BMJ Open

Variable trajectory: a systematic review, analytic synthesis and construct domain consolidation of international measures of competence in doctors and medical students

Kirsty L Hodgson , Daniel J Lamport , Allán Laville

To cite: Hodgson KL, Lamport DJ, Laville A. Variable trajectory: a systematic review, analytic synthesis and construct domain consolidation of international measures of competence in doctors and medical students. *BMJ Open* 2021;**11**:e047395. doi:10.1136/bmjopen-2020-047395

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online. (http://dx.doi.org/10.1136/bmjopen-2020-047395).

Received 26 November 2020
Accepted 05 August 2021

School of Psychology and Clinical Language Sciences, University of Reading, Reading, UK

Correspondence to
Dr Allán Laville;
a.laville@reading.ac.uk

ABSTRACT

Background Competence is assessed throughout a doctor's career. Failure to identify and manage impaired competence can have critical consequences. Consistent conceptualisation and accurate measurement of this construct is imperative. Therefore, the objective of this review was to identify and evaluate measures used to assess competence in doctors and medical students.

Methods A systematic search of the published literature was undertaken between December 2019 and February 2020 for articles reporting on the measurement of competence in doctors and/or medical students. Searches were conducted in the PsychSOURCE, US National Library of Medicine National Institutes of Health, MEDLINE (PubMed), The Cochrane Central Register of Controlled Trials and Web of Science electronic databases. Citation screening and forward citation tracking of included studies were carried out to identify any further relevant papers for inclusion. One thousand one hundred and thirty-six potentially relevant articles were screened. An analytic synthesis approach was implemented to the identification, organisation and interpretation of homogenous study and measure characteristics.

Results Twelve competence domains were identified from the 153 identified measures. Knowledge and procedural competence domains were the dominant focus of publications reporting current medical practice, but less so in research-based studies which more frequently assessed interpersonal, psychological, physiological and ethical competencies. In the 105 included articles, the reporting of measurement instrument quality was varied, with comprehensive reporting only present in 53.6% of measures; validation for some of the measures was particularly limited.

Discussion While this review included a considerable number of publications reporting the measurement of competence in doctors and medical students, the heterogeneity of the measures and variation of findings limit the ability to evaluate their validity and generalisability. However, this review presents a resource for researchers and medical educators which may inform operational practice and future research.

PROSPERO registration number CRD42020162156.

Strengths and limitations of this study

- ► To our knowledge, this is the first contemporary systematic review to consolidate globally recognised domains relevant to competence assessment of doctors and medical students in both research and in-practice measures.
- ► This PROSPERO registered review employed an extensive and rigorous search strategy of four databases and forward citation searches using defined selection and abstraction criteria calibrated by three researchers, undertaken in alignment with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement, and guided by the synthesis without meta-analysis reporting guidelines.
- ► The review identified a broad range of measures used to assess competence, and identifies the homogeneity and heterogeneities using a textual and conceptual analytic synthesis approach.
- ► A limitation of this review is that English language restrictions were placed on the searches, and although research studies were derived from 18 countries, the inclusion of further languages may yield unidentified studies, and subsequent differences in worldwide assessments of competence.

INTRODUCTION

Competence is assessed throughout doctors' professional lifespan, from selection for entry to medical school to medical student to junior doctor, to senior doctor until retirement. The timely identification, support and remediation of doctors who do not demonstrate sufficient competence to undertake their duties safely is imperative. However, the proportion of doctors who are incompetent, or dyscompetent, and failing to uphold adequate practice standards is unclear.[1] Robust assessment of competence necessitates accurate and standardised measures that comprehensively examine all domains relevant to the

core competencies required to undertake a job effectively. It has been argued that competencies are learnt, complex context-specific dispositions, but the complexity and diversity inherent in doctors' responsibilities make measurement of competence a complex task.[2]

Despite numerous definitions,[3] an agreed explicit, amalgamated, and quantifiable definition of the construct of competence, and in many cases a consistent and transparent chain of inferences to validate the observed assessment and measurement of competence are lacking.[4] It has been argued,[5] that it is challenging to identify reliable and valid measures of competence as many of the skills are considered difficult to quantify. Contemporary competence-based medical education is predominantly driven by patient safety, efficiency and clinical outcomes.[6]

In order to explore these challenges, and provide an overview of the diverse facets of competence, and their dominance within current assessment measures; the objective of this systematic review of peer-reviewed scientific literature was to identify and evaluate specific measures of competence relevant to doctors and medical students. As the accurate measurement of competence is paramount to patient safety, a systematic review of the measures employed and domains of competence assessed in medical student selection, medical students and doctors was considered important.

METHOD

The review protocol was registered with PROSPERO, adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement,[7] and guided by the 'synthesis without meta-analysis reporting guidelines'.[8]

Literature search

Data sources

A systematic search of the literature was conducted in the PsychSOURCE, US National Library of Medicine National Institutes of Health, MEDLINE (PubMed), The Cochrane Central Register of Controlled Trials (CENTRAL) and Web of Science electronic journal databases. Citation screening and forward citation tracking of included studies was carried out to identify any further relevant papers; these were retrieved from the journal site or academic library databases.

