Best Practice in the Assessment of Competence: Appendices

School of Medical Education
Newcastle University

September 2018

Bryan Burford
Charlotte Rothwell
Gillian Vance
School of Medical Education

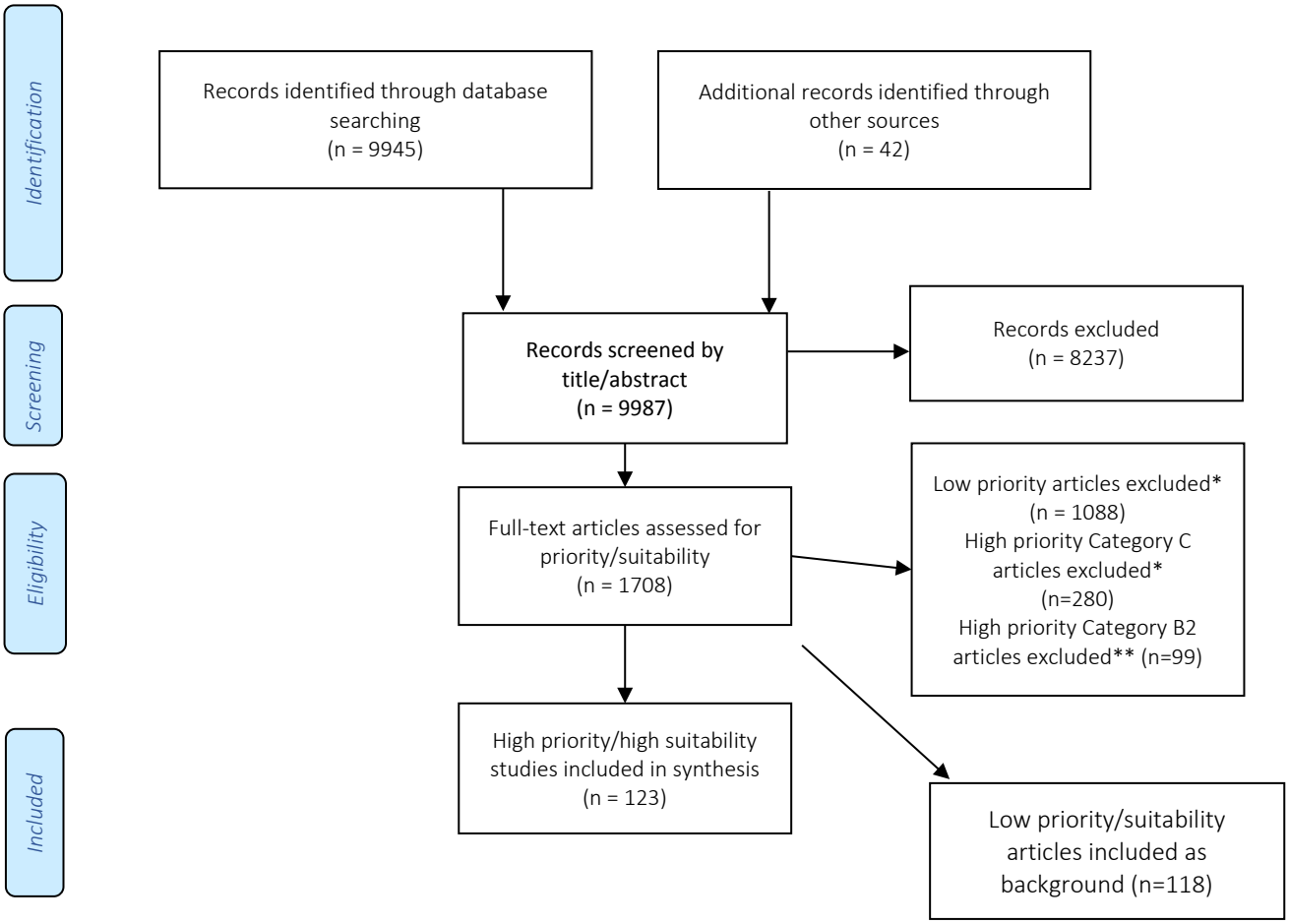
Fiona Beyer
Louise Tanner
Evidence Synthesis Group, Institute for Health and Society

List of Appendices

Appendix A. Details of literature review

This evidence synthesis project followed principles of rigorous and transparent systematic review methods (Centre for Reviews & Dissemination 2009) in terms of a comprehensive literature search, robust screening methods, critical appraisal of the included studies, and transparent categorisation and synthesis of the evidence. Given that our topic of interest covers a wide scope of literature and is reported using many different methodologies, our synthesis adopted an 'evidence mapping' approach. A mapping review examines a broad question, maps features and gaps and identifies opportunities for further review/research (Althuis and Weed 2013). A brief summary of each of the steps in the process is given below. Figure A1 provides a flow chart illustrating the process of the search.

Figure A1. Flow chart of literature search process (after Moher et al 2009)



* A sample of 'low priority' papers were included in the final report as background and to give some representation of these areas. Similarly, some category C and B2 articles were included for background.

** These papers are listed in Appendix D.

A.1 Initial search

The topics of assessment we wished to capture in our search are ill-defined, and so our initial search emphasised sensitivity in order to capture all potentially relevant papers. Subsequent stages of screening and review increased specificity so that only papers that would address our research questions were included.

Searches were conducted in medical and non-medical databases:

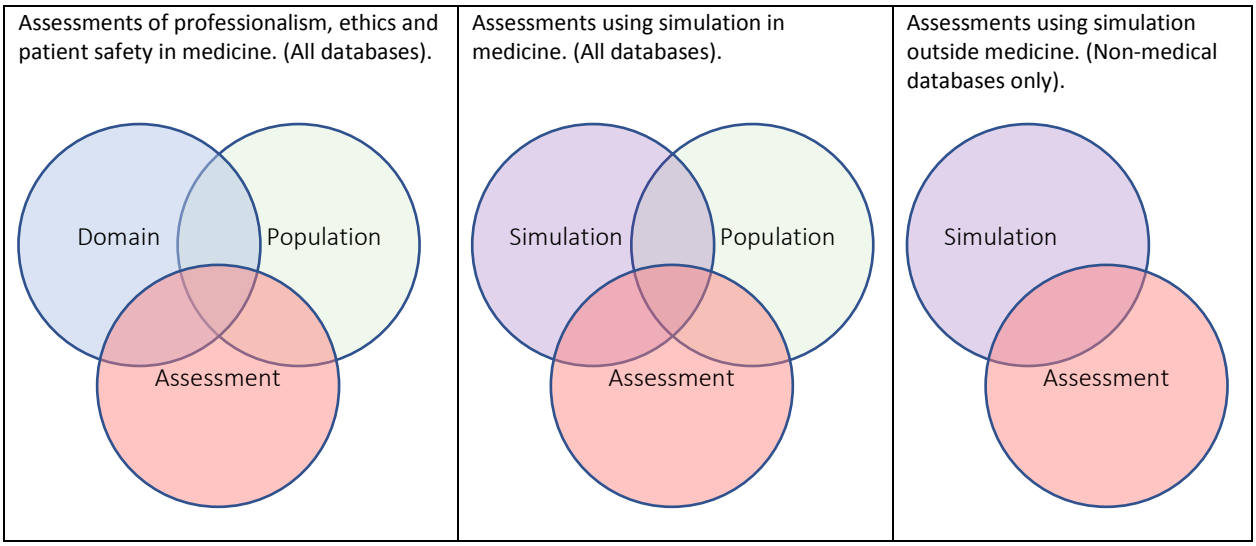
- Medical literature: MEDLINE and EMBASE (both via OVID)
- Psychological literature: PsycINFO (via OVID)
- Educational literature: ERIC (via EBSCO)

- Multidisciplinary literature: Science Citation Index, Social Science Citation Index, Conference Proceedings Index – Science, Conference Proceedings Index – Social Science (all via Web of Science)

Different searches were developed for each research question, as illustrated in figure A1. Each circle represents a key concept operationalised using database index terms and keyword string matching. Searches were restricted to the period 2007-2017, inclusive, to minimise papers that may have been superseded by changes in curricula and/or regulation.

