous renal access? World Journal of Urology, 34: 733-739

O'Connor K, King R, Malone KM, Guerandel A. (2014) Clinical examiners, simulated patients, and student self-assessed empathy in medical students during a psychiatry objective structured clinical examination. Academic Psychiatry, 38: 451-457

Ogle J, Bushnell JA, Caputi P. (2013) Empathy is related to clinical competence in medical care. Medical Education, 47: 824-831

Okrainec A, Vassiliou M, Kapoor A, et al. (2013) Feasibility of remote administration of the Fundamentals of Laparoscopic Surgery (FLS) skills test. Surg Endosc, 27: 4033-4037

Oliven A, Nave R, Gilad D, Barch A. (2011) Implementation of a web-based interactive virtual patient case simulation as a training and assessment tool for medical students. Studies in Health Technology & Informatics, 169: 233-237

Oza SK, Boscardin CK, Wamsley M, et al. (2015) Assessing 3rd year medical students' interprofessional collaborative practice behaviors during a standardized patient encounter: A multi-institutional, cross-sectional study. Medical Teacher, 37: 915-925

Parikh PP, Brown R, White M, et al. (2015) Simulation-based end-of-life care training during surgical clerkship: assessment of skills and perceptions. J Surg Res, 196: 258-263

Park YS, Lineberry M, Hyderi A, et al. (2016) Differential weighting for subcomponent measures of integrated clinical encounter scores based on the USMLE Step 2 CS examination: effects on composite score reliability and pass-fail decisions. Academic Medicine, 91: S24-S30

Pedersen P, Palm H, Ringsted C, Konge L. (2014) Virtual-reality simulation to assess performance in hip fracture surgery. Acta Orthopaedica, 85: 403-407

Pell G, Fuller R, Homer M, Roberts T. (2013) Advancing the objective structured clinical examination: sequential testing in theory and practice. Medical Education, 47: 569-577

Peters JH. (2004) Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery, 135: 21-27

Ponton-Carss A, Hutchison C, Violato C. (2011) Assessment of communication, professionalism, and surgical skills in an objective structured performance-related examination (OSPRE): a psychometric study. American Journal of Surgery, 202: 433-440

Ponton-Carss A, Kortbeek JB, Ma IW. (2016) Assessment of technical and nontechnical skills in surgical residents. American Journal of Surgery, 212: 1011-1019

Posner G, Nakajima A. (2011) Assessing residents' communication skills: disclosure of an adverse event to a standardized patient. J Obstet Gynaecol Can, 33: 262-268

Pugh D, Hamstra SJ, Wood TJ, et al. (2015) A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. Advances in Health Sciences Education, 20: 85-100

Raison N, Ahmed K, Fossati N, et al. (2017) Competency based training in robotic surgery: benchmark scores for virtual reality robotic simulation. BJU International, 119: 804-811

Raper SE, Resnick AS, Morris JB. (2014) Simulated disclosure of a medical error by residents: development of a course in specific communication skills. Journal of Surgical Education, 71: e116-126

Reason J. (2000) Human error: models and management. BMJ, 320: 768–770

Reed S, Kassis K, Nagel R, et al. (2015) Breaking bad news is a teachable skill in pediatric residents: A feasibility study of an educational intervention. Patient Education & Counseling, 98: 748-752

Reinert A, Berlin A, Swan-Sein A, et al. (2014) Validity and reliability of a novel written examination to assess knowledge and clinical decision making skills of medical students on the surgery clerkship. American Journal of Surgery, 207: 236-242

Reising DL, Carr DE, Tieman S, et al. (2015) Psychometric testing of a simulation rubric for measuring interprofessional communication. Nurs Educ Perspect, 36: 311-316

Reiter HI, Eva KW, Rosenfeld J, Norman GR. (2007) Multiple mini-interviews predict clerkship and licensing examination performance. Medical Education, 41: 378-384

Roberts WL, Solomon M, Langenau E. (2011) An investigation of construct validity of humanistic clinical skills on a medical licensure examination. Patient Education & Counseling, 82: 214-221

Rodriguez E, Siegelman J, Leone K, Kessler C. (2012) Assessing professionalism: summary of the working group on assessment of observable learner performance. Academic Emergency Medicine, 19: 1372-1378

Rutherford JS, Flin R, Irwin A, McFadyen AK. (2015) Evaluation of the prototype Anaesthetic Non-technical Skills for Anaesthetic Practitioners (ANTS-AP) system: a behavioural rating system to assess the non-technical skills used by staff assisting the anaesthetist. Anaesthesia, 70: 907-914