Search strategy

A pilot search was undertaken to determine relevant electronic search engines, refine the search strategy, define search keywords, identify relevant MeSH terms, determine search restrictions and select Boolean operators. English and American-English language terms were used. Search terms employed the keywords; 'doctor', 'competence' and 'performance', incorporating the Boolean operator ('and'), in addition to MeSH terms (physicians, medical students). Restrictions on the databases included human studies, published in English language and full text availability. No date restrictions were applied. Searches were undertaken between December 2019 and February 2020. Further searches in response to the literature were performed through reference lists of reviewed articles in forward citation searches. Results were downloaded to reference management software and de-duplicated.

Study identification and selection

Inclusion/exclusion criteria

Studies were included if they described the measurement of competence in medical students and doctors, with full-text availability in peer-reviewed journals, published in English language. Commentaries, editorials, letters, book chapters, conference papers, books, doctoral theses, dissertations and articles which did not adequately report competence assessment measures were excluded.

Calibration of findings

One reviewer undertook title and abstract screening, and selection methods were supervised by two reviewers who verified the extraction of data, checked consistency and methodological clarity. Study selection disagreements and subsequent queries about inclusion following full-text review were resolved by discussion between researchers to reach consensus.

Study selection

From searches, 1136 papers were identified; after duplicates were removed 648 original articles remained. Following assessment for relevance to the objectives and selection criteria 105 articles were selected for systematic analysis. Figure 1 reports the PRISMA (2009) flow diagram of study selection, and specific search engine results are presented in online supplemental file 1.

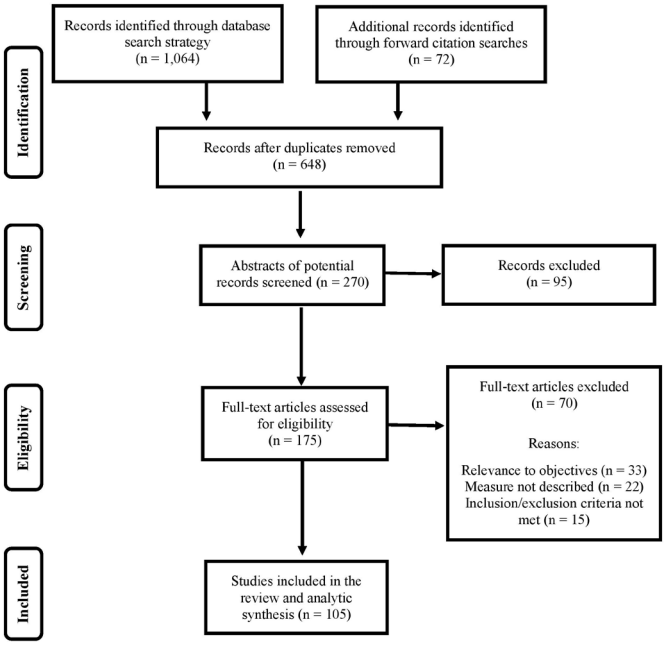

Figure 1 Flow diagram of the review search strategy and study selection based on PRISMA (2009) guidelines.

Data extraction

Selected articles were subjected to three preliminary levels of screening: title, abstract and full text. The following data (when available) from articles was extracted using a predesigned and piloted data-extraction pro-forma which included the following categories and subcategories: (1) referencing information, (2) quality assessment, (3) study characteristics (geographical origin, sample, objectives, design), (4) type of measure, (5) assessment to examine whether the properties of each measurement tool (construct validity, content validity, criterion validity, internal consistency, test–retest reliability, inter-rater reliability, etc) had been reported, (6) main results/outcomes, (7) additional further information relevant to the objectives. Results were recorded in an Excel spreadsheet and each study was assigned a reference identifier code (online supplemental file 2).

Study assessment of methodological quality

Quality inclusion criteria were restricted by whether sufficient detail was provided to enable analysis of methodological quality and the evaluation of study measures. The five 'appraisal prompts for informing judgements about quality of papers' defined by critical interpretive synthesis (CIS)[9] methodology were evaluated and tabulated for all included studies. CIS yielded quality appraisal nominal data scores for each question (1 yes, 0 no), with a cumulative score ranging from 0 to 5; scores ranged from 2 to 5. Ninety-four of the 105 studies achieved the maximum (5/5) cumulative rating score, six scoring (4/5), four scoring (3/5) and one scoring (2/5).

Data evaluation and synthesis

As the protocol anticipated, due to the heterogeneity of measurement tools in this review, and inconsistencies in methodology and reporting of results, a meta-analysis was not possible. Preliminary categorisation of the studies included data extraction of study characteristics, types of competence measures and measurement properties. A conceptual analytic synthesis of identified dimensions of similarity and difference in defining the construct of competence in this population was undertaken.

Patient and public involvement

Patients or the public were not directly involved in the design or planning of this study.

RESULTS

The search strategy generated 105 articles selected for full data abstraction: 13 from PsychSOURCE, 19 MEDLINE/PubMed, 24 Cochrane Library—CENTRAL, 21 Web of Science and 28 from specific paper retrieval forward citation searches at the journal site. Selected reviewed papers in the results section are referenced in the format 'P:1,2,3,4,5', and individual measures are referred to as 'M:1,2,3,4,5' (see online supplemental file 2 for assigned reference identifier codes and corresponding references).