This was supplemented by iterative searching based on references lists and Google searches for unindexed journals. Grey literature (unpublished papers, reports) was identified by searching the websites of organisations that are known to provide summative assessments, such as the Medical Royal Colleges and relevant international bodies. Any relevant examples have been integrated in the report where appropriate.

Figure A2. Venn diagrams illustrating search strategies



Papers of interest were located in the intersections between all sets in each Venn diagram, ie, those that were present in each of the sets and so were related to each of the concepts.

The search strategy was registered with the Prospero register of systematic reviews (https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42018086773).

A.2 Initial screening of titles and abstracts

Screening on the basis of titles and abstracts provides an efficient way to identify relevant papers. Inclusion and exclusion criteria were agreed and refined in discussion between the authors following pilot screening of a sample of papers. An adequate level of inter-rater reliability was achieved with this pilot sample, indicating that criteria were clearly and robustly defined.

Each paper was then reviewed individually against the final criteria summarised in table A1, following the decision path illustrated in figure A3. Screening was conducted using the Rayyan online systematic review tool (rayyan.qcri.org).

The criteria for 'simulation' and 'domain' were considered such that an assessment involving simulation was included regardless of any domain-level inclusions, and vice versa. Key exclusions included papers which used assessments purely as outcome measures in teaching evaluations or to compare groups. Conversely, while knowledge tests were generally excluded on the basis that the research questions are based on behaviours, tests addressing application of knowledge rather than recall were included.

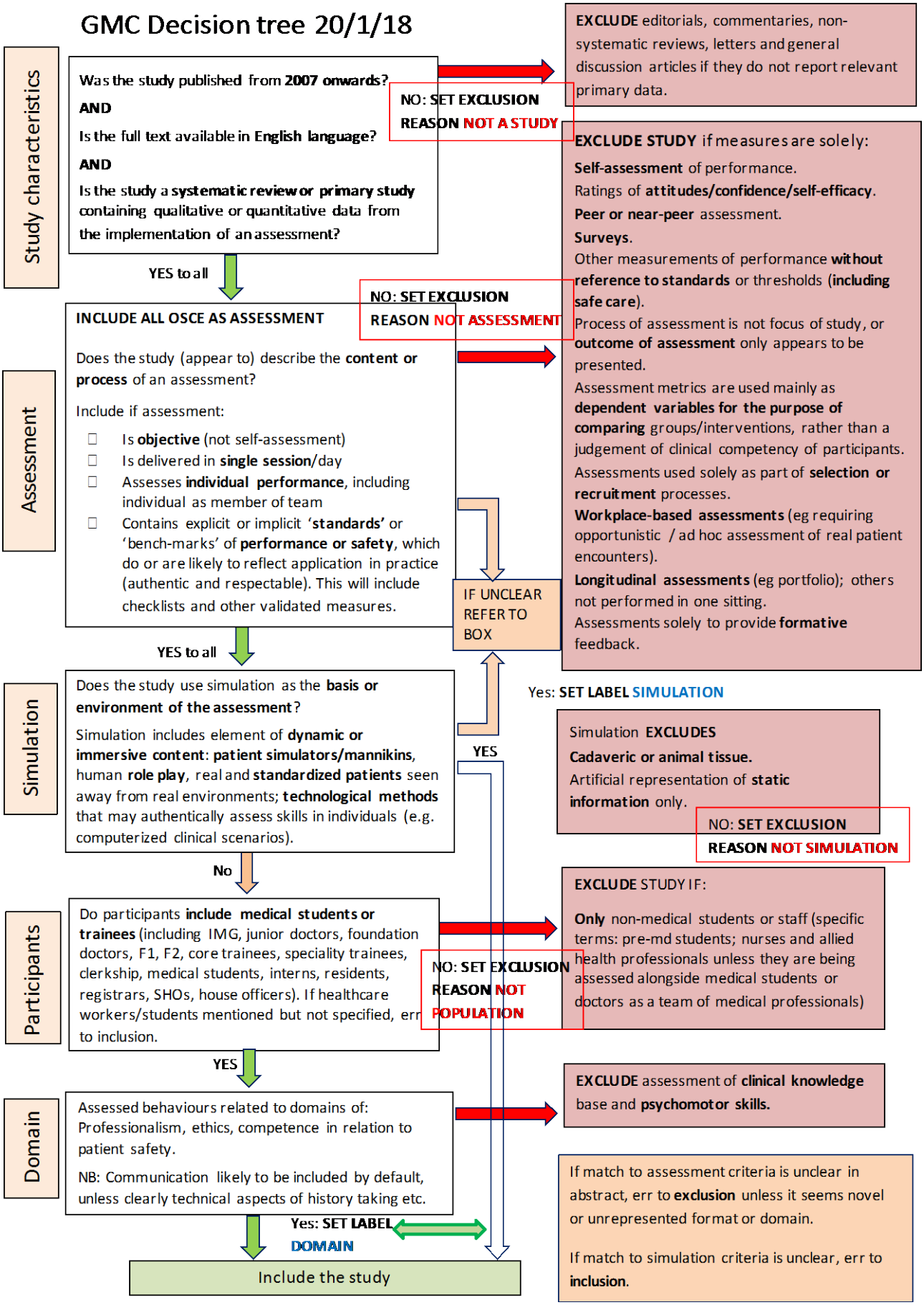
Simple measures of attitudes were largely excluded. While some self-reported attitudes may be associated with professionalism (for example, attitudes to marginalised patient groups), they do not represent practice in a way that can be assessed robustly. Apart from philosophical concerns, such assessments may be open to challenge as they represent what *may* be done, rather than what *is* done.

Overall, the application of criteria was pragmatic, and where domains or methods were potentially under-represented, we erred towards inclusion for review of the full text. Additionally, not all abstracts provided sufficient detail for definitive screening at this stage, and so these were included as borderline cases for further screening. Conference abstracts which met criteria, provided detailed information, and were not duplicated in a subsequent print paper, were retained. Those which provided little information were not.

Table A1. Inclusion and exclusion criteria for screening

Study characteristics	
Inclusion	Paper describes an empirical study reporting primary data, or a systematic review.
Exclusion	Paper is a commentary, editorial or other non-empirical piece, or reference is a non-peer reviewed thesis
Assessment	
Inclusion	The study describes the content or process of assessment, with assessment being defined as: • objective • delivered in single session/day • assessing individual performance, including individual as member of team • containing explicit or implicit standards or bench-marks of performance or safety, which do, or are likely to reflect application in practice (authentic and respectable). This will include checklists and other validated measures.
Exclusion	Self-assessment of performance. Ratings of attitudes/confidence/self-efficacy. Peer or near-peer assessment. Surveys. Measurements of performance without reference to standards or thresholds. Studies where assessment outcomes alone are the object of study, rather than process or content. For instance, the assessment is used as a dependent variable for comparison of groups/interventions, rather than a judgement of learner performance. Assessments used solely as part of selection or recruitment processes. Workplace-based assessments (eg requiring opportunistic / ad hoc assessment of real patient encounters). Longitudinal assessments (eg learning portfolios), or others not performed in one sitting. Assessments providing only formative feedback, where there is no clear reference to standards or benchmarks.
Population	
Inclusion	Participants include medical students or doctors (including IMGs, junior doctors, foundation doctors, F1, F2, core trainees, specialty trainees, clerkship, medical students, interns, residents, registrars, SHOs, house officers).
Exclusion	Participants consist only of non-medical students or staff (including pre-med students, nurses and allied health professionals) unless they are being assessed alongside medical students or doctors as a team of medical professionals
Domain – relevant only when simulation is excluded	
Inclusion	Assessed behaviours relate to domains of professionalism, ethics, competence in relation to patient safety.
Exclusion	Assessment of clinical knowledge base and psychomotor skills.
Simulation – relevant only where domain is excluded	
Inclusion	Simulation includes elements of dynamic or immersive content: patient simulators/manikins, human role play, real and standardized patients seen away from real environments; technological methods that may authentically assess skills in individuals (eg computerized clinical scenarios).
Exclusion	Activities using cadaveric or animal tissue only. Artificial representation of static information only, eg ECG or X-ray, which in isolation does not contribute to an immersive assessment, but rather forms part of a test of decontextualised knowledge.

