Exploring the perceived credibility of assessment in medical education: A scoping review

Stephanie Long

Department of Family Medicine

McGill University, Montreal

August 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Family Medicine

Table of Contents

Abstract

Background: In medical education, one of the goals of assessment is to support learning and improve performance. This goal is best achieved when trainees are receptive to and actively engage with the assessment process. Many factors contribute to whether learners will consider assessment-generated feedback as valuable, such as the perceived credibility of an assessment process and the feedback source. If assessment-generated feedback is not perceived as credible, it is unlikely to be integrated into a trainee's knowledge base, leaving the goal of improved performance unmet.

Methods: Applying a scoping study methodology, I searched the medical education scholarly literature for research relevant to assessment of medical trainees, and the perceived credibility of assessment, in five databases: OVID MedLine, ERIC, Scopus, PsycInfo, and EMBASE.

Results: The search strategy identified 3101 unique articles, which were coded for inclusion by myself and my principal supervisor, with a third reviewer adjudicating disagreements. Full-text review yielded 114 articles to be included in the synthesis. I identified three overarching themes that described perceptions of the assessment process and credibility: (i) characteristics of good assessment, (ii) consequences of assessment being perceived as credible (or not), and (iii) factors that affect the perceived credibility of assessment.

Conclusion: Medical trainees make complex judgments regarding the credibility of assessments and assessment-generated feedback to determine what information will be used to improve their performance and what information will be ignored. This review has identified factors that affect trainee perceptions of credibility of assessment and assessment-generated feedback. Findings from

this review can be used to inform assessment development and administration, and the provision of assessment-generated feedback to improve the likelihood of supporting trainee performance.

Résumé

Introduction : Un des buts des évaluations en l'éducation médicale est d'informer l'enseignement et améliorer la performance. Ce but est mieux atteint quand les étudiants sont réceptifs et bien engagés avec le processus d'évaluation. Plusieurs facteurs contribuent à la valorisation de la rétroaction par les apprenants. Par exemple, si ceux-ci ne considèrent pas crédibles les outils d'évaluations et les superviseurs, la rétroaction risque de ne pas être intégrée et alors leur apprentissage n'en bénéficiera pas.

Méthodes : A l'aide d'une étude de la portée (« *scoping review* »), j'ai cherché dans la littérature en pédagogie médicale « évaluation des apprenants » et « crédibilité perçue » en utilisant une combinaison de ces mots-clés, ainsi que des « *subject headings* » dans cinq bases de données : OVID MedLine, ERIC, Scopus, PsycInfo, et EMBASE.

Résultats : La stratégie de recherche a donné 3101 articles uniques. Après la révision de texte, 114 articles sont inclus. J'ai identifié trois thèmes globaux qui décrivant le processus d'évaluations et de jugements de crédibilité découlent de l'analyse performé: (i) caractéristiques des bonnes évaluations, (ii) conséquences des évaluations qui sont perçues crédibles (ou non), et (iii) les facteurs qui affectent les perceptions de crédibilité.

Conclusion : Les résultats de cette revue contribueront aux écrits scientifiques en pédagogie médicale en plus de faire progresser les pratiques évaluatives en lien avec la rétroaction. Cette revue pourra ainsi contribuer à l'amélioration du rôle que peut avoir la rétroaction sur l'apprentissage des apprenants en pédagogie médicale.

Acknowledgements

Over the past two years, several individuals have been instrumental to this thesis, and my growth and development as a researcher. First and foremost, I would like to thank my principal supervisor, Dr. Meredith Young, for viewing my non-straight path as a good thing and for seeing the potential in me, even when I didn't. Without her, this thesis would not have been possible. I am ever grateful for her endless support, reassurance, and guidance. Secondly, I want to thank my co-supervisor, Dr. Charo Rodriguez for always having her door open to me in times of need, providing endless support, and for pushing me to improve my qualitative methodology.

Special thanks also go to my thesis committee members who also played a pivotal role in the development of this thesis. Thank you to Dr. Christina St-Onge for her expertise in assessment, and to Dr. Pierre-Paul Tellier for his clinical input. I greatly appreciate all the valuable feedback you provided during the write-up of my thesis. Also thank you to Nazi Torabi for helping me craft my search strategy, and all library-related support. I must also acknowledge the funding I received to conduct my research from a SSHRC Insight Grant and a Family Medicine Education Research Group Grant.

I am also grateful to Dr. Ian Shrier for sharing wisdom about life, research, and medicine. Thank you for all the chats over beers.

Thank you to my boyfriend, Alexandre Piché, for being my rock, for keeping me focused and pushing me to be the best I can be.

Most importantly, thank you to my parents, Francis and Charlotte Long, for providing me with all the resources to be successful. Thank you for supporting me all these years and giving me

encouragement when I needed it the most. Also thank you to my sister, Jennifer, for all the gourmet meals, fun trips, and distractions from my work during stressful times.

1.0 Introduction

The purpose of this research was to map the currently available medical education literature on the perceived credibility of assessment and assessment-generated feedback. One of the primary means of determining whether a learner is progressing as anticipated through an educational program is the collection and interpretation of data generated by educational assessments. The *Standards for Educational and Psychological Testing* has defined assessment as "any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs" (p. 172) [1]. In medical education, assessment is generally described as any strategy involving testing, measuring, collecting, and combining information to make judgments about trainees' achievement of specific goals of learning [2, 3]. Assessments in medical education are commonly used to accomplish three main goals [4]: (i) to support and provide direction for future learning [2, 5], (ii) to provide a basis for selecting applicants for advanced training, and (iii) to protect the public by ensuring those entering practice are competent.

According to Norcini et al. [2] the first goal, i.e. effective orientation of student's future learning, is best supported when an assessment has a 'catalytic effect' - an ability to drive future learning forward via feedback that encourages improvement. For this *catalytic effect* to be achieved, a learner must participate in the assessment process by being receptive to and actively engaging with assessment-generated feedback [2]. When learners fail to engage with assessment-generated feedback as an opportunity for learning, they miss out on valuable opportunities to learn and to improve performance.

Several factors have been identified as contributors to whether medical trainees will engage with the assessment process and integrate assessment-generated feedback for the purposes of

improving performance. Among those that will encourage future learning are the learner's perceived credibility of the assessment itself, and how credible they perceive their supervisor to be [6-8]. Regarding the latter, i.e. the supervisor, it has been reported that medical trainees (e.g. medical students, residents, or fellows) value specific, clinically relevant, timely, and actionable feedback based on observable behaviour from a trusted and respected clinical supervisor [6-9]. The findings of these studies [6-9] have suggested that if medical trainees do not view their feedback or supervisors as credible, they will ignore or disengage from the assessment-generated feedback. The assessment itself, must also be perceived as credible by learners to successfully drive learning [10]. Therefore, if an assessment or assessment-generated feedback is not perceived as credible, the major learning goal of assessment - to improve performance and to support positive development by driving learning – will not be met.

In the context of this thesis, I define *assessment-generated feedback* as any information resulting from an assessment (i.e. scores, comments, ratings). However, it is important to note that most current definitions of feedback are not limited to mere information, but also encompass the process for provision. For assessment-generated feedback to be perceived as *credible*, it must be used in a way that offers motivation and guidance for future learning, which allows trainees to make sense of their experiences and provides direction for next steps [11]. Literature on *effective* feedback in medical education suggests that feedback should provide trainees with specific information on their performance, emphasizing both their strengths and weaknesses. The provision of feedback allows for communication of "the dissonance between the intended result and the actual result, thereby providing impetus for change" (p. 777) [12]. Feedback can only contribute to improved performance if it is actually provided, Ende [12] warned that "[w]ithout feedback,

mistakes go uncorrected, good performance is not reinforced, and clinical competence is achieved empirically or not at all" (p. 778).

When feedback is not regularly provided to medical trainees, it may lead to overreliance on internal cues ("self-assessment") – to judge the quality and adequacy of their performance as they develop clinical skills – and disregard of important external cues [12]. Depending exclusively on internal cues can impede progress and learning as research has demonstrated that individuals in general, including medical trainees, are quite poor at self-assessment [13-17]. A literature review on self-assessment in medical education found: (i) little to no relationship between actual performance and self-rated performance, and that (ii) most individuals overestimate their performance, with this phenomena most starkly present in the lowest performers [18]. These findings suggest the importance of providing adequate feedback to trainees to ensure they do not depend solely on internal cues, but also on external information provided by an assessment and, when relevant, their supervisors.

All forms of assessment can conceivably be used as a form of feedback by medical trainees, and, therefore, have the potential to cause a *catalytic effect*. However, most research investigating the educational value and perceived credibility of feedback has focused on supervisor-provided feedback in the context of workplace-based assessment [9, 10]. While workplace-based assessment represents a critical context within medical education, it remains only one of many different forms of assessment used in medical training. If perceived credibility of feedback (and the feedback provider) is a key factor in the educational utility of workplace-based assessment, then we may speculate that perceived credibility is also an important consideration for the educational value of other forms of assessment and assessment-generated feedback.

The literature on medical trainees' perceptions of the credibility of assessment is expanding, but what is still missing is a clear understanding of how judgments of credibility are made by medical trainees across various forms of assessment, extending the current work focused on supervisor-provided feedback during workplace-based assessment.

In short, despite its relevance, current work regarding the perceived credibility of assessment-generated feedback is disparate in the medical education field of inquiry, necessitating a careful synthesis of a broad body of research works.

2.0 Literature review

2.1 Premise

In this section, I will situate the context within which an assessment takes place and perceptions of credibility occur. I will first summarize current frameworks of medical education and then describe assessment approaches commonly used in the domain.

2.2 Frameworks in medical education

Medical education has historically adopted a structure- and process-based curriculum of education which defined training experience by exposure to specific content for certain lengths of time [19] and readiness for independent practice by specified years of training [20]. Despite being widely adopted at the beginning of the 20th century, this curricular model has received significant criticisms [21]. First, the structure- and process-based curriculum has been scrutinized for an apparent overemphasis on the demonstrated acquisition of knowledge over the development of skills and attitudes required for medical practice [22]. This focus on knowledge-based objectives may result in learning that is not integrated with the curriculum [21]. Additionally, it may create difficulties for clinician educators in identifying and assisting trainees who have issues with the curriculum and who are at risk of being left behind [23]. Second, because the structure- and process-based framework focuses specifically on time spent in training, it may not fully take into account the learning trajectory of an individual learner [21]. Common criticisms include lack evidence to support the notion that number of years of training predicts competent medical practice [20].

These criticisms contributed to calls for the implementation of a competency-based framework of medical education, which has been defined as:

> *"… an approach to preparing physicians for practice that is fundamentally oriented to graduate outcome abilities and organized around competencies derived from an analysis of societal and patient needs. It de-emphasizes time-based training and promises greater accountability, flexibility, and learner-centeredness"* (p. 636) [24].

Through this definition it is apparent that this framework attempts to increase accountability to patients and society, improve curricular objectives, implement assessments that promote positive trainee development, and facilitate more learner-centred approaches. This framework focuses on the documented achievement of competence in a domain and decouples time in training from the attainment of educational targets [25, 26].

2.2.1 Defining competency

For the purpose of assessment and training, basic elements of physician roles are translated into desired and measurable outcomes of training (i.e. "competencies") [25]. These end-of-training competencies are used to shape the educational content and to further the educational process [19, 25, 26]. Competence has been defined as "the possession of a required skill, knowledge, qualification, or capacity" [27]. However, in the context of medical education, there is significant variation in the definitions of competence used across different contexts. In this domain, competence has been described as "a complex set of behaviours built on the components of knowledge, skills, and attitudes" (p. 362) [19], "the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice

for the benefit of the individuals and communities being served" (p. 226) [28], and "knowledge, skills, attitudes and personal qualities essential to the practice of medicine" [29].

2.2.2 Competency-based frameworks adopted in medical education

In 1996, the Canadian Medical Education Direction for Specialists (CanMEDS) was tasked with defining key competencies necessary for residency programs, which must be achieved by trainees in accordance with the Royal College of Physicians and Surgeons of Canada (RCPSC) [30]. The initial framework was described in 1996 and updated most recently in 2015. The CanMEDS framework has viewed *competence* as "the ability to successfully apply professional knowledge, skills, and attitudes to new situations as well as to familiar tasks" [31]. This framework consists of core competencies grouped thematically within seven roles. A physician must be a Medical Expert, as well as a good: Communicator, Professional, Collaborator, Leader, Health Advocate, and Scholar [30] (Figure 1).

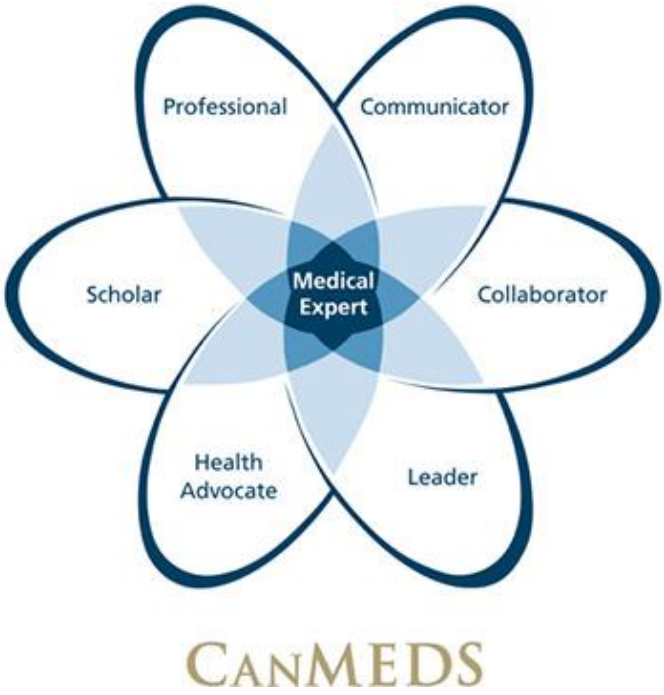


Figure 1 CanMEDS diagram illustrating the seven CanMEDS roles [30].

Similar competency-based frameworks such as *The Outcome Project* have been developed in the United States by the American Council for Graduate Medical Education (ACGME), which defined six core competencies: patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice [32]. The General Medical Council (GMC) in the United Kingdom has also defined outcomes in *Tomorrow's Doctor*, including doctor as a scholar and a scientist, doctor as a practitioner, and doctor as a professional [33].

2.3 Competency in the family medicine context

In 2009, the College of Family Physicians of Canada (CFPC) adapted the 2005 CanMEDS framework to suit the needs of family medicine education creating CanMEDS-Family Medicine (CanMEDS-FM) [34] (Figure 2). CanMEDS-FM was designed to assist in the development of family medicine residency programs and to create a common vocabulary between the RCPSC and CFPC for educational purposes and for the evaluation of medical trainees [34, 35].

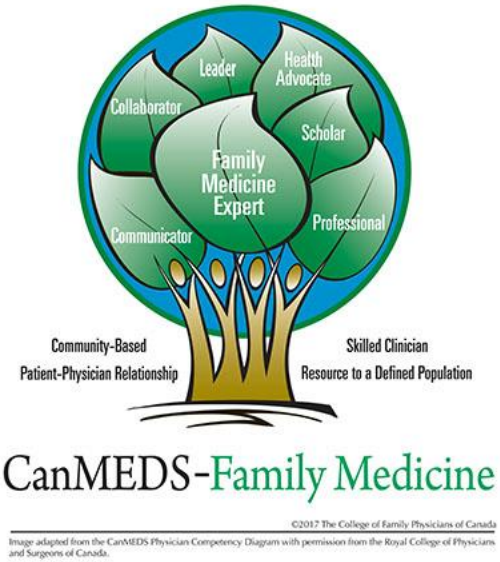


Figure 2. CanMeDS-Family Medicine 2017 diagram illustrating the seven physician roles adapted for Family Medicine [36] .

The seven physician roles of CanMEDS-FM are very similar to those outlined in CanMEDS 2005, with the notable replacement of the Royal College's "Medical Expert" role with "Family Medicine Expert" [35]. This new "Family Medicine Expert" role has been revised from the original CanMEDs' "Medical Expert" role, and now encompasses many of the key competencies necessary in day-to-day practice of primary care in family medicine, such as health promotion and disease prevention, diagnosis, acute treatment, chronic disease management, rehabilitation, supportive care, and palliation [35].

The expected competencies outlined by the CanMEDS-FM roles coexist with the Four Principles of Family Medicine: (i) doctor-patient relationship, (ii) family physician is a skilled clinician, (iii) family medicine is community-based, and (iv) family physician is a resource to a defined practice [35]. Because family physicians engage in primary care, and therefore, interact across a variety of health care issues and interventions, their clinical skills must include the ability to understand the patient's perspective and experience of illness [35].