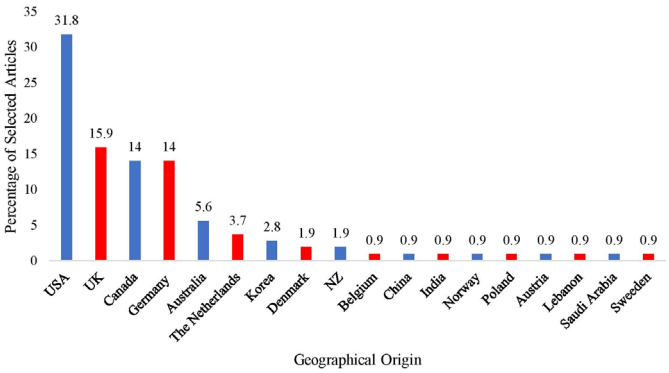

Figure 2 Percentage distribution of all articles' geographic origin.

Characteristics of selected studies

Included articles were published between 1990 and 2020. There was a progressive increase in the number of publications per year over this period (mean annual publications per year 0.3 from 1990 to 1999, 1.9 from 2000 to 2009 and 8.2 from 2010 to 2019), consistent with increasing research interest and societal awareness of the extent and consequences of medical error.[P:105] The study designs included: 2 intervention, 22 measurement development, 26 measurement evaluation, 17 curriculum development, 40 clinical assessment, 21 training evaluation and 11 experiential assessment; 26 studies comprised multiple categories.

Studies were restricted to English language, but without restriction on geographic origin and were derived from 18 countries. One study[P:26] undertook data collection in two countries (USA and Canada), therefore data were counted in both. The majority of research was from the USA (n=34 articles), UK (n=17), Germany (n=15) and Canada (n=15). There were five studies from Australia,[P:3,35,64,80,102] four from the Netherlands[P:23,49,58,69] and three from Korea.[P:13,61,73] Two studies were from New Zealand[P:5,12] and Denmark,[P:45,66] Sweeden,[P:10] Lebanon,[P:36] Saudi Arabia,[P:46] Austria,[P:48] Belgium,[P:50] Norway,[P:59] Poland,[P:60] India[P:79] and China[P:103] were each represented by a single study. The percentage distribution of all articles geographic origin is presented in figure 2.

Competence assessment measures

From the 105 articles, 153 measures of competence were identified, 98 of which were unique to that study (a comprehensive list of all measures is presented in online supplemental data file 3). The competence assessment measures were categorised as either established 'in-practice measures' (IM) currently implemented for the assessment of doctors and medical students, or 'research-measures' (RM) developed or applied to assess competence for research evaluation, but not at the time of publication employed for occupational competence assessment. Thirty-eight studies examined IM, 45 examined RM and 22 explored both RM and IM. Most (n=76) of the selected studies were single-institution assessments—limiting the generalisability of findings, 25 were

multicentric (based at multiple institutions) and 4 were not institution-based.

Measures were categorised by their procedural comparability, and seven dominant operational characteristics were identified: (1) clinical performance assessment—was applied in 30.7% of measures (n=38 IM, n=9 RM), typically assessed using simulated patient ratings to assess clinical competencies such as communication skills and diagnostic reasoning in response to presented symptoms; (2) anatomic model simulation—providing simulations of human anatomy in order to test knowledge and used in 6.5% of measures (n=9 IM, n=1 RM); (3) psychometric assessment—predominantly evaluated personality, aptitude, reasoning and emotionality, and conducted in 7.8% of measures (n=2 IM, n=10 RM); (4) patient records—to assess the medical record keeping accuracy and completeness, were used in 1.3% (n=2 IM); (5) computerised testing—included a variety of computerised assessments, used in 9.8% of measures (n=1 IM, n=14 RM) and (6) written testing—included multiple-choice and short-answer questions in traditional pen-and-paper testing in 31.4% of measures (n=19 IM, n=29 RM). The prevalence of each measurement category demonstrated a preference for the application of written testing and clinical performance assessment in the assessment of competence. Fifty-six (36.6%) measures applied a unidimensional procedure to assess the target domains, and 97 (63.4%) operated multidimensional procedures.

Established IM

Competence was assessed through a variety of IM methods. Objective structured clinical examinations (OSCEs) typically involve patients or simulated patients and the most duplicated measure (in 20 articles). OSCEs are frequently considered[P:53] the 'gold standard' for assessing competence skills, as they are high-fidelity, can be reliable, and are regarded as valid because of their ability to differentiate between occupational levels of aptitude and specialisms. The use of standardised patients (SPs) endeavours to replicate doctor–patient consultations can identify variation in clinical practice between doctors in a realistic simulated setting. SP single-case assessments were used to measure specific competencies in nine of the studies.

Anatomic model simulation testing, employed in 10 measures, is widely used as a safe method to develop procedural competence in early medical education. However, whether the anatomic-model patients are the most valid means of assessing procedural competence is widely debated.[P:38] Workplace-based assessment (WBA) such as mini-clinical evaluations (Mini-CEX) and direct observation of procedural skills (DOPS) are heavily integrated in medical education. However, the Mini-CEX and DOPS lack standardisation, cannot be used with large numbers of participants simultaneously, are costly, time consuming, and transferability of skills is not always guaranteed in a simulated assessment setting; particularly in a workplace with high levels of social, psychological and physiological stress.[P:20] Four studies[P:20,29,79,89] examined 360-degree assessments (ie, rating evaluations from colleagues within the workplace).