Figure A3. Decision tree for screening of titles and abstracts



A.3 Full paper screening: eligibility and data extraction

With a large number of papers passing initial screening, two further stages were included in order to better identify papers that addressed our specific research questions.

Firstly, papers which referred to content or types of assessment of importance to the research questions were identified by searching for key words in titles and abstracts. These 'high *priority'* papers referred explicitly to elements of professionalism, ethics or patient safety, or described novel technological approaches. Low priority papers were those judged to be more tangentially related to the domains, or describing more commonplace technologies.

Secondly, each of the high priority papers was categorised based on the extent to which it clearly described an assessment *suitable* for summative use. Four categories were used:

- Category A: Paper describes current summative assessment
- Category B1 (borderline): Paper does not describe current summative assessment, but could be easily adapted. Includes pilot studies, novel approaches to assessment process, and descriptions of formative assessments that indicate standards analogous to pass/fail.
- Category B2 (borderline): Paper does not describe current summative assessment, with no evidence of novelty or that the approach is transferable or adaptable.
- Category C: Papers excluded entirely on the basis of information which was not apparent in the abstract.

This categorisation does not represent a judgement of quality of the paper, nor of the reported assessment, but rather its 'suitability' (ie potential for summative use) in terms of how the assessment has been used.

These stages allowed those papers of greatest relevance (*priority* subject areas and *suitable* for summative assessment use) to be identified. Detailed data extraction was then focused on papers that were most aligned with the research questions.

For completeness, a sample of low priority papers was also selected for review and categorisation of suitability in the domains of 'basic communication' and 'simulation'. This was to ensure representation of data from these broad areas in the mapping review. These papers are presented in appendix C and section 3.7 of the report.

Category A and B1 papers only were included in the full review. Salient data – location and context of the study, its aims and main findings, and any details of authenticity and validation evidence – from these papers were summarised on a proforma.

The distinction between B1 and B2 was not always clear-cut even in full papers, and a pragmatic judgement of relevance and novelty was applied.

A.3.1 Critical appraisal

As this was a mapping review aiming to provide an overview of a broad literature, critical appraisal considered classes of assessment – defined by content or type – overall, rather than focusing on individual papers. We considered a global judgement of the amount of evidence presented, analogous to the GRADE system (Grading of Recommendations Assessment, Development and Evaluation initiative http://www.gradeworkinggroup.org/), which is used in systematic reviews of healthcare interventions to assess the overall confidence in the body of evidence under review. For each class of assessment we assigned a level of confidence as described in table A2.

Table A2. 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 also considered the quality of evidence with reference to the five sources of assessment validity described by Downing (2003). This framework was also used by Archer et al (2015) in their review of licensing examinations. These are summarised in table A3, with operational definitions reflecting our interpretation of these sources.

Table A3. 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.

A.4 Synthesis

Proformas for high priority Category A and B1 papers were sorted and summarised on the basis of the type and content of the assessments, and narratives developed as presented in the results section of the main report. The Downing (2003) approach to assessment validity was used as a lens for determining good practice. An Excel spreadsheet was used to track inclusions and cross-check that papers had been excluded consistently.

Appendices C and D summarise those papers relating to existing licensing examinations, and 'basic' communication skills, respectively (see A3 above). Appendix E lists the 'high priority' B2 papers, which were not eligible for the final synthesis.

A.5 References for Appendix A

Althuis MD and 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

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, :

Centre for Reviews & Dissemination. (2009) Systematic reviews: CRD's guidance for undertaking reviews in health care. York: University of York, University of York. Centre for Reviews and Dissemination:

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

Moher D, Liberati A, Tetzlaff J,et al. (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement.. PLoS Med, 6: e1000097

Appendix B. List of Project Advisory Group members

Name	Place / Organisation	Practice exemplar and/or practice discussion
Alan Jaap	Edinburgh Medical School; Medical Schools Council (MSC)	Yes
Dyfrig Hughes	Sheffield Medical School; MSC	
Marina Anderson	Liverpool Medical School; MSC	
Neil Kennedy	Queen's University Medical School, Belfast; MSC	Yes
Rachel Williams	Cambridge University Medical School; MSC	Yes
Elizabeth Fistein	Cambridge University Medical School; UK Council of Teachers of Professionalism	Yes
Lindsey Pope	Glasgow Medical School; UK Council of Teachers of Professionalism	
Rachel Westacott	Leicester Medical School; Medical Schools Council Assessment Alliance (MSC-AA)	Yes
Thomas Gale	Plymouth Medical School; MSC-AA	Yes
Andrew Elder	Academy of Medical Royal Colleges	
Isobel Braidman	Manchester; UK Council of Teachers of Professionalism	Yes
Hilary Neve	Plymouth; UK Council of Teachers of Professionalism	
Wing May Kong	London; Institute of Medical Ethics	
Bob McKinley	Keele; ASME	
Jennifer Hallam	Leeds; ASME	
Richard Fuller	Leeds; AMEE	Yes
Jiv Gosai	Research lead, ASPiH	Yes
Nick Spittle	East Midlands, UK Foundation Programme Office (UKFPO)	
Angela Carragher	N Ireland, UKFPO	Yes
Dennis Okolo	PLAB part 2	
Julian Hancock	Oxford; PLAB part 1	
Jon Scott	Northern region, UKFPO	

Appendix C. Summary of assessments of basic communication skills

In the main review we identified 'complex' communication skills as reflecting professionalism, because the manner of communication in those situations is central to good clinical practice, rather than the procedural elements of communication. We contrasted these with 'basic' communication skills which are more routine and protocol-driven, such as history taking. Assessments of basic communication were therefore classed as 'low priority'. In this appendix we summarise some of the literature in this area to illustrate similarities, and differences, in the way this aspect of communication is assessed.

Performance in history taking and patient interviews has been assessed using a number of measures. The context of these tended to be OSCEs, with ratings provided by standardised patients (SPs), faculty or expert examiners. A common approach was based on the Kalamazoo Essential Elements Communication Checklist (and variants), based on a consensus statement of essential communication skills (Makoul 2001). This consists of 24 competencies in seven domains (Builds the Relationship, Opens the Discussion, Gathers Information, Understands the Patient's and Family's Perspective, Shares Information, Reaches Agreement, and Provides Closure) and was based on a consensus statement which identified key elements of communication (Makoul 2001). Versions have been used which assess performance on a 3-point checklist rating (not done, needs improvement, done well), the 'adapted' version which uses a 5-point scale, and the Gap-Kalamazoo Communication Skills Assessment Form which uses multiple response formats and data from multiple source, including self-assessment. Its origins predate this review, but it was referred to several times (Amaral et al 2016, Baker-Genaw et al 2016, Brown et al 207, Porcerelli et al 2015, Peterson et al 2014, Joyce et al 2010). Validity evidence for the Kalamazoo class of tools was high.

Similar approaches have been described by others. The Common Ground Assessment Scale (Lang et al 2004) is another approach derived from the Kalamazoo consensus statement, which consists of a binary checklist, frequency checklist and rating scale responses. It was referred to in three papers in our search (Hess et al 2016, Van Nuland et al 2012, Gorniewicz et al 2014).

Boissonault et al (2016) described a similar tool for the assessment of patient interviewing skills in physiotherapists. The 'ECHOWS' (Establishing rapport, Chief complaint, Health history, Obtain psychosocial perspective, Wrap-up, Summary) tool consisted of a binary checklist of 22 items across the 'ECHOW' phases, plus a 3-point response to 10 items for the summary. Inter-rater reliability was high, but discrimination between novice and expert interaction was low.

Generic communication competencies have been addressed by established licensing examinations. For example, the USMLE Step 2 Clinical Skills identified four domains: 'interviewing and collecting information', 'skills in counselling and delivering information', 'rapport' and 'personal manner' (eg van Zanten et al 2007a, 2007b) while the COMLEX distinguished six, albeit very similar: 'ability to elicit information', 'listening skills', 'giving information', 'respectfulness', 'empathy' and 'professionalism' (professionalism here was defined as the candidates' ability to appear 'both appropriately confident and therapeutic', demonstrate altruism and ensure patient confidentiality', Weidner et al, 2010). Performance in these domains was rated by standardised patients on global Likert-type scales. The USMLE used a four-point scale from poor to excellent, while the COMLEX used a nine-point scale in three classifications 'unacceptable', 'acceptable' and 'superior'.