2.3.1 Triple C Curriculum

In 2010, the CFPC incorporated the CanMEDS-FM roles into a new curriculum for family medicine residency programs called the Triple C Competency-Based Curriculum [37]. This curriculum has the following characteristics:

- **Competency-based curriculum**. This type of curriculum has origins in the outcomes-oriented approach, which emphasizes learner and program outcomes, as opposed to strict focus on the specifics of the pathways and processes to attain them [21]. The CFPC has developed competency material for residency programs that suggest educational assessments for residents, including guidelines on how to best track and document residents' achievement of the CanMEDS-FM competencies [37].

- **Comprehensive care and education.** Family medicine residency programs must ensure that their residents are capable of meeting community needs by providing comprehensive patient care [37, 38]. This is best achieved when residents are embedded in a comprehensive curriculum that allows them to develop the full range of competencies outlined by the CanMEDS-FM framework.
- **Continuity of education and patient care.** Continuity of resident education and patient care are imperative to the development of physicians who provide comprehensive care. Continuity of education is comprised of three elements: supervision, learning environment, and curriculum. Teaching and assessment by a small group of primary preceptors aims to support trusting relationships between residents and supervisors, with the hopes of contributing to a safe and supportive learning environment. Continuity of curriculum involves authentic learning experiences that promote integrated learning and continuous development of competencies shaped through ongoing feedback and assessment. Continuity of patient care is an essential part of family medicine, and is found at the core of the required competencies; its value lies in its ability to strengthen physician-patient relationships and improve patient outcomes [37].
- **Centred in family medicine.** The context of learning must be primarily situated within family medicine settings. The content should, thus, be relevant to the needs of family medicine residents and allow them to develop their identity as family physicians, while achieving the necessary competencies. Residents should be educated by family physicians as well as other specialists in order to achieve the full range of competencies.

2.4 Assessment

Regardless of which framework for education is applied, a variety of assessment approaches can be used to garner information about a trainee's performance including: structured observations, workplace-based assessment, written exams, and case presentations [39]. Through a comprehensive assessment process, medical educators are able to make value judgments about a trainee's progress [39].

Assessment is a multifaceted process that can be used to determine the nature and extent of trainee learning and development [2, 39]. Assessment is most effective when: (i) administrators clearly define what is being assessed (the target construct), (ii) its format is relevant to the construct or characteristics being assessed, (iii) it is comprehensive (e.g. uses a variety of approaches), and (iv) a given assessment is used with considerations of its limitations [39]. Additionally, an assessment should only be used when there is a clearly-defined purpose, that is well-communicated to trainees. As it is an irresponsible use of time and resources to collect data on trainees for no defined purpose (i.e. an assessment for the sake of an assessment) [39]. More importantly, a trainee is more likely to engage with an assessment as a means to support their learning if it has the above characteristics.

2.4.1 Formative and summative assessment

A comprehensive approach to assessment includes a description of *what* is being assessed, *why* it is being assessed, *how* it is being assessed, and consideration for its *usefulness* in supporting and driving learning [28]. When the purpose of an assessment is well-communicated to trainees, it is more likely to be effective in fostering trainee engagement and maximizing educational value. A key consideration regarding the development of assessment is its purpose – including whether an assessment is intended to be used formatively or summatively. Formative assessment is

typically informal and low-stakes in nature [2] and is intended to stimulate and guide learning [4, 40]. Assessment can be used for formative purposes to provide evidence regarding a learners' progress towards academic goals and to deliver feedback outlining actionable steps for future development [3, 39]. According to Norcini et al. [2], formative assessment is most effective when it is embedded in the instructional process, provides specific and actionable feedback, is ongoing, and is provided in a timely manner.

In contrast, summative assessment is "designed to determine the extent to which the instructional goals have been achieved and is used primarily for assigning course grades or for certifying student mastery of the intended learning outcomes" (p. 39) [39]. Summative assessment can, therefore, be used to provide evidence of performance or readiness to practice [41] and can help address the need for accountability [2]. Because it frequently consists of traditional test material [2] (e.g. written examinations comprised of multiple choice questions and short answer questions) taken by trainees at important academic milestones [42], summative assessment also acts as a barrier to further professional practice or training, if certain levels of achievement are not reached [4]. Indirect feedback via assessment scores is provided to learners reflecting their performance on the assessment, indicating their current progress which enables them to adapt their learning or studying habits in order to improve knowledge and future performance [43].

2.4.2 Assessment approaches

Most of the published work on perceived credibility of assessment has focused on performance-based assessment. However, performance-based assessment represents only one of many forms of assessment used in medical training. Each assessment approach has different characteristics and associated strengths and weaknesses, which may influence how a trainee perceives its credibility.

van der Vleuten [44] outlined five criteria for evaluating the utility of a given assessment method: reliability; validity; educational impact to the trainees and examiners; acceptability to learners, faculty, and stakeholders; and costs (of the assessment, to the trainees, the institution, and society). These criteria [44] can be considered in light of a widely known framework for assessments in medical education, George Miller's [45] pyramid of competence (Figure 3). Miller's pyramid has been used to conceptualize the essential elements towards the achievement of clinical competence: beginning with factual knowledge ("*knows*"), followed by applied knowledge ("*knows how*"), to *in vitro* demonstration of performance ("*shows how*"), culminating in the translation of knowledge and skills into *in vivo* performance ("*does*"). Because some assessments are better suited than others at assessing certain skill domains and/or competencies, each level of Miller's pyramid provides a framework to consider the form of assessment best suited to the targeted level of performance. The higher the skills in Miller's pyramid, the more clinically authentic the assessment is likely to be [46]. My discussion of different assessment approaches will be scaffolded within Miller's pyramid, as a means to highlight the purposes, uses, and characteristics of an assessment that may influence how a trainee perceives the credibility of a given assessment and/or assessment-generated feedback.

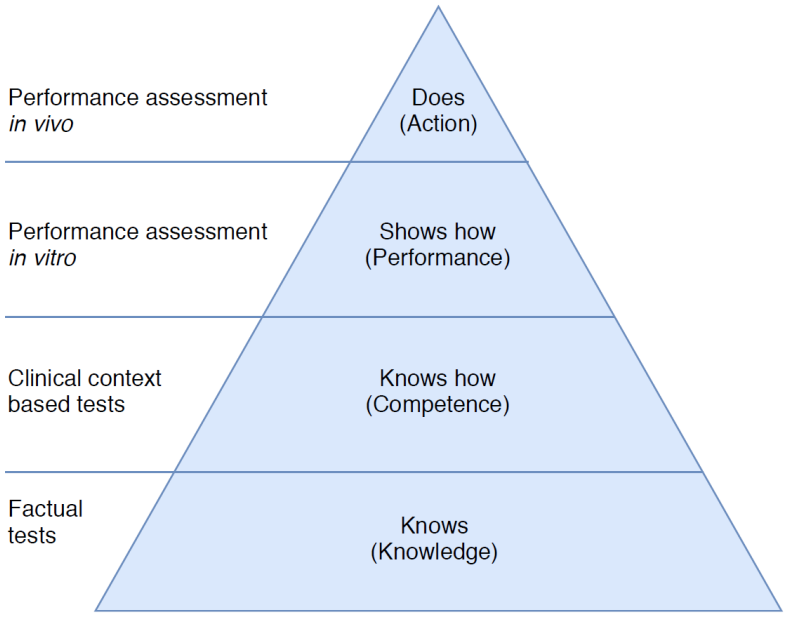


Figure 3: Miller's pyramid of competence [45], adapted from [47].

The base of the pyramid represents the *"knows"* (i.e. factual recall of knowledge) and *"knows how"* (i.e. application of knowledge and problem-solving skills in clinical setting) levels, which are typically the focus of assessments in the earlier stages of medical training [47]. At the base of the pyramid, it is expected that a trainee may *know* what is required in the effective provision of clinical skills (i.e. *knowledge* base) [45]. Assessment of the knowledge base is often done through traditional test methods such as written and oral examinations [47, 48].

Two of the most commonly used forms of written assessment are multiple-choice questions and open-ended questions [49, 50]. Multiple-choice questions are the most widely used [28], due to their high reliability [51] resulting from the large number of testable items that can be used to sample knowledge from a single content area. Multiple-choice questions can also be administered in a short period of time and are computer-gradable [4, 47] - factors that contribute to the feasibility of their use. However, exam administrators may have difficultly constructing high-quality questions regarding complex issues such as ethical dilemmas or cultural ambiguities [52].

Multiple-choice questions may also produce a cueing effect, which occurs when an examinee answers a question correctly by recognizing the correct response, but may not have been able to spontaneously generate the answer in the absence of options [49, 53].

Open-ended questions such as essays and short-answer questions do minimize cueing [49] and well-designed essay questions require trainees to process, summarize, and apply information [4, 49]. However, essay questions are difficult to generate and challenging to mark consistently [54]. Despite having comparatively lower reliability than multiple-choice questions, essays have been argued to have strong validity evidence supporting their use, and are considered to be superior at assessing higher-order cognitive skills [50]. The lower reliability of essays is due to the limited amount of essay questions that can be asked during a single assessment, and the lack of predictive power (of performance) across essays [47, 49]. Short-answer questions are not more reliable but are less expensive to produce and easier to correct than essays. The use of short answer questions is, thus, well-suited when the goal is to evaluate spontaneous generation of a response, rather than the identification of the 'correct' answer among a list of response options [49].

Assessment of the "*knows how*" level of Miller's pyramid can be accomplished through the previously mentioned assessment methods, however, this stage of clinical competence is typically evaluated with oral examinations [47]. Due to criticisms of traditional oral examinations concerning lack of consistency and uniformity of questions and level of difficulty, structured oral examinations were introduced to address these issues [55, 56]. Unlike the traditional format, structured oral examinations use predetermined questions and marking schemes to ensure consistency in exam delivery from one trainee to the next [56, 57]. Proponents of oral examinations believe they are well adapted to assess communication skills, professional attitudes, and the integration of the knowledge, skills, and attitudes acquired during training [55, 57].

The assessment of trainees at the "*shows how*" level (level 3) typically involves practical clinical examinations made *in vitro* (i.e. controlled settings). Several assessment methods currently exist, including: traditional long cases [58], Objective Structured Clinical Examinations (OSCE) [59, 60], Mini-Clinical-Evaluation Exercise (mini-CEX) [61, 62], and assessments completed by standardized patients [63]. Long cases and the mini-CEX were developed to increase the frequency of direct observation of trainees [4]. During these assessments, a supervising physician typically observes a trainee taking history, conducting a physical examination, and presenting a diagnosis, in a 10- to 20-minute period [4, 62]. The supervising physician then scores the trainee and may provide written or verbal feedback.

OSCEs are an assessment approach in which trainees rotate through a series of timed stations, each focused on a different clinical task [4] embedded within a clinical context [47]. OSCEs often involve the use of standardized or simulated patients [28, 63]. Following completion of the stations, the observing physician (or assessor) scores the trainee's performance using a global rating form [64] or checklist of specific behaviours [28, 47]. The U.S. Medical Licensing Examination (USMLE) mandated that medical students' must pass an OSCE (i.e. USMLE step 2 Clinical Skills) in order to be licensed [65], suggesting sufficient evidence supporting its use in assessing clinical skills. However, OSCEs are expensive in terms of time and resources – to achieve an adequate level of reliability, trainees must undergo a minimum of 10 stations (3 to 4 hours) – and its administration requires case development, simulated patient training, and reliance on trained assessors, often physicians [2].

Miller's "*does*" level (level 4) has traditionally been a challenging target for assessment. This level builds on the previous level, except performance is evaluated *in vivo* (i.e. in real life practical settings) [48]. These practice-setting based assessments have gained increased attention

due to their use in determining the competency and fitness to practice of trained physicians. Direct observation in the work environment is a core means to assess performance [4, 66] (e.g. workplace-based assessments [67]) and is a key feature of competency-based medical education [68, 69]. In this level, trainee assessment typically involves direct observation by an assessor that scores trainee performance, which often includes written comments from a variety of other sources (i.e. other physicians, residents, nurses) [4, 66]. The major concern with the use of direct observation is feasibility – both trainees and clinical faculty report a lack of time to complete a sufficient number of assessments [2, 70].

2.5 Feedback

Earliest reported use of 'feedback' was in the beginning of the 20th century and has since extended within and beyond the sciences to the humanities. Despite widespread use within and among those disciplines, authors have yet to reach a consensus regarding a consistent definition. Feedback was first described in engineering as "information that a system uses to make adjustments in reaching a goal" (p. 777) [12]. An early social science definition stated, "…feedback [signifies] that the behaviour of an objective is controlled by the margin of error at which the object stands at a given time with reference to a relatively specific goal" (p.2) [71]. This definition portrays feedback as cycle that connects the input and output, providing impetus for change. The notion of feedback as a cycle (i.e. input and output) spread over time and has evolved into "information" and "reaction" [72]. As demonstrated by a popular definition of feedback in the education literature, "information provided by an agent (e.g. teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (p.81) [73].

The importance and value of feedback has been well-documented and established across educational domains as well as in health professions education [73-75]. Feedback is most effective

in improving trainee performance when educators and trainees work together to create an actionable plan for trainee improvement [75]. In many academic settings, feedback is typically unidirectional from the evaluator to the student. In fact, an analysis of feedback interactions in medical education demonstrated that feedback dialogue was primarily teacher-centred and greatly underemphasized the role of the learner [9]. In order for learners to reap presumed educational benefits, feedback should be a phenomenon including new instruction from the evaluator, rather than the evaluator merely informing the trainee about correctness [76]. This notion suggests that an assessor should provide actionable steps for future improvement, not just highlight correct or incorrect performance.

Feedback is most effective when "information about previous performance is used to promote positive and desirable development" (p. 102) [74]. However, from what is known in the literature, several steps must be taken for trainees to find value in their feedback. First, trainees must be receptive to the feedback delivered by an assessment or a supervisor [77]. Second, trainees must understand the message being conveyed, such that it aligns with their learning objectives and curricular outcomes [77]. Lastly and arguably most importantly, trainees must deem the source of feedback (i.e. supervisors, evaluators, assessment) as credible [8, 78, 79]. This final element is the focus of this thesis and will be discussed in greater detail in the following section.

As previously stated, assessment can fill formative or summative roles depending on the educational goals of the assessment, which may result in differences in how feedback is received and processed by trainees [2]. Formative assessment is more naturally aligned with the provision of feedback as the objective of this form of assessment is to enhance and support learning in a typically low-stakes manner [2, 4, 80]. However, given the purpose of summative assessment, i.e.

to determine whether trainees have attained sufficient competence to advance to the next level, there may be fewer efforts and opportunities to provide specific feedback to trainees [74, 77].

2.6 Medical trainees' perceptions of credibility of supervisor-provided feedback

Perceived credibility of assessment and feedback influences how assessment-generated feedback is perceived, understood, and used by medical trainees. In 1997, Bing-You, Paterson and Levine [9] first explored the effect of a feedback provider's (a.k.a. sender) credibility in a study examining the factors that influence a trainees' receptivity to feedback. Thematic analysis of semi-structured interviews with 12 residents revealed four themes pertaining to a supervisor's and feedback's credibility and its subsequent use: (i) residents' perceptions of supervisor characteristics (e.g. trust and respect, clinical experience, supervisory status), (ii) residents' observations of supervisor behaviour (e.g. inattention, lack of interpersonal skills, lack of observation), (iii) content of feedback (e.g. non-specific, irrelevant, incongruent with self-perceptions), and (iv) method of delivering feedback (e.g. judgmental, occurs in group setting) [9]. Findings from this study suggest that if a trainee does not perceive their supervisor-provided feedback as credible, it is unlikely to be used to improve performance, leaving the goal of assessment leading to improved performance unmet. These findings have since been explored, replicated, and expanded upon by Watling et al. [8, 10, 81].

Watling, Driessen, van der Vleuten, and Lingard [10] presented a model of clinical learning that suggested only supervisor-provided feedback judged as credible by learners will be influential in shaping their learning. In this context, credibility judgments occur when learners organize, weigh, and allocate value to the learning cues presented to them, deciding which information should be integrated into their learning and which should be dismissed [10]. Medical trainees' judgments of the credibility of feedback are influenced by many factors. First, the feedback

message must be aligned with the learner's own self-assessment – if the learner believes that the evaluator wrongly provided negative feedback, it will be discredited and ignored [10]. Second, the provider of feedback (i.e. supervisor, professor, observing physician, assessor) must have personal and professional values that align with the learner's – if the values of the learner and supervisor are misaligned, the trainee will be unlikely to take his/her feedback into account [10]. Lastly, credibility judgments are influenced by medical trainees' reported amount of respect for the source of the feedback (i.e. supervisor, evaluator, observing physician); respect appears to be related to the supervisor's perceived performance as a clinician, rather than their teaching abilities [10]. These findings reinforce the results of Bing-You, Paterson, and Levine [9] and can be used to support the importance of considering perceived credibility of assessment and assessment generated feedback in medical education. Ultimately indicating that if trainees do not perceive the feedback as credible, it is unlikely to influence their learning or support their improvement. But, if the feedback is perceived as credible, it is more likely to be incorporated into their learning and support future development.