Written tests provide reliable assessment of several domains, particularly knowledge and psychometric assessment. Three studies[P:2,95,97] used the Jefferson Scale of Physician Empathy, one[P:3] employed the Jefferson Scale of Physician Empathy-Student Version. Written tests measuring knowledge included two studies using the US Medical Licensing Examinations, and three studies that included the Medical College Assessment Test—an assessment used by many medical schools as part of the entry-selection process. However, written tests are considered insufficient to measure behavioural clinical competence;[P:20] therefore, there has been a traditional reliance on the judgement of competence by educators or seniors in demonstration-observation tests of clinical performance skills. The combination of a knowledge test (written assessment) with an applied test of clinical skill (OSCE) was shown to increase predictive validity versus either alone.[P:12] Vignettes (five studies) and written case simulations have been widely used to assess competence in both clinical training and experimental research, because of their ease of administration, negligible cost and objective quantifiable results. However, they do not assess important social interaction skills such as history taking, nor assess competence in a real-world situation including occupational stressors. Among IM measures there was significant variability in testing procedures, content domains and inconsistent assessment outcomes, therefore these were not grouped for evaluation purposes.

Contemporary experimental research measures

Electronic-technology has an increasing role in healthcare, but raises new challenges for competence assessment; one study presented an electronic platform[M:149] that automatically assesses a doctor's competence from online textual consultations with patients using a novel machine that provides auto-evaluation based on patient satisfaction following online consultation. In addition to patient evaluations, innovative simulation-based medical education and assessment provide alternatives to practice and assessment without risk to patients.[P:26]

In an era where doctors' retirement age has extended, further cognitive competence considerations have emerged. Forty-two measures in 37 of the studies assessed cognitive competence, publication dates indicating that research interest in this area has markedly risen in recent years (7.6% of the studies between 2000 and 2009, compared with 26.7% from 2010 to 2020). Neuropsychological screening found[P:81] that many of the doctors who fell significantly below expected competence assessment levels had cognitive impairment sufficient to explain their diminished competence. Another study developed 'the Mini-NeuroCart',[M:80] a cognitive and psychomotor test battery for assessment of subjective and objective measures of alertness, mood, concentration and self-assessed procedural competence, and showed that the

competence of postcall surgeons was similar to, or worse than that of ethanol-intoxicated surgeons.

Competence domains

Studies were categorised according to the competence assessment domains. The list of measures in each domain is presented in online supplemental material file 4. Twelve ascendant domain categories based on their prevalence in the selected measures were identified, with descriptors of these in table 1.

Consideration of each domain within the literature

Knowledge (*Domain 1*), deemed as the primary proximal determinant of competence[P:11] was the most commonly assessed competency in IM, historically assessed through written examinations,[P:31] contemporary measures endeavour to combine knowledge with procedural assessment. Concordantly, only 27.2% of measures conformed to traditional written-test approaches, though 6.4% of computerised test measures were a comparable format.

Within *Domain 2* (procedural competence), 54.6% of the measures employed a clinical performance test; other measures included patient vignettes, written or computerised tests (30.2%) and anatomic model simulation (15.6%).

Domain 3 (judgement and bias) was addressed in 41.8% of measures and incorporated competence with diversity, and awareness of bias in an occupational context. Cultural competence was found to be lacking in one study[P:8] as a result of an interaction between race and treatment outcomes as there was a higher likelihood of doctors escalating the use of opioids for African American patients. Sexual and gender minority individuals experience high rates of discrimination when seeking healthcare, contributing to patient care disparities.[P:68] Furthermore, doctors reported discomfort when treating people who have disabilities,[P:96] as assessed by the Scale of Attitudes Toward Disabled Persons.[M:141]

Domain 4 (communication): competent doctor–patient and inter-professional communication is included within this domain. Ineffective communication with colleagues has been associated with reduced patient safety and risk to care quality,[P:71] and the most frequent cause of complaints against doctors is related to poor communication.[P:42]

Though doctors are expected to uphold the customary standards of professionalism (*Domain 5*), there was no consensus on its definition,[P:77] and an absence of explicit standardised factors measured within this domain.[P:78] Most (73.8%) of the measures used subjective rating scales; many within OSCE and WBA. Remarkably, although 36 IM measured professionalism, only 9 RM assessed this domain.

