The validity and reliability of clinical judgement and decision-making skills assessment in nursing: A systematic literature review

ABSTRACT

Objectives: To appraise the validity and reliability of approaches to assessing the clinical decision-making skills of nurses, and use findings to inform the assessment of students as they transition to newly qualified nurses.

Design: The preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines were used to conduct the review.

Data sources: Medline, CINAHL and the British Nursing Index were searched from inception to November 2019.

Review methods: Studies were grouped according to their assessment approach following a competency framework with findings presented as a narrative synthesis.

Results: 38 articles were included in the review which assessed clinical decision-making in a variety of settings; clinical practice, simulation, written examinations and self-assessment. Multi-level rubric and checklist approaches demonstrated good validity and reliability in practice and simulation settings, and the former was effective at differentiating between students at different stages of their training. Written, case study examinations were also effective at assessing clinical decision-making, although an optimum structure for their presentation was not possible to discern. Students tended to score themselves more highly than faculty staff when undertaking rubric-based self-assessments.

Conclusions: Findings suggest that the best approach to assess clinical decision-making for final year students is to use several low-stakes, snap-shot summative assessments in practice environments, which are marked using a multi-level observational rubric. To assure reliability, it is recommended that a small team of expert practice assessors undergo regular training and peer review, have protected time to complete their assessor role and are appropriately supported.

Keywords

Clinical judgement

Clinical competence

Clinical decision-Making

Competency assessment

Nurses

1. Introduction

Patients' lives can depend on a nurse's ability to respond to clinical deterioration with competent decision-making skills (Banning, 2008; Thompson et al., 2009). Clinical decision-making skills represent an "evolving process, where data are gathered, interpreted, and evaluated in order to select an evidence-based choice of action" (Tiffen et al., 2014: p399). The skills identified for good clinical judgement and decision-making are recognised internationally as being fundamental to critical thinking within nursing practice (Scheffer and Rubenfeld, 2000) and reflect a standard expected of degree level graduates (Seec, 2016). However, there remains considerable variation in the quality of decisions that nurses make (Thompson et al., 2013) with some lacking the ability to clinically reason using a hypothesis driven approach to inform their practice (Andersson, Klang, and Petersson, 2012). This is evident when nurses judge whether critical events are likely to happen based on the same clinical information (Thompson et al., 2009), or in the management and care planning of a patient's functional status and self-caring abilities, and when delivering patient education (Doran et al., 2006).

Inconsistent clinical decision-making is of concern as nurses assume higher levels of responsibility and accountability for patient care in healthcare environments that are increasingly demanding and complex (Simmons, 2010; Chan, 2013). It is critical that providers of nurse education are able to determine whether a student nurse has met the standards of proficiency for registration, including their competency to make safe clinical decisions (Nursing and Midwifery Council [NMC] 2018). In 2013 Thompson et al. noted that measuring clinical judgement and decision-making are unanswered questions in clinical decision-making research. Since then there remains variation in how these competencies are assessed with no formal appraisal of these approaches.

1.1. Objectives

This literature review had two objectives :

- to appraise the validity and reliability of approaches to assessing the clinical decision-making skills of nurses, and;
- use findings to inform the assessment of students as they transition to newly qualified nurses.

2. Methods

This review was conducted in accordance with the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement (Moher et al., 2009).

2.1. Search strategy

Systematic strategies were developed to search across three bibliographic databases from their inception to November 2019: MEDLINE, Cumulative Index to Nursing and Allied Health Literature (CINAHL) and the British Nursing Index (BNI). Search strings were tailored to each database according to Medical Subject Heading (MeSH) and keywords, and were applied using the Boolean operators AND/OR. The key search terms across each database were 'clinical decision-making', 'competency assessment', 'validity/reliability' and 'nurses'. The reference lists of included articles were hand searched for potential information sources not captured through database searches. An example of the search strategy for MEDLINE is presented in Table 1, which returned 45 citations.

Table 1 MEDLINE search strategy

clinical decision-making.mp. OR exp decision-making/ OR exp clinical decision-making/ OR clinical judgement.mp OR clinical reasoning.mp.
AND
exp clinical competence/ OR competency assessment.mp.
AND
nurse.mp. OR exp nurses/
AND
alidity.mp. OR exp social validity research/ OR exp reproducibility of results/ OR exp psychometrics/ OR reliability.mp. OR exp psychometrics/ OR psychometric testing.mp.

2.2. Eligibility criteria

Included studies were those published in the English language that examined nurse or student nurse-patient interactions, and which reported the validity, reliability or psychometric properties of their assessment tools. Studies were excluded if they did not use a competency assessment framework, or if used, the framework was not related to nurse-patient interactions. Studies that applied a competency framework to assess discrete skills, tasks or patient conditions without more general applicability were also excluded.

The title and abstracts of retrieved papers were screened for eligibility by two reviewers (XX, XX [blinded for peer review]). After removing duplicates, the full texts of all potentially relevant articles were read independently by both reviewers with any inclusion uncertainties resolved through discussion. Conference abstracts without a title were excluded from the review. However, when PhD

topics appeared relevant to the research aim, their authors were searched online to determine whether any findings were available in the public domain.

2.3. Quality assessment

The Critical Appraisal Skills Programme (CASP) 12-item tool for diagnostic tests was adapted to appraise the quality of included studies (CASP, 2018). Four items not relevant to this review's context were removed e.g. whether the disease status of the tested population had been clearly described. The wording of two items was changed minimally to more accurately reflect the review context e.g. the substitution of 'test' with 'assessment', and the remaining six items were unchanged. To ascertain the quality of eligible studies, two reviewers (XX, XX) independently rated the articles with any disagreements resolved through discussion. The two items that specify use of an assessment reference standard and sufficient description of the assessment method were given more weighting in this validity and reliability review. Articles that met both these criteria and overall had a good quality profile were considered low risk of bias.

2.4. Data extraction and synthesis

Tabulation and grouping techniques guided data extraction and a narrative synthesis. A template was used to extract key methodological detail for each paper including the assessment tool used, setting, participants, and key findings in terms of validity and reliability. Because the included studies featured heterogeneous methods and contexts a meta-analysis could not be conducted. As such, studies were grouped according to their assessment method following Miller's pyramid of competence (Miller, 1990):

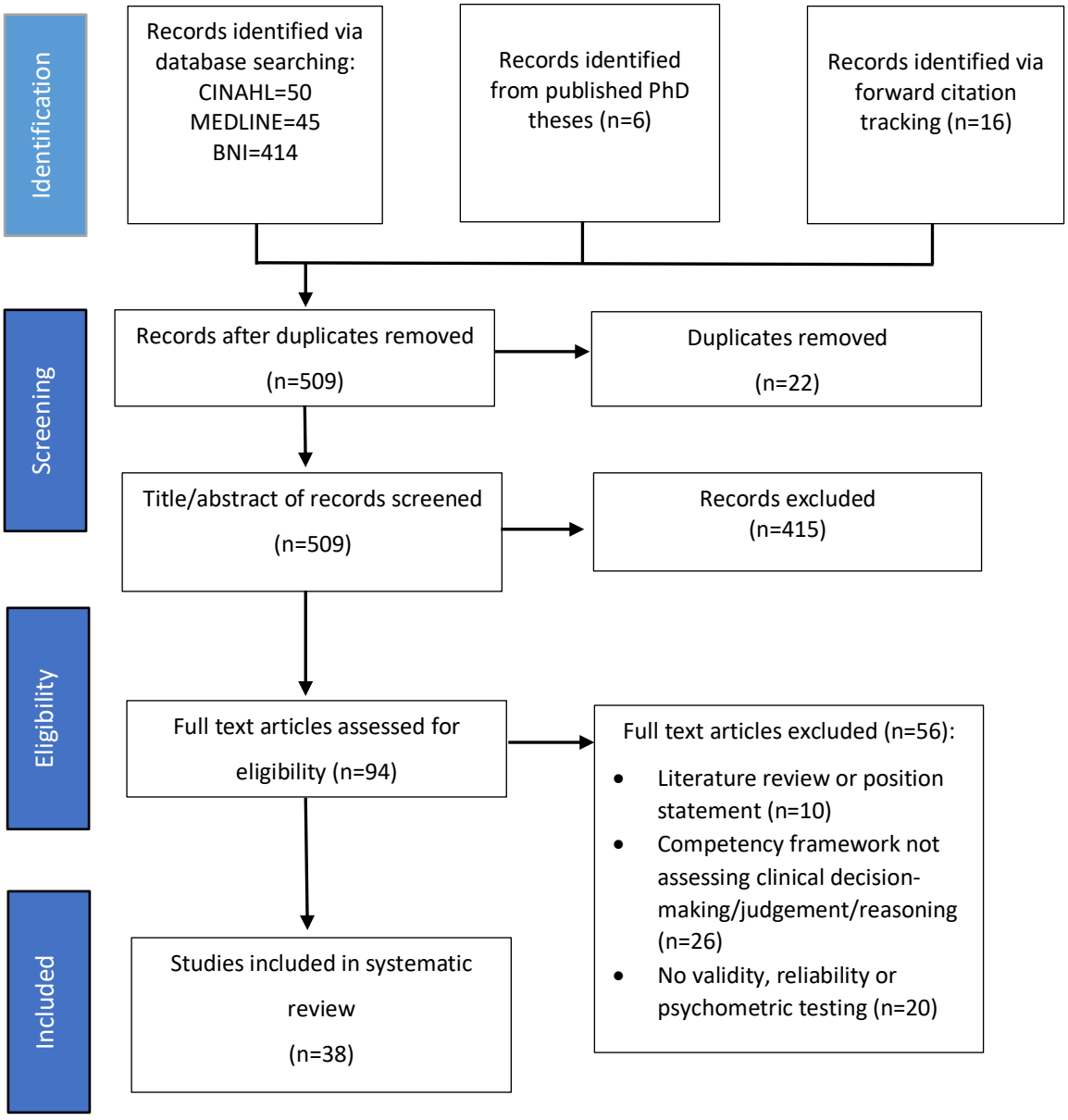
- Does: assessment of direct patient care;
- Shows how: assessment in simulation;
- Knows how: oral and written examinations.