Scores in such generic or global communication assessments can vary systematically between clinical cases/OSCE stations (Weidner et al 2010, Baig et al 2009, Boulet et al 2007), and between OSCE and workplace measures (Baig et al 2009). This has been inferred to mean communication skills are context-dependent, but could also imply differences in difficulty. Performance can also vary with candidate age, and SP and candidate sex (eg Weidner et al 2010). Such variability notwithstanding, statistically generalisable findings have been found with up to 11 cases (Weidner et al 2010).

Basic communication also includes written communication, whether paper-based or online. The USMLE and COMLEX both contain specific assessments of students' skills in writing patient notes (eg Park et al 2013, 2016; Langenau et al 2010). These have been assessed with checklists or rubrics. A similar approach to assessing simulated email communication was also described by Mittal et al (2010). Furthermore, communication today must also consider online systems, such as decision support systems and electronic patient records (Biagoli et al 2017). Communication with patients online (which is encompassed in the GMC's *Outcomes for Graduates* within 'communicate by… electronic methods'; GMC 2017) has also been simulated (Christner et al 2010).

A subset of papers looked at communication alongside other related constructs in assessments of 'non-technical skills' (NTS). This phrase was adopted from aviation, and implies a conceptual distinction between technical skills based on clinical knowledge and practice, and those based in communication. While several assessments here address these skills together, a subset of papers explicitly position themselves as assessments of NTS. Three main tools were used: NOTECHS, which was derived directly from similar use in aviation, the Non-Technical Skills for Surgeons Scale (NOTSS) and the Anaesthetists' Non-Technical Skills Scale (ANTS). NOTECHS is ostensibly generic, although can involve case-specific items, while the other two are designed for particular specialties. However, there is substantial overlap in their content and all use scaled checklists as assessment data.

NOTECHS measures four domains - communication and interaction, cooperation and team skills, leadership and management skills and decision-making. NOTSS and ANTS use similar domains of situational awareness, decision-making and communication/teamwork, while NOTSS also assesses leadership, while ANTS task management, reflecting the different roles of surgeons and anaesthetists in theatre. Rutherford et al (2015) described an adaptation of ANTS to non-medical anaesthetic practitioners, which omitted the decision-making domain. All use scaled checklists, and have implicit thresholds of competence, although we did not find examples of these being used summatively. While subscale scores can be calculated, they generally produce a single score, with internal consistency generally reported to be acceptable.

Some construct validity evidence is offered in this literature. Black et al (2010) reported consultants achieving higher NOTECHS scores than junior trainees. Doumouras et al (2017) found that NOTSS and ANTS scores varied in surgical crisis scenarios, which suggests they assess volatile constructs, which may not be desirable in assessment. However, construct validity was suggested by their observation that while NOTSS scores were lower in a haemorrhage scenario than one based on airway management, ANTS scores remained high, and this may reflect the lesser role of the surgeon in the airway scenario.

Lambden et al (2013) and Brunckhorst et al (2015) examined correlations between NTS measures and technical skills. While Lambden found a moderate correlation between NOTECHS and technical skills in a paediatric emergency scenario, Brunckhorst et al (2015) found a strong correlation between NOTTS and technical skills in a uteroscopy procedure. These were different assessments, different procedures and different populations (specialty trainees in Lambden et al, medical students in Brunckhorst et al), but the difference indicates that measures, and underlying constructs, are not consistently related.

While these approaches are based on scaled checklists, global rating scales (GRS) have been used. Nunnink et al (2014) and Jirativanont et al (2017) both compared ANTS with the Ottawa GRS measure of non-technical skills. Validity evidence was similar for both scales, but Jirativanont et al (2017) reported the GRS as being more 'user friendly'. Other GRS-based NTS assessments were reported by Pugh et al (2015) and Rovamo et al (2011). Both reported acceptable validity evidence.

In summary, this area is well-established, and there are several approaches which have good evidence of validity. There is evidence however that further work is necessary even here. Sensitivity of assessments to context is not inherently problematic, but must be understood at the specific level. The type of assessment is not unambiguous. Consequently, even for relatively formulaic responses, detailed descriptors may be preferred to a basic checklist.

References for Appendix C

Amaral AB, Rider EA, Lajolo PP, et al. (2016) Development of a Brazilian Portuguese adapted version of the Gap-Kalamazoo communication skills assessment form. Int J Med Educ, 7: 400-405

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

Baker-Genaw K, Kokas MS, Ahsan SF, et al. (2016) Mapping Direct Observations From Objective Structured Clinical Examinations to the Milestones Across Specialties. Journal of Graduate Medical Education, 8: 429-434

Biagioli FE, Elliot DL, Palmer RT, et al. (2017) The Electronic Health Record Objective Structured Clinical Examination: Assessing Student Competency in Patient Interactions While Using the Electronic Health Record. Academic Medicine, 92: 87-91

Black SA, Nestel DF, Kneebone RL and 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

Boissonault JS, et al. (2016) Reliability of the ECHOWS Tool for Assessment of Patient Interviewing Skills. Physical Therapy, 96: -443

Boulet JR, McKinley DW, Rebbecchi T, Whelan GP (2007) Does composition medium affect the psychometric properties of scores on an exercise designed to assess written medical communication skills?. Advances in Health Sciences Education, 12: 157-167

Brown SD, Rider EA, Jamieson K, et al. (2017) Development of a Standardized Kalamazoo Communication Skills Assessment Tool for Radiologists: Validation, Multisource Reliability, and Lessons Learned. AJR. American Journal of Roentgenology, 209: 351-357

Brunckhorst O, Shahid S, Aydin A, et al. (2015) The relationship between technical and nontechnical skills within a simulation-based ureteroscopy training environment. Journal of Surgical Education, 72: 1039-1044

Christner JG, Stansfield RB, Schiller JH, et al. (2010) Use of simulated electronic mail (e-mail) to assess medical student knowledge, professionalism, and communication skills. Academic Medicine, 85: S1-4

Doumouras AG, Hamidi M, Lung K, et al. (2017) Non-technical skills of surgeons and anaesthetists in simulated operating theatre crises. British Journal of Surgery, 104: 1028-1036

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

Hess R, Hagemeier NE, Blackwelder R, et al. (2016) Teaching Communication Skills to Medical and Pharmacy Students Through a Blended Learning Course. Am J Pharm Educ, 80: 64

Jirativanont T, Raksamani K, Aroonpruksakul N, et al. (2017) Validity evidence of non-technical skills assessment instruments in simulated anaesthesia crisis management. Anaesthesia and Intensive Care, 45: 469-475

Joyce BL, Steenbergh T, and Scher E (2010) Use of the Kalamazoo Essential Elements Communication Checklist (Adapted) in an Institutional Interpersonal and Communication Skills Curriculum. Journal of Graduate Medical Education, 2: 165-169

Lambden S, DeMunter C, Dowson A, et al. (2013) The Imperial Paediatric Emergency Training Toolkit (IPETT) for use in paediatric emergency training: development and evaluation of feasibility and validity. Resuscitation, 84: 831-836

Lang F, McCord R, Harvill L, Anderson DS. (2004) Communication assessment using the common ground instrument: psychometric properties. Family Medicine, 36: 189-198

Langenau EE, Dyer C, Roberts WL, et al. (2010) Five-year summary of COMLEX-USA level 2-PE examinee performance and survey data. J Am Osteopath Assoc, 110: 114-125

Makoul G. (2001) Essential elements of communication in medical encounters: the Kalamazoo consensus statement. Academic Medicine. 76: 390-393.