Cognitive competence (*Domain 6*) denotes the specific cognitive resources required for the selection and application of skills such as diagnostic and clinical reasoning. Wechsler Adult Intelligence Scale (WAIS-IV) assessments have found cognitive resource differences between doctors who experience difficulties, and those who do

Table 1 Domain descriptors in descendent order of prevalence in all measures

Number	Domain	Prevalence (%)	Descriptor
1	Knowledge	55.5	The acquisition and retention of learning.
2	Procedural competence	45.1	The ability to perform specific technical skills/procedures required for the occupational obligations.
3	Judgement and bias	41.8	An ability to nullify or negate judgement biases to competently work in the benefit of patients and colleagues.
4	Communication	40.5	The ability to engage patients and colleagues in effective interaction.
5	Professionalism	29.4	To uphold the expected traditional occupational behavioural standards of the profession, and regulatory code of conduct.
6	Cognitive competence	27.4	To effectively use the cognitive resources required for the selection and application of skills.
7	Environmental competence	23.5	This domain relates to the ability to maintain acuity in environmental adversity.
8	Coping competencies	21.5	Competence to psychologically manage the complex occupational exigencies.
9	Self-assessment accuracy	20.9	Self-assessment and accurate self efficacy in continuous self-monitoring of competence.
10	Empathic competence	20.2	Empathic Competence describes the awareness and responsiveness to patients' and colleagues' feelings and experiences.
11	Ethics and patient safety	16.9	A commitment to exemplar medical practice by prioritising patient safety and ethical standards.
12	Physiological competence	6.5	Physiological competence describes a state of optimal functioning that supports the physical and psychological workplace demands.

not.[P:70] The cognitive concept of 'mental workload' perceives the brain having a limited capacity to process stimuli, research has found that high levels of workload are associated with error and poor competence.[P:19] Traditional measurement of task performance does not necessarily predict competence when a combination of tasks must be undertaken under time pressure resulting in cognitive overload, and the 'task-shedding' of key competencies resulting in risk of impaired competence.

Environmental competence (*Domain 7*) employs measures that account for (or simulate) the context of occupational pressures in assimilating and responding to environmental context[P:72] using real-time information management.[P:54] One study[P:19] examined the effect of multitasking on competence in medical students using a secondary task method employing a venepuncture procedure[M:32] alongside an electronic application[M:31] which demanded a tapping response when the device vibrated, mental workload was measured in time delay and effects were observed on communication deterioration. Measures that included time (chronometry)[P:15,19,26,34,46] highlighted that timed test increased the reliability and validity of measuring diagnostic accuracy,[P:15] as time-pressure has been found to disrupt the dynamic interaction between reasoning systems influencing diagnostic tasks.[P:46] Conclusions about competence may be flawed unless contextual environmental factors are considered.[P:72]

Doctors are expected to handle limited resources, pressure, and life or death performance outcomes that often carry an emotional load; 'coping competence' (*Domain 8*) is therefore paramount.[P:72] Distress and inability to cope has been associated with medical errors.[P:105] Therefore, the ability to cope with environmental, cognitive, physical and emotional workplace demands is key to competence and occupational fulfilment.[P:104]

The ability to accurately 'self-assess' and work at a level of occupational aptitude. *Domain 9* is key to the identification of difficulties, and safe practice. RM focused more heavily on self report measures (n=27) than IM (n=5). Concerningly, one study[P:3] found that self-rated empathy and observer ratings were not associated, and another[P:7] found no correlation between confidence scores and clinical or knowledge abilities. A doctor's ability at any career stage to accurately self-assess their competence acquisition provides assurance of working within scope of practice.[P:14]

Medical educators and professional regulators recognise the importance of empathic competence (*Domain 10*); strong associations were found between observer ratings of empathy and OSCE scores.[P:3] The prominence of empathic competence within the selected measures was comparable in RM (n=16) and IM (n=15). In medical education research,[P:95] empathic competence has been demonstrated to increase in male trainees and decrease in females during the course of training. Importantly, empathic competence has been found to be strongly associated with patient satisfaction.[P:97]

Domain 11 relates to adherence to ethical standards and commitment to patient safety. This domain was addressed in regulatory assessments,[eg, M:129,130,138] rating scales assessed by senior doctors,[eg, M:1,9,91,117,126] WBA's,[eg, M:146,147] patient evaluations[M:36] and values-based measures such as a questionnaire survey about notions of a 'good doctor'.[M:8] Neuropsychological testing[M:123] has been found to be a strong predictor of risk to competence[P:86]; a consistent decline in competence has also been identified in recertification examinations in America related to years since graduation.[P:82] Another study[P:75] developed and tested an OSCE designed to directly assess sociocultural dimensions of patient safety competency,[M:117] demonstrating sufficiently reliable station scores in this domain. Engagement with continuous competence improvement and collaboration with colleagues (rather than isolated practice) has been identified as a protective factor.[P:28]

Physiological competence (*Domain 12*) emerged as a domain rarely considered in assessment. RM measures within this domain included psychometric measures of burnout,[M:18,52] the Epworth Sleepiness Scale[M:151] and the Medical Outcomes Study Short Form (SF8) psychological and physical health survey.[M:48] Traditional training and assessment typically occurs in optimum daytime working conditions, which arguably[P:44] neglects nocturnal circadian effects. One study[P:44] found that training delivered during a simulated night shift in an emergency department was effective in significantly increasing medical students' self-efficacy. Another study[P:48] used electroencephalographic recordings and measured the competence of doctors in the course of 24 hours shifts with and without afternoon rest, demonstrating nocturnal cognitive deactivation without rest, and a stimulating vigilance promoting effect of the resting period. Only 2/10 psychophysiological measures[M:129,130] were IM, both were regulatory assessments.