These stages of competency development relate to nurses' knowing, functioning and behaviours (know-how, show how and does), thus enabling assessment of blended theory and practice; components required of nursing programmes internationally (World Health Organisation, 2009). Information not captured by Miller's (1990) competency framework was inductively grouped into key themes of interest. A synthesis of competency findings and additional themes allowed for a meaningful narrative of evidence to meet the review's objectives. Data were extracted and thematically grouped by one reviewer (XX), with verification undertaken by the second reviewer (XX).

3. Results

The search strategy returned a total of 509 records, from which 38 studies were included in the review after title/abstract screening and application of inclusion/exclusion criteria (Figure 1).

Figure 1 PRISMA flow diagram (Moher et al., 2009)



3.1. Study characteristics

Table 2 summarises the characteristics of the 38 included studies which used a range of assessment tools in a variety of contexts including nurse training facilities and hospitals. The majority of papers were from the United States of America (n=21), two each were from Egypt, Canada, Taiwan, Singapore, South Korea and Australia, and one each from the United Kingdom, the

Netherlands, Malaysia, Sweden and Japan. Study samples included nursing students on Bachelor of Nursing Courses (BSN), Advanced Diploma in Nursing course (ADN) and Master of Nursing Courses (MN). Seven studies assessed clinical decision-making in practice settings, 22 undertook assessments in simulated settings (using real time assessments, recordings of practical simulations and/or virtual simulations), 10 reported written examinations and nine considered nurses' own perception of their clinical decision-making. Among this total of 48, eight studies assessed students in two different settings, and one took place over three settings (n=38 studies).

3.2. Quality appraisal

CASP quality appraisals are presented in Table 3. Eight studies met both defined criteria on reference standards and method detail, and were considered low risk of bias (Adamson & Kardong-Edgren, 2012; Ball & Kilger, 2016; Gorton & Hayes, 2014; Liaw et al., 2018; Liou et al., 2016a; Prion et al., 2015; Selim et al., 2012; Vreugdenhil & Spek, 2018). Many other studies provided insufficient information to adequately rate their quality and were considered uncertain risk of bias (n=25). The remaining five studies were considered high risk of bias because they met only one or none of the criteria (Murcott & Clarke, 2017; Randolph et al., 2012; Reyes & Rodriguez, 2016; Robins & Hoke, 2008; Starkweather et al., 2017).

3.3. Does: practical assessments of direct patient care

The seven studies conducted in practice settings used a variety of approaches: numeric scales, checklists, multi-level rubric, or a clinical observation and oral viva (SOAP approach). The 1-10 scale poorly differentiated between students (Gorton & Hayes, 2014), whereas observational multi-level rubrics were able to effectively correlate performance based on practice assessors' feedback with clinical placement exposure (Vreugdenhil & Spek, 2018; $r=0.62$, $p<0.001$) and level of study (Prion et al., 2017; $r=0.83$, $p<0.05$). Well-defined rubric descriptors set out student expectations, minimised subjective bias between assessors and promoted student self-assessment (Vreugdenhil & Spek, 2018). Multi-level rubrics were perceived to be objective by staff (Nielsen et al., 2016) and had a good level of agreement between nursing faculty and practice coaches (Vreugdenhil & Spek, 2018). There appeared to be most support for using the Lasater Clinical Judgement Rubric (LCJR), which demonstrated good psychometric properties, with one low risk of bias paper supporting its use (Vreugdenhil & Spek, 2018). However, being an exclusively observational assessment tool, it may not be sensitive enough to assess students who work with complex patients in their final year of study (Vreugdenhil & Spek, 2018) and there is some criticism that it is too lengthy and cumbersome to be used in clinical practice (Prion et al., 2017). The majority of students also perceived alternative

methodologies, such as the subjective, objective, assessment, plan (SOAP) model, to be reflective of their clinical practice (Levett-Jones et al., 2011).

3.4. Shows how: practical assessments within simulation settings

The 22 studies conducted in the simulation setting used a variety of techniques: multi-level rubrics, numeric scales or checklists approaches. Among these 22 simulation studies, one used virtual reality (Georg et al., 2018), indicating the potential to modify assessment strategies for such use, although this was underexplored. When compared directly, there were no significant differences between rubric and checklist approaches to assessment (Adamson & Kardong-Edgren, 2012), with scores significantly correlated (Liaw et al., 2018) and both approaches showed good validity across studies (Table 2). However, there was some indication that checklists may be less able to differentiate between students when they are marked against areas easy for them to demonstrate for their level of study compared to rubrics (Randolph et al., 2012). Two studies with a low risk of bias considered the checklist approach but they did not attempt to discriminate between the mark awarded and stage of training however analysis of the C-SEI (Checklist) found that 38 assessors could differentiate between students' performance at three different levels (Adamson & Kardong-Edgren, 2012).

Rubric approaches could generally differentiate between students at different stages of training: 1st and 2nd year students (Ball & Kilger, 2016; Prion et al., 2017), 2nd and 3rd year students (Liaw et al., 2018) or further apart in their careers. However, none directly considered the final year student with a newly qualified nurse. The CREST tool used by Liaw et al (2018) also explored students' underlying thought processes alongside a multi-level observational rubric that scored the rationale for clinical decisions, which enabled differentiation between 2nd and 3rd year students. This may be a useful measure of student performance in complex clinical situations, although analysis of the individual sections of the CREST model indicated areas where the examiner asked students to explain their thought process, which did not differentiate between 2nd and 3rd year students but observing the student's actions in the scenario did (Liaw et al., 2018) This suggests it was no more useful than on observational rubric alone.

3.5. Knows how: oral, written and online assessments

The 10 studies that assessed examinations used a variety of techniques: script concordance test, accuracy of clinical diagnosis, and short answer responses using either rubric marking schemes or criterion marking. None of the articles considered single best answer multiple choice questions (MCQs) or written coursework assessments, which are popular in nurse training programmes.

Responses to short answer and workbook questions in relation to unfolding cases-studies within an exam situation were considered in five studies (Fenske et al., 2013; Lasater et al., 2015; Liou et al., 2016b; O'Rourke & Zerwic, 2016; Reyes and Rodriguez, 2016; Selim & Dawood, 2015), and where reported, clinical experience and level of training impacted student performance (Fenske at al., 2013; Lasater et al., 2015; Liou et al., 2016b). Additionally, one study found that a third of nurses who scored poorly on their unfolding case studies raised concerns for managers at 9-month follow-up (Lasater et al., 2015). This demonstrates the link between performance in unfolding case studies and clinical practice experience. However, there was no indication of which method was better to structure marking for unfolding case study assessments.

A script concordance test (SCT) was evaluated in two papers (Dawson et al, 2014; Deschênes et al, 2011) where students indicated on a Likert scale whether the preceding information impacted on their interpretation of patient data or plan of nursing care. Answers to the same questions were obtained from an 'expert panel' of qualified nurses, awarding one mark for the modal responses and a partial credit for other responses (Dawson et al, 2014; Deschênes et al, 2011). It was argued that introducing SCT early in nurse education helps students develop mental scripts that can be expanded to promote the development of hypo-deductive reasoning processes, with consistencies in answers given by the expert panel compared to first year nursing students (Dawson et al, 2014; Deschênes et al, 2011). However, the validity of the test is based on the competence of nurses on the 'expert panel' and further evaluation is required.

Two articles examined the ability of qualified nurses to formulate appropriate differential and nursing diagnoses on information provided in case studies (Gorton and Hayes, 2014; Hasegawa et al., 2007). Their diagnoses related to levels of clinical experience and clinical decision-making responsibilities (Hasegawa et al, 2007).

3.6. Self-assessment

Eight articles considered self-assessment by asking individuals to self-assess aspects of clinical decision-making on a Likert scale using a tool such as the CDMNS (Gorton & Hayes, 2014; Ludin, 2018; Liou et al, 2016a), or by a snapshot evaluation to self-evaluate a specific episode of care using a rubric tool such as the LCJR (Fenske et al., 2013; Jensen, 2013; Shin et al., 2015; Strickland et al., 2017; Vreugdenhil & Spek, 2018). Findings indicated that Likert scores effectively differentiated between nurses with different levels of experience (Liou et al., 2016a; Ludin, 2018). Students' perception of their decision-making behaviours remained stable for up to a month (Liou et al., 2016a; Ludin, 2018) and enabled students to identify their weaknesses. When using a rubric approach for a snapshot of care students tended to score themselves higher than faculty ratings

(Jensen, 2013; Strickland et al., 2017; Vreugdenhil & Spek, 2018) and were generally less effective at differentiating levels of performance for those with less experience (Jensen, 2013; Strickland et al., 2017).

3.7. Resource and time implications

Time and resource implications were evaluated in direct patient care and simulation settings. For assessments of direct patient care the student was required to work 1-to-1 with their assessor ranging from six hours when the SOAP model was used (Levett-Jones et al., 2011), or one morning to observe two students using the LCJR (Vreugdenhil & Spek, 2018). Both methods used an assessment of directly observed episode(s) of care. Findings showed that the SECC-35 was quicker to complete (10 minutes) but utilised the assessor's perceptions of achievement over a longer period rather than a structured one-off assessment (Prion et al., 2015). The usability of clinical grading tools to assess nurses in clinical practice was perceived as more challenging in comparison to use in the simulated environments because of the time required (Hayden et al., 2014). However, this was only compared in one study.