However, until now, research examining judgements of perceived credibility have been focused on the credibility of an individual (e.g. supervisor, assessor, examiner) who provides the feedback. Most work on feedback focuses on feedback provided by a teacher or supervisor, however, in the context of this thesis, feedback is also considered to be information generated by an assessment or assessment process (e.g. scores, ratings, numerical grades, comments). Thus, if perceived credibility of a feedback provider is an important factor that influences trainee interactions with feedback, I speculate that perceived credibility will also influence trainees' perceptions of an assessment and increase the likelihood that assessment-generated feedback will support future practice improvement.

2.7 Aim, research question, and objectives

Most of the research exploring notions of credibility focuses on assessor's characteristics and the assessor-assessee relationship [9, 10, 81], uses disparate terminology, and little work has identified the effect of a trainee's perceived credibility of assessment and assessment-generated feedback beyond workplace-based assessment. Here, I propose to explore the perceived credibility of assessment-generated feedback across the medical education literature. Therefore, my aim in conducting this scoping review was to map the published literature on trainee perceptions of credibility of assessment-generated feedback in medical education. I addressed the following research question:

> What do we know about trainees' perceived credibility of assessment-generated feedback in the published medical education literature?

3.0 Methodology

Scoping reviews are a form of knowledge synthesis that are used to map key concepts underlying a research domain using relevant sources and evidence [82, 83].

3.1 Scoping reviews vs. Systematic reviews

Scoping reviews share some methodological similarities to systematic reviews; however, they contain synthesis approaches suited to different research questions. Both review forms have rigorous, transparent, and replicable methods to identify, analyze, and summarize literature relevant to a research objective [84, 85]; however, they differ in several ways. First, one of the primary uses of a scoping review is to broadly map a body of literature on a given research topic [82, 86], whereas as a systematic review aims to systematically appraise and synthesize the best research evidence in order to inform policy or evidence-informed decision making [87]. Thus, a scoping review intends to present a diversity of findings from a large body of literature pertaining to a broad topic and/or research question [88], while a systematic review concentrates on empirical evidence from a smaller number of studies pertaining to a focused research question [87].

Second, to address the goal of broadly mapping the literature, scoping reviews can include a wider range of research and non-research materials (e.g. informal or formal commentary from professional meetings) [88] as compared to systematic reviews, which traditionally only include randomized controlled trials [82] or empirical studies [89].

Third, scoping reviews can be used to provide a descriptive overview of the included records without the necessity of critically appraising the individual studies [85], whilst in systematic reviews there must always be an assessment of the risk of bias in the included studies [90].

3.2 Scoping review methodology

There are four common reasons for undertaking a scoping review: (i) examination of the extent of a research activity, (ii) determination of the value of undertaking a systematic review, (iii) summarization and dissemination of research findings, and (iv) identification of research gaps in the existing literature [82]. The present literature on learner perceptions of credibility of assessment in medical education is disparate and comprised of articles that are highly variable in methodology and focus. This, in combination with the recognition that the area of focus is an emerging area of research within medical education is why I selected a scoping review methodology.

This study progressed iteratively with guidance from the methodological framework proposed by Arksey and O'Malley [82], which is comprised of five stages.

3.2.1 Step one: Identify research question

This review was guided by the research question, "What do we know about the perceived credibility of assessment-generated feedback in the published medical education literature?"

3.2.2 Step two: Identifying relevant studies

Due to the scope of the topic under exploration, I decided to focus the search of relevant studies on works published in peer-reviewed journals. I did not include grey literature, as I believed that the topic of interest would be reflected in the formal literature, is amenable to formal study, and therefore, likely to be contained within the formal research literature. Thus, it was believed that the value of the grey literature would not be sufficiently high enough to justify the "increased costs of securing these difficult-to-locate studies" (p. 257) [91].

In collaboration with an academic medical librarian (NT), a search strategy was developed encompassing three key concepts: assessment, credibility, and medical education. These three key concepts were combined with AND to find all relevant research encompassing them together. Furthermore, relevant controlled vocabularies (i.e. MeSHs, Emtree, subject headings, index terms) and keywords (i.e. synonyms, related terms, and/or spelling variations) were identified for each key concept and were combined with OR, to ensure relevant studies were identified.

The search strategy was adapted and implemented on five databases: MEDLINE (Ovid), PsycInfo (Ovid), Scopus, EMBASE (Ovid), and ERIC (EBSCO) (See appendices A – E for full search strategy adapted to each database). I limited the search to studies published between 2000 to June 22, 2017 (date of the search execution). I chose to anchor to 2000 as this time period represented a shift from literature discussing assessments as a means to measure performance to a discussion of the educational benefits of assessment (i.e. assessment *of* learning to assessment *for* learning) [21].

To best answer the research question, eligibility of studies was not restricted by methodology (e.g. qualitative designs, quantitative designs (observation studies, randomized controlled trials, cohort studies, cross-sectional studies, longitudinal studies, etc.)) or by publication type (e.g. commentaries and conference proceedings, etc.). However, only articles published in English or French were included due to the linguistic competencies of myself and my thesis advisory committee. To ensure all relevant articles were identified, citation tracking of key articles was employed.

3.2.3 Step three: Study selection

In close collaboration with my principal supervisor (MY), I developed inclusion and exclusion criteria, based on the research question, at the outset of the project (Table 1: Eligibility

criteria). We established *a priori* that a 90% threshold for agreement rate between myself and my principal supervisor (MY) must be met after every batch of 250 abstracts screened, otherwise we would re-evaluate the eligibility criteria and discuss how these criteria were being applied.

Table 1: Eligibility Criteria

Inclusion Criteria	**Exclusion Criteria**

- Setting: medical education
- Population: medical trainees (e.g. medical students, residents, fellows)
  - If studies include medical trainees as well as other health professions (e.g. dentists, optometrists, nurses, etc.), it will be included.
- Content:
  - Must discuss assessment of individual medical trainees by faculty members, clinicians, residents, patients, peers, medical boards.
  - The perception of credibility must be related to the assessment (e.g. written exam, performance-based assessment, etc.).
	- Exclusively reporting on a non-medicine trainee population (e.g. dentistry, veterinary, optometry, pharmacy, nursing, physical/occupation therapy, etc.).
	- Studies examining assessment of patients, medical programs, hospital policy, physician workload, stress, hospital administration, medical program, teaching, etc.
	- Studies of assessment of physicians not undergoing training (i.e. continuing professional development).
	- Studies focused on program evaluation (i.e. participant's perceptions were about the programs rather than the assessment itself).

Also, under her close supervision, we performed a pilot test of the eligibility criteria by screening 100 of the same articles independently. Following the pilot test, my principal supervisor (MY) and I determined that our eligibility criteria did not adequately map back to our key concepts. Therefore, more flexibility was added into the inclusion criteria of "perception of credibility" such that we included synonymous terms (e.g. valuable, useful, helpful, etc.) that we identified in the abstracts reviewed in the pilot test.

My principal supervisor (MY) and I, then, independently screened all titles and abstracts applying the eligibility criteria outlined above (Table 1) [86]. Title and abstract screening was conducted on the web-based screening application, Rayyan (http://rayyan.qcri.org). In cases of disagreement, a member of my thesis advisory committee (CSTO) acted as a third reviewer to resolve discrepancies. I calculated raw percent agreement after every batch of 250 articles was screened. This was used as a measure of inter-rater reliability [92].

I exported articles that passed the first stage of screening from Rayyan to EndNote X8.0.2, a citation management software. I, then, independently screened all full-text articles. As I screened the articles for inclusion or exclusion, I sorted them into three groups (i.e. "Include in synthesis", "Exclude from synthesis", or "Maybe include in synthesis"). My principal supervisor (MY) screened all articles classified as "maybe" to assist in final decision making. To further increase transparency and scientific rigor, after I screened each batch of 100 full-text articles for inclusion, my principal supervisor (MY) also screened 10% of these full-text articles; discrepancies between us, once again, were resolved with the support of another member of my thesis advisory committee (CSTO).

3.2.4 Step four: Charting the data

Because this was a scoping review, this step involved extracting relevant key pieces of information from the research material under review. Given the focus of this review investigated how perceived credibility was described in the literature, relevant material was extracted from the entire article i.e. from the abstract to the conclusion.

A first iteration of the data extraction form was developed, and pilot tested with five articles. Following this exercise, my principal supervisor (MY) and I reviewed the extracted data and adapted the extraction form to better capture the richness of the data. The second iteration of

the data extraction form recorded the following information for each included article: first author; journal; year of publication; geographical location of study (continent); study design; methodology (e.g. qualitative, quantitative, mixed methods); data collection method; data analysis; type of participants; research objectives; types of assessment; whom provided the assessment; type of feedback provided; use of term "credibility" (yes/no), if "credibility" was not used which term was used to refer to the construct; definition of credibility; factors that affect credibility; and relevant quotes.

To ensure consistency of the data extraction, my principal supervisor (MY) independently extracted and recorded the data for a subset of the articles (n = 5). We then met to discuss the extracted data. After this exercise, we accepted this iteration as the final version of the data extraction form. From this point forward, I extracted data from all the remaining articles, and my principal supervisor (MY) co-extracted 5% of the articles and verified the extracted data. To further ensure scientific rigor, my co-supervisor (CR) and one other member of my thesis advisory committee (PT) extracted a subset of articles (n = 4) that varied in methodology, type of data collected, and country of publication. All four members (SL, MY, CR, PT) then met to discuss the extracted data and to ensure that the findings were congruent. Disagreements were rare and consensus was achieved through discussion and re-extraction of the original articles.

3.2.5 Step five: Collating, summarizing and reporting the results

The data synthesis method was comprised of three steps: data analysis (quantitative bibliometric descriptive and qualitative thematic analysis), reporting of findings, and discussion of the implications of the findings.

3.3 Data analysis

3.3.1 Quantitative analysis

This descriptive bibliometric analysis described the nature and distribution of the studies (number of studies, study design, year of publication, continent and country, study population, and methodology). The data gathered from the analysis was presented graphically and with descriptive statistics.

3.3.2 Qualitative thematic analysis

All articles that passed full-text screening were uploaded verbatim onto QSR's NVivo Software to facilitate qualitative data analysis. I applied the methodological framework for thematic analysis described by Thomas and Harden [93]. This method was appropriate as the research objective was to broadly map the medical education literature on the perceived credibility of assessment-generated feedback and to determine how notions of credibility of assessment have been represented in the literature. This type of analysis is used to identify, analyze, and report patterns (themes) within and across the data [95]. I defined a theme as a recurrent pattern of information related to the overarching research question (e.g. perceived credibility of assessments in medical education).

I independently coded all the articles using an inductive iterative approach, creating new codes as they arose. This method was comprised of three steps, which proceeded iteratively. In the first step, I freely coded the information in each article line-by-line, generating new codes as necessary, while ensuring to keep these initial codes very close to the original data [93]. To ensure rigor and congruency of findings, my two supervisors (MY, CR) and a third member of my thesis advisory committee (PT) independently coded a subset (n = 3) articles. Following this exercise, I met with the members of my thesis advisory committee (MY, CR, PT) to discuss the codes and

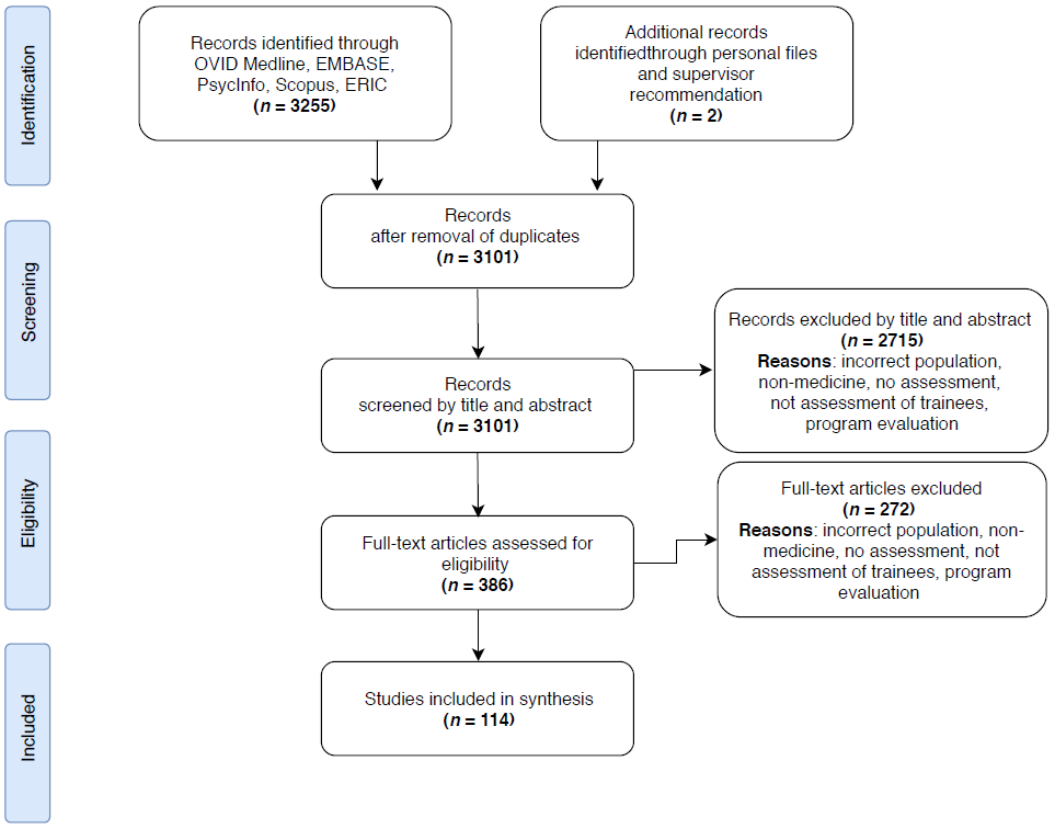
confirm whether they reflected the core research objectives. For the second step, I organized the codes identified in step one into descriptive themes. In this stage, I identified similarities and differences between the unique codes and categorized them hierarchically – resulting in the creation of descriptive themes. In the final step, I generated analytical themes by inferring elements of assessment, feedback, and credibility from the views expressed by medical trainees in different forms of assessment captured by the descriptive themes.

4.0 Results

4.1 Search results

The search strategy was adapted and executed in: OVID Medline, EMBASE, PsycInfo, Scopus, and ERIC (See appendices A-E for full search strategies as adapted to relevant databases). The search strategy, in combination with articles from my personal files and recommendations from my supervisors, yielded a total of 3255 articles. Upon removal of duplicates, 3101 unique articles remained to be screened by title and abstract by my principal supervisor (MY) and myself. There was a 92.3% agreement rate between myself and my principal supervisor (MY) for title and abstract screening. After applying the eligibility criteria, 2715 records were excluded leaving 386 records to undergo further full-text review. Of these, 269 records were excluded due to incorrect population (i.e. not medical trainees), non-medical context, and lack of focus on assessment. The remaining 114 articles met the inclusion criteria and were included in the synthesis (Figure 4).

Figure 4: PRISMA diagram



4.2 Characteristics of included literature

All articles included in the synthesis described medical trainees' perspectives of assessment. Details such as author, journal, year of publication, and research objectives of all included articles can be found in Appendix F.