Figure 3 reports the distribution of domains of competence assessed in selected studies reporting in-practice assessments of competence (% of IM), and those from research studies (% of RM). There was notably some variation between the IM and RM in domain prevalence. There was less focus on the 'knowledge' domain in the RM studies. However, the largest discrepancy between the RM and the IM papers was for the 'professionalism' domain, which was rarely included (n=9) in RM compared with IM (n=36) (χ^2=26.6; p<0.001).

Figure 4 presents the percentage of the procedures that were used in the measures within each of the 12 domains. Therefore, the outcome of the construct domain consolidation of measures of competence was the identification of 12 domains. There were significant differences in the predominant domains assessed in IM versus RM studies, with less focus on professionalism and knowledge-based domains in the latter.

Instrument properties

Instrument properties were evaluated for their attention to the reporting of measurement properties such

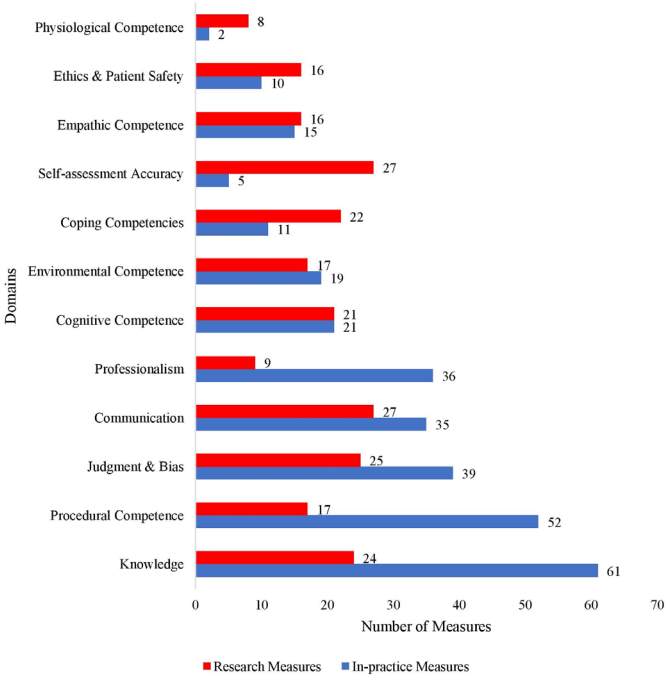

Figure 3 Distribution of domains of competence assessed in selected studies reporting in-practice assessments of competence, and those from research studies.

as reliability, validity and interpretability. Despite the preponderance of identified measures, there is a dearth of evidence for their reliability and validity; 53.5% of the identified measures provided reliability and validity data

Figure 4 Donut charts representing the percentage of the procedures that were used in the measures within each of the 12 domains.

(37/73 IM and 45/80 RM), whereas 21.5% of studies described measures without reference to either their reliability or validity (16/73 IM and 17/80 RM).

Measurement content validity was predominantly based on citation in several of the measures[eg, M:46,47,48,50,52,64,65,114] rather than validation against an agreed standard measure. With regard to internal consistency, a Cronbach's alpha (α) value was reported for 39 reviewed measures. Cross-cultural validity was addressed in 29.4% of measures. Predictive validity was addressed by a small number of studies to measure the predictive value of premedical school assessment on later performance, and the predictive value of cognitive assessment in predicting subsequent impaired competence in ageing doctors.[P:81,M:123] Computerised, written and oral measures were typically single-domain measures of knowledge but reported higher levels of validity.

Details on measurement interpretability were found for 119 reviewed measures. As many established measurement methods were institution-specific there was substantial variability in testing procedures, content domains and subsequent outcomes. However, practice and test–retest effects in longitudinal studies received minimal consideration.[P:42, 81] Testing measures of competence in real practice using double-blind observations are evidently unfeasible for logistical and ethical reasons; however, workplace-based 360-degree feedback from a doctor or trainees colleagues provides an alternative whereby varied sources of in-practice observations with real patients and there is evidence for the validity of this method.[M:33,43]

Measurement assessment context

Increasing effort has been dedicated to enhancing the realism of technical skills training by introducing SP's and role-plays[P:51]; 45.4% of the reported measures (53/73 IM and 30/80 RM) were designed to be assessed in a simulated context, and just 20.9% with real patients (only 18 of which were IM), 33.7% of measures used neither. The majority assessed competence in either a WBA (n=19), or (more commonly) a simulated assessment (n=67). However, the acceptability and feasibility of real patient measures rather than simulated context assessment encounters has significant cost, time and ethical barriers. However, there has been increased emphasis in more recent studies on assessments which attempt to mimic real world clinical situations.

Positional orientations

Studies assessed competence measurement from two positional orientations; assessor and assessee. Measurement of assessee competence was categorised by four populations; medical school applicants (n=3), medical students (n=60), junior doctors (n=19) and fully trained (senior) doctors (n=20). The assessment of medical school applicants for core competencies received minimal attention (three studies) in the selected literature; one research study[P:9] used a video-based situational judgement test[M:16] assessing social competencies, the second[P:58] addressed

an assessment framework measure[M:91] and the third[P:60] employed a coping responses inventory[M:94] in addition to a state entrance examination.[M:95] Medical students were the target population in the majority of studies, as a competency-based curriculum is becoming a dominant organising framework for medical education and assessment[P:47], and its emphasis was reflected in its research dominance in the literature. Only three studies examined multiple assessee populations; one[P:60] studied medical school applicants, medical students, junior doctors and fully qualified doctors using psychometric burnout[M:94] and knowledge test[M:95] measures, and found that the assessment of coping competence at medical school enrolment was a predictor of longitudinal professional competence and career development. The second[P:72] (medical students, junior doctors and fully qualified doctors) used a group evaluation[M:113] to explore the contextual-environmental factors that affect competence assessment, and the third study[P:92] found that vignettes,[M:135] SP's[M:136] and chart abstraction[M:137] measures were a valid and comprehensive method to measure competence in junior doctors and fully qualified doctors.