Saleh GM, Gauba V, Sim D, et al. (2008) Motion analysis as a tool for the evaluation of oculoplastic surgical skill: evaluation of oculoplastic surgical skill. Arch Ophthalmol, 126: 213-216

Saylor J, Vernoony S, Selekman J, Cowperthwait A. (2016) Interprofessional Education Using a Palliative Care Simulation. Nurse Education, 41: 125-129

Schenarts PJ, Love KM, Agle SC, Haisch CE. (2008) Comparison of surgical residency applicants from US medical schools with US-born and foreign-born international medical school graduates. Journal of Surgical Education, 65: 406-412

Schildmann J, Kupfer S, Burchardi N, Vollmann J. (2012) Teaching and evaluating breaking bad news: a pre-post evaluation study of a teaching intervention for medical students and a comparative analysis of different measurement instruments and raters. Patient Education & Counseling, 86: 210-219

Schubert S, Ortwein H, Dumitsch A, et al. (2008) A situational judgement test of professional behaviour: development and validation. Med Teach, 30: 528-533

Sennekamp M, Gilbert K, Gerlach FM, Guethlin C. (2012) Development and validation of the "FrOCK": Frankfurt observer communication checklist. Zeitschrift fur Evidenz Fortbildung und Qualitat im Gesundheitswesen, 106: 595-601

Setyonugroho W, Kropmans T, Kennedy KM, et al. (2016) Calibration of communication skills items in OSCE checklists according to the MAAS-Global. Patient Education & Counseling, 99: 139-146

Shen L, Li F, Wattleworth R, Filipetto F. (2010) The promise and challenge of including multimedia items in medical licensure examinations: some insights from an empirical trial. Academic Medicine, 85: S56-59

Shirazi M, Labaf A, Monjazebi F, et al. (2014) Assessing medical students' communication skills by the use of standardized patients: emphasizing standardized patients' quality assurance. Academic Psychiatry, 38: 354-360

Sim JH, Abdul Aziz YF, Mansor A, et al. (2015) Students' performance in the different clinical skills assessed in OSCE: what does it reveal? Medical Education Online, 20: 26185

Sternbach JM, Wang K, El Khoury R, et al. (2017) Measuring Error Identification and Recovery Skills in Surgical Residents. Annals of Thoracic Surgery, 103: 663-669

Stroud L, McIlroy J, Levinson W. (2009) Skills of internal medicine residents in disclosing medical errors: a study using standardized patients. Academic Medicine, 84: 1803-1808

Sullivan WM. (2000) Medicine under threat: professionalism and professional identity. CMAJ, 162: 673-675

Szmuilowicz E, el-Jawahri A, Chiappetta L, et al. (2010) Improving residents' end-of-life communication skills with a short retreat: a randomized controlled trial. Journal of Palliative Medicine, 13: 439-452

Thijssen AS, Schijven MP. (2010) Contemporary virtual reality laparoscopy simulators: quicksand or solid grounds for assessing surgical trainees? American Journal of Surgery, 199: 529-541

Tiffin PA, Finn GM, McLachlan JC. (2011) Evaluating professionalism in medical undergraduates using selected response questions: findings from an item response modelling study. BMC Medical Education, 11: 43

Till H, Ker J, Myford C, et al. (2015) Constructing and evaluating a validity argument for the final-year ward simulation exercise. Advances in Health Sciences Education, 20: 1263-1289

Tsai TC, Chen DF, Lei SM. (2012) The ethics script concordance test in assessing ethical reasoning. Medical Education, 46: 527

Tsai TC, Harasym PH, Coderre S, et al. (2009) Assessing ethical problem solving by reasoning rather than decision making. Medical Education, 43: 1188-1197

Turner TR, Scerbo MW, Gliva-McConvey GA, Wallace AM. (2016) Standardized patient encounters: periodic versus postencounter evaluation of nontechnical clinical performance. Simulation in Healthcare, 11: 164-172

Tweed M, Wilkinson T. (2009) A randomised controlled trial comparing instructions regarding unsafe response options in a MCQ examination. Medical Teacher, 31: 51-54

Tweed M, Thompson-Fawcett M, Schwartz P, Wilkinson TJ. (2013) Determining measures of insight and foresight from responses to multiple choice questions. Medical Teacher, 35: 127-133

van Mook WN, Gorter SL, O'Sullivan H, et al. (2009) Approaches to professional behaviour assessment: tools in the professionalism toolbox. Eur J Intern Med, 20: e153-157

van Mook WN, Van Luijk SJ, Fey-Schoenmakers MJ, et al. (2010) Combined formative and summative professional behaviour assessment approach in the bachelor phase of medical school: a Dutch perspective. Medical Teacher, 32: e517-531

van Mook WNKA, van Luijk SJ, O'Sullivan H, et al. (2009) General considerations regarding assessment of professional behaviour. European Journal of Internal Medicine, 20: e90-e95