4.3 Bibliometric analysis

Table 2 presents detailed statistics on bibliometric data describing the corpus of literature included in this review. Studies included were published between 2000 to June 22, 2017, with a recent increase in the number of publications (Figure 5). Included literature was drawn from 54 journals, with the most articles published in *Medical Education* (n=21; 39%) and *Medical Teacher*

(n=17; 31%). There was diverse geographical representation of the included studies: Europe (n=42), North America (n=34), Asia (n=24), Oceania (n=10), Africa (n=3) and South America (n=1). The majority of the records were primary research studies (n=111) with one commentary [96], one literature review [97], and one systematic review [98]. The synthesized records included 57 quantitative studies, 28 qualitative studies, 11 mixed methods studies, and 17 studies using multiple methods. A variety of qualitative data collection methods were applied including: semi-structured interviews (n=13) [99-112], focus groups (n=26) [108, 113-138] and free-text comments from surveys or questionnaires (n=9) [139-148]. Several quantitative data collection methods were applied including: questionnaires [55, 111, 116, 117, 122, 125, 127, 128, 131, 133, 135, 138, 139, 141, 142, 146, 147, 149-185], surveys [118, 119, 124, 136, 140, 143, 144, 186-206], pile-sorting activity [103] and psychometric analyses of assessment data [109, 202, 207]. Many different assessment approaches were represented in the database including: performance-based assessment [102, 107, 111, 119, 130, 133, 137, 142, 147, 148, 152, 153, 155, 157-159, 161-163, 166, 167, 169, 171, 172, 175, 177-180, 184-186, 188-190, 192, 198-200, 202, 204, 206, 208], workplace-based assessment [98, 100, 101, 105, 106, 108, 110, 112, 116, 117, 121, 125-127, 131, 139, 140, 150, 156, 176, 181, 183, 191, 194], and written assessment [96, 97, 103, 104, 118, 135, 141, 143, 149, 151, 152, 155, 160, 163-165, 190, 195, 197, 203, 209]. Participants of the included studies were: medical students (n=72; 64%), residents (n=19; 17%), fellows (n=4; 3.5%), and specialist trainees (n=18; 16%). The majority of assessments were provided by an assessor or an examiner, and assessment-generated feedback was primarily presented as scores [78, 97, 101, 106, 109, 131, 138, 143, 149, 150, 152, 155, 160, 165, 171, 172, 176, 178, 180, 183, 184, 186, 188, 190, 192, 195, 202, 207] or checklists [102, 105, 108, 116, 142, 152, 153, 166, 167, 169, 189, 199, 200, 206].

Figure 5: Number of publications included in my study, presented per year across time.

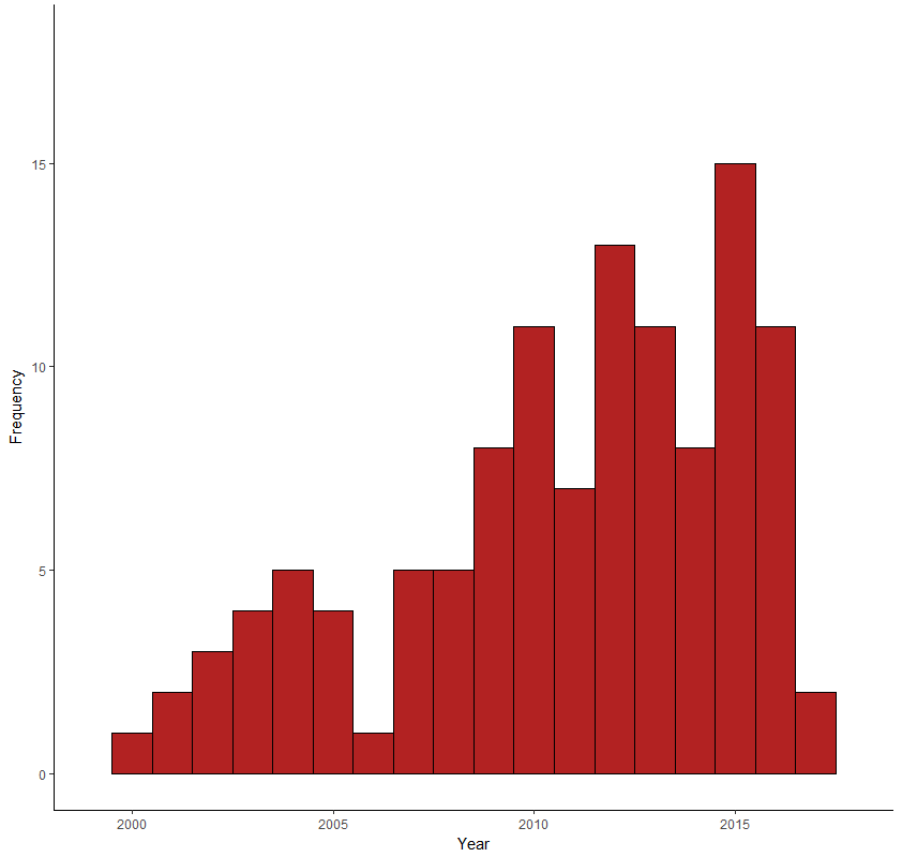


Table 2. Bibliometric details of studies publications included in this review	
Characteristic	**No. of publications (%)**
Continent of publication:	
Europe	42 (37)
North America	34 (30)
Asia	24 (21)
Oceania	10 (8.7)
Africa	3 (2.6)
South America	1 (0.9)
Type of study:	
Quantitative	57 (50)
Qualitative	28 (25)
Mixed methods	11 (9.7)
Multiple methods	17 (15)
Participants: (N_{total} = 14319)	
Medical student (N_{total} = 11088)	72 (64)
Resident (N_{total} = 1204)	19 (17)
Fellow (N_{total} = 83)	4 (3.5)
Specialist trainee (N_{total} = 1944)	18 (16)
Assessment type:	
Performance-based assessment:	
Not specified	25 (15)
OSCE	29 (17)
Simulated patient-based assessment	2 (1.2)
Workplace-based assessment:	
Not specified	5 (2.9)
Mini-CEX	7 (4.1)
Multisource feedback	6 (3.4)
In-training record	5 (2.9)
Supervisor observation	4 (2.3)
Supervised learning events	2 (1.2)
Logs	2 (1.2)
Field notes	2 (1.2)
Case-based discussion	4 (2.3)
Direct observation of procedure skills	4 (2.3)
Written assessment:	
Not specified	10 (5.8)

Multiple choice questions	15 (7.4)
Short answer questions	2 (1.2)
Essay questions	7 (3.4)
Script concordance test	3 (1.5)
Purpose of assessment:	
Formative assessment	8 (4.0)
Summative assessment	1 (0.5)
Other forms of assessment:	
Self-assessment	3 (1.5)
Peer assessment	6 (3.0)
Programmatic assessment	3 (1.5)
Oral assessment	10 (5.8)
Team-based assessment	1 (0.5)
Assessment completed by whom:	
Assessor	66 (56)
Examiner	17 (15)
Medical program	11 (9.4)
Peer	6 (5.1)
Self	5 (4.3)
Tutors	5 (4.3)
Simulated patient	4 (3.4)
Consultant	2 (1.7)
Nursing staff	1 (0.8)
Assessment-generated feedback:	
Type:	
Not specified	16 (16)
Scores	36 (35)
Supervisor-provided	9 (8.8)
Checklist	17 (17)
Ratings	11 (11)
Forms	5 (5.0)
Comments	4 (3.9)
Report	3 (2.9)
Peer-provided	1 (0.1)
Format:	
Verbal	23 (53)
Written	16 (37)
Web-based	2 (4.7)
Audio	1 (2.3)
Video	1 (2.3)

4.3 Conceptualization of credibility

Of the 114 records included in the synthesis, 26 articles used the term ‘credibility’, however, none provided a definition or conceptualization. Despite limited use of ‘credibility’, I identified 29 other terms that were used to describe the same phenomenon (i.e. “perceived credibility of assessment”). The most frequently used terms were: useful, fair, helpful, valuable, valid, appropriate, satisfaction, realistic, and acceptable (Table 3).

To ensure that my findings did not simply mirror the terminology included in the search strategy, I compared the original 17 search terms with the 29 terms found in the literature for ‘credibility’. Table 4 provides a list of all key words used in the search strategy to probe at the notion of ‘credibility’. A total of 17 search terms were used in the search strategy and only six were found in the literature included in this review (Table 3). Therefore, the terms found to conceptualize ‘credibility’ in the literature were not solely the terms used in the search strategy.

Table 3: Other terms for 'credibility'		
Term	**No. (%)**	**Used as search term**
Useful	26 (20)	No
Fair	24 (19)	Yes
Helpful	14 (11)	Yes
Valuable	9 (7.0)	Yes
Valid	8 (6.3)	No
Appropriate	6 (4.7)	Yes
Satisfaction	5 (3.9)	No
Realistic	4 (3.2)	No
Acceptable	4 (3.2)	Yes
Positive	3 (2.3)	No
Authentic	2 (1.6)	No
Effective	2 (1.6)	No
Adequate	2 (1.6)	No
Objective	2 (1.6)	No
Comprehensive	2 (1.6)	No
Accurate	2 (1.6)	No
Relevant	1 (0.8)	Yes
Sufficient	1 (0.8)	No
Appreciated	1 (0.8)	No
Beneficial	1 (0.8)	No
Supportive	1 (0.8)	No
Transparency	1 (0.8)	No
Reliable	1 (0.8)	No
Constructive	1 (0.8)	No
Preferred	1 (0.8)	No
Practical	1 (0.8)	No
Unbiased	1 (0.8)	No
Reasonable	1 (0.8)	Yes

Table 4: List of all search terms used to describe 'credibility'	
Term	**Found in publications included in this review (Yes/No)**
Helpful	Yes
Relevant	Yes
Value	Yes
Acceptable	Yes
Appropriate	Yes
Fair, fairness	Yes
Reasonable	No
Receptivity	No
Counterproductive	No
Justifiable	No
Defensible	No
Legitimate	No
Influence	No
Productive	No
Trust, trustworthy, trustworthiness	No
Merit	No
Applicable	No

4.4 Thematic analysis

The first stage of the thematic analysis, line-by-line coding, generated 994 descriptive codes. These descriptive codes were then organized hierarchically and categorized into 51 descriptive themes. To gain a richer portrait of the data, these descriptive themes were further organized into three overarching themes: (i) characteristics of good assessment, (ii) consequences of an assessment being perceived as credible (*or not*), and (iii) factors that affect the perceived credibility of assessment.

4.5 Theme 1: Characteristics of good assessment as identified by trainees

Through synthesis of included records, I identified three trainee-generated subthemes describing characteristics of good assessment. These subthemes indicate that good assessment should: (i) identify weaknesses, (ii) identify good future physicians, and (iii) have educational value.

4.5.1 Identify weakness

Trainees valued assessments perceived to accurately identify their weaknesses and strengths, as they perceived it to be a useful guide for their improvement. This perception was identified across a broad range of assessment approaches including portfolios [210]; workplace-based assessment [101, 131, 181], ‘‘[i]f you do a mini-CEX it may expose significant deficiencies in knowledge or other professional attributes, … it is good because it’s better to expose those weaknesses than to cover them up and never to improve” (Trainee 4, p. 1349) quoted from [101]; performance-based assessment [142, 147, 154, 159, 172, 192, 200], “90.8% of the examinees reported that OSCE provided an ample opportunity to learn and compensate for areas of clinical weakness despite a huge stress factor” (p. 206) [159]; and written assessment [135], “[a]cross all years students at School B, where feedback is

given on [Progress Test] performance by subject area, were significantly more likely to agree that the [Progress Test] helped then improve their knowledge and monitor how it was improving" (p. 579) [135].

4.5.2 Identify good future physicians

Assessments were also valued if trainees believed it could identify good future physicians. From the trainee perspective, an assessment should be able to distinguish between performers [127], recognize borderline trainees [156, 198], and to predict a trainee's future performance as a physicians [78, 155, 172], ultimately identifying excellence and achievement [125, 131]. For instance, "[s]tudents felt in control if feedback from low-stakes assessments appeared predictive of future performance in high-stakes assessments" (p. 279) [78].

4.5.3 Educational value

Trainees also perceived educational value to be an important element of good assessment, "[a]ny assessment process should also have educational impact" (p. 579) [172]. An assessment was perceived to have educational value if it could drive learning, improve trainee performance, encourage positive development of a trainee as a learner and as a future physician [194], and be used as a gauge for readiness for independent practice, "I didn't get the epidural right and I didn't pass. But it's ridiculous that I did 3 epidurals alone on the same call... you should not be allowed to do anything on your own before you have passed" (Trainee 11, p. 772) quoted from [110]. When assessment was perceived to lack these qualities, trainees dismissed and disengaged with the assessment process, "[w]ithout a perceived educational benefit experienced by trainees, [performance-based assessments] simply become a paper exercise" (p. 447) [161].

4.6 Theme 2: Consequences of an assessment being perceived as credible (or not)

Results from this review also identified a number of consequences that arose when an assessment was perceived as credible and when an assessment was *not* perceived as credible. These consequences were consistent across assessment approaches, level of training, and geographical location.

4.6.1 Consequences of an assessment being perceived as credible

The literature included in this analysis suggests three consequences that arise when trainees perceive an assessment to be credible: (i) positive effect of assessment, (ii) positive effect of feedback, and (iii) positive views and effect of scores,

Positive effect of assessment. Assessment perceived as credible by trainees are best able to drive learning and improve performance; documented across assessment of a wide variety of skills including clinical skills [98, 118, 161, 173], communication skills [182, 200, 202], professionalism [183], self-directed learning [129, 182], and consultation skills [172]. When perceived as credible, assessment lead to positive effects in trainees, such as development of competence [155, 196], positive skills development [101, 116, 117, 119], increased knowledge [99, 135, 167, 184], and improved self-reflection [109, 210]. As one medical student participant demonstrated, "[a]lthough my initial goal was to become skilled in the use of our portfolio system, I now view this as a tool that fostered my development toward becoming a reflective practitioner" (Medical student 3, p. 224) [210].

Positive effect of assessment-generated feedback. When assessment is perceived as credible, feedback can have positive effects such as improved performance [100, 111, 120, 128, 130, 134, 188] or behaviour change [124, 132]. Under these circumstances, trainees are more likely to view assessment-generated feedback as ongoing appraisal of behaviour that highlights their strengths

and weaknesses suggesting areas for improvement [103, 124, 177, 181]. As stated by one trainee, "…got feedback from an observed consultation that I needed to work on certain areas of history taking, this focus has enabled me to improve skills…" (Trainee ID 840, p. 717) quoted from [128].

Positive views and effect of scores. Trainees will also have positive views of their scores when they perceive an assessment credible. Under these circumstances, trainees will consider scores to be evidence of mastery [78, 201], a measurable representation of progress [106, 128, 201, 210], motivation for improvement [106], and/or an indicator of how much one has or could improve [106, 151]:

> *"I prefer to have grades so I can see what my abilities are at the moment and what I need to improve on. Grades give me reassurance if they are good or motivate me to work harder if they need improvement" (Trainee F14, p. 316) as quoted in [106].*

4.6.2 Consequences of an assessment not being perceived as credible

This review also identified three consequences of an assessment *not* perceived as credible were also identified: (i) assessment perceived as 'hoop to jump through', (ii) gaming and manipulative behaviour, and (iii) negative views and effects of scores.

'Hoop to jump through'. When perceived as not credible, trainees tended to view assessments as 'hoops to jump through' [99, 101, 117, 132] or 'checkbox exercises' [105, 116, 156, 161] rather than a robust system for learning and feedback provision [131]. This view arose from assessments that were perceived to have no clear educational purpose, as evidenced by trainees, "…general view amongst the trainees and consultants is that the current system of [workplace-based assessment] is a relatively pointless, 'tick-box' exercise…education and training is a complete afterthought in the current system…" (Unidentified trainee, p. 579) quoted from [116] and felt to be "just another hoop that the College has established for us all to jump through" (Trainee 9, p.

1348) as quoted from [101].

Gaming and manipulative behaviour. The ability of trainees to engage in gaming and manipulative behaviour during an assessment lead to it being perceived as not credible. This type of behaviour broadly encompassed behaviour that undermined an assessment or interpretability of an assessment score. For instance, during workplace-based assessments [100, 101, 105, 127, 131, 156], trainees would engage in gaming behaviour by purposefully selecting easier-scoring assessors [100, 101, 105, 127, 131] or less difficult cases on which to be assessed [101, 131]. Trainees felt that, "[t]he whole [assessment] tool is completely flawed because you choose your assessors" (Trainee F/ST2/A, p. 95) as quoted from [131] because "[m]ost trainees are going to be able to game the system in order (to) find consultants (supervisors) who give them an easy ride" (Trainee 2, p. 1349) as quoted in [101].

Negative effects and views of scores. Assessments not perceived as credible led trainees to develop negative views of scores [106, 112, 131]. Under these circumstances, trainees viewed scores as demoralizing [106] and harmful to their self-confidence, "… if I'm graded badly, rather than seeing it as a reason to try harder, I'll be demoralised and unwilling to try the skill at all" (Trainee M22, p. 49) as quoted in [106]. Scores will be unlikely to be used to support performance improvement.

4.7 Theme 3 Factors that affect the credibility of assessment

Three major subthemes were identified regarding factors that affect the perceived credibility of assessment: (i) assessment process, (ii) trainee characteristics, and (iii) contextual factors.

4.7.1 Assessment process

Synthesis of included studies identified four elements of the assessment process that influenced trainees' perceptions of credibility: (a) assessor and feedback provider, (b) procedures of an assessment, (c) scoring, and (d) inferences based on scores.