Assessor perspectives are important in the interpretation of competence outcomes. Assessor-performed analogue rating scales were used in 47 measures (31 IM and 16 RM). Evaluative standpoints of assessors varied between the selected studies and were described as follows: Fifty measures (54.8% of IM, 12.5% of RM) employed supervisory occupational appraisal of a junior or student. Six measures (4.1% of IM, 3.8% of RM) used peer assessment. Twenty-two measures (19.2% of IM, 10% of RM) used patients or SP's who can provide a key role in evaluating competence as judged by their direct experience of the doctor or trainee. Four measures (2.7% of IM, 2.5% of RM) used disciplinary/regulatory assessment procedures designed to assess a comprehensive range of domains. Finally, 37 measures (8.2% of IM, 38.8% of RM) involved self-assessment of competence. Therefore, a diverse range of assessor standpoints influence competence assessment and providing feedback is arguably[P:23] a complex affective process and the assessment is determined by the assessor's cognition, beliefs and emotions. Inter-rater reliability is pertinent to the interpretability of selected measures in this review as the majority of multi-domain IM evaluations of clinical performance, such as OSCE's, and the Mini-CEX rely extensively on rater-assessment. Inter-rater reliability was reported for a small number of studies[eg, P:51,67,75,77,89]; however, many were monocentric studies.

DISCUSSION

This systematic review sought to identify and evaluate measures of competence in doctors, medical students and those applying for entry to medical school. One hundred and five articles were identified that included a broad range of 153 competence assessment measures. The principal findings were as follows: very few studies conceptualised competence through definition or operationalising the construct. Therefore, although competence is considered paramount for safe practice, no consensus in the selected literature exists regarding its conceptualisation, nor is there measurement standardisation. Twelve domains used to conceptualise and measure competence were identified. The deconstruction of each domain and its relevance yielded strong operational purpose in 11 of the domains but identified inconsistencies and procedural subjectivities regarding professionalism (Domain 5). Assessment of professionalism in the studies lacked reliable measurement evaluation tools and was open to implicit judgement biases; this review therefore questions whether further objective assessment of professionalism may be identified and validated. In contrast, judgement and bias (Domain 3) included a range of specific measures applied to assess competence with diversity and judgement orientations; the findings that patient characteristics were commonly associated with treatment outcomes requires continuous examination from a training and assessment perspective to ensure non-discriminatory ethical treatment of all patients.[10] However, it is important to acknowledge that the competence domain categories are not necessarily indicative or proportionally reflective of all globally accepted assessment tools. There are likely varying conceptualisations of competence in different cultures, and in different clinical, social and economic contexts, therefore further research on geographic and cultural construct analysis is advocated to explore potential differences in worldwide assessments.

Interestingly, the distribution of competence domains differed between IM and RM; the RM reflecting more explicit and focused consideration of the contribution of behaviouristic (skills and performance), systemic pressures, psychophysiological (cognitive, emotional, health) and vantage-point (beliefs, perceptions and attitudes) factors as conceptually relevant to competence. A further follow-up exploration to examine which of the RM are being implemented in systemic operational practice would be worthwhile. It is noteworthy that findings related to the validation of the measures were often limited or entirely absent, therefore further validation of existing measures is recommended. There is a broadly held assumption that assessment formats with singular competence domains are limited[11] whereas using comprehensive multidimensional assessment measures will minimise deficiencies.[12] Furthermore, the majority (n=76) of selected studies were single-centre assessments. To address this, we encourage multicentric testing of competence assessment to offer greater generalisability of the data. Correspondingly, the sharing of assessment measures between institutions may encourage greater homogeneity in practices.

The assessment of some domains such as coping competence is considered by many medical schools to be crucial in identifying those with the necessary attributes to train to be a doctor. These domains may therefore be considered as part of the continuum of assessment

of competence throughout a medical career. These were therefore included in this systematic review. This continuum approach is consistent with the increasing recent interest in programmatic longitudinal assessment of both students and doctors using multiple assessment techniques, in which individual assessment elements are part of a process of gathering information contributing to the overall assessment rather than being used on their own to make decisions. In this model these assessments contribute to the learning process, breaking down the traditional dichotomy between formative and summative assessment.[13] The findings of this systematic review support this, because it is clear that individual assessments taken in isolation each have their limitations. Furthermore, the inclusion of both formative and summative components in this longitudinal assessment enhances the learning process.

A wide range of global competence measures were reported in the selected studies. These included written assessments, OSCEs, anatomic simulations and WBAs. The combination of a written test with a clinical skill OSCE test has been demonstrated to increase the predictive validity of either alone. In research studies assessing competence there is considerable interest in simulation-based assessments and some have focused on cognitive assessment which may be of particular importance with many doctors working beyond retirement.