3.8. Consistency of assessment

One study considered the inter-rater reliability of data in practice settings, which found consistent LCJR ratings between nurses trained to mentor students and nurse educators (Vreugdenhil & Spek, 2018). This was more widely researched in simulation settings with positive results reported for test-retest reliability and inter-rater-reliability when multi-level rubrics were used (Adamson & Kardong-Edgren, 2012; Adamson et al., 2012; Liaw et al., 2018; Prion et al., 2017; Shin et al., 2014), or when clear descriptors of performance were checked off (Adamson & Kardong-Edgren, 2012; Liaw et al., 2018; Randolph et al., 2012; Starkweather et al., 2017). There was an indication that the checklist approach had marginally better inter-rater reliability, though both had similar test-retest reliability (Table 2). Assessor consistency on a multi-level rubric did not appear to be influenced by subjective biases (Adamson, 2016), and intra-class correlation for the LCJR of 0.908 compared to 0.883 for the CCEI checklist when the same assessor watched the same student at a different time points (Adamson & Kardong-Edgren, 2012).

Within the examination setting, marking of case study driven examinations also had excellent inter-rater reliability where reported (Lasater et al, 2015; O'Rourke & Zerwic, 2016) and statistically significant correlations when the same student was reassessed at a later time point (O'Rourke & Zerwic, 2016).

3.9. One-off or ongoing assessment

Assessors perceived one-off assessments of an episode of care would not be effective at evaluating students' overall performance in practice (Hayden et al., 2014) but students and staff perceived snap-shot practice assessment and feedback as positive to their learning and professional development (Levett-Jones et al., 2011; Nielson et al., 2016). However, no studies in the review directly compared the use of one-off to ongoing assessment or recommended an ideal duration of a snap-shot assessment. The utilisation of a standardised assessment framework in clinical practice of direct patient care was believed to track nurses' performance on qualifying (Nielsen et al., 2016) and develop student nurses' and newly qualified nurses' clinical judgement skills (Levett-Jones et al., 2011; Nielsen et al., 2016) and therefore may support the utilisation of multiple assessment points.

4. Discussion

This review has considered a variety of assessment strategies to determine clinical decision-making, which cover all aspects of Miller's pyramid of competence (1990). Findings indicate that multi-level rubrics can provide valid data on students' performance in both simulation and direct-patient-care settings, and demonstrate the performance of professional skills that are essential components of nursing programmes (Pitt et al., 2012). For summative practice assessments, well-defined rubric descriptors can set out student expectations, minimise subjective bias between assessors and promote student self-assessment (Vreugdenhil & Spek, 2018). Direct-patient-care environments are an ideal context for developing student's clinical judgement skills because the assessor can consider a variety of real-world patient scenarios, and evaluate students' holistic approach to patient care and associated professional skills (Wu et al., 2015). These rubrics also reduced bias and increased objectivity in the simulation setting (Adamson & Kardong-Edgren, 2012; Adamson et al., 2012; Liaw et al., 2018; Prion et al., 2017; Shin et al., 2014).

Providing feedback in multi-level rubric assessments had a positive effect on the development of learning and clinical judgement (Levett-Jones et al., 2011; Nielsen et al., 2016), specifically when a common language was used to provide feedback (Nielson et al., 2016). This is supported by the wider literature in which Hughes et al. (2019) reported that 87.9% of practice assessors and mentors found student performance improved following feedback. These findings provide evidence to recommend the use of multi-level rubrics for the assessment of students as they transition between their final year and qualified practice. However, a number of caveats were raised in the review.

Measures currently used to assess student performance in clinical practice are non-criterion referenced (Wu et al., 2015). Promoting critical thinking through meaningful feedback is also dependent on a supportive environment to bridge theory and practice (Henderson & Eaton 2013;

Kaddoura, 2013). However, nurses in practice may not be adequately trained in the delivery of feedback (Wu et al., 2015) which can unwittingly affect student performance. The appropriate training and support of assessors is therefore of paramount importance, particularly as assessments are known to be influenced by subjective biases (Helminen et al., 2016) and the lenience of assessors (Daly et al., 2017). Even when training is offered, one study found that 44.8% of practice assessors did not believe there were enough safeguards in place to ensure consistency in marking and moderation processes (Hughes et al., 2019). To address this, moderation or peer review of assessors is required to ensure the reliability of assessments. From this perspective, assessments within simulation settings may provide more safeguards as student assessments can be recorded.

Vreugdenhil & Spek (2018) also raised the need for further development of multi-level observational rubrics to assess more complex decisions and to assess the transition from student nurse to registered practice. However, when Liaw et al. (2018) included student questioning to explain thought processes, they found it was no more useful than an observational rubric alone. Nevertheless, the utilization of assessment rubrics may be better suited to assign summative grades to final year students than a checklist approach, especially when achieved / not achieved criteria are given for areas that can be easily achieved for their level of experience (Randolph et al., 2012). Since qualified nurses require clinical judgement and reasoning skills to navigate complex and unpredictable healthcare environments (Johansen & O'Brien, 2016) an assessment that is able to demonstrate these skills, such as a multi-level rubric, is necessary to support their transition to registered practice.

Snap-shot practice assessments and feedback were positive to professional development (Levett-Jones et al., 2011; Nielson et al., 2016). However, previous research indicates one-off practical assessments encourage students to focus on their need to pass rather than seeking feedback to improve practice (Harrison et al., 2014) and may not be representative of overall clinical performance due to the stress and anxiety associated with this type of assessment (Wu et al., 2015). Thus, a series of assessments with rich qualitative feedback are necessary to enable deeper learning to occur (Harrison and Wass, 2016).

Given the resource intense nature of practice-based and simulated assessments, more traditional approaches including written exams were an important feature of the review. Findings indicated that written, case study driven examinations were effective at assessing clinical decision-making (Fenske at al., 2013; Lasater et al., 2015; Liou et al., 2016b). However, it was not clear how best to write and mark case studies, although there was support for the use of a rubric to mark short answer responses to unfolding case studies (Fenske et al., 2013; Lasater et al., 2015; O'Rourke, & Zerwic, 2016; Selim & Dawood, 2015). The potential for short answer case study responses to

predict clinical performance at 9-month follow-up (Lasater et al., 2015) indicates their potential for assessing students' transition to registered practice. Studies by Lasater et al. (2015) and O'Rourke & Zerwic (2016) also demonstrated that the inter-rater reliability of multilevel rubrics for marking purposes was excellent. However, written case study-based examinations do not capitalise on the positive impact clinical practice has on professional development and associated opportunities to evaluate a student's holistic approach to patient care (Wu et al., 2015). Therefore, this assessment strategy should be supplemented with some form of practice-based learning.

The review identified that self-assessment of clinical judgement and decision-making skills was helpful to students (Gorton & Hayes, 2014; Ludin, 2018; Liou et al., 2016a) and enabled them to reflect on their care experiences to identify learning needs (Jenson, 2013). This approach is widely used in undergraduate and postgraduate nurse education, with self-reflection having a positive effect on longitudinal learning in terms of critical thinking, performance and communication skills (Kim et al., 2018). Thus, self-assessment tools could be used as a catalyst for reflection on-action rather than summative assessments as students overinflate their own performance (Jensen, 2013; Strickland et al., 2017; Vreugdenhil & Spek, 2018). However, the assessment of written reflections of student's decision making based on their self-assessment was not covered in any of the included studies.

Interestingly, none of the reviewed articles considered single best answer multiple choice questions (MCQs) or written coursework, both of which are traditionally used in nurse education. The exclusion of MCQs and written coursework to assesses clinical reasoning and decision-making skills is not a new finding. A study by DulBeno (2005) found that over half of newly qualified nurses who entered onto the nursing register using this method did not meet the expected standard when multiple choice options of nursing care were not provided. An inability to recognise subtle changes in a patient's condition in the real world is thought to be related to the theory-practice gap, with nurse educators preparing students to pass their MCQ examinations rather than preparing them for clinical practice (Huston et al., 2018). However, in medicine single best answer MCQs, when developed involving a panel of experts, are considered reliable, credible, and cost-effective form of assessment (Okubuiro et al., 2019). However, this review cannot support using single best answer MCQs alone to assess nursing students' clinical decision-making.

4.1. Limitations

This review focused on assessing clinical decision-making and did not include holistic assessments of students' competence. The review also omitted the validity and reliability of

assessment rubric tools used in medical education and other health disciples, which may have provided alternative frameworks to assess nurses' clinical decision-making ability.

The overall risk of bias associated with the evidence in this review was appraised as uncertain. Many of the studies provided insufficient information to adequately judge their quality. However, eight of the 38 papers were appraised as low risk of bias and provided a small evidence base upon which to forge conclusions. More robust research that comprehensively describes the methodologies used, and which incorporates reference standards for comparison purposes, is needed to develop knowledge in this field.

5. Conclusion

A range of valid and reliable methods to assess clinical decision-making have been identified, although existing strategies need further development to ensure they fully assess student management of complex situations as they make the transition to registered practice. Findings suggest that the best approach to assess clinical decision-making for final year students is to use several low-stakes, snap-shot summative assessments in direct patient care and simulated practice, which are marked using a multi-level observational rubric, such as the LCJR. To ensure such assessment are reliable, it is recommended that a small team of expert practice assessors work collaboratively across academic and practice environments, undergo regular training and peer review, have protected time to complete their assessor role and are appropriately supported. This approach would ensure students they have the appropriate clinical reasoning skills on qualifying to consistently, safely and holistically manage evolving patients' clinical situations, and prevent unwanted variations in practice which have a negative impact on patient safety and the delivery of effective evidence-based care. Where educational resources may be limited, an alternative valid and reliable assessment strategy for transitioning students would be a written short answer examination of unfolding case scenarios, which require the nurse to clinically reason in complex decision-making situations, supplemented by self-assessments of clinical performance.

References

Adamson K.A., Kardong-Edgren S., (2012) A method and resources for assessing the reliability of simulation evaluation instruments. Nursing Education Perspectives. 33 (5), 334-9.

Adamson K.A., Gubrud P., Sideras S., Lasater K., (2012) Assessing the Reliability, Validity, and Use of the Lasater Clinical Judgment Rubric: Three Approaches Journal of Nursing Education. 51 (2), 66-73.

Adamson K.A., (2016) Rater Bias in Simulation Performance Assessment: Examining the Effect of Participant Race/Ethnicity. Nursing Education Perspectives. 37 (2), 78-82.