Assessor & feedback provider.

Assessor and feedback provider included aspects of: (a) trusting relationship with supervisor, (b) interest in long-term trainee progress, (c) lack of experience/training with assessment, and (d) respect.

Trusting relationship with supervisor. Most trainees, irrespective of their level of training or speciality, perceived an assessment and assessment-generated feedback as credible if they had a strong and trusting relationship with the individual who provided it [106, 112, 121]. This favourable view was not limited to supervisors, but also extended to assessments made by peers [130]. This finding was clear across all forms of assessment and indicates that medical trainees were accepting and highly responsive to any form of assessment-generated feedback, be it positive or negative, if there was a trusting relationship with the individual assessing their performance. For instance, "[y]ou look for assessors that you know are knowledgeable and where you get something out of it, a good dialogue or really learn something. Not just marks on a sheet of paper" (Trainee 12, p. 773) quoted from [110] and:

> *"It was someone who had observed me for three weeks.... He saw me with families . . . and there were ethical issues that crept up and he saw me handle those, and so he could make an accurate interpretation… I respected him as an evaluator because he took the time to do it. And you can tell who takes the time to get to know you and observe you. And you can also tell who doesn't, and who's just filling these in at the end of the day with no comments, just bubble marks. And so that was a good evaluation I think because you knew the person and made an effort. And so you respect what they told you in the end" (Resident 7, p. S99) as quoted in [112].*

The inverse was also true, trainees regularly ignored and discounted feedback from individuals who were less familiar with them or their skills [116, 121, 127, 170].

Interest in trainee long-term progress. Additionally, trainees' valued supervisors who they felt were genuinely interested in their long-term progress [120] and familiar with their performance [103, 120]. This preference encompassed supervisors who were engaged and enthusiastic [161] about their supervisory activities and actively observed their trainees:

> *"I do not have the impression that my supervisor is well informed on how I'm progressing in my training. I find that a supervisor should be interested in his trainees and should be well informed on their progress and which competencies they have" (Trainee 5-D) as quoted in [120].*

Supervisors were perceived as not credible when they did not actively observe their trainees or made judgments about performance based on insufficient observations [105, 116, 121, 127]. Supervisors who provided a time and space for personalized [114, 121, 124], specific [100, 102, 103, 121, 124], and actionable assessment-generated feedback [114, 128, 177] were valued, as evidenced by trainee statements: "[t]here should be a dialogue between my supervisor and myself about my performance on the activity" (Trainee SIU, p. 260) quoted from [121] and "[i]n general, the comments I found unhelpful were vague – there was nothing the student could take away, no examples of how the student was doing well" (Unidentified medical student, p. 5) quoted from [103].

Lack of experience/training with assessment. Trainees' perceptions of credibility were negatively affected when an assessor lacked training and/or experience with the assessment process. This perception was present when an assessor was unfamiliar with how to implement the assessment process [114] and unsure about how to properly evaluate competence [125]. This was

most apparent in performance-based assessment [102], workplace-based assessment [100, 110, 125, 127, 136], and portfolios [125, 131, 194]:

> *"…trainees felt that in order to generate accurate scores, assessor training was required, comparable with examiner training for College Fellowship examinations. They felt assessor scoring needed calibration, if consistent scores were to be obtained. They noted some specialists rated the mini-CEX items 'good, good, good,' (T) without justifying their score or discussing their feedback with the trainee." (p. 526) [136].*

Respect. Trainees' perception of the supervisor providing the assessment and/or feedback was another important feature in the perceived credibility of assessment and assessment-generated feedback. Trainees reported valuing and preferring assessment-generated feedback from a physician they respected – this respect arose from both the physician's clinical skills [114] and teaching abilities [120, 134]. Trainees also stressed the importance of supervisors who wanted to improve their own teaching skills [120, 134], as trainee 5-C quoted from [120] "[y]ou can notice which supervisors are really teaching-minded: they tend to do teach the teacher courses, prepare themselves and give structured feedback."

Procedures of an assessment. The major factors that affected trainee perceptions of credibility of the procedures of an assessment were: (a) standardization, (b) purpose, (c) clinical relevance and authenticity, and (d) timing.

Standardization. Trainees perceived standardized assessment and assessment-generated feedback as more credible than non-standardized forms [78]. Trainees raised concerns regarding the lack of standardization and structure of assessment methods such as workplace-based assessment [166] or performance-based assessment [192], but lesser so for written forms of assessment. For instance, trainees stressed the importance of being assessed in a uniform manner [193] and having their performance evaluated against explicit standards [78, 130, 134, 136, 181]: "[p]erformance relative to your peers is very important ... it gives you something to sort of work

at ... that's actually quite a powerful motivator" (Unidentified trainee, p. 527) as quoted in [136], and:

> *"It was quite interesting to be able to compare how well you thought you were with the rest of the people in your group because otherwise you have no sort of standard apart from the doctors to measure yourself against, because obviously you're not going to be as good as doctors." (Medical student, year 4, p. 872) as quoted in [130]*

Trainees also felt that unstructured assessments were unfair [140] and less representative of their performance, "[s]tudents whose assessors used the Structured Question Grid believed that the assessment result was less representative of their ability than students whose assessors did not" (p. 51) quoted from [207].

Clear purpose. The perceived credibility of an assessment was also largely dependent on the clarity of its purpose. Trainees perceived assessments to be more meaningful when they understood its purpose [102, 194], which lead them to engage more with the assessment process, "[w]hen the learner understood its purpose, he or she would buy into it and, consequently, the element would become meaningful to learning" (p. 495) [104]. However, when trainees were confused or unclear about the purpose of an assessment, they tended to dismiss its value [116]: "…their [structured learning events] role is unclear. Trainers or trainees don't seem to [*sic*] able to clearly define what constitutes SLEs…" (Unidentified trainee, p. 580) [116], and supported by additional findings, "…[Trainees] felt that [mini-CEX's] success as an educational tool was limited by lack of understanding of its contents and purpose" (p. 4) [98].

Clinical relevance. The clinical relevance of an assessment was another important influence on the perception of its credibility. Trainees valued clinically relevant assessments because they provided opportunities for practicing clinical skills in authentic scenarios [133, 142, 167, 192] that replicated real-life clinical care [114, 118, 198]. These assessments were viewed as

opportunities to demonstrate clinical competence: "[m]ake the primary exams more clinically relevant as they seem very IRRELEVANT when you are sitting them and would rather be learning about information and procedures you need in your daily practice" (Provisional trainee, p.543) as quoted in [118].

Timing of assessment. Lastly, the timing of an assessment also affected how a trainee perceived its credibility. This subtheme encompasses both the time at which an assessment is given during training and time constraints of an assessment itself. Perceived credibility of assessment increased when the assessment was believed to be relevant and appropriate to the curriculum [129, 142, 169, 173, 184] and level of training [142, 148, 194]. Kalet al. [194] reported that trainees felt it was a poor use of time to be assessed on skills they had not yet been exposed to:

> *"There is not enough exposure to issues where professionalism comes up during first and second year to warrant that much amount of reflection...I honestly think it's a good thing to want to address the issue of professionalism; but that the portfolio is just not a good way to do so. It doesn't actually help assess us based on our professional behavior; that will mostly come from being on the wards in third and fourth year" (Unidentified trainee, p. 1071) as quoted in [194].*

In addition, certain performance-based assessments (e.g. OSCE, simulated clinical examination) [148] were requested earlier in training to optimize learning potential and identify areas for improvement: "after talking to residents who did not get this opportunity (end-of-life care assessment), they were thrilled that medical schools are starting to teach this, as they (residents) had to experience it in real-life without this kind of training" (Unidentified medical student, p. 261) as quoted in [202], and "[t]he timing of the SCE at the beginning of PGY1 training was also described by some residents as a helpful refresher because they hadn't 'been around patients for a number of months' since completing their undergraduate medical education" (p. 407) [119].

Several studies also reported trainees raising concerns about the time allotted for certain

assessments [118, 133]. Most studies exploring trainee perceptions of the OSCE found that trainees felt there was insufficient time [102, 142, 153, 169, 192, 200, 203, 204], which lead to augmented levels of stress and pressure. For example, a Brazilian study reported "70% of respondents were discontent with the time available at each station" (p. 13) [153], a Jamaica study found, "most (70%) felt that they needed more time to complete the stations" (p. 4) [142], and from an Iranian study, "more than 64% were not satisfied with the time allocation for each station" (p. 190) [169].

Scoring. Two factors were found to influence perceived credibility of scoring: (a) standardized scoring, and (b) variability across assessors.

Standardized scoring. Regardless of assessment method, trainees responded most favourably to scores that were standardized as they were believed to be most representative of their performance [142, 192]. Lack of standardized scoring was an issue primarily raised with performance-based assessments [142, 192] and workplace-based assessments [136, 139, 140] as evidenced by a participant statement "[t]he main problem is the numerical marking. There is no consistency between doctors, some give all 10s, others refuse to give more than a 6. I think they should be changed so the only grades are fail, pass, clear pass" (Unidentified medical student, p. 402) quoted from [140]; however, one study identified similar concerns on a written assessment (in-training examination) [139]. For performance-based and workplace-based assessments, this concern was strongly linked to perceived biased introduced by trainees selecting their own assessors:

> *"[b]eing honest, you do select people that you get on with. If I'd had a problem with somebody I wouldn't give them a form and whether that makes them valid ... well it doesn't make them valid does it because that person's opinion might be quite important as part of the process"(Trainee 5, p. 1000) [100].*

Variability across assessors. Concerns were also raised regarding variability in assessor's scores due to their own unique standards of evaluation [110, 154, 161, 207], "[i]f you have a different assessor each time they might not have the same standard" (Unidentified trainee, p. 526) as quoted from [136]:

> *"I've had a change of tutor since the first and second weeks and I feel like there would be a lack of consistency which would be reflected in my grades so I would rather in this case not have grades than have a grade which isn't necessarily reflective of how I may have improved. But if I had had the same tutor, I would have chosen to have grades" (Trainee F19, p. 313) quoted from [106],* and
>
> *"They've got all different opinions about what's good. So then one counsellor comes and says, 'No, you should change this.' And then next you get another counsellor who's also going to check your portfolio and then suddenly it's all wrong and you've got to change it back" (Trainee FG2, p. 281) quoted from [78].*

Inferences based on scores. Two factors were found to influence perceived credibility of assessment: (a) scores explained with feedback, and (b) consequences of suboptimal performance.

Scores explained with feedback. Trainees preferred scores accompanied by feedback that explained or contextualized their performance [131], and ideally provided guidance on how to improve, as evidenced by two trainees from two different studies, "I think the facilitator should spend time to explain to the student regarding [problem-based learning] assessment, because the marks given in [problem-based learning] assessments are just marks" (Unidentified trainee, p. 397) quoted from [141] and:

> *"Everyone has to reach proficiency so I wouldn't mind having grades, but I don't find grades useful unless there is detailed feedback. I can see how they are useful in terms of knowing where you are with respect to the exit ... without feedback I can't use the grade. It merely demotivates me. I don't expect to be perfect but I need the feedback to explain the bad grade" (Trainee F21, p.313) as quoted in [106].*

Consequences of suboptimal performance. Furthermore, assessments were perceived to be more credible when there were clear consequences of suboptimal performance [113]. Some trainees felt assessments with no consequences limited potential for learning, "[t]he only thing is, if you are being assessed with the purpose to stimulate learning and the result of the assessment is without consequences, the impact will be disappointing" (Trainee 6-C, p. e1399) as quoted from [120]. This perception was true not only of supervisor-based assessment, but also peer assessment, "... [m]aybe a student would realize that someone's actually paying attention to their behavior . . . that affects their grade and that would be the motivation for them" (Trainee 3-A, p. 821) as quoted from [113].

4.7.2 Trainee characteristics

Three elements pertaining to trainee-related characteristics were identified to influence perceived credibility of an assessment: (a) trainees' traits and preferences, (b) trainee experience, and (c) social pressures and influences.

Trainees' traits and preferences. This subtheme identified two factors: (a) trainee preferences, and (b) trainee's level of training.

Trainee preferences. Each trainee has their own individual personality that may affect how they perceive the credibility of assessment. For instance, mastery-oriented trainees preferred formative assessment and self-assessment due to provision of feedback that guides improvement, "[t]o me, all feedback is valuable; I think you can use all information one way or another on your way to medical expertise" (Trainee PG2-P, p. S68) quoted from [134]. On the other hand, performance-oriented trainees preferred rigorous assessment methods with clear consequences, as Trainee PG3-P stated, "I would prefer good old knowledge exams: Clear study materials, clear pass/fail standards and clear consequences. It helps me to start studying and in this way I know

once I have mastered a subject" (p. S68) as quoted in [134].

Trainee's level of training. A trainee's level of training also influenced their perceived credibility of an assessment and their subsequent receptivity to assessment-generated feedback [128, 135]. A clear difference was found between how junior and senior medical trainees perceived assessment and assessment-generated feedback. For instance, junior trainees did not like peer assessment, as it was felt to be less reliable than feedback from a supervisor, "…the feedback they gave (academic) was what I took away rather than my class mate's" (Medical student 12, p. 205) as quoted from [115]. Additionally, junior trainees felt that it may be difficult for their peers to be truly objective when evaluating their skills, "…[a peer would be]…probably not as reliable …especially if I didn't know them well, because you don't want to be harsh and you don't want to upset them…" (Trainee A1h, p. 718) quoted from [128]. Senior trainees, however, more often did find value in peer assessment as it was perceived to be helpful to hear from someone who could truly empathize with their feelings [127, 130]:

> *"I valued what they [peers] said as much, if not a bit more sometimes than what the other person [GP], because they knew exactly what I was going, you know, you're going through the same things, so for them to say something which is good, you know, it was quite good, you had to say something positive first or whatever and then, but then even like the things they said you could improve on you actually took on board" (Medical student 5, p. 872) quoted from [130].*

Senior trainees also appreciated peer assessment for its immediacy and the ability to follow-up with in-depth discussion:

> *"…if you didn't know it you could ask your partner and maybe learn it together …I think it reinforces self-reflection or like peer reflection and then if you both are in touch with your tutor... it reinforces… if you have a problem and talk about it aloud then it seems to make more sense just to get somebody receptive to it and they give you a way to think it through…" (Trainee A5a, p. 718) quoted from [128].*

As trainees progressed from junior to senior trainees, a developmental shift may be occurring from passive reception of feedback (e.g. expecting supervisors to inform if they are meeting standards) to active reception, in which the feedback is used to adapt learning strategies in order to improve performance [120, 128]. This shift was documented in trainee perceptions of feedback, such that junior trainees wanted positive feedback to affirm their performance and were demoralized by negative feedback, "…if I'm doing something and if someone gives me positive feedback that makes me try harder and motivates me more, if someone gives me negative feedback I sort of get downhearted" (Trainee A1c, p.718) quoted from [128]. On the contrary, senior trainees saw greater value in negative feedback as it could be used to improve performance [128, 131, 211], for example: "[i]t's nice to know what you did wrong so you can do better the next time… Tell us what we're doing wrong, just about anything…cause there must always be room for improvement… 'Do feel free to be harsh!'" (Trainee B5, p. 4) as quoted in [114].

In fact, senior trainees felt that positive feedback was less meaningful because "…positive feedback can make you complacent" (Trainee A3a, p. 718) quoted from [128] and it did not always provide actionable steps for improvement, "[i]t means very little to me to always get these 'great job, great job, great job' versus someone who is trying to find ways to help me get better" (Trainee FG2, p. 281) quoted from [78].

Trainee experience. Each trainee has their own unique experience with an assessment and that influences how they perceive an assessment's credibility. Four factors describing the trainee experience were identified to affect credibility perceptions of an assessment: (a) influences on trainee behaviour, (b) stress, (c) trainee preferences of assessment and feedback, and (d) interaction with assessment.

Influences on trainee behaviour. Certain characteristics of an assessment influence how

a trainee will engage and whether it will be perceived as credible. For example, during observation-based assessments such as performance-based assessment [78], workplace-based assessment [126, 136], or team-based assessment [145], trainees report altering their behaviour because they were being assessed, one trainee reports, "[y]ou do it the way they [Specialists] would do it to make them happy" (Unidentified trainee, p. 527) quoted from [136]. Additionally, familiarity with assessment format or content sometimes led trainees to strategically study for the assessment, as opposed to studying for learning and improving performance [99, 105, 130, 149, 173, 210], as stated by one medical student, "[i]n previous grade-based systems, I found myself slavishly studying material that was assigned by the professor in order to get a high letter grade or percentage on the next test" (Medical student 2, p. 223) quoted from [210]. Trainees also tended to dismiss assessment-generated feedback when it was inconsistent with their self-assessment [112, 132].