Mittal MK, Dhuper S, Siva C, et al. (2010) Assessment of email communication skills of rheumatology fellows: a pilot study. J Am Med Inform Assoc, 17: 702-706

Nunnink L, Foot C, Venkatesh B, et al. (2014) High-stakes assessment of the non-technical skills of critical care trainees using simulation: feasibility, acceptability and reliability. Crit Care Resusc, 16: 43440

Park YS, Hyderi A, Bordage G, et al. (2016) Inter-rater reliability and generalizability of patient note scores using a scoring rubric based on the USMLE Step-2 CS format. Advances in Health Sciences Education, 21: 761-773

Park YS, Lineberry M, Hyderi A, et al. (2013) Validity evidence for a patient note scoring rubric based on the new patient note format of the United States Medical Licensing Examination. Academic Medicine, 88: 1552-1557

Peterson EB, Calhoun AW, Rider EA (2014) The reliability of a modified Kalamazoo Consensus Statement Checklist for assessing the communication skills of multidisciplinary clinicians in the simulated environment. Patient Education & Counseling, 96: 411-418

Porcerelli JH, Brennan S, Carty J, et al. (2015) Resident Ratings of Communication Skills Using the Kalamazoo Adapted Checklist. Journal of Graduate Medical Education, 7: 458-461

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

Rovamo L, Mattila MM, Andersson S, Rosenberg P (2011) Assessment of newborn resuscitation skills of physicians with a simulator manikin. Arch Dis Child Fetal Neonatal Ed, 96: F383-389

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

Van Nuland M, Van den Noortgate W, van der Vleuten C and Jo G (2012) Optimizing the utility of communication OSCEs: omit station-specific checklists and provide students with narrative feedback. Patient Education & Counseling, 88: 106-112

van Zanten M, Boulet JR and McKinley D (2007) Using standardized patients to assess the interpersonal skills of physicians: six years' experience with a high-stakes certification examination. Health Communication, 22: 195-205

van Zanten M, Boulet JR, McKinley DW, et al. (2007) Assessing the communication and interpersonal skills of graduates of international medical schools as part of the United States Medical Licensing Exam (USMLE) Step 2 Clinical Skills (CS) Exam. Academic Medicine, 82: S65-68

Weidner AC, Gimpel JR, Boulet JR and Solomon M (2010) Using standardized patients to assess the communication skills of graduating physicians for the comprehensive osteopathic medical licensing examination (COMLEX) level 2-performance evaluation (level 2-PE). Teaching & Learning in Medicine, 22: 42217

Appendix D. Studies of medical licensing examinations

D.1 Existing licensing examinations around the world

Medical licensing examinations are well established around the world. Archer and colleagues conducted a systematic review for the GMC in 2011 (Archer et al 2015, 2016) which described licensing examinations in countries 'similar to the UK', as defined on the basis of a United Nations index. They described 22 examples in three levels: those only for graduates of the country of intended practice, those only for doctors who graduated in another regulatory area (IMGs) and a universal exam for all wishing to practise. Gillis et al (2015) surveyed 37 international medical licensing bodies in 30 countries, of which 11 were identified as developing countries, in contrast to Archer et al's sample. However they do not indicate the representation of the 29 bodies in 22 countries which responded. Of these, most (n=17) tested language proficiency, but few (n=5) conducted other communication skills testing. All undertook knowledge testing, but most (n=25) assessed clinical and technical skills by review of credentials or written test rather than direct practical testing. Four bodies reported a period of mentored practice as part of assessment.

Most prominent in the literature is the United States Medical Licensing Examination (USMLE), which must be passed by all medical graduates and IMGs wishing to practise in the USA. It consists of two stages of knowledge-based multiple-choice exams (Step 1 and 3) and a clinical knowledge (CK) test and clinical skills (CS) practical exam (Step 2 CK and Step 2 CS). While registration in the United States is regulated by Medical Boards at State level, the USMLE is a national examination [http://www.usmle.org/]. The USA also has a separate regulatory process for 'osteopathic medicine', which qualifies individuals to practise medicine and surgery. These practitioners must pass the Comprehensive Osteopathic Medical Licensing Examination (COMLEX-USA, [https://www.nbome.org/exams-assessments/comlex-usa/]), which has a similar structure to the USMLE.

In Canada, licensing is also at State/Territory level, but at a national level the Medical Council of Canada provides evidence of eligibility through the Licentiate of the Medical Council of Canada (LMCC; which is concurrent with registration). A core requirement for LMCC is passing the Medical Council of Canada Qualifying Examinations (MCCQE). Undergraduate and postgraduate medical education in Canada is based around the CanMEDS competencies, initially developed by the Royal College of Physicians of Canada, but now widely adopted (Frank et al 2015).

Alongside licensing examinations, some countries have processes for the assessment of international medical graduates in addition to the requirement for a recognised medical qualification. In the UK, the PLAB provides such an assessment, while in the USA the Educational Commission for Foreign Medical Graduates (ECFMG) provides certification based on completing the USMLE parts 1 and 2. In Canada, IMGs must pass the Medical Council of Canada Evaluating Examination (MCCEE) in order to enter postgraduate training, and to take the MCCQE. Passing the MCCEE provides eligibility for an 'educational' licence, not a full licence for which the MCCQE is necessary.

The situation within Europe is simplified by the European Union (EU) Recognition of Professional Qualifications Directive (2005/36/EC) which ensures mutual recognition of professional qualifications between member states of the EU and European Economic Area (EEA). For individual practitioners this ensures freedom of movement, but regulations established by national regulators still apply, and language testing is permitted. For those entering the EEA from elsewhere, the situation is more complicated and individual regulators take a range of approaches (Herfs et al 2007). The situation regarding doctors moving between the UK and EU following the UK's planned exit from the EU in 2019 remains uncertain.

Licensing examinations necessarily include scientific and clinical knowledge, but also address the professional domains of practice that are the focus of this report. For example, the USMLE includes 'Communication and Interpersonal Skills', 'Professionalism, including Legal and Ethical Issues' and 'Systems-based Practice, including Patient Safety' (FSMB/NBME 2014). In Canada the CanMEDs framework includes competencies organised under roles, which include 'leader', 'professional' and 'communicator' (Frank et al 2015), and in Switzerland the Federal Licensing Examination includes communication (Guttormsen et al 2013).

Our review found some papers around licensing assessments, and some of these are referred to in sections focusing on particular domains. These studies are a disparate group and an overall evaluation is hard to provide, however it seems for high stakes licensing examinations there is some, but limited, evidence of predictive validity with regard to subsequent ratings. IMGs appear to perform less well in some areas, which may be a reflection of validity, or may indicate an implicit bias. Further and closer analysis may be necessary.

Some studies considered the predictive validity of licensing examinations for performance in practice. Cuddy et al (2016, 2017) found some association between USMLE Step 2 scores and first year residency performance and longer term likelihood of disciplinary action. Similarly, Langenau et al (2011) found that COMLEX scores relating to biomedical/biomechanical areas were moderately positively correlated with competency ratings of first year residency performance, but humanistic scores were not. There is therefore some evidence of predictive, outcome-based validity for these assessments. However, even for these it is not absolute, and these studies reported significant relationships only for some measures.

In another application, two papers provided data on the use of examinations for those wishing to return to medical practice. These are not necessarily high stakes examinations per se, but inform regulators on whether doctors can return to practice. DeMaria et al (2013) described a simulation-based assessment programme for anaesthesiologists using checklists, rating scales and specific tools as appropriate for the doctor, followed by remediation and re-retesting in some circumstances. Grace et al (2010) described a different approach using 90 minute interviews and 2 or 3 simulated patient encounters. Neither of these studies described a discrete assessment, and the numbers of cases were small, but they illustrate a particular approach to assessment.

Many studies of licensing exams have considered their role in the assessment and licensing of IMGs, including differential attainment between international and home candidates. For example, Schenarts et al (2008) found demographics of IMGs taking the USMLE were slightly different to US graduates, being slightly older and more often male, and that while Step 1 scores were not significantly different, Step 2 scores were significantly lower. For both steps, IMGs overall required more attempts. Guttormsen et al (2013) in their description of the introduction of a new licensing exam in Switzerland found that only half of IMGs passed the exam, compared to nearly 100% of Swiss graduates. This failure rate was not linked to language, but to the application of clinical knowledge. Holtzman et al (2014) found differences in performance on different elements of the USMLE clinical knowledge test based on candidates' nationality.

A small effect of English as a first language (EFL) was reported by van Zanten, Boulet, McKInley et al (2007) with EFL speakers scoring more highly on the communication and interpersonal skills element of the USMLE CS (communication skills). However, there was no indication that this effect was associated with a practical difference in performance, and the means for both groups were acceptable. Nayer & Rothman (2013) found a relationship between countries of origin that are English-speaking, in western Europe or South America, and higher scores on a 12-station OSCE, including one dedicated ‘ethics and communication’ station. There was little correlation between this exam and the written MCCQE.