Andersson N., Klang B., Petersson G., (2012). Differences in clinical reasoning among nurses working in highly specialised paediatric care. Journal of Clinical Nursing. 21 (5-6), 870-9.

Ball L.S., Kilger L., (2016) Analysing Nursing Student Learning Over Time in Simulation. Nursing Education Perspectives. 37 (6), 328-330.

Banning M., (2008). A review of clinical decision making: models and current research. Journal of Clinical Nursing. 17 (2), 187-195.

Bujack L., McMillan M., Dwyer J., Hazelton M., (1991) Assessing comprehensive nursing performance: the objective structured clinical assessment (OSCA). Part 2--Report of the evaluation project. Nurse Education Today. 11(4), 248-55.

Critical Appraisal Skills Programme (2018). CASP - Diagnostic Test Study Checklist. [Online] Available at: https://casp-uk.net/wp-content/uploads/2018/01/CASP-Diagnostic-Checklist.pdf. Accessed: Date Accessed. 13/04/2020

Chan Z., (2013) A systematic review of critical thinking in nursing education. Nurse Education Today 33 (3), 236–240.

Daly M., Salamonson Y., Glew P.J., Everett B (2017) Hawks and doves: The influence of nurse assessor stringency and leniency on pass grades in clinical skills assessments Collegian 24 (5) 449-454

Dawson T., Comer L., Kossick M.A., Neubrander J., (2014) Can SCRIPT Concordance testing be used in nursing education to accurately assess clinical reasoning skills? Journal of Nurse Education 53(5), 281-6.

Deschênes M.F., Charlin B., Gagnon R., Goudreau J., (2011) Use of a script concordance test to assess development of clinical reasoning in nursing students. Journal of Nurse Education. 50(7), 381-7.

Doran, D., Harrison, M.B., Laschinger, H., Hirdes, J., Rukholm, E., Sidani, S., Hall, L.M., Tourangeau, A.E., Cranley, L., (2006) Relationship between nursing interventions and outcome achievement in acute care settings. Research in Nursing & Health. 29 (1), 61–70.

Del Bueno D., (2005) A crisis in critical thinking. Nurse Education Perspectives. 26(5), 278-82.

Fenske C.L., Harris M.A., Aebersold M.L., Hartman L.S., (2013) Perception versus reality: a comparative study of the clinical judgment skills of nurses during a simulated activity. Journal of Continuing Education in Nursing. 44(9), 399-405.

Gantt L.T., (2010) Using the Clark Simulation Evaluation Rubric with associate degree and baccalaureate nursing students. Nurse Education Perspectives. 31(2), 101-5.

Georg C., Karlgren K., Ulfvarson J., Jirwe M., Welin E., (2018) A Rubric to Assess Students' Clinical Reasoning When Encountering Virtual Patients. The Journal of Nursing Education. 57(7), 408-415.

Gorton K.L., Hayes J., (2014) Challenges of assessing critical thinking and clinical judgement in nurse practitioner students. Journal of Nursing Education. 53(3), S26-9.

Harrison C., Wass V., (2016) The challenge of changing to an assessment for learning culture. Medical Education. 50(7), 704-6.

Harrison C.J., Könings K.D., Schuwirth L., Wass V., Van der Vleuten C., (2014) Barriers to the uptake and use of feedback in the context of summative assessment . *Advances in Health Sciences Education.* 20(1), 229-45.

Hasegawa T., Ogasawara C., Katz E.C., (2007) Measuring Diagnostic Competency and the Analysis of Factors Influencing Competency Using Written Case Studies. International Journal of Nursing Terminology and Classification. 18(3), 93-102.

Hayden J., Keegan M., Kardong-Edgren S., Smiley R.A., (2014) Reliability and validity testing of the Creighton Competency Evaluation Instrument for use in the NCSBN national simulation study. Nursing Education Perspectives. 35(4), 244-52.

Helminen K., Coco K., Johnson M., Turunen H., Tossavainen K., (2016) Summative assessment of clinical practice of student nurses: A review of the literature. International Journal of Nursing Studies. 53 (1), 308-19.

Henderson A., Eaton E., (2013) Assisting nurses to facilitate student and new graduate learning in practice settings: what 'support' do nurses at the bedside need? Nurse Education in Practice. 13, 197-201.

Hughes L.J, Mitchell M.L, Johnston A.N.B., (2019) Just how bad does it have to be? Industry and academic assessors' experiences of failing to fail - A descriptive study. Nurse Education Today. 76 (1), 206-215.

Huston C.L., Phillips B., Jeffries P., Todero C., Rich J., Knecht P., Sommer S., Lewis M.P., (2018). The academic-practice gap: Strategies for an enduring problem. Nursing Forum. 53(1), 27-34.

Jensen R., (2013). Clinical reasoning during simulation: comparison of student and faculty ratings. Nurse Education in Practice. 13(1), 23-8.

Johansen M.L., O'Brien J.L. (2016) Decision Making in Nursing Practice: A Concept Analysis. Nursing Forum. 51(1), 40-8

Kaddoura M., (2013) The effect of preceptor behaviour on the critical thinking skills of new graduate nurses in the intensive care unit. Journal of Continuing Education. Nurse 44 (11), 488-495

Kim, Y.H., Min, J., Kim, S.H. *et al.* Effects of a work-based critical reflection program for novice nurses. *BMC Medical Education.* 18, 30 (2018). https://doi.org/10.1186/s12909-018-1135-0

Lasater K., Nielsen A.E., Stock M., Ostrogorsky T.L., (2015) Evaluating the clinical judgement of newly hired staff nurses. Journal of Continuing Education in Nursing. 46(12), 563-71.

Levett-Jones T., Gersbach J., Arthur C., Roche J., (2011) Implementing a clinical competency assessment model that promotes critical reflection and ensures nursing graduates' readiness for professional practice. Nurse Education in Practice. 11(1), 64-9.

Liaw S.Y., Scherpbier A., Klainin-Yobas P., Rethans J.J., (2011) Rescuing A Patient In Deteriorating Situations (RAPIDS): An evaluation tool for assessing simulation performance on clinical deterioration. Resuscitation. 82(11), 1434-9.

Liaw S.Y., Rashasegaran A., Wong L.F., Deneen C.C., Cooper S., Levett-Jones T., Goh H.S., Ignacio J., (2018) Development and psychometric testing of a Clinical Reasoning Evaluation Simulation Tool (CREST) for assessing nursing students' abilities to recognize and respond to clinical deterioration. Nurse Education Today. 62 (3), 74-79.

Liou S.R., Liu H.C., Tsai H.M., Tsai Y.H., Lin Y.C., Chang C.H., Cheng C.Y.,(2016a) The development and psychometric testing of a theory-based instrument to evaluate nurses' perception of clinical reasoning competence. Journal of Advanced Nursing. 72(3), 707-17.

Liou S.R., Liu H.C., Tsai S.L., Cheng C.Y., Yu W.C., Chu T.P., (2016b). Development of the Computerized Model of Performance-Based Measurement System to Measure Nurses' Clinical Competence. Computers Informatics, Nursing. 34(4), 159-68.

Ludin S.M., (2018) Does good critical thinking equal effective decision-making among critical care nurses? A cross-sectional survey. Intensive and Critical Care Nursing. 44 (2), 1-10.

Miller E.G., (1990) The assessment of clinical skills/competence/performance. Academic Medicine. 65 (9), s63-67. DOI: 10.1097/00001888-199009000-00045

Moher D., Liberati A., Tetzlaff J., Altman D. and the PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine* 151 (4), 264-269.

Murcott, W., Clarke, N., (2017) Objective structured clinical exam: a successful approach to pre-registration mental health nurse assessment. The Journal of Mental Health Training, Education and Practice. 12 (2), 90-97.

Nielsen A., Lasater K., Stock M., (2016) A framework to support preceptors' evaluation and development of new nurses' clinical judgment. Nurse Education in Practice. 19:84-90

Nursing and Midwifery Council (2018) Future nurse: Standards of proficiency for registered nurses. Available at: https://www.nmc.org.uk/globalassets/sitedocuments/education-standards/future-nurse-proficiencies.pdf (accessed 15 April 2020)

Okubuiro E.O., Ebirim L. N., Okoli C.E (2019) Utility of Single Best Answer Questions as a Summative Assessment Tool in Medical Education: A review. International Journal of Recent Innovations in Academic Research. 3 (1), 1-12.

O'Rourke J., Zerwic J., (2016) Measure of Clinical Decision-Making Abilities of Nurse Practitioner Students. Journal of Nursing Education. 55(1), 18-23.

Park H., Park J., Kim C., Song J., (2017) Development and validation of simulation teaching strategies in an integrated nursing practicum. Collegian. 24 (5), 479–486.

Pitt V, Powis D, Levett-Jones T, Hunter S. (2012) Factors influencing nursing students' academic and clinical performance and attrition: an integrative literature review. Nurse Educ Today. 2012 Nov;32(8):903-13.

Prion S., Berman A., Karshmer J., Van P., Wallace J., West N., (2015) Preceptor and self-evaluation competencies among new RN graduates. Journal of Continuing Education in Nursing. 46(7), 303-8.

Prion, S. K., Gilbert, G. E., Adamson, K. A., Kardong-Edgren, S., Quint, S., (2017). Development and testing of the Quint Leveled Clinical Competency Tool. Clinical Simulation in Nursing. 13(3), 106-115.

Randolph P.K., Hinton J.E., Hagler D., Mays M.Z., Kastenbaum B., Brooks R., DeFalco N., Miller K., Weberg D., (2012) Measuring competence : Collaboration for safety. Journal of Continuing Education in Nursing. 43 (12), 541-7.

Reyes I., Rodriguez J, (2016) Conducting Objective Structured Clinical Exams in a Pediatric Nurse Practitioner Program Using Google Tools. The Journal for Nurse Practitioners. 12 (8), 566-573.

Robbins L.K., Hoke M.M., (2008) Using Objective Structured Clinical Examinations to meet Clinical Competence Evaluation Challenges with Distance Education Students. Perspectives in Psychiatric Care. 44 (2), 81-88.