Stress. Trainees perceived certain methods of assessment to be more stressful than others. Assessments involving an observation component such as performance-based assessment (e.g. OSCE), workplace-based assessment (e.g. mini-CEX [126], and case based discussions [140]) were most often reported as stressful. Oral examinations were also perceived as stressful [55], but to a lesser extent. Medical students from 10 separate studies [142, 152, 153, 157, 159, 166, 167, 169, 171, 200, 204] found that the OSCE to be a stressful experience. The OSCE was also reported to be more stressful than other traditional assessment methods such as written assessment, as one medical student participant described, "...with the nervousness, I think it's just looking or feeling stupid in front of the patient or in front of the doctor which in a written paper nobody knows what you're writing" (Trainee S6, p. 769) quoted from [171].

Trainee preferences for assessment approach. In addition, certain methods of assessment were preferred over others. As previously mentioned, trainees tended to prefer

assessments with clinical relevance that allowed them to practice and demonstrate their clinical skills; thus, despite being perceived as the most stressful, OSCEs were the most preferred assessment compared to other assessment methods such as written assessment [166, 167], oral examinations [167], and long cases [157, 166].

Trainees also reported preferences regarding form of feedback. In general, written feedback [100, 103, 128, 131, 141, 145, 148, 151] was the most preferred form of feedback and most often cited as helpful, following any method of assessment. Trainees reported a lack of written feedback to be frustrating [105]. However, some trainees felt written feedback was not enough and that supervisors should also provide a verbal explanation of their performance, "[j]ust written feedback is not complete. The supervisor should write down the feedback and provide an oral explanation" (Trainee FVMU, p. 260) quoted from [121]. Verbal feedback was also felt to be more memorable, "[y]ou remember it much longer than when someone tells you" (Unidentified trainee, p. 179) quoted from [107].

Social pressures and influences. Certain social pressures and influences were also found to affect how a trainee interacted with assessment [104, 171]. Trainees sometimes found themselves in uncomfortable situations where they felt social pressures to pass on information about an assessment to their peers:

> *"You know, when you bear in mind that you're going to be working with these people and you're going to rely on these people for favours in the future you know, with getting days off or swapping days, and you know you're going to have to work with them... My year is quite cohesive, we all get on really well and I don't think anyone wants to be seen as someone who doesn't help the good of the year sort of thing"(Medical student, S7, p. 771) as quoted in [171].*

Unfortunately, trainees who refused to pass information were disapproved:

> *"Some people I guess may feel a bit, maybe betrayed, or maybe that's too strong a word but they might feel well, you know, you have that information, we're all doing the same exam, we're all in the same boat, we're all nervous, why can't you pass it on?" (Medical student, S4, p. 772) as quoted in [171].*

4.7.3 Contextual factors

Two contextual factors were identified that influence perceived credibility: (a) context of medical education and (b) cultural element. These factors differ from the previously identified factors that affect the perceived credibility of assessment, because they reflect issues at the level of the program or institutions, and therefore may be more difficult to amend, adapt, or adjust to support the perceived credibility of assessment-generated feedback.

Context of medical education. From the synthesized medical education literature, I identified two factors that encompass the context of medical education: (a) safe learning environment and (b) feedback consistencies. These factors are also important in other educational contexts, outside of medical education.

Safe learning environment. Across assessment methods, trainees perceived assessment occurring in a safe learning environment as credible because it fostered learning [121, 132] and self-reflection [108] and facilitated engagement with assessment and assessment-generated feedback. More specifically, a safe learning environment was described as a learning climate in which trainees felt comfortable to seek help, to admit knowledge gaps, and to openly discuss mistakes; as evidenced by two trainees, "[p]art of a safe environment is also that you are able to mention having difficulties with a supervisor" (Trainee UMCU, p. 260) quoted from [121], and:

> *"Because there is that level of trust within the group, I don't mind my peers knowing that I might not know the answer to something. Because I don't feel they would judge me by the fact I don't know the answer to something that comes up as part of this" (Trainee M2, p. 1217) quoted from [132].*

Consistency of assessment-generated feedback. The consistency of assessment-generated feedback also contributed to trainee perceptions of credibility. Some trainees reported infrequent feedback [128, 136, 141, 168]: "[o]ne thing that's totally lacking in medical training across the board is feedback, and knowing where you are in relation to your colleagues and also what your specialist actually really [thinks]" (Unidentified trainee, p. 527) quoted from [136]. Of the feedback provided, most was judged as unhelpful as it was general and primarily directive:

> *"It is no problem to get some advice of a supervisor on a patient problem, however, usually I get a very directive answer, without him seeing the patient, while I really would like to get some structured feedback after being observed with the patient" (Trainee 3-A, p. e1399) quoted from [120].*

However, other trainees felt feedback content and provision was improving, becoming more specific and clinically focused, "[f]eedback is more focused now, it's better i.e. towards clinical things and being a doctor rather than in previous years where it was more general and theoretical" (Trainee 852, p. 718) quoted from [128]. These incongruent findings suggest that each medical institution has its own culture that affects the provision of feedback and subsequent receptivity by trainees, but one trainee disagreed reporting that feedback was limited across educational sites, "[w]hat feedback? Consistent over several hospitals" (Unidentified trainee, p. 542) quoted from [118].

Additionally, feedback appears to vary by course and year of training, making it difficult to integrate and use for further development as it may not be applicable in future training. These feedback inconsistencies have been identified by trainees as due to a limited of feedback culture within medical education [136].

Cultural element. The final contextual factor that influences trainee interaction with the perceived credibility of assessment and assessment-generated feedback relates to the trainee's cultural background. This element was only reflected in one study included in this review, but I felt it was important to include as a theme as cultural differences have received limited focus in medical education. Suhoyo, Van Hell and Kerdijk [181] found differences in feedback preferences between trainees originating from collectivist countries (Eastern nations that emphasize family and group goals over individual needs) and individualist countries (Western nations that promote individual needs above all else) [181]. For instance, because collectivist nations have a tendency towards modesty, compliment-based feedback is felt to have low educational value, whereas feedback with the goal of correcting errors and behaviour is perceived as more useful to learning [181].

4.8 Concept map

The findings of this scoping review have contributed to the development of a concept map of factors that affect and downstream consequences of the perceived credibility of assessment and assessment-generated feedback by medical trainees (Figure 6). This concept map helps put shape to the findings of this review and suggests that as trainees undergo assessment they come with (i.e. trainee characteristics) and are exposed to factors (i.e. assessment process and contextual factors) that influence their perceptions of credibility of the assessment and assessment-generated feedback. Assessment and assessment-generated feedback perceived as credible lead to different consequences than those perceived as *not* credible. When trainees perceive assessment and assessment-generated feedback as credible, they were more likely to hold positive views of assessment, assessment-generated feedback, and scores. However, when trainees did not perceive an assessment as credible, they were more likely to have negative views of assessment and scores,

and more likely to engage in gaming or manipulative behaviour. In general, how a trainee perceives the credibility of an assessment or assessment-generated feedback influences the likelihood of the assessment supporting learning and performance improvement in addition to supporting positive and proactive student engagement with assessment.

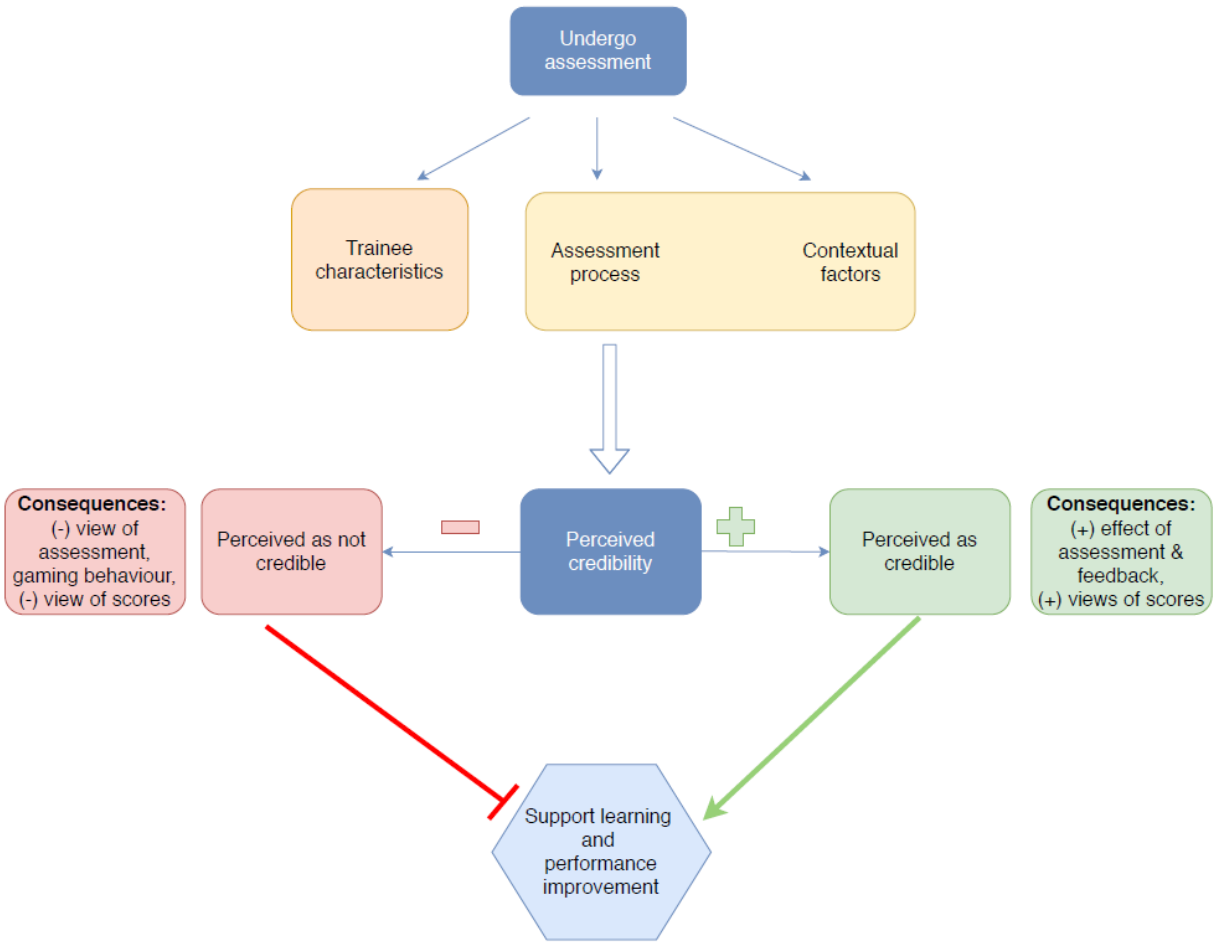


Figure 6: Concept map of perceived credibility of assessment and assessment-generated feedback by medical trainees.

4.9 Summary of findings

Table 5 summarizes the design-related factors (i.e. assessment process and scoring) that influence the perceived credibility of three assessment approaches (written assessment, performance-based assessment, workplace-based assessment) and assessment-generated feedback identified in this review.

Table 5: Design factors that affect the perceived credibility of assessment

		Written Assessment	Performance-based assessment	Workplace-based assessment
Assessment process	**Factors increase perceived credibility:** • + relationship with assessor	N/A	No evidence.	[106, 110, 112, 121]
	• Standardized structure	No evidence.	[130, 136, 159, 166, 192]	[78, 181]
	• Clear purpose	No evidence.	[102, 194]	[98, 116]
	• Clinical relevance & authenticity	[118]	[133, 142, 167, 192, 198]	[114]
	Factors decrease perceived credibility: • Assessors unfamiliar with process	No evidence.	[102]	[100, 110, 125, 127, 136]
Scoring	**Factors increase perceived credibility:** • Standardized	No evidence.	[142]	[106, 136, 139, 140, 207]
	Factors decrease perceived credibility: • Self-selection of assessor	N/A	No evidence.	[100, 127, 131]
	• Scores not explained to learner in any way	[141]	[141]	[78, 105, 106, 131]

5.0 Discussion

The purpose of this scoping review was to map and meaningfully synthesize current published research regarding trainees' perceived credibility of assessment and assessment-generated feedback in the medical education literature. In this section, I will discuss my findings and provide a comprehensive overview of what is currently known about the perceived credibility of assessment and assessment-generated feedback from the trainees' viewpoint. In addition, I will draw attention to knowledge gaps within the literature that were identified through this review, where relevant.

5.1 What defines 'good' assessment in medical education?

Findings from this review seem to suggest that medical learners consider an assessment to be 'good' when it is able to highlight their weaknesses and strengths, forecast their competency as future physicians, and provide educational value - ultimately driving their learning processes and improving their performance. Although criteria for good assessment have been previously defined by other authors [2, 212, 213], the learner perspective rarely takes centre stage. Norcini et al. [2] outlined seven criteria for good assessment, some of which highlight the learner perspective: validity, reproducibility, equivalence, feasibility, educational effect, catalytic effect, and acceptability. Similarities can be found between 'educational effect', which states that an "assessment motivates those who take it to prepare in a fashion that has educational benefit" [2] and the third subtheme identified in this review, i.e. 'have educational value'. As both emphasize the ability of an assessment to drive learning and provide educational benefits to trainees. Additional parallels can be drawn between 'catalytic effect' (i.e. "the assessment provides results and feedback in a fashion that creates, enhances, and supports education; it drives future learning forward" [2]) and the first subtheme of this review, i.e. 'able to identify weaknesses', which refers

to trainee's use of assessment-generated feedback to identify their deficiencies, and be used to improve performance.

The underrepresentation of the trainee perspective in assessment frameworks is at odds with current trends in medical education that foster student-centred learning approaches such as competency-based medical education [214]. The learners' perspective should be considered during the design and implementation of assessment, if the goal of the assessment process is to support learning and improve performance. The characteristics of good assessment identified in this review can be used as a guide in the conception of new assessment frameworks. Future research could explore other characteristics of assessment that trainees perceive to be important and meaningful contributors to their learning.

5.2 Perceived 'credibility' in medical education

The findings of this scoping review indicate that the concept of 'perceived credibility' has become more present in the medical education literature - as reflected in the recent growth in the number of publications. Very few included studies actually used the term 'credibility'; and no record included an explicit definition, instead several terms were identified that referenced the concept of credibility – including terms such as fair, valid, helpful, useful, and valuable. With several terms being used interchangeably, and no explicit definitions, this suggests that perceived credibility may be an emerging concept with work to be done to facilitate more consistent and clear communication around trainees' perceptions of assessment and assessment-generated feedback.

Through the analysis of the articles included in this review, it became apparent that *credibility* and *validity* are terms that may refer to similar concepts, but the responsibility for score interpretation has traditionally been 'housed' in different educational stakeholders. On the one

hand, *validity* focuses on collected evidence that supports the interpretation of scores, which is typically housed in the role of assessment administrators [215]. This means that evidence is collected to support the interpretation of scores as measures of competence, knowledge, skills, etc., and to support the decisions that result from the interpretation of those scores – typically pass/fail decisions or judgments of competence. On the other hand, *credibility* places the 'responsibility' of score interpretation in the hands of individual trainees, who are responsible for interpreting their scores or assessment results as indicators of their own performance or standing, then to identify areas of further study or performance improvement.

In this regard, Linn et al. [212] highlighted the importance of considering the intended and unintended 'consequences' of an assessment that may unduly influence interpretation, use, and response to and validity of its results. Similar views were expressed by trainees in many of the studies included in this review. For instance, trainees reported that certain aspects of assessment, such as its format [130, 149], or ability to assess learning [135, 139], affected how they engaged with it. Additional parallels can be drawn between another of Linn et al., [212]'s criteria, 'fairness', which emphasized the importance of considering the cultural backgrounds of trainees involved and equity of scoring practices, and the dimension, 'cultural influences', within the category 'contextual factors' – some evidence suggests that a trainee's cultural background may influence how they engage with an assessment and assessment-generated feedback [181].

5.3 Downstream consequences of assessment perceived as credible or *not* credible

This review has also identified reported consequences of assessment perceived as credible or *not* credible. Although some scholars [9, 10] have previously explored trainees' credibility perceptions, their focus was more so on influential factors in specific learning contexts (i.e. supervisors providing feedback to trainees). This review helps expand our considerations for the

downstream consequences of perceived credibility of assessment and assessment-generated feedback, beyond the acceptance or dismissal of feedback and/or learning cues [9, 10]. In fact, trainees make judgments about the credibility of aspects of the assessment itself, the provider of the assessment, and the results of the assessment (i.e. scores, feedback). When an assessment is perceived as credible, it is more likely to have a positive effect (e.g. drive learning forward) [101, 116, 117, 119], trainees will be more receptive to scores whether they be positive or negative [78, 106, 128, 201, 210], and feedback will be used to improve their performance and development as a physician [100, 111, 120, 124, 128, 130, 132]. When an assessment is *not* perceived as credible, trainees: (i) could perceive the assessment as a 'hoop to jump through' with limited educational value [99, 101, 117, 132], (ii) may be more likely to engage in gaming behaviour in hopes of advancing their scores with limited focus on actual learning [100, 101, 105, 127, 131, 156], and (iii) develop negative views of resultant assessment scores [106, 112, 131].