Also in Canada, MacLellan et al (2010) found that IMGs scored lower than Canadian graduates on the regional qualifying examination OSCE in Quebec, although not directly linked to language. Vallevand and Violato (2012) reported predictive validity for a small sample (n-39) of an OSCE for IMGs in western Canada, with accurate prediction of pass/fail outcome on a three month placement to determine if the candidate was ‘practice ready’. They also reported construct validity.

In a substantial analysis of over 60,000 cardiac cases where the attending physician was an IMG, Norcini et al (2014) found that performance on the USMLE Step 2 CK examination had an inverse relationship with mortality, when corrected for covariates. Predictive validity of the examination was thereby inferred.

A smaller-scale and weaker examination of the NZREX clinical examination reported by Lillis and Roblin (2014) found that nearly 90% of IMG residents who passed the exam had no unsatisfactory reports from their four supervisors during their first year of practice. An apparent association between the number of attempts at the exam and unsatisfactory reports was referred to, although with no statistical analysis. Most adverse reports came from the first two ‘runs’ or placements (42% in the first, 35% in the second), suggestive of further workplace learning taking place. Lillis and Roblin concluded that the exam had acceptable criterion validity.

Takahashi et al (2012) described a 12-station OSCE piloted with 846 IMGs in five Canadian cities simultaneously. Construct validity was explored by comparing scores from samples of third year medical students and first year residents (Canadian

graduates only). The residents scored higher than the students, with the mean for IMGs falling between these groups. Takahashi et al concluded that the exam provided discrimination, and allowed the appropriate level of practice for IMGs to be identified.

References for Appendix D

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, :

Cuddy MM, Winward ML, Johnston MM, et al. (2016) Evaluating Validity Evidence for USMLE Step 2 Clinical Skills Data Gathering and Data Interpretation Scores: Does Performance Predict History-Taking and Physical Examination Ratings for First-Year Internal Medicine Residents?. Academic Medicine, 91: 133-139

Cuddy MM, Young A, Gelman A, et al. (2017) Exploring the relationships between usmle performance and disciplinary action in practice: a validity study of score inferences from a licensure examination. Academic Medicine, 92: 1780-1785

DeMaria S, Samuelson ST, Schwartz AD, et al. ( ) Simulation-based assessment and retraining for the anesthesiologist seeking reentry to clinical practice a case series. Anesthesiology, 119: 206-217

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

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 , :

Gillis A, Weedle R, Morris M and Ridgway P (2016) An international survey of medical licensing requirements for immigrating physicians, focusing on communication evaluation. International journal of medical education, 7: 44-47

Grace ES, Korinek EJ, Weitzel LB and Wentz DK (2010) Physicians reentering clinical practice: characteristics and clinical abilities. J Contin Educ Health Prof, 30: 180-186

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

Herfs PGP, Kater L and Haalboom JRE (2007) Non-EEA doctors in EEA countries: doctors or cleaners?. Medical teacher, 29: 383-389

Langenau EE, Pugliano G and Roberts WL (2011) Relationships between high-stakes clinical skills exam scores and program director global competency ratings of first-year pediatric residents. Medical Education Online, 16: 7362

Lillis S and Roblin H (2014) Progress of successful New Zealand Registration Examination (NZREX Clinical) candidates during their first year of supervised clinical practice in New Zealand. The New Zealand medical journal, 127: 36-42

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

Nayer M and Rothman A (2013) IMG candidates' demographic characteristics as predictors of CEHPEA CE1 results. Can Fam Physician, 59: 170-176

Norcini JJ, Boulet JR, Opalek A and Dauphinee WD (2014) The relationship between licensing examination performance and the outcomes of care by international medical school graduates. Academic Medicine, 89: 1157-1162

Schenarts PJ, Love KM, Agle SC and Haisch CE (2008) Comparison of Surgical Residency Applicants from U.S. Medical Schools with U.S.-Born and Foreign-Born International Medical School Graduates. Journal of Surgical Education, 65: 406-412

Takahashi SG, Rothman A, Nayer M, et al. (2012) Validation of a large-scale clinical examination for international medical graduates. Can Fam Physician, 58: e408-417

Vallevand A and Violato C (2012) A predictive and construct validity study of a high-stakes objective clinical examination for assessing the clinical competence of international medical graduates. Teaching & Learning in Medicine, 24: 168-176

van Zanten M, Boulet JR, McKinley DW, et al. (2007) Assessing the communication and interpersonal skills of graduates of international medical schools as part of the United States Medical Licensing Exam (USMLE) Step 2 Clinical Skills (CS) Exam. Academic Medicine, 82: S65-68

Appendix E. List of high priority, low suitability paper

These 100 papers were identified as describing assessments of 'high priority' areas – professionalism, ethics or patient safety – or using novel technologies, but fell short of the criterion of suitability for inclusion in the final synthesis, meaning they did not describe a summative assessment, or an assessment which was clearly translatable to summative use.

Casas RS, Xuan Z, Jackson AH, et al. (2017) Associations of medical student empathy with clinical competence. Patient Education & Counseling, 100: 742-747

Cooper S, Cant R, Porter J, et al. (2013) Managing patient deterioration: assessing teamwork and individual performance. Emergency Medicine Journal, 30: 377-381

Daupin J, Atkinson S, Bedard P, et al. (2016) Medication errors room: a simulation to assess the medical, nursing and pharmacy staffs' ability to identify errors related to the medication-use system. J Eval Clin Pract, 22: 907-916

DeCesare J and Jackson J. (2016) Advocacy skills in resident doctors. Clinical Teacher, 13: 48-51

Dikici MF, Yaris F and Cubukcu M. (2009) Teaching medical students how to break bad news: a Turkish experience. Journal of Cancer Education, 24: 246-248

Dong T, Kelly W, Hays M, et al. (2017) An investigation of professionalism reflected by student comments on formative virtual patient encounters. BMC Medical Education, 17: 3

Dudas RA and Barone MA. (2015) Can medical students identify a potentially serious acetaminophen dosing error in a simulated encounter? a case control study. BMC Medical Education, 15: 13

Dworetzky B, Peyre S, Bubrick E, et al. (2011) Medical simulation of sentinel events: VAlidation and implementation of a team training curriculum for patient safety in the epilepsy monitoring unit. (EMU). Epilepsy Currents, 11:

Emmert MC and Cai L. (2015) A pilot study to test the effectiveness of an innovative interprofessional education assessment strategy. Journal of Interprofessional Care, 29: 451-456

Espey E, Baty G, Rask J, et al. (2017) Emergency in the clinic: a simulation curriculum to improve outpatient safety. American Journal of of Obstetrics and Gynecology, 217: 699 e691-699 e613

Fanous A, Rappaport J, Young M, et al. (2017) A longitudinal simulation-based ethical-legal curriculum for otolaryngology residents. Laryngoscope, 127: 2501-2509

Ford H, Cleland J and Thomas I. (2017) Simulated ward round: reducing costs, not outcomes. Clinical Teacher, 14: 49-54

Foster A, Chaudhary N, Kim T, et al. (2016) Using Virtual Patients to Teach Empathy: A Randomized Controlled Study to Enhance Medical Students' Empathic Communication. Simulation in Healthcare, 11: 181-189

Foster A, Chaudhary N, Murphy J, et al. (2015) The Use of Simulation to Teach Suicide Risk Assessment to Health Profession Trainees-Rationale, Methodology, and a Proof of Concept Demonstration with a Virtual Patient. Academic Psychiatry, 39: 620-629

Friedrich O, Hemmerling K, Kuehlmeyer K, et al. (2017) Principle-based structured case discussions: do they foster moral competence in medical students? - A pilot study. BMC Medical Ethics, 18: 21

Gettman MT, Pereira CW, Lipsky K, et al. (2009) Use of high fidelity operating room simulation to assess and teach communication, teamwork and laparoscopic skills: initial experience. Journal of Urology, 181: 1289-1296