Scheffer B.K., Rubenfeld M.G., (2000) A consensus statement on critical thinking in nursing. Journal of Nursing Education, 39, 352-359.

Seec (2016) Credit level descriptors for Higher Education. [Online] Available at: https://seec.org.uk/

Selim A.A., Dawood E., (2015) Objective Structured Video Examination in Psychiatric and Mental Health Nursing: A Learning and Assessment Method. Journal of Nursing Education. 54 (2), 87-95.

Selim A.A., Ramadan F.H., El-Gueneidy M.M., Gaafer M.M., (2012) Using objective structured clinical examination (OSCE) in undergraduate psychiatric nursing education is it reliable and valid? Nurse Education Today. 32(3), 283-8.

Shin H., Shim K., Lee Y., Quinn L., (2014) Validation of a new assessment tool for a pediatric nursing simulation module. Journal of Nursing Education. 53(11), 623-9.

Shin H., Park C.G., Shim K., (2015) The Korean version of the Lasater Clinical Judgment Rubric : A validation study. Nurse Education Today. 35 (1), 68-72.

Simmons B., (2010) Clinical Reasoning: Concept Analysis. Journal of Advanced Nursing. **65** (5), 1151-1158.

Stacey D., Taljaard M., Drake E.R., O'Connor A.M., (2008) Audit and feedback using the brief Decision Support Analysis Tool (DSAT-10) to evaluate nurse–standardized patient. Encounters. Patient Education and Counselling. 73(3), 519-25.

Starkweather A., Sargent L., Nye C., Albrecht T., Cloutier R., Foster A., (2017) Progressive Assessment and Competency Evaluation Framework for Integrating Simulation in Nurse Practitioner Education. Journal for Nurse Practitioners. 13(7), e301-e310.

Strickland H.P., Cheshire M.H., March A.L., (2017) Clinical Judgment during Simulation: A Comparison of Student and Faculty Scores. Nursing Education Perspectives. 38(2), 85-86.

Thompson C., Bucknall T., Estabrooks C., Hutchinson A., Fraser K., De Vos R., Binnecade J., Barrett G., Saunders J. (2009) Nurses' critical event risk assessments : a judgement analysis. Journal of Clinical Nursing. 18 (4), 601–612.

Thompson C., Aitken L., Doran D., Dowding D., (2013) An agenda for clinical Decision-making and judgement in nursing research and education. International Journal of Nursing Studies. 50 (12), 1720-6.

Tiffen J., Corbridge S.J., Slimmer L., (2014) Enhancing clinical Decision-making: development of a contiguous definition and conceptual framework. Journal of Professional Nursing. 30 (5), 399-405.

Vreugdenhil J., Spek B., (2018) Development and validation of Dutch version of Lasater clinical judgment rubric in hospital practice an instrument design study. Nurse Education Today. 62, 43-51.

World Health Organisation (2009) Global Standards for the initial education of professional nurses and midwives [available online] http://www.who.int/hrh/nursing_midwifery/en/ *[Last accessed 30/01/2020]*

Wu V.X., Enskär K., Lee C.C.S.,Wang W., (2015) A systematic review of clinical assessment for undergraduate nursing students. Nurse Education Today 35(2), 347–359.

Table 2: Study characteristics

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
1	Adamson (2016), USA	Lasater Clinical Judgement Rubric (LCJR)*: Rubric*	Simulation (video)	Randomised Control Trial assessing impact of ethnicity and gender on marking	Simulation assessors (*n*=68)	One of four randomly assigned videos. Same 'script' with different ethnic/gendered actors.	**Content:** Based on theoretical framework **Concurrent:** Non-significant differences in assessor scores	**Inter-rater:** Non-significant differences in assessor scores
2	Adamson and Kardong-Edgren (2012), USA	LCJR: *Rubric* Seattle University Evaluation Tool (SUET) *: 0-5 for competency* Creighton-Simulation Evaluation Instrument (C-SEI) *: Checklist*	Simulation (video)	Single blind experimental study assessing videos of different abilities using three different techniques.	Simulation assessors (*n*=29 to 38 depending on the number that fully completed the study)	Six videos of different levels of nursing proficiency assessed in random order.	**Content:** All based on theoretical framework **Discriminant:** Significant differences between proficiency levels using each of the tools (One-way ANOVA *p*<.005) **Concurrent:** Consistent scores for different levels of proficiency using all three tools	**Test-retest:** (Intraclass correlation coefficient [ICC], Pearson [r], Spearman [p]) LCJR: ICC = .908, r = .908, p = .910 SUET: ICC = .907, r = .907, p = .900 C-SEI: ICC = .883, r = .883, p = .849 **Inter-rater:** (ICC) LCJR = .889, SUET = .858, C-SEI = .952 **Internal consistency:** (Cronbach's α) LCJR = .974, SUET = .965, C-SEI = .979

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
3	Adamson et al (2012)	LJCR *: Rubric*	Simulation (video)	Single blind experimental study assessing videos of different abilities	Simulation assessors (*n*=29)	Three videos of different levels of nursing proficiency	**Discriminant:** raters accurately identified known levels of scenarios	**Inter-rater:** Intraclass correlation coefficient = .889
			Simulation (real time)	Cross-sectional study to assess assessor agreement.	Simulation assessors (*n*=2)	*n*=36-year 2 associate degree nursing students	**Discriminant:** raters accurately identified known levels of students	**Inter-rater:** % agreement = 92% - 96%

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
			Simulation (real time)	Cohort study to compare different levels of students.	Simulation assessors (*n*=4)	*n*=22 junior students *n*=25 senior students	**Discriminant:** raters accurately identified progress of students with significant differences across all four aspects of the LJCR: • Noticing *t* = -2.54, *p*=.015 • Interpreting *t* = -3.15, *p*=.003 • Responding *t* = -2.77, *p*=.008 • Reflecting *t* = -3.14, *p*=.003 **Content:** All based on theoretical framework	**Inter-rater:** % agreement = 57% - 100%

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
4	Ball and Kilger (2016), USA	Sweeny-Clark Simulation Performance Rubric (SCSPR) *: Rubric*	Simulation (real time)	Longitudinal study to assess student performance over time	Simulation assessors (*n*=2)	*n*=86 associate degree nursing students assessed in 2nd, 3rd and 4th semester	**Discriminant:** The odds of increasing scores over each succeeding semester were statistically significant in all areas of the SCSPR (*p*<.0001): • Communication (odds ratio [OR]=41.38) • Clinical Judgment (OR=27.61) • Patient Assessment (OR=18.73) • Nursing Interventions (OR=16.84) • Patient Teaching (OR=15.28) • History Gathering (OR=9.66) • Lab and Diagnostics (OR=8.20) • Safety (OR=6.99)	**Internal consistency:** (Cronbach's α) Range from .86 to .96 across all areas of the SCSPR
5	Bujack et al (1991), Australia	Objective Structured Clinical Assessment (OSCA)	Simulation (real time)	Cross-sectional study to correlate OSCA performance and other assessments, with qualitative surveys.	Simulation assessors (*n*=3)	Student nurses (*n* not reported)	**Concurrent:** Low correlation between OSCA scores and other assessment methods used in the course unit (statistical test scores were not provided)	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
6	Dawson et al (2014), USA	Script Concordance Test (SCT)*: Likert scale on importance of proceeding information*	Written exam	Cross-sectional study, with comparisons to mark scheme of RN panel	Marking scheme derived from registered nurse (RN) panel	44 student nurses (yr1)	**Discriminant:** Student mean scores lower than RN panel ($p<0.05$).	**Internal consistency:** (Cronbach's α) .855
7	Deschênes et al (2011), Canada	SCT: *Likert scale*	Written exam	Cross-sectional study, with comparisons to mark scheme of RN panel.	Marking scheme derived from registered nurse (RN) panel	30 student nurses (yr1)	**Discriminant:** Student mean scores lower than RN panel ($p<0.05$).	**Internal consistency:** (Cronbach's α [α], = .86, significant linear relationship between the different human caring assessment dimensions
8	Fenske et al (2013), USA	LCJR *: Rubric*	Written responses to video scenario	Cross-sectional study, with retrospective analysis based on clinical experience and age.	Faculty (n=1)	73 Registered Nurses (0-41 yrs experience)	**Discriminant:** Significant differences between nurses with less than and more than 1 years' experience	**Internal consistency:** (Cronbach's α [α], = between .934 and .97

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
			Self-assessment of own performance in same scenario	Cross-sectional study, with retrospective analysis based on clinical experience and age. Findings compared to written responses in video scenarios.		73 Registered Nurses (0-41 yrs experience)	**Discriminant:** Significant differences between nurses with less than and more than 1 years' experience in responding category only. **Content:** All based on theoretical framework **Concurrent**: Similar scores between self-perception and assessed performance**.**	**Internal consistency:** (Cronbach's α) [between .62 and .749
9	Gantt (2010), USA	SCSPR: *Rubric*	Simulation (real time)	Cohort study using patient scenarios relevant to module studying (obstetrics for year 1, and medical-surgical for final year).	Simulation assessors (*n* not reported)	*n*=69-year 1 associate degree nursing students *n*=109 graduating baccalaureate nursing students	**Content:** Based on theoretical framework **Discriminant:** Better scores for graduating baccalaureate students (mean=74) than associate degree students (mean=39.1 to 50 depending on scenario)	**Inter-rater:** authors were able to establish consistent grading practices after review and discussion of approximately 8 to 10 student rubric scores for the same scenario

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
10	Georg et al (2018), Sweden	Virtual Patient Lasater Clinical Judgement Rubric (vpLCJR): *Rubric*	Simulation (virtual reality)	Iterative panel evaluation of vpLCJR using pre-existing virtual simulation data	Educators (*n*=4) Practicing nurses (*n*=4)	n/a	**Face:** relevant to capture clinical reasoning. Minor modifications recommended and added to final version **Content:** based on theoretical framework	
				Cross-sectional study with deductive coding of student responses on the vpLCJR.	Educators (n=4)	N=28 nursing students	**Construct:** students assessments distributed over all rubric (mean score 29.75+/-6.2, range 15 to 44)	**Internal consistency:** (Cronbach's α) .892