These findings also highlight that the educational value of assessment-generated feedback is best supported, when the assessment is standardized [78, 118, 130, 134, 136, 166, 181, 192, 193], clinically relevant and authentic [114, 118, 133, 142, 167, 192], and in the case of observation-based assessment (e.g. workplace-based assessment), provided by attentive and trusted supervisors [106, 110, 112, 121]. These features and its alignment with trainee's self-perception [112, 132], will increase the likelihood that trainees will engage with the potential educational value of the assessment process.

5.5 Contribution of concept map

The findings of this review also lead to the development of a concept map that linked how trainees make judgments about the credibility of an assessment itself, those who provide it, and the feedback generated from it. The concept map presented here differs from a previously

published model of clinical learning by Watling et al. [10], which was based on work focused on trainee perceptions of feedback provided by a supervisor, not directly linked to an assessment. My concept map extends beyond supervisor-provided feedback, identifying factors that affect the perceived credibility of assessment and assessment-generated feedback. These factors influence trainee engagement with the assessment process and highlight the potential positive and/or negative downstream consequences of such credibility perceptions.

Parallels can be drawn between the previously identified factors that affect the perceived credibility of feedback and factors that affect the perceived credibility of assessment and assessment-generated feedback. Watling et al., [10] determined that a trainee's perception of the source of feedback affects the perceived credibility of that feedback. When trainees trust and respect the source of feedback, they are more likely to perceive it as credible. My findings also suggest that perceived credibility of assessment-generated feedback was strengthened when it was provided by an individual the trainee respected and with whom they had a trusting relationship [106, 112, 121]. Contrary to previous works [10], respect appeared to originate from both the supervisor's teaching abilities [120, 134] and their clinical skills [114], not specifically their skills as a clinician. In addition, trainee's perceptions of the content of feedback were also found to influence perceived credibility [10]. Similarly, this review suggests that in the context of observation-based assessments (e.g. performance-based assessment, workplace-based assessment), trainee perceptions of assessment-generated feedback were strengthened when it was in line with their own self-assessment [112, 132]. This determination is supported by other studies which identified learners as having difficulties with accepting and using feedback that is incongruent with their self-appraisal [216]. The similarities and differences in these findings suggest that there are many elements of an assessment process – the assessment itself, the

individual administrating the assessment, and the assessment-generated feedback – that a trainee will judge to determine whether it has value and should be integrated into their professional development or *not*.

In summary, this review revealed several important aspects of assessment that trainees perceive as important contributors to their education. These include design- and implementation-related aspects of the assessment process, factors that influence how trainees perceive the credibility of an assessment and assessment-generated feedback, and the downstream consequences of those perceptions of credibility. Themes identified across this review suggest that there are factors that influence how trainees engage with an assessment and whether they are receptive to the assessment-generated feedback. Depending on whether a trainee perceives an assessment process as credible or not results in different downstream consequences, such as positive or negative views of assessment and scores, and the likelihood of the assessment contributing to improved performance.

5.4 Strengths and limitations

Scoping reviews are an increasingly popular knowledge synthesis technique that provide "a unique opportunity to retrieve and scan a broad range of literature to answer a research question" (p. e62) [217]. Although seen as a methodological advantage, the broad range of literature identified by scoping reviews may be perceived as a limitation if the relevant literature identified are diverse and difficult to synthesize using traditional synthesis approaches. Citation tracking of key articles was employed in hopes of identifying all relevant articles, but this review did not perform handsearching of the relevant journals, and thus, some pertinent articles may not have been included. However, there was a reasonable amount of diversity of findings, study types, and geographic representation, and sources were retrieved from 54 different journals.

Recognizing that this work was focused on a relatively emerging area of research within medical education and that current literature is disparate, consisting of articles highly variable in terms of methodology and focus, I did not limit the inclusion of any article based on type of publication (e.g. original research, commentary, literature reviews) or methodology (i.e. qualitative, quantitative, mixed methods). This was a purposeful decision to ensure the research question was comprehensively addressed by all relevant records available in the literature.

Additionally, the articles included in this review were not appraised for their quality nor were judgments made regarding their scientific rigor, as scoping reviews typically do not require critical appraisals of evidence. As the research objective was to gain an understanding of a specific concept and not to evaluate the effectiveness of an intervention or provide evidence-based recommendations, I felt it was appropriate to include, summarize, and report the overall findings without a formal appraisal process. I recognize the strength of my findings are affected by the reporting practices of each individual study; however, the identification of recurrent themes across studies remains a valuable contribution to the medical education literature.

Lastly, it is important to address the potential limitations of any studies examining learner perceptions. By definition, learner perception is subjective in nature and is not equivalent to a comprehensive evaluation of an assessment's quality, impact, or influence. Trainees perceiving an assessment to be credible (*or not*) only indicates that the assessment was well-received, I cannot make any conclusions regarding the validity or efficacy of these assessments in evaluating a learner. Although findings from this synthesis suggest positive and negative downstream consequences of assessments perceived to be credible (*or not*), these are not definitive outcomes and may still arise regardless of trainee perceptions.

6.0 Conclusion

To my knowledge this is the first review to focus on trainees' perceived credibility of assessment in medical education. This scoping review identified three major overarching themes that describe the perceived credibility of assessment, the factors influence perceived credibility of assessment and assessment-generated feedback, and the consequences of an assessment being perceived as credible or *not*. These themes outline elements of assessment that we can infer trainees perceive to be important in order for assessment to contribute meaningfully to their learning and future performance improvement. At the institutional level, the findings of this review could be used to inform assessment development, implementation, and monitoring, in addition to contributing to assessor training in order to increase the likelihood that an assessment would be perceived as credible, and therefore, maximize the educational potential of assessment. This review also identified contextual factors of assessment that are less amenable to change (e.g. cultural influences and context of medical education), which could be a target for future research investigating elements that could mitigate the influence of these contextual factors or improve the learning context or conditions in which assessment practices are embedded.

Although valuable, the findings from this synthesis could be further refined or expanded upon for more in-depth consideration of specific assessment approaches, levels of training, or educational contexts. These findings constitute a solid foundation for future empirical studies that capture trainee's perceived credibility of assessment or assessment-generated feedback in action. These studies could be used to refine our understanding of how these judgments occur and potentially unveil other factors that influence them within various assessment environments.

From this review, it is clear that medical trainees make judgments about the credibility of assessment based on many contextual, process, and format aspects of assessment – including

assessors, the assessment itself, and feedback – to determine what information they will dismiss and what they will integrate and use for future performance improvement. The aspects of assessment that resonate most with medical trainees tended to reflect in the clinical work to which they have dedicated their training. In order to maximize the potential value of assessment and assessment-generated feedback as a means to support and drive learning, it may be meaningful for assessment developers, assessment administrators, and medical educators to consider adopting a learner-centred approach and include medical trainees in the development of learning assessment strategies and tools for assessment.

References:

1. American Education Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological testing*. 1999, Washington, DC: American Educational Research Association.
2. Norcini, J., et al., *Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference.* Med Teach, 2011. **33**(3): p. 206-14.
3. Harlen, W., *Criteria for Evaluating Systems for Student Assessment.* Studies in Educational Evaluation, 2007. **33**(1): p. 15-28.
4. Epstein, R., *Assessment in medical education.* N Engl J Med, 2007. **356**(4): p. 387-397.
5. Epstein, R.M., *Assessment in medical education.* The New England Journal of Medicine, 2007. **356**(4): p. 387-397.
6. Watling, C., et al., *Learning from clinical work: the roles of learning cues and credibility judgements.* Medical Education, 2012. **46**(2): p. 192-200.
7. Watling, C., *Cognition, culture, and credibility: deconstructing feedback in medical education.* Perspect Med Educ, 2014. **3**(2): p. 124-8.
8. Watling, C., et al., *Beyond individualism: professional culture and its influence on feedback.* Med Educ, 2013. **47**(6): p. 585-94.
9. Bing-You, R.G., J. Paterson, and M.A. Levine, *Feedback falling on deaf ears: residents' receptivity to feedback tempered by sender credibility.* Medical Teacher, 1997. **19**(1): p. 40-44.
10. Watling, C., et al., *Learning from clinical work: the roles of learning cues and credibility judgements.* Med Educ, 2012. **46**(2): p. 192-200.

11. Watling, C.J., *Unfulfilled promise, untapped potential: feedback at the crossroads.* Medical teacher, 2014. **36**(8): p. 692-7.

12. Ende, J., *Feedback in Clinical Medical Education.* JAMA, 1983. **250**: p. 777-781.

13. Ward, M., L. Gruppen, and G. Regehr, *Measuring self-assessment: Current state of the art.* Adv Health Sci Educ, 2002. **7**: p. 63-80.

14. Eva, K. and G. Regehr, *Self-Assessment in the Health Professions: A reformulation and research agenda.* Acad Med, 2005. **80**(10 Suppl): p. S46-S55.

15. Eva, K. and G. Regehr, *Knowing when to look it up: A new conception of self-assessment ability.* Acad Med, 2007. **82**(10 Suppl): p. S81-S84.

16. Eva, K.W., *Assessing Tutorial-Based Assessment.* Advances in Health Sciences Education, 2001. **6**(3): p. 243-257.

17. Eva, K.W. and G. Regehr, *Exploring the divergence between self-assessment and self-monitoring.* Adv Health Sci Educ Theory Pract, 2011. **16**(3): p. 311-29.

18. Regehr, G. and K. Eva, *Self-assessment, self-direction, and the self-regulating professional.* Clin Orthop Relat Res, 2006. **449**: p. 34-8.

19. Carraccio, C.W., S D.; Englander, R.; Ferentz, K.; Martin, C, *Shifting paradigms: From Flexner to competencies.* Academic Medicine, 2002. **77**(5): p. 361.

20. Long, D., *Competency-based residency training: The next advance in graduate medical education.* Academic Medicine, 2000. **75**: p. 1178-1183.

21. Frank, J.R., et al., *Competency-based medical education: theory to practice.* Med Teach, 2010. **32**(8): p. 638-45.

22. Talbot, M., *Monkey see, monkey do: a critique of the competency model in graduate medical education.* Med Educ, 2004. **38**(6): p. 587-92.

23. Fraser, A.B., E.J. Stodel, and A.J. Chaput, *Curriculum reform for residency training: competence, change, and opportunities for leadership.* Can J Anaesth, 2016. **63**(7): p. 875-84.

24. Frank, J.R., et al., *Toward a definition of competency-based education in medicine: a systematic review of published definitions.* Med Teach, 2010. **32**(8): p. 631-7.

25. Leung, W., *Competency based medical training: review.* BMJ, 2002. **325**: p. 693-696.

26. Hawkins, R.E., et al., *Implementation of competency-based medical education: are we addressing the concerns and challenges?* Med Educ, 2015. **49**(11): p. 1086-102.

27. Dictionary.com. *Competence*. 2017 [cited 2017 December 5].

28. Epstein, R. and E. Hundert, *Defining and assessing professional competence.* JAMA, 2002. **287**(2): p. 226-236.

29. Albanese, M.A., *The rise of competencies*, in *Presentation on Receipt of the Jack L Maatsch Visiting Scholar in Medical Education Award*. 2005: Lansing, MI: Michigan State University.

30. Royal College of Physicians and Surgeons of Canada. *CanMEDS: Better standards, better physicians, better case*. 2011 [cited 2017 December 5].

31. Bhatti, N. and C. Cummings, *Competency in surgical residency training: Defining and raising the bar.* Academic Medicine, 2007. **82**(6): p. 569-574.

32. Educational Commission for Foreign Medical Graduates. *ACGME Core Competencies*. 2012 [cited 2012 December 5].

33. General Medical Council, *Tomorrow's Doctors: Outcomes and standards for undergraduate medical education*. 2009, General Medical Council,.

34. Ng, V., C. Burke, and A. Narula, *Knowledge of CanMEDS–Family Medicine roles: Survey of Canadian Family Medicine residents.* Canadian Family Physician, 2013. **59**: p. e428-434.

35. Tannenbaum, D., et al., *CanMeds - Family Medicine*, T.C.o.F.P.o. Canada, Editor. 2009.

36. Shaw, E., I. Oandasan, and N. Fowler, *CanMEDS-FM 2017: A competency framework for family physicians across the continuum*. 2017, The College of Family Physicians of Canada: Mississauga, ON.

37. Tannenbaum, D., et al., *Triple C competency-based curriculum. Report of the Working Group on Postgraduate Curriculum Review - Part 1*. 2011, College of Family Physicians of Canada: Mississauga

38. Oandasan, I., *Advancing Canada's family medicine curriculum: Triple C.* Canadian Family Physician, 2011. **57**: p. 739-740.

39. Miller, M., R. Linn, and N. Gronlund, *Measurement and Assessment in Teaching*. 10th ed. 2009, Upper Saddle Rive, New Jersey: Pearson.

40. Hays, R., *Assessment in Medical Education: Roles for clinical teachers.* The Clinical Teacher, 2008. **5**: p. 23-27.

41. Ferris, H.A. and D. O' Flynn, *Assessment in Medical Education; What Are We Trying to Achieve?* International Journal of Higher Education, 2015. **4**(2).

42. O'Shaughnessy, S.M. and P. Joyce, *Summative and Formative Assessment in Medicine: The Experience of an Anaesthesia Trainee.* International Journal of Higher Education, 2015. **4**(2).

43. Rudolph, J.W., et al., *Debriefing as formative assessment: closing performance gaps in medical education.* Acad Emerg Med, 2008. **15**(11): p. 1010-6.

44. Van der Vleuten, C., *The assessment of professional competence: Developments, research and practical implications.* Adv Health Sci Educ, 1996. **1**: p. 41-67.

45. Miller, G., *The assessment of clinical skills/competence/performance.* Academic Medicine, 1990. **65**(9): p. S63-S68.

46. van der Vleuten, C., *Validity of final exams in undergrad med training.* BMJ, 2000. **321**: p. 1217-1219.

47. Wass, V., et al., *Assessment of clinical competence.* The Lancet, 2001. **357**(9260): p. 945-949.

48. Glavin, R. and N. Maran, *Development of a scoring system for assessment of clinical competence.* Br J Anaesth, 2002. **88**(3).

49. Schuwirth, L.W. and C.P. van der Vleuten, *Different written assessment methods: what can be said about their strengths and weaknesses?* Med Educ, 2004. **38**(9): p. 974-9.

50. Hift, R., *Should essays and other open-ended-type questions retain a place in written summative assessment in clinical medicine?* BMC Medical Education, 2014. **14**: p. 249-267.

51. Norcini, J., et al., *A comparison of knowledge, synthesis, and clinical judgment: Multiple-choice questions in the assessment of physician competence.* Eval Health Prof., 1984. **7**(4): p. 485-500.

52. Frederiksen, N., *The real test bias: influences of testing on teaching and learning.* Am Psychol, 1984. **39**: p. 193-202.

53. Schuwirth, L., C. van der Vleuten, and H. Donkers, *A closer look at cueing effects in multiple-choice questions.* Med Educ, 1996. **30**: p. 44-49.

54. Frijns, P., et al., *The effect of structure in scoring methods on the reproducibility of tests using open ended questions*, in *Teaching and assessing clinical competence*, W. Bender, et al., Editors. 1990, Bockwerk: Gromingen. p. 466-471.

55. Shenwai, M.R. and B.P. K, *Introduction of Structured Oral Examination as A Novel Assessment tool to First Year Medical Students in Physiology.* Journal of Clinical and Diagnostic Research JCDR, 2013. **7**(11): p. 2544-7.

56. Anastakis, D., R. Cohen, and R. Reznik, *The Structured Oral Examination as a method for assessing surgical residents.* The American Journal of Surgery, 1991. **162**: p. 67-71.

57. Jayawickramarajah, P., *Oral examinations in medical education.* Med Educ, 1985. **19**: p. 290-293.

58. Norman, G., *The long case versus objective structured clinical examination.* BMJ, 2002. **324**: p. 748-749.

59. Harden, R.M., et al., *Assessment of clinical competence using objective structured examination.* British Medical Journal, 1975. **1**: p. 447-451.

60. Harden, R.M. and F.A. Gleeson, *Assessment of clinical competence using an objective structured clinical examination (OSCE).* Med Educ, 1979. **13**(1): p. 41-54.