Grosseman S, Novack DH, Duke P, et al. (2014) Residents' and standardized patients' perspectives on empathy: issues of agreement. Patient Education & Counseling, 96: 22-28

Groves PS, Bunch JL, Cram E, et al. (2017) Priming Patient Safety Through Nursing Handoff Communication: A Simulation Pilot Study. West J Nurs Res, 39: 1394-1411

Guglielmo R, Gustafson S, Fong K, et al. (2017) Using simulation to assess a the competency of graduating medical students to perform the entrustable professional activities of a pediatric intern. Academic Pediatrics, 17: e59

Joyner BD and Vemulakonda VM. (2007) Improving professionalism: making the implicit more explicit. J Urol, 177: 2287-2290; discussion 2291

Karpa KD, Hom LL, Huffman P, et al. (2015) Medication safety curriculum: enhancing skills and changing behaviors. BMC Medical Education, 15: 234

Kerfoot BP, Conlin PR, Travison T and McMahon GT. (2007) Patient safety knowledge and its determinants in medical trainees. Journal of General Internal Medicine, 22: 1150-1154

Kilminster S, Roberts T and Morris P. (2007) Incorporating patients' assessments into objective structured clinical examinations. Education for Health, 20: 6

Kim CW, Myung SJ, Eo EK and Chang Y. (2017) Improving disclosure of medical error through educational program as a first step toward patient safety. BMC Medical Education, 17: 52

Kim S, Brock D, Prouty CD, et al. (2011) A web-based team-oriented medical error communication assessment tool: development, preliminary reliability, validity, and user ratings. Teaching & Learning in Medicine, 23: 68-77

Klamen DL, Reynolds KL, Yale B and Aiello M. (2009) Students learning handovers in a simulated in-patient unit. Medical Education, 43: 1097-1098

Klemenc-Ketis Z and Vrecko H. (2014) Development and validation of a professionalism assessment scale for medical students. Int J Med Educ, 5: 205-211

Krajewski A, Rader C, Voytovich A, et al. (2008) Improving surgical residents' performance on written assessments of cultural competency. J Surg Educ, 65: 263-269

LaCoss J, Pagan-Ferrer J, Hagiwara Y, et al. (2016) The Early Bird Gets the Worm: An Initial Intervention to Gain Palliative Care Communication Skills among Pre-Medical Students. Journal of the American Geriatrics Society, 64: S177-S177

Lam CK, Sundaraj K, Sulaiman MN and Qamarruddin FA. (2016) Virtual phacoemulsification surgical simulation using visual guidance and performance parameters as a feasible proficiency assessment tool. BMC Ophthalmology, 16: 88

Larkin AC, Cahan MA, Whalen G, et al. (2010) Human Emotion and Response in Surgery (HEARS): a simulation-based curriculum for communication skills, systems-based practice, and professionalism in surgical residency training. J Am Coll Surg, 211: 285-292

Layat Burn C, Hurst SA, Ummel M, et al. (2014) Telling the truth: medical students' progress with an ethical skill. Medical Teacher, 36: 251-259

Le TD, Adatia FA and Lam WC. (2011) Virtual reality ophthalmic surgical simulation as a feasible training and assessment tool: results of a multicentre study. Can J Ophthalmol, 46: 56-60

LeBlanc J, Hutchison C, Hu Y and Donnon T. (2013) A comparison of orthopaedic resident performance on surgical fixation of an ulnar fracture using virtual reality and synthetic models. J Bone Joint Surg Am, 95: e60, S61-65

Lee JY, Mucksavage P, Kerbl DC, et al. (2012) Validation study of a virtual reality robotic simulator--role as an assessment tool?. J Urol, 187: 998-1002

LeFlore JL, Sansoucie DA, Cason CL, et al. (2014) Remote-Controlled Distance Simulation Assessing Neonatal Provider Competence: A Feasibility Testing. Clinical Simulation in Nursing, 10: 419-424

Lehmann KS, Holmer C, Gillen S, et al. (2013) Suitability of a virtual reality simulator for laparoscopic skills assessment in a surgical training course. Int J Colorectal Dis, 28: 563-571

Lie D, Bereknyei S, Braddock CH, 3rd, et al. (2009) Assessing medical students' skills in working with interpreters during patient encounters: a validation study of the Interpreter Scale. Academic Medicine, 84: 643-650

Lifchez SD, Cooney CM and Redett RJ, 3rd. (2015) The Standardized Professional Encounter: A New Model to Assess Professionalism and Communication Skills. Journal of Graduate Medical Education, 7: 230-233

Lim BT, Moriarty H and Huthwaite M. (2011) Being-in-role: A teaching innovation to enhance empathic communication skills in medical students. Medical Teacher, 33: e663-669

Lindsley JE, Morton DA, Pippitt K, et al. (2016) The Two-Stage Examination: A Method to Assess Individual Competence and Collaborative Problem Solving in Medical Students. Academic Medicine, 91: 1384-1387

Loukas C, Nikiteas N, Kanakis M and Georgiou E. (2011) Deconstructing laparoscopic competence in a virtual reality simulation environment. Surgery, 149: 750-760

Luigi Ingrassia P, Ragazzoni L, Carenzo L, et al. (2015) Virtual reality and live simulation: a comparison between two simulation tools for assessing mass casualty triage skills. Eur J Emerg Med, 22: 121-127

Lypson ML, Gosbee JW and Andreatta P. (2008) Assessing the patient safety knowledge and experience of trainees. Medical Education, 42: 1133-1134

Maagaard M, Sorensen JL, Oestergaard J, et al. (2011) Retention of laparoscopic procedural skills acquired on a virtual-reality surgical trainer. Surg Endosc, 25: 722-727

Mansour M, Skull A and Parker M. (2015) Evaluation of World Health Organization Multi-Professional Patient Safety Curriculum Topics in Nursing Education: Pre-test, post-test, none-experimental study. J Prof Nurs, 31: 432-439

Merckaert I, Lienard A, Libert Y, et al. (2013) Is it possible to improve the breaking bad news skills of residents when a relative is present? A randomised study. Br J Cancer, 109: 2507-2514

Middleton RM, Alvand A, Garfjeld Roberts P, et al. (2017) Simulation-Based Training Platforms for Arthroscopy: A Randomized Comparison of Virtual Reality Learning to Benchtop Learning. Arthroscopy, 33: 996-1003

Mirza DM. (2010) Assessing professionalism in undergraduates using an RCGP oral membership examination. Medical Education, 44: 509-510

Mittal MK, Morris JB and Kelz RR. (2011) Germ simulation: a novel approach for raising medical students awareness toward asepsis. Simulation in Healthcare, 6: 65-70

Mohamadipanah H, Parthiban C, Nathwani J, et al. (2016) Can a virtual reality assessment of fine motor skill predict successful central line insertion?. American Journal of Surgery, 212: 573-578 e571

Moon MR, Hughes MT, Chen JY, et al. (2014) Ethics skills laboratory experience for surgery interns. J Surg Educ, 71: 829-838

Morgan P, Tregunno D, Brydges R, et al. (2015) Using a situational awareness global assessment technique for interprofessional obstetrical team training with high fidelity simulation. J Interprof Care, 29: 13-19

Naleem S and Chakravorty I. (2013) Measuring patient safety practices in a simulated environment to enhance postgraduate medical training. European Respiratory Journal. Conference: European Respiratory Society Annual Congress, 42:

Naleem S, Holme V, Gosling N and Chakravorty I. (2013) Impact of simulation based learning on patient safety in postgraduate medical training: Design and pilot of a new assessment tool. American Journal of Respiratory and Critical Care Medicine. Conference: American Thoracic Society International Conference, ATS, 187:

Neves Feitosa H, Rego S, Unger Raphael Bataglia P, et al. (2013) Moral judgment competence of medical students: a transcultural study. Advances in Health Sciences Education, 18: 1067-1085

Nguyen N, Watson WD and Dominguez E. (2016) An Event-Based Approach to Design a Teamwork Training Scenario and Assessment Tool in Surgery. J Surg Educ, 73: 197-207

Nomura T, Mamada Y, Nakamura Y, et al. (2015) Laparoscopic skill improvement after virtual reality simulator training in medical students as assessed by augmented reality simulator. Asian J Endosc Surg, 8: 408-412