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
11	Gorton and Hayes (2014), USA	7X 1-10 statements: *1-10 for competency* Differential Diagnosis from information provided CDMNS: *rate on five-point Likert scale "strongly agree" to "strongly disagree" for each item*	Clinical Practice Exam Self-assessment	Cross-sectional study to correlate three different methods to assess clinical decision making.	Practice Assessor (*n* not reported)	50 Registered Nurses (with MSc)	**Concurrent:** non-significant correlations with critical thinking (CCTST) for clinical practice assessment and self-assessment **Construct:** student's assessments in clinical practice distributed over most of scale (range 24-69 out of possible scores 10-70).	**Internal consistency:** (Cronbach's α [α]) Clinical practice: α =.917 Self-assessment: α =.67
12	Hasegawa et al (2007), Japan	NANDA: *state nursing diagnosis, evidence and causes or nursing diagnosis and risk factors for case studies*	Exam	Cross sectional study of nursing diagnostic abilities, with retrospective analysis based on participant characteristics.	16 experts "wrote" answers	376 Registered Nurses (+3yrs experience)	**Discriminant:** Clinical experience positively associated with diagnostic capability for Case Study 1 (p<.0001) and Case Study 2 (p= .022). However, not associated with ability to identifying risk factors or underlying causes.	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
13	Hayden et al (2014), USA	Creighton Competency Evaluation Instrument (CCEI) *: Checklist*	Simulation (video)	Single blind experimental study assessing videos of different abilities.	Educators (n=31)	Three videos of different levels of nursing proficiency	**Content:** developed and evaluated by group of 35 nurse educators	**Inter-rater:** 79.4% agreement with "expert" assessor. **Internal consistency:** Cronbach's α [α], = between .974 and .979
			Simulation (real time)	Cohort study to compare different levels of students.	Educators (n=?)	3 baccalaureate nursing programmes, 2 associate degree nursing programmes (n=?)	**Discriminant:** BSN had higher scores than ADN (mean score 83.3% vs 74.2%,	
			Clinical Practice & Simulation (real time)	Survey of educator's experience in simulation and clinical practice.	Educators (n=8)	(n=?)	**Construct:** Faculty evaluated more favourably when in simulation than clinical practice	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
14	Jensen (2013), USA	LCJR*: Rubric*	Simulation (real time) Self-assessment	Cohort study to compare different levels of students, with comparisons between self-assessment and faculty ratings.	Simulation assessors (*n* not reported)	26 baccalaureate nursing students 62 associate degree nursing students	**Discriminant:** BSN had higher scores than ADN (34.33 v's 30.9, p= .01) **Concurrent:** Students tended to score themselves higher than faculty ratings (33.04 +/- 3.8 vs 31.08 +/- 6.9) **Construct:** Faculty identified some students under-performing due to extreme anxiety.	**Internal consistency:** (Cronbach's α [α]) Simulation: α = 0.95
15	Lasater et al (2015), USA	LCJR – adapted for Newly Hired Nurses: *Rubric*	Exam – written case study and short answer responses	Cross sectional longitudinal study of nurses' performance in exam, with retrospective analysis based on participant characteristics and longitudinal follow-up.	1-4 assessors	Registered Nurses (n=202)	**Content:** Based on theoretical framework **Discriminant:** RN with <1 year experience scored lower than those with >1 year experience (11.7+/-2.37 vs 13.01+/-2.18, p<.05) **Predictive:** 2/9 RN achieving "beginning level" were on probation within their clinical practice 9 months later.	**Inter-rater:** 4 markers rated 1-case studies and achieved 90%+ reliability after training.

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
16	Levett-Jones et al (2011), Australia	Structured Observation and Assessment of Practice (SOAP) : *Observation & Viva*	Clinical Practice	Survey of student experience and perception of SOAP assessment using 5 open ended questions and 46 questions scored on a 5-point Likert scale	Clinical Assessors from Nursing Faculty (*n* not reported)	Final year nursing students (n=654)	**Content:** Developed following literature review in consultation with practice and academic assessors **Concurrent:** 86% of students agreed SOAP consistent with general clinical performance. Correlation between SOAP and academic results (no data shown) **Construct:** 63% of students agreed SOAP assessment made them feel anxious. However, findings indicated they were able to overcome this quite quickly and it did not affect overall performance.	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
17	Liaw et al (2011), Singapore	RAPIDS (Checklist) Global assessment (1-10)	Simulation (video)	Cohort study to compare different levels of students using two techniques.	Simulation assessors (n= 3)	Student nurses (n = 15 year 2 & 15 year 3)	**Discriminant:** Third year higher than 2nd in ABCDE domain (20.31+/- 3.48 vs 6.63+/-2.07, p<.001) & SBAR domain (31.42+/-4.06 vs 11.36 +/- 2.95, p<.001) **Concurrent:** RAPIDS correlated with global assessment for ABCDE & SBAR (r=0.94, p<.01)	**Inter-rater:** Interclass correlation [ICC] across 3 assessors on all 30 videoed performances RAPIDS: ICC = .97- .99 Global assessment: ICC = .80 - .85
18	Liaw et al (2018), Singapore	Clinical Reasoning Evaluation Simulation Tool (CREST) : Rubric	Simulation (video)	Cohort study to compare different levels of students using two techniques.	Simulation assessors (n= 2) Different assessor (n=1) assessed same video on RAPIDS)	Student nurses (n = 15 year 2 & 15 year 3)	**Content:** Expert panel (15) assessed content validity **Discriminant:** year 3 had higher scores than year 2 (median 33 vs median 25, p<.01). **Concurrent:** CREST correlated with RAPIDs (r=0.71, p<.01),	**Inter-rater:** Interclass correlation 0.88 **Internal consistency:** (Cronbach's α) .92
19	Liou et al (2016a), Taiwan	Nurses Clinical Reasoning Scale (NCRS): *five-point Likert scale "strongly agree" to "strongly disagree" for each item*	Self-assessment	Cohort study to compare different levels of students, including retest 2 weeks later.	Self	Student Nurses (n=47 final year & n=50 2nd year students)	**Content:** Based on theoretical framework. Expert panel of 3 assessed content. **Discriminant:** 3rd year students had higher scores than 2nd year (53+/-7.3 vs 44.2+/-3.1, p<.001)	**Test-retest reliability:** Interclass correlation between baseline to results 2 weeks later = .87, p<.001 **Internal consistency** (Cronbach's α) = 0.93

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
				Cohort study to compare registered nurses with final year nurses, including retest 2 weeks later.		Registered Nurses (n=100) & final year student nurses (n=151)	**Discriminant:** RN had higher scores than final year students (55.1+/-7.8 vs 52.6+/-7.0, p<.01)	**Test-retest reliability:** Interclass correlation between baseline to results 2 weeks later = .85, p<.001 **Internal consistency** (Cronbach's α) = 0.94
20	Liou et al (2016b), Taiwan	Computerised model of performance-based measurement (CMPBM): *Case based MCQ and short answer questions*	Exam	Cohort study to compare registered nurses with final year nurses, including retest 2-4 weeks later.	Unreported	Final year student nurses (n=30) & experienced registered nurses (n=30)	**Content:** developed by 4 senior clinical experts. **Discriminant:** Student nurses had lower scores than RN (t=-4.63, p<0.001), with significant differences across all three aspects of the assessment.	**Test-retest reliability:** Correlation with repeated results 2-4 weeks later (r=0.70, p<0.01).

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
				Cohort study to compare registered nurses with student nurses.	Unreported	Student Nurses (n=157), RN (n=52)	**Discriminant:** RN scored more than student nurses (t=0.302, p=0.03). This remained significant for each aspect of the scale. • Collect and manage information (t=3.08, p=.003), • Diagnose and differentiate problem urgency (t = 2.5, p= .01) Solve problems (t=2.55, p=.01).	**Internal consistency:** (Kuder-Richardson formula 20) = 0.9
21	Ludin (2018), Malaysia	CDMNS: *Likert scale*	Self-assessment	Cross sectional study of self-assessment, with correlation to critical thinking and retrospective analysis based on participant characteristics.	Self	Critical care registered nurses (n=113)	**Discriminant:** CDMNS positively related to years worked as RN (f=2.090, p<0.004) but was not related to education level. **Concurrent:** Positive correlation with critical thinking score on SF-CTDI-CV (r=0.637, p=0.001)	**Internal consistency** (Cronbach's α)**:** = .797

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
22	Murcott and Clarke (2017), UK	OSCE (3 stations, lasting 30 minutes)	Simulation (real time)	Cross sectional study of OSCE performance with standardised patient (actor) and discussion with colleague.	Simulation assessors (n= 5)	Student Nurses (42, 2nd year mental health)	**Face:** based on module content mapped to learning outcomes, nursing process and NMC standards **Content:** Developed from multiple reviews by academic team Student feedback External examiner feedback	**Inter-rater**: Two independent markers at each station, discussed to agree mark. Data from initial marks and changes not shown.
23	Nielson et al (2016), USA	LCJR: Rubric	Clinical Practice	Focus group discussions of experience and perception of using LCJR in clinical practice	Experienced preceptors supporting newly qualified nurses (n= 7)	Newly qualified nurses (n not reported)	**Content:** Perceived LCJR objective means of assessment by 7 staff and can develop clinical judgement skills in newly qualified staff	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
24	O'Rourke and Zerwic (2016), USA	Written responses to unfolding case-studies (Rubric for each question, not shown)	Exam	Cohort study to compare qualified and student advanced nurse practitioners, including retest 4 weeks later.	3 (2 assessors looked at n=15, 25% of cases)	Newly qualified advanced nurse practitioners (n=15) and advanced nurse practitioner students (n=37), re-test taken by 21 (40%) at one month	**Content:** Unfolding case studies and rubrics developed based on Tiffin theory of decision making. 3 nurse consultants judged relevance of the questions and grading rubrics. Most items received 100% agreement but, one item was unable to be revised following feedback so was removed.	**Internal consistency** (Cronbach's α)**:** case study 1 = .211, case study 2 = .535 **Test-retest:** Correlated with results 1 month later - case study 1: r= .9, p<0.01, case study 2: r=0.88, p<0.01 **Inter-rater:** 25% double marked, Interclass correlation between assessors in case study 1 = .967, case study 2: = .955