Many assessments of competence are contingent on assessor subjectivity. While some studies demonstrated attempts to assess inter-rater reliability, this is an important issue that requires further attention. Assessor bias is therefore a worthwhile area of further research due to concerning discrepancies in assessment accuracy that have been attributed to unreliable self-assessment,[14] inconsistent applications of organisational standards resulting in a lack of reliability, and unclear rating scale criteria. Assessments involving direct observations inherently include automatic judgement processes, and do not necessarily provide objective assessment,.[15] However, it has been argued,[16] that standardisation may be a constraint on authentic assessment interpretations and delineations. Approaches using 360-degree global rating evaluations by multidisciplinary multilevel colleagues who interact with a doctor or trainee routinely may improve objectivity through a comprehensive range of perspectives compared with single-assessor perspectives.[17] This review highlighted the inconsistent judgement acuity of assessors as paramount to competence assessment.[eg, P:1, 5,7,17,18,20,28,36,47,52,69,101] When evaluating instrument properties, the consideration of feasibility is paramount, as all measures require resources for systemic implementation. Furthermore, the number of individuals being assessed, the time constraints, financial resources, the interpretability, the validity and reliability of measures are key considerations. When reviewing these points, many experimental measures are unfeasible as routine IM due to resource availability and/or ethical and time constraints.

There was unanimity in the literature that competence assessment throughout a doctor's career is necessary to protect patients and the profession.[1] Therefore, continued exploration of evidence-based measures to assess competence throughout doctors' training and career is advocated. This review highlights the absence of consensus on the domains relevant to competence, and a lack of consistency in existing clinical measurement standards. In conclusion, the findings of this study strongly support the use of multidomain evidence-based measurement-selection, and attending to developing measures that replicate the unique systemic context through the evaluation of all domains relevant to competence is recommended as a continued focus of research in this area.

Contributors All authors developed the protocol for this systematic review. KLH conducted the screening of studies, data extraction and critical appraisal. DJL and AL reviewed each stage of study selection and calibrated findings. All authors assisted in the interpretation and write up of results. All authors approved the final manuscript prior to submission.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Institutional review board clearance is not required for this systematic review as all data are publicly available.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data availability statement Additional data are available upon reasonable request.

ORCID iDs
Kirsty L Hodgson http://orcid.org/0000-0002-9680-0445
Daniel J Lamport http://orcid.org/0000-0002-4592-0439
Allán Laville http://orcid.org/0000-0001-9678-9269

REFERENCES

1 Grace ES, Wenghofer EF, Korinek EJ. Predictors of physician performance on competence assessment. *Academic Medicine* 2014;89:912–9.
2 Giesler M, Forster J, Biller S, *et al*. Development of a questionnaire to assess medical competencies: reliability and validity of the questionnaire. *GMS Z Med Ausbild* 2011;28:31.
3 Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002;287:226–35.
4 Prediger S, Schick K, Fincke F, *et al*. Validation of a competence-based assessment of medical students' performance in the physician's role. *BMC Med Educ* 2020;20:6.

5 Lehmann F, Côté L, Bourque A, *et al*. Physician-patient interaction: a reliable and valid check-list of quality. *Can Fam Physician* 1990;36:1711–6.
6 Smith C, Likourezos A, Schiller J. Focused teaching improves medical student professionalism and data gathering skills in the emergency department. *Cureus* 2019;11:e5765.
7 Moher D, Liberati A, Tetzlaff J, *et al*. Reprint--preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Phys Ther* 2009;89:873–80.
8 Campbell M, McKenzie JE, Sowden A, *et al*. Synthesis without meta-analysis (swim) in systematic reviews: reporting guideline. *BMJ* 2020;368:16890.
9 Dixon-Woods M, Cavers D, Agarwal S, *et al*. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Med Res Methodol* 2006;6:35.
10 Burgess DJ, Dovidio J, Phelan S, *et al*. The effect of medical authoritarianism on physicians' treatment decisions and attitudes regarding chronic pain1,*. *J Appl Soc Psychol* 2011;41:1399–420.
11 Radabaugh CL, Hawkins RE, Welcher CM, *et al*. Beyond the United States medical licensing examination score: assessing competence for entering residency. *Acad Med* 2019;94:983–9.
12 Thomas MR, Beckman TJ, Mauck KF, *et al*. Group assessments of resident physicians improve reliability and decrease halo error. *J Gen Intern Med* 2011;26:759–64.
13 Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, *et al*. Twelve tips for programmatic assessment. *Med Teach* 2015;37:641–6.
14 Morgan PJ, Cleave-Hogg D. Comparison between medical students' experience, confidence and competence. *Med Educ* 2002;36:534–9.
15 Oudkerk Pool A, Govaerts MJB, Jaarsma DADC, *et al*. From aggregation to interpretation: how assessors judge complex data in a competency-based Portfolio. *Adv in Health Sci Educ* 2018;23:275–87.
16 Ginsburg S, McIlroy J, Oulanova O, *et al*. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Acad Med* 2010;85:780–6.
17 Joshi R, Ling FW, Jaeger J. Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. *Acad Med* 2004;79:458–63.