61. Norcini, J., et al., *The Mini-CEX- A Method for Assessing Clinical Skills.* Ann Intern Med, 2003. **138**: p. 476-481.

62. Norcini, J., et al., *The Mini-CEX (Clinical Evaluation Exercise): A Preliminary Investigation.* Ann Intern Med, 1995. **123**: p. 795-799.

63. Swanson, D.B. and C.P. van der Vleuten, *Assessment of clinical skills with standardized patients: state of the art revisited.* Teach Learn Med, 2013. **25 Suppl 1**: p. S17-25.

64. Hodges, B. and J. McIlroy, *Analytic global OSCE ratings are sensitive to level of training.* Med Educ, 2003. **37**: p. 1012-1016.

65. Papadakis, M., *The Step 2 clinical skills examination.* N Engl J Med, 2004. **350**(17): p. 1703-1706.

66. Carr, S.J., *Assessing clinical competency in medical senior house officers: how and why should we do it?* Postgraduate Medical Journal, 2004. **80**(940): p. 63-66.

67. Norcini, J.J., *Work based assessment.* BMJ, 2003. **326**(7392): p. 753-5.

68. Holmboe, E.S., et al., *The role of assessment in competency-based medical education.* Med Teach, 2010. **32**(8): p. 676-82.

69. Iobst, W.F., et al., *Competency-based medical education in postgraduate medical education.* Med Teach, 2010. **32**(8): p. 651-6.

70. McQueen, S.A., et al., *Examining the barriers to meaningful assessment and feedback in medical training.* Am J Surg, 2016. **211**(2): p. 464-75.

71. Rosenblueth, A., N. Wiener, and J. Bigelow, *Behaviour, purpose, and teleology.* Philos Sci, 1943. **10**(1): p. 18-24.

72. Clement, D. and K. Frandsen, *On conceptual and empirical treatments of feedback in human communication.* Commun Monogr, 1976. **43**: p. 11-28.

73. Hattie, J. and H. Timperley, *The Power of Feedback.* Review of Educational Research, 2007. **77**(1): p. 81-112.

74. Archer, J.C., *State of the science in health professional education: effective feedback.* Med Educ, 2010. **44**(1): p. 101-8.

75. Schartel, S.A., *Giving feedback - an integral part of education.* Best Pract Res Clin Anaesthesiol, 2012. **26**(1): p. 77-87.

76. Kulhavy, R., *Feedback in written instruction.* Review of Educational Research, 1977. **47**(1): p. 211-232.

77. Harrison, C.J., et al., *Barriers to the uptake and use of feedback in the context of summative assessment.* Adv Health Sci Educ Theory Pract, 2015. **20**(1): p. 229-45.

78. Harrison, C.J., et al., *Factors influencing students' receptivity to formative feedback emerging from different assessment cultures.* Perspect Med Educ, 2016. **5**(5): p. 276-284.

79. Watling, C., *The uneasy allience of assessment and feedback.* Perspect Med Educ, 2016. **5**: p. 262-264.

80. Palmer, E. and P. Devitt, *The assessment of a structured online formative assessment program: a randomised controlled trial.* BMC Medical Education, 2014. **14**(8): p. 1-10.

81. Watling, C., et al., *Learning culture and feedback: an international study of medical athletes and musicians.* Med Educ, 2014. **48**(7): p. 713-23.

82. Arksey, H. and L. O'Malley, *Scoping studies: towards a methodological framework.* International Journal of Social Research Methodology, 2005. **8**(1): p. 19-32.

83. Thomas, A., et al., *Knowledge Syntheses in Medical Education: Demystifying Scoping Reviews.* Acad Med, 2016. **92**(2): p. 161-166.

84. Pham, M.T., et al., *A scoping review of scoping reviews: advancing the approach and enhancing the consistency.* Res Synth Methods, 2014. **5**(4): p. 371-85.

85. Grant, M.J. and A. Booth, *A typology of reviews: an analysis of 14 review types and associated methodologies.* Health Info Libr J, 2009. **26**(2): p. 91-108.

86. Peters, M.D., et al., *Guidance for conducting systematic scoping reviews.* Int J Evid Based Healthc, 2015. **13**(3): p. 141-6.

87. van Tulder, M., et al., *Updated Method Guidelines for Systematic Reviews in the Cochrane Collaboration Back Review Group.* Spine, 2003. **28**: p. 1290.

88. Rumrill, P.D., S.M. Fitzgerald, and W.R. Merchant, *Using scoping literature reviews as a means of understanding and interpreting existing literature.* Work, 2010. **35**(3): p. 399-404.

89. Munn, Z., et al., *What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences.* BMC Med Res Methodol, 2018. **18**(1): p. 5.

90. *Cochrane Handbook for Systematic Reviews of Interventions: Version 5.1.0*, J. Higgings and S. Green, Editors. 2011, The Cochrane Collaboration.

91. Conn, V., et al., *Grey literature in meta-analyses.* Nursing Research, 2003. **52**(4).

92. Kastner, M., et al., *What is the most appropriate knowledge synthesis method to conduct a review? Protocol for a scpoing review.* BMC Med Res Methodol, 2012. **12**(114): p. 1-10.

93. Thomas, J. and A. Harden, *Methods for the thematic synthesis of qualitative research in systematic reviews.* BMC Med Res Methodol, 2008. **8**: p. 45.

94. Abbasi, M.H., et al., *Comparative study of dactylography among the students of avicenna medical college Lahore.* Pakistan Journal of Medical and Health Sciences, 2012. **6**(2): p. 362-365.

95. Braun, V. and V. Clarke, *Using thematic analysis in psychology.* Qualitative Research in Psychology, 2006. **3**(2): p. 77-101.

96. Parslow, G.R., *Commentary: Decaying Numerical Skills. "I Can't Divide by 60 in My Head!".* Biochemistry and Molecular Biology Education, 2010. **38**(1): p. 46-47.

97. McCoubrie, P., *Improving the fairness of multiple-choice questions: A literature review.* Medical Teacher, 2004. **26**(8): p. 709-712.

98. Miller, A. and J. Archer, *Impact of workplace based assessment on doctors' education and performance: a systematic review.* BMJ (Clinical research ed.), 2010. **341**: p. c5064.

99. Al-Kadri, H.M., et al., *Self-assessment and students' study strategies in a community of clinical practice: A qualitative study.* Medical Education Online, 2012. **17**(1).

100. Brown, J.M., et al., *An investigation into the use of multi-source feedback (MSF) as a work-based assessment tool.* Medical teacher, 2014. **36**(11): p. 997-1004.

101. Castanelli, D.J., et al., *Perceptions of purpose, value, and process of the mini-Clinical Evaluation Exercise in anesthesia training.* Canadian Journal of Anesthesia, 2016. **63**(12): p. 1345-1356.

102. Green, A., et al., *Designing and implementing a cultural competence OSCE: lessons learned from interviews with medical students.* Ethnicity & disease, 2007. **17**(2): p. 344-50.

103. Gulbas, L., W. Guerin, and H.F. Ryder, *Does what we write matter? Determining the features of high- and low-quality summative written comments of students on the internal medicine clerkship using pile-sort and consensus analysis: a mixed-methods study.* BMC medical education, 2016. **16**: p. 145.

104. Heeneman, S., et al., *The impact of programmatic assessment on student learning: Theory versus practice.* Medical Education, 2015. **49**(5): p. 487-498.

105. Ingram, J.R., E.J. Anderson, and L. Pugsley, *Difficulty giving feedback on underperformance undermines the educational value of multi-source feedback.* Medical teacher, 2013. **35**(10): p. 838-46.

106. Lefroy, J., et al., *Grades in formative workplace-based assessment: A study of what works for whom and why.* Medical Education, 2015. **49**(3): p. 307-320.

107. Nestel, D., et al., *Formative Assessment of Procedural Skills: Students' Responses to the Objective Structured Clinical Examination and the Integrated Performance Procedural Instrument.* Assessment & Evaluation in Higher Education, 2011. **36**(2): p. 171-183.

108. Nikendei, C., et al., *An innovative model for teaching complex clinical procedures: integration of standardised patients into ward round training for final year students.* Medical teacher, 2007. **29**(2): p. 246-52.

109. Plant, J.L., et al., *Understanding self-assessment as an informed process: residents' use of external information for self-assessment of performance in simulated resuscitations.* Advances in health sciences education : theory and practice, 2013. **18**(2): p. 181-92.

110. Ringsted, C., et al., *Educational impact of in-training assessment (ITA) in postgraduate medical education: a qualitative study of an ITA programme in actual practice.* Medical Education, 2004. **38**(7): p. 767-77.

111. Spanager, L., et al., *Comprehensive feedback on trainee surgeons' non-technical skills.* International journal of medical education, 2015. **6**: p. 4-11.

112. Watling, C.J., et al., *Rules of engagement: residents' perceptions of the in-training evaluation process.* Academic medicine : journal of the Association of American Medical Colleges, 2008. **83**(10): p. S97-100.

113. Arnold, L., et al., *Medical students' views on peer assessment of professionalism.* Journal of General Internal Medicine, 2005. **20**(9): p. 819-24.

114. Bleasel, J., et al., *Feedback using an ePortfolio for medicine long cases: quality not quantity.* BMC medical education, 2016. **16**(1): p. 278.

115. Burgess, A. and C. Mellis, *Receiving feedback from peers: medical students' perceptions.* The clinical teacher, 2015. **12**(3): p. 203-7.

116. Cho, S.P., D. Parry, and W. Wade, *Lessons learnt from a pilot of assessment for learning.* Clinical medicine (London, England), 2014. **14**(6): p. 577-584.

117. Cottrell, E., et al., *Assessing academic clinical fellows in general practice: square pegs in round holes?* Education for primary care : an official publication of the Association of Course Organisers, National Association of GP Tutors, World Organisation of Family Doctors, 2013. **24**(4): p. 266-73.

118. Craig, S., et al., *Assessment and feedback in emergency medicine training: Views of Australasian emergency trainees.* EMA - Emergency Medicine Australasia, 2010. **22**(6): p. 537-547.

119. Curran, V.R., et al., *Evaluation of the usefulness of simulated clinical examination in family-medicine residency program.* Medical teacher, 2007. **29**(4): p. 406-7.

120. Dijksterhuis, M.G.K., et al., *A qualitative study on trainees' and supervisors' perceptions of assessment for learning in postgraduate medical education.* Medical teacher, 2013. **35**(8): p. e1396-402.

121. Duijn, C.C.M.A., et al., *Am I ready for it? Students' perceptions of meaningful feedback on entrustable professional activities.* Perspectives on medical education, 2017.

122. Foucault, A., et al., *Learning medical professionalism with the online concordance-of-judgment learning tool (CJLT): A pilot study.* Medical Teacher, 2015. **37**(10): p. 955-960.

123. Harrison, C.J., et al., *Factors influencing students' receptivity to formative feedback emerging from different assessment cultures.* Perspectives on Medical Education, 2016. **5**(5): p. 276-84.

124. Harrison, C.J., et al., *How we give personalised audio feedback after summative OSCEs.* Medical teacher, 2015. **37**(4): p. 323-326.

125. Johnson, G., et al., *Feedback from educational supervisors and trainess on the implementation of curricula and the assessment system for core medical training.* Clinical Medicine, Journal of the Royal College of Physicians of London, 2008. **8**(5): p. 484-489.

126. Malhotra, S., R. Hatala, and C.A. Courneya, *Internal medicine residents' perceptions of the mini-clinical evaluation exercise.* Medical Teacher, 2008. **30**(4): p. 414-419.

127. McKavanagh, P., A. Smyth, and A. Carragher, *Hospital consultants and workplace based assessments: How foundation doctors view these educational interactions?* Postgraduate Medical Journal, 2012. **88**(1037): p. 119-124.

128. Murdoch-Eaton, D. and J. Sargeant, *Maturational differences in undergraduate medical students' perceptions about feedback.* Med Educ, 2012. **46**(7): p. 711-21.

129. Papinczak, T., L. Young, and M. Groves, *Peer Assessment in Problem-Based Learning: A Qualitative Study.* Advances in Health Sciences Education, 2007. **12**(2): p. 169-186.

130. Rees, C., C. Sheard, and A. McPherson, *Communication skills assessment: The perceptions of medical students at the University of Nottingham.* Medical Education, 2002. **36**(9): p. 868-878.

131. Sabey, A. and M. Harris, *Training in hospitals: What do GP specialist trainees think of workplacebased assessments?* Education for Primary Care, 2011. **22**(2): p. 90-99.

132. Sargeant, J., et al., *Conditions influencing informed self-assessment and use of feedback.* Medical Education, Supplement, 2011. **45**: p. 40-41.

133. Shafi, R., K. Irshad, and M. Iqbal, *Competency-based integrated practical examinations: Bringing relevance to basic science laboratory examinations.* Medical Teacher, 2010. **32**(10): p. e443-7.

134. Sharma, S., et al., *Formative assessment in postgraduate medical education - Perceptions of students and teachers.* International journal of applied & basic medical research, 2015. **5**: p. S66-70.

135. Wade, L., et al., *Student perceptions of the progress test in two settings and the implications for test deployment.* Advances in Health Sciences Education, 2012. **17**(4): p. 573-583.

136. Weller, J.M., et al., *Investigation of trainee and specialist reactions to the mini-Clinical Evaluation Exercise in anaesthesia: Implications for implementation.* British Journal of Anaesthesia, 2009. **103**(4): p. 524-530.

137. Winkel, A.F., et al., *Notes from the Field: Residents' Perceptions of Simulation-Based Skills Assessment in Obstetrics and Gynecology.* Evaluation & the health professions, 2016. **39**(1): p. 121-5.

138. Wu, V., et al., *Development and Validation of the McMaster Prescribing Competency Assessment for Medical Trainees (MacPCA).* Journal of Population Therapeutics & Clinical Pharmacology, 2015. **22**(2): p. e173-8.

139. Kim, H., et al., *Evaluating the quality, clinical relevance, and resident perception of the radiation oncology in-training examination: A national survey.* Practical Radiation Oncology, 2016. **6**(1): p. 44-9.

140. Nesbitt, A., et al., *Student perception of workplace-based assessment.* Clinical Teacher, 2013. **10**(6): p. 399-404.

141. Perera, J., et al., *Formative feedback to students: the mismatch between faculty perceptions and student expectations.* Medical teacher, 2008. **30**(4): p. 395-9.

142. Pierre, R.B., et al., *Student evaluation of an OSCE in paediatrics at the University of the West Indies, Jamaica.* BMC Medical Education, 2004. **4**.

143. Rudland, J.R., P. Schwartz, and A. Ali, *Moving a formative test from a paper-based to a computer-based format. A student viewpoint.* Medical Teacher, 2011. **33**(9): p. 738-43.

144. Sharma, N., *Medical students' perceptions of the situational judgement test: A mixed methods study.* British Journal of Hospital Medicine, 2015. **76**(4): p. 234-238.

145. Sharma, N., et al., *Team-based assessment of medical students in a clinical clerkship is feasible and acceptable.* Medical Teacher, 2012. **34**(7): p. 555-561.

146. Tweed, M. and J. Cookson, *The face validity of a final professional clinical examination.* Medical Education, 2001. **35**(5): p. 465-73.

147. Vanlint, A., et al., *Evaluation of the introduction of the OSCE to the fifth-year Geriatric Medicine Teaching Programme.* Australasian Journal on Ageing, 2016. **35**(4): p. 285-288.

148. Wiener-Ogilvie, S. and D. Begg, *Formative assessment of GP trainees' clinical skills.* Education for Primary Care, 2012. **23**(2): p. 101-6.

149. Baerheim, A. and E. Meland, *Medical students proposing questions for their own written final examination: Evaluation of an educational project.* Medical Education, 2003. **37**(8): p. 734-738.

150. Burford, B., et al., *User perceptions of multi-source feedback tools for junior doctors.* Medical education, 2010. **44**(2): p. 165-76.

151. Carter, E. and H. Pascoe, *Feedback to medical students: Do we give enough?* Medical Education, Supplement, 2010. **44**: p. 163.

152. Dadgar, S.R., et al., *OSCE as a tool for evaluation of practical semiology in comparison to MCQ & Oral examination.* Journal of the Pakistan Medical Association, 2008. **58**(9): p. 506-507.

153. de Almeida Troncon, L.E., *Clinical skills assessment: Limitations to the introduction of an "OSCE" (Objective Structured Clinical Examination) in a traditional Brazilian medical school.* Sao Paulo Medical Journal, 2004. **122**(1): p. 12-17.

154. Dowling, S., et al., *The acceptability, feasibility and educational impact of a new tool for formative assessment of the consultation performance of specialty registrars in an Irish general practice training scheme.* Education for