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
25	Park et al (2017), Korea	Skill performance *Calculated from scoring 20 core functions from Korean NB	Simulation (real time)	Quasi-experimental study evaluating pre-and post- test performance after simulation intervention, with evaluation of relationship to critical thinking, self-efficacy and learning motivation.	Simulation assessor (*n* not reported)	69 BSN (4th yr) (85% female), 3 students excluded as data incomplete.	**Content:** Skill performance based on Korean Nursing Board Guidance, each scored 0 (deficient) to 2 (good) **Concurrent:** Skill performance after 30-hour simulation programme (82.43±5.54 out of 100) correlated with critical thinking (r=0.349, p= .03), self-efficacy (r=0.316, p= .008) and learning motivation (r=0.246, p= .042)	
26	Prion et al (2015), USA	SECC-35 (rate beginning (1) / developing (2) / accomplished (3) for each item)	Clinical practice	Retrospective analysis of preceptor ratings on the SECC-35.	Preceptors supporting newly qualified nurses (*n* not reported)	Newly qualified nurses (n=193)	**Face validity:** Reviewed by Multi-site subject matter experts (n=6).	**Internal consistency** (Cronbach's α): .92

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
			Self-reported	Cohort study to compare registered nurse and student nurses self-perception.	Self	Student nurses (n=94), registered nurses from academic staff (n=17)	**Discriminant:** Student nurses scored mean 2.27 +/- .29 vs faculty 2.86 +/- .27 for each item	**Internal consistency** (Cronbach's α): .82
27	Prion et al (2017), USA	Quint Leveled Clinical Competency Tool (QLCCT): Rubric	Simulation (real time)	Focus group discussions of experience and perception of tool by faculty members following these simulations.	Simulation assessors (n not reported)	Student nurses (n=67)	**Content:** Based on Tanner theory, developed as found existing LCJR too lengthy and cumbersome. **Face:** Reviewed by Multi-site subject matter experts from 11 programmes following trail on student nurses	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
			Clinical practice	Cohort study to compare different stages of students.	Unreported	Student nurses from advanced diploma in nursing programmes (year 1 and year 2 students)	**Discriminant:** year 2 scored higher than year 1 (27.6+/-5.4 vs 19.3+/-4.4 out of 36). None in Year 1 showed behaviour of "graduate nurse", 12% did in Year 2. Correlation between students score and level of study (r=.83, p not reported) **Content:** same criticism of LCJR as too long and cumbersome for use in clinical practice	
			Simulation (video)	Single blind experimental study assessing videos of different proficiency	Simulation assessors (n= 29)	3 standards of video (below, expected, above expectations) for different scenarios	**Discriminant:** Able to discriminate between video standards - Below (average 11, Standard error [SE] .21) - Expected (average 25, SE 1.1) - Above (average 33, SE .95)	**Inter-rater: interclass correlations** = 0.87, (95% CI: .62-1.00)

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
			(after using QLCCT for several years within the department)	Participants rated each item to determine if its useful.	11 subject matter experts asked to rate tool	n/a	**Content**: (Content validity index) = .72	
28	Randolph et al (2012), USA	TERCAP-41: Checklist (competent / incompetent)	Simulation (video)	Single blind experimental study assessing videos of different abilities, and reassessing 1 week later.	Simulation assessors (n= 5)	Videos of student performance (n unreported)	**Content:** Developed by 5 subject matter experts from nursing faculty	**Test-retest:** Consistency between weeks 1 & 2, intra-ratter reliability 92% (range 85-97%) **Inter-rater:** inter-rater reliability 92%, experienced nurses working clinically rate performance more critically than educators. **Internal consistency:** (Cronbach's α): .93
				Cross sectional study of registered nurses, with retrospective analysis of clinical experience	Simulation assessors (n= 3)	63 videos of "registered nurse performers" with some coached to make errors,	**Discriminant:** RN with 1+ experience performed better than RNs <1yr in 6/9 categories ($p<0.05$)	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
29	Reyes and Rodriguez (2016), USA	OSCE, using interpretation of results / videos and case study based written exam style stations	Exam	Reporting experience of using written exam style OSCE stations. Including survey of faculty and students.	Faculty staff (n=2)	Advanced practice Student nurse) (n unreported)	**Content:** Developed following Ottawa Conference recommendations on OSCE best practice. **Face**: OSCE marking criteria related to course objectives and topics.	
30	Robbins and Hoke (2008), USA	OSCE, using standardized patients	Simulation (real time)	Reporting experience of using standardised patients, clinical documentation, and self-reflection stations.	(unreported)	Advanced practice student nurses (n unreported)	**Content:** Developed by multiple faculty members to meet course objectives.	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
31	Selim et al (2012), Egypt	OSCE: checklist (3 out of 11 stations were standardised patient [S.P] stations),	Simulation (real time)	Cross sectional study of student nurse's performance in OSCE correlated to other assessments.	Simulation assessors (n= 2) examined S.P stations	Student nurses, final year (n=76)	**Face:** OSCE marking criteria related to course objectives **Concurrent:** Correlation between OSCE exams and other assessments - clinical evaluation r^s= .536 (p<0.001), - viva exam r^s= .337 (p=0.003) - written exam r^s= .593 (p<0.001).	**Inter-rater:** Correlation between ratter's on S.P stations - General assessment: r^s= .672 (p<0.001) - Assessing suicidal patient: r^s= .708 (p<0.001), - Assessing hallucinations: r^s= .581 (p<0.001), **Internal consistency** (Cronbach's α)**:** Varied between stations range from .29 to .802.
32	Selim and Dawood (2015), Egypt	OSCE, using video & written scenarios (Model answers with rubric marking for each question, not shown)	Exam	Cross sectional study of student nurse's performance in OSCE correlated to other assessments, with survey of student perceptions	(unreported)	Student nurses enrolled in psychiatric and mental health course (n=87)	**Concurrent:** Correlation with final MCQ exam (r=0.6, p<0.001) **Face:** 58.5% of students agreed the OSCE was fair, with only 6.9% of students thinking it did not eliminate personal bias of instructor towards a student.	**Internal consistency** (Cronbach's α)**:** .714

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
33	Shin et al (2014), USA	LCJR: Rubric (modified to define areas expected in paediatric case & in Korean)	Simulation (real time)	Cross sectional study of student nurse's performance in paediatric simulation, correlated to critical thinking assessment.	n= 3	Student nurses (n=250) from 3 nursing schools	**Content:** 7 experts evaluated rubric. content validity index = .9 **Concurrent:** Scores correlated to Yoon's Critical thinking inventory, - Noticing (r=.13, p<0.05) - Not significant correlations for interpreting, responding or reflecting.	**Internal consistency (**Cronbach's alpha): = .863
34	Shin et al (2015), USA	LCJR: Rubric (modified to define areas expected in paediatric case & in Korean)	Self-assessment	Cross sectional study of student nurse's performance in simulation	Self	Student Nurses (n=152) from 3 nursing schools		**Internal consistency (**Cronbach's alpha): = .910

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
35	Stacey (2008), Canada	Decision Support Analysis Tool (DSAT-10: Checklist	Simulation (verbal triaging of standardised patients)	Single blind experimental study assessing audio-recordings of standardised patients undergoing nursing triage in nurses who have and have not undergone training on using a decision support tool.	5 trained coders analysed responses	n=18 registered nurse after online decision support training and 3 hour workshop, n= 58 registered nurse no training		**Inter-rater:** After training coders achieved agreement of 85% or higher for the DSAT-10 on three consecutive audio-recordings (ICC 0.96; 95% CI: 0.943–0.973). However, this varied if nurses had received specific training to in triaging patients. - for trained nurses (91.1% agreement, Intra-class correlation coefficients .96 (95% CI: .943, .973) - for untrained nurses (74.3% agreement, Intra-class correlation coefficients .564 (95% CI: .415, .564)
36	Starkweather et al. (2017), USA	Progressive assessment and competency evaluation (PACE) Framework	Simulation (real time)	Reporting experience of using PACE framework	Faculty Evaluators (n unclear)	Advanced Nurse Practitioner Students (n unclear)	**Content:** Mapped course companies onto simulation evaluation criteria / mark scheme.	**Inter-rater:** Faculty evaluators reached interclass correlation = .96 after training

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
37	Strickland et al (2017), USA	LCJR: Rubric	Sim (real time)	Cross sectional study of student nurses	Faculty	Student nurses (3rd year on a four-year course) (n=94)	**Content:** Based on Tanner theory. Previously evaluated.	**Internal consistency** (Cronbach's alpha): = .82
			Self-assessment of simulation	Cross sectional study of student nurses compared to faculty ratings.	Self-reported		**Concurrent:** Self-rating scores by students higher than faculty ratings (33.48+/-3.7 vs 31.19 +/- 3.2 out of 43). Sig. Correlation between scores (r=0.314, p=0.03)	

No.	Author/country	Tools	Settings	Method	Assessors	Assessed	Validity	Reliability
38	Vreugdenhil and Speck (2018), Netherlands	LCJR: rubric (modified into Dutch)	Clinical Practice	Delphi technique to review content LCJR Non-blind experimental study comparing nurse educator, clinical nurse coaches and self-assessment, with retrospective analysis of clinical experience and qualitative survey.	N = 2 (assessed each student, one nurse educator, one nurse coach).	Student nurses (n=52) • year 1 = 9 • Year 2 = 9 • Year 3 = 23 • Year 4 = 11	**Content:** Based on Tanner theory Translation and context reviewed by 5 using Delphi technique. Subject Matter Experts (n=7) reviewed to determine Content Validity Index = 85%. **Discriminant:** Student nurse experience (0-40 months) correlated with LCJR (r=0.62, 95% CI .51- .71, p<0.001).	**Internal consistency** (Cronbach's alpha): = .93 **Inter-rater:** ICC between nurse coach and nurse educator = 0.78 (95% CI .64-.86). Faculty scores correlate to coaches scores (30.32 +/- 6.56 vs 30.93 +/- 6.31). This gives a bias of 0.69 points (2.1%, p = 0.68) and limits of agreement of−9.14 to 7.77.
			Self-assessment (after same morning of clinical observations)		Self		**Concurrent:** Self-evaluation corelates with faculty and coaches scores (r=0.78, 95% CI = .64- .87). There was no difference in mean scores between student self-ratings to faculty and coaches (p=0.137, 95% CI - .54 – 3.89).	**Inter-rater:** Student Nurses vs nurse educator = 2-point bias (32.34 +/- 5.29 vs 30.32 +/- 6.56, p=0.02), St. Nurs vs nurse coaches = 1.3-point bias (32.34 +/- 5.29 vs 30.93 +/- 6.31, p=0.07).

Table 3: Quality appraisals

Author (date)	Assessment Tool	Was there a clear question for the study to address?	Was there a comparison with an appropriate reference standard?	Did all students get the " new assessment" and reference standard?	Could the results of the test have been influenced by the results of the reference standard? (N is positive)	Were the methods for performing the test described in sufficient detail?	How sure are we about the results?	Can the results be applied to your population of interest?	Can the " assessment" be applied to your patient or population of interest?
Papers considering assessment in Clinical practice (n=6)									
Gorton & Hayes (2014)	Formation of Nursing Diagnosis in Practice	Y	Y	Y	N	Y	N	Y	?
Hayden J et al (2014)	Creighton Competency Evaluation Instrument (CCEI)	Y	?	N/A	N/A	?	Y	Y	?
Levett-Jones T et al (2011)	Structured Observation and Assessment of Practice (SOAP)	Y	N	N/A	N/A	Y	Y	?	?
Nielsen et al (2016)	Lasater Clinical Judgment Rubric (LCJR)	Y	?	N/A	N/A	Y	Y	?	?
Prion et al (2015)	35-item competency score (SECC-35)	Y	N	N/A	N/A	Y	N	?	Y
Prion et al (2017)	Quint Levelled Clinical Competency Tool	N	Y	N	N	N	Y	Y	Y
Vreugdenhil & Spek (2018)	Lasater's clinical judgment rubric (LCJR), Dutch version	Y	Y	Y	N	Y	Y	Y	Y
Papers considering assessment in Simulated setting (n=22)									
Adamson & Kardong-Edgren (2012)	Three methods (LCJR; the Seattle University Evaluation Tool; C-SEI)	Y	Y	Y	N	Y	N	Y	Y

Author (date)	Assessment Tool	Was there a clear question for the study to address?	Was there a comparison with an appropriate reference standard?	Did all students get the " new assessment" and reference standard?	Could the results of the test have been influenced by the results of the reference standard? (N is positive)	Were the methods for performing the test described in sufficient detail?	How sure are we about the results?	Can the results be applied to your population of interest?	Can the " assessment" be applied to your patient or population of interest?
Adamson (2016)	LCJR using video archives	?	?	N	N	Y	Y	Y	Y
Adamson, et al (2012)	LCJR in both simulated area and using video archives	Y	N	N	N/A	Y	Y	Y	Y
Ball & Kilger (2016)	Sweeney-Clark Simulation Performance Rubric (SCSPR)	Y	Y	Y	N	Y	N	Y	?
Bujack et al (1991)	Objective Structured Clinical Assessment (OSCA)	Y	Y	Y	N	N	N	?	N
Gantt (2010).	Sweeney-Clark Simulation Performance Rubric (SCSPR)	N	N	N/A	N/A	Y	N	Y	Y
Georg et al (2018)	Virtual Patient Lasater Clinical Judgment Rubric (vpLCJR)	Y	N	N/A	N/A	Y	N	Y	Y
Hayden J et al (2014)	Creighton Competency Evaluation Instrument (CCEI)	Y	?	N/A	N/A	?	Y	Y	?
Jensen, (2013)	The Lasater Clinical Judgment Rubric (LCJR)	?	?	Y	N	N	Y	Y	Y
Liaw et al (2011)	Rescuing A Patient In Deteriorating Situations (RAPIDS)	Y	N	N/A	N/A	Y	Y	Y	Y
Liaw et al (2018)	Clinical Reasoning Evaluation Simulation Tool (CREST)	Y	Y	Y	N	Y	Y	Y	Y
Murcott, & Clarke (2017)	Objective Structured Clinical Examination	N	N	N/A	N/A	N	N	Y	N
Park et al (2017)	Scoring 20 core functions from Korean Nursing Board	Y	N	N/A	N/A	Y	?	Y	N
Prion et al (2017)	Quint Levelled Clinical Competency Tool	N	Y	N	N	N	Y	Y	Y

Author (date)	Assessment Tool	Was there a clear question for the study to address?	Was there a comparison with an appropriate reference standard?	Did all students get the " new assessment" and reference standard?	Could the results of the test have been influenced by the results of the reference standard? (N is positive)	Were the methods for performing the test described in sufficient detail?	How sure are we about the results?	Can the results be applied to your population of interest?	Can the " assessment" be applied to your patient or population of interest?
Randolph et al (2012)	TERCAP-41	N	?	N/A	N/A	N	N	?	N
Reyes & Rodriguez (2016)	Objective Structured Clinical Examination	N	N	N/A	N/A	?	N	?	N
Robbins & Hoke (2008)	Objective Structured Clinical Examination	?	N	N/A	N/A	N	N	Y	N
Selim et al (2012)	Objective Structured Clinical Examination	Y	Y	Y	N	Y	Y	Y	N
Shin, et al (2014)	Modified version of LCJR (for paediatric nursing in Korean)	Y	?	Y	N	Y	Y	?	?
Stacey (2008)	Decision Support Analysis Tool (DSAT-10)	Y	N	N/A	N/A	Y	Y	N	Y
Stalkweather (2017)	Progressive assessment and competency evaluation (PACE)	Y	N	N/A	N/A	N	N	?	N
Strickland, et al (2017)	The Lasater Clinical Judgment Rubric (LCJR)	Y	?	Y	N/A	Y	Y	Y	Y
Papers considering assessment during written assessments (n=10)									
Dawson, et al (2014)	Script Concordance Test (SCT)	Y	?	N	Y	Y	Y	Y	N
Deschênes, et al (2011)	Script Concordance Test (SCT)	Y	?	N	Y	Y	Y	Y	N
Fenske, et al (2013)	Short answers to unfolding cases-studies	N	?	Y	N	Y	?	Y	?

Author (date)	Assessment Tool	Was there a clear question for the study to address?	Was there a comparison with an appropriate reference standard?	Did all students get the " new assessment" and reference standard?	Could the results of the test have been influenced by the results of the reference standard? (N is positive)	Were the methods for performing the test described in sufficient detail?	How sure are we about the results?	Can the results be applied to your population of interest?	Can the " assessment" be applied to your patient or population of interest?
Gorton & Hayes (2014)	Formation of Nursing Diagnosis in Exam	Y	Y	Y	N	Y	N	Y	?
Hasegawa, et al (2007)	Formation of Nursing Diagnosis in Exam	Y	N	N/A	N/A	Y	N	?	Y
Lasater, et al (2015)	Short answers to unfolding cases-studies	Y	N	N/A	N/A	Y	Y	Y	?
Liou et al (2016b)	Short answers to unfolding cases-studies	Y	?	N	N	Y	y	?	N
O'Rourke, & Zerwic, (2016)	Short answers to unfolding cases-studies	Y	N	N/A	N/A	Y	Y	Y	N
Reyes & Rodriguez (2016)	Short answers to unfolding cases-studies	N	N	N/A	N/A	?	N	?	Y
Selim & Dawood, (2015)	Short answers to unfolding video cases-studies	Y	?	Y	N	Y	Y	Y	N
Papers considering self-assessment of clinical decision making (n=9)									
Fenske, et al (2013)	Short answers to unfolding cases-studies	N	?	Y	N	Y	?	Y	?
Gorton & Hayes (2014)	Clinical Decision-making in Nursing Scale (CDMNS)	Y	Y	Y	N	Y	N	Y	?
Jensen, (2013)	The Lasater Clinical Judgment Rubric (LCJR)	?	Y	Y	N	N	Y	Y	Y
Liou et al (2016a)	Nurses Clinical Reasoning Scale (NCRS) - Self-assessment tool.	Y	Y	N	N	Y	Y	Y	Y

Author (date)	Assessment Tool	Was there a clear question for the study to address?	Was there a comparison with an appropriate reference standard?	Did all students get the " new assessment" and reference standard?	Could the results of the test have been influenced by the results of the reference standard? (N is positive)	Were the methods for performing the test described in sufficient detail?	How sure are we about the results?	Can the results be applied to your population of interest?	Can the " assessment" be applied to your patient or population of interest?
Ludin SM. (2018)	Clinical Decision-making in Nursing Scale (CDMNS)	Y	?	Y	N	Y	Y	Y	Y
Prion, et al (2015)	35-item competency score (SECC-35)	Y	Y	N/A	N/A	Y	N	?	Y
Shin et al (2015)	Modified version of LCJR (for paediatric nursing in Korean)	Y	N	N/A	N/A	Y	Y	?	?
Strickland, et al (2017)	The Lasater Clinical Judgment Rubric (LCJR)	Y	?	Y	N/A	Y	Y	Y	Y
Vreugdenhil & Spek (2018)	Lasater's clinical judgment rubric (LCJR), Dutch version	Y	Y	Y	N	Y	Y	Y	Y
Y = Yes, N = No, ? = Unclear, N/A = Not applicable									