Van Zantan M, Boulet JR, McKinley DW. (2004) The influence of ethnicity of patient satisfaction in a standardized patient assessment. Academic Medicine, 79: S15-S17

Varkey P, Natt N. (2007) The Objective Structured Clinical Examination as an educational tool in patient safety. Jt Comm J Qual Patient Saf, 33: 48-53

Vassiliou MC, Dunkin BJ, Fried GM, et al. (2014) Fundamentals of endoscopic surgery: creation and validation of the hands-on test. Surg Endosc, 28: 704-711

Veloski JJ, Fields SK, Boex JR, Blank LL. (2005) Measuring professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. Academic Medicine, 80: 366-370

Waldmann UM, Gulich MS, Zeitler HP. (2008) Virtual patients for assessing medical students - Important aspects when considering the introduction of a new assessment format. Medical Teacher, 30: 17-24

Walzak A, Bacchus M, Schaefer JP, et al. (2015) Diagnosing technical competence in six bedside procedures: comparing checklists and a global rating scale in the assessment of resident performance. Academic Medicine, 90: 1100-1108

Wass V, Van der Vleuten C, Shatzer J, Jones R. (2001) Assessment of clinical competence. The Lancet, 357: 945-949

West CP, Shanafelt TD. (2007) The influence of personal and environmental factors on professionalism in medical education. BMC Medical Education, 7: 29

Whelan P, Church L, Kadry K. (2009) Using standardized patients' marks in scoring postgraduate psychiatry OSCEs. Academic Psychiatry, 33: 319-322

Wilkinson TJ, Wade WB, Knock LD. (2009) A blueprint to assess professionalism – results of a systematic review. Academic Medicine, 84: 551-558

Willaert WI, Cheshire NJ, Aggarwal R, et al. (2012) Improving results for carotid artery stenting by validation of the anatomic scoring system for carotid artery stenting with patient-specific simulated rehearsal. J Vasc Surg, 56: 1763-1770

Williams K, Wryobeck J, Edinger W, et al. (2011) Assessment of competencies by use of virtual patient technology. Academic Psychiatry, 35: 328-330

Wiskin CM, Elley K, Jones E, Duffy J. (2013) Clinician and simulated patient scoring - the psychometrics of a national programme recruiting dentists to DF1 training posts. Br Dent J, 215: 125-130

Wong BM, Coffey M, Nousiainen MT, et al. (2017) Learning through experience: influence of formal and informal training on medical error disclosure skills in residents. Journal of Graduate Medical Education, 9: 66-72

Wong ML, Fones CS, Aw M, et al. (2007) Should non-expert clinician examiners be used in objective structured assessment of communication skills among final year medical undergraduates? Medical Teacher, 29: 927-932

Wouda JC, van de Wiel HB. (2012) The communication competency of medical students, residents and consultants. Patient Education & Counseling, 86: 57-62

Wouda JC, van de Wiel HB. (2013) Inconsistency of residents' communication performance in challenging consultations. Patient Education & Counseling, 93: 579-585

Wright B, McKendree J, Morgan L, et al. (2014) Examiner and simulated patient ratings of empathy in medical student final year clinical examination: are they useful? BMC Medical Education, 14: 199

Wright MC, Segall N, Hobbs G, et al. (2013) Standardized assessment for evaluation of team skills: validity and feasibility. Simulation in Healthcare, 8: 292-303

Yang RL, Hashimoto DA, Predina JD, et al. (2013) The virtual-patient pilot: testing a new tool for undergraduate surgical education and assessment. J Surg Educ, 70: 394-401

Yang YY, Lee FY, Hsu HC, et al. (2013) Validation of the behavior and concept based assessment of professionalism competence in postgraduate first-year residents. J Chin Med Assoc, 76: 186-194

Zabar S, Adams J, Kurland S, et al. (2016) Charting a Key Competency Domain: Understanding Resident Physician Interprofessional Collaboration (IPC) Skills. Journal of General Internal Medicine, 31: 846-853

Zanetti M, Keller L, Mazor K, et al. (2010) Using standardized patients to assess professionalism: a generalizability study. Teaching & Learning in Medicine, 22: 274-279

Zhang X, Roberts WL. (2013) Investigation of standardized patient ratings of humanistic competence on a medical licensure examination using Many-Facet Rasch Measurement and generalizability theory. Advances in Health Sciences Education, 18: 929-944