O'Sullivan AJ and Toohey SM. (2008) Assessment of professionalism in undergraduate medical students. Medical Teacher, 30: 280-286

Ohm F, Vogel D, Sehner S, et al. (2013) Details acquired from medical history and patients' experience of empathy--two sides of the same coin. BMC Medical Education, 13: 67

Ohta K, Kurosawa H, Shiima Y, et al. (2017) The Effectiveness of Remote Facilitation in Simulation-Based Pediatric Resuscitation Training for Medical Students. Pediatric Emergency Care, 33: 564-569

Orientale E, Jr., Kosowicz L, Alerte A, et al. (2008) Using web-based video to enhance physical examination skills in medical students. Family Medicine, 40: 471-476

Oza SK, Sznewajs A, Wamsley MA, et al. ( ) Development and implementation of an inter professional standardized patient assessment. Journal of General Internal Medicine, 1: S458

Paige JT, Garbee DD, Kozmenko V, et al. (2014) Getting a head start: high-fidelity, simulation-based operating room team training of interprofessional students. J Am Coll Surg, 218: 140-149

Palmer RT, Biagioli FE, Mujcic J, et al. (2015) The feasibility and acceptability of administering a telemedicine objective structured clinical exam as a solution for providing equivalent education to remote and rural learners. Rural Remote Health, 15: 3399

Papademetriou M, Perreault G, Gillespie C, et al. (2016) Reducing Medical Errors: Using OSCEs to Assess Fellows' Performance in System Based Practice Milestones. Gastroenterology, 150: S146-S146

Paxton JH and Rubinfeld IS. (2009) Medical errors education for students of surgery: a pilot study revealing the need for action. J Surg Educ, 66: 20-24

Paxton JH and Rubinfeld IS. (2010) Medical errors education: A prospective study of a new educational tool. American Journal of Medical Quality, 25: 135-142

Pernar LI, Shaw TJ, Pozner CN, et al. (2012) Using an Objective Structured Clinical Examination to test adherence to Joint Commission National Patient Safety Goal--associated behaviors. Jt Comm J Qual Patient Saf, 38: 414-418

Perron NJ, Perneger T, Kolly V, et al. (2009) Use of a computer-based simulated consultation tool to assess whether doctors explore sociocultural factors during patient evaluation. J Eval Clin Pract, 15: 1190-1195

Phaosavasdi S, Taneepanichsakul S, Witoonpanich P, et al. (2010) Assessment of medical ethics of fourth-year medical students. J Med Assoc Thai, 93: 1115-1118

Poirier TI, Pailden J, Jhala R, et al. (2017) Student Self-Assessment and Faculty Assessment of Performance in an Interprofessional Error Disclosure Simulation Training Program. Am J Pharm Educ, 81: 54

Prescher H, Grover E, Mosier J, et al. (2015) Telepresent intubation supervision is as effective as in-person supervision of procedurally naive operators. Telemed J E Health, 21: 170-175

Rai AS, Rai AS, Mavrikakis E and Lam WC. (2017) Teaching binocular indirect ophthalmoscopy to novice residents using an augmented reality simulator. Can J Ophthalmol, 52: 430-434

Rawlings A, Knox AD, Park YS, et al. (2015) Development and evaluation of standardized narrative cases depicting the general surgery professionalism milestones. Academic Medicine, 90: 1109-1115

Reid J, Stone K, Brown J, et al. (2012) The Simulation Team Assessment Tool (STAT): development, reliability and validation. Resuscitation, 83: 879-886

Reynolds P and Martindale J. (2014) Development and Validation of the Medical Professionalism Behavior Assessment Tool. Journal of General Internal Medicine, 29: S68-S68

Rhienmora P, Haddawy P, Khanal P, et al. (2010) A virtual reality simulator for teaching and evaluating dental procedures. Methods of Information in Medicine, 49: 396-405

Riesen E, Morley M, Clendinneng D, et al. (2012) Improving interprofessional competence in undergraduate students using a novel blended learning approach. J Interprof Care, 26: 312-318

Roy M and Smee S. ( ) Flagging professionalism during a high-stakes objective structured clinical examination (OSCE). Medical Education, Supplement, 1: 68

Saad AH, Sweet BV, Stumpf JL, et al. (2007) Pharmacist recognition of and adherence to medication-use policies and safety practices. American Journal of Health Syst Pharm, 64: 2050-2054

Sabin M, Weeks KW, Rowe DA, et al. (2013) Safety in numbers 5: Evaluation of computer-based authentic assessment and high fidelity simulated OSCE environments as a framework for articulating a point of registration medication dosage calculation benchmark. Nurse Education in Practice, 13: e55-65

Saleh GM, Wawrzynski JR, Saha K, et al. (2016) Feasibility of Human Factors Immersive Simulation Training in Ophthalmology: The London Pilot. JAMA Ophthalmol, 134: 905-911

Selvander M and Asman P. (2013) Cataract surgeons outperform medical students in Eyesi virtual reality cataract surgery: evidence for construct validity. Acta Ophthalmol, 91: 469-474

Shavit I, Keidan I, Hoffmann Y, et al. (2007) Enhancing patient safety during pediatric sedation: the impact of simulation-based training of nonanesthesiologists. Arch Pediatr Adolesc Med, 161: 740-743

Shaw TJ, Pernar LI, Peyre SE, et al. (2012) Impact of online education on intern behaviour around joint commission national patient safety goals: a randomised trial. BMJ Qual Saf, 21: 819-825

Shrader S, Dunn B, Blake E and Phillips C. (2015) Incorporating Standardized Colleague Simulations in a Clinical Assessment Course and Evaluating the Impact on Interprofessional Communication. Am J Pharm Educ, 79: 57

Stephenson L, Gold JA, Mohan V and Gorsuch A. (2014) Simulation to improve use of electronic health records and improve patient safety. American Journal of Respiratory and Critical Care Medicine. Conference: American Thoracic Society International Conference, ATS, 189:

Stocker M, Menadue L, Kakat S, et al. (2013) Reliability of team-based self-monitoring in critical events: a pilot study. BMC Emerg Med, 13: 22

Supiot S and Bonnaud-Antignac A. (2008) Using simulated interviews to teach junior medical students to disclose the diagnosis of cancer. Journal of Cancer Education, 23: 102-107

Taglieri CA, Crosby SJ, Zimmerman K, et al. (2017) Evaluation of the Use of a Virtual Patient on Student Competence and Confidence in Performing Simulated Clinic Visits. Am J Pharm Educ, 81: 87

Thomsen ASS. (2017) Intraocular surgery - assessment and transfer of skills using a virtual-reality simulator. Acta Ophthalmol, 95: 44562

Tobler K, Grant E and Marczinski C. (2014) Evaluation of the impact of a simulation-enhanced breaking bad news workshop in pediatrics. Simulation in Healthcare, 9: 213-219

Tromp F, Rademakers JJ and Ten Cate TJ. (2007) Development of an instrument to assess professional behaviour of foreign medical graduates. Medical Teacher, 29: 150-155

van der Sijs H, van Gelder T, Vulto A, et al. (2010) Understanding handling of drug safety alerts: a simulation study. Int J Med Inform, 79: 361-369

van Dulmen S, Tromp F, Grosfeld F, et al. (2007) The impact of assessing simulated bad news consultations on medical students' stress response and communication performance. Psychoneuroendocrinology, 32: 943-950

Verma A, Griffin A, Dacre J and Elder A. (2016) Exploring cultural and linguistic influences on clinical communication skills: a qualitative study of International Medical Graduates. BMC Medical Education, 16: 162

Wagner DP, Hoppe RB and Lee CP. (2009) The patient safety OSCE for PGY-1 residents: a centralized response to the challenge of culture change. Teaching & Learning in Medicine, 21: 41852

Wang XJ, Sim J, Lucero C, et al. (2013) The 1:00 AM Consult: Assessing Communication with Primary Providers as a Clinical Skill in Gastroenterology Fellowship Training. American Journal of Gastroenterology, 108: S488-S488

Wiest KM, Farnan JM, Byrne E, et al. (2017) Use of Simulation to Assess Incoming Interns' Recognition of Opportunities to Choose Wisely. Journal of Hospital Medicine, 12: 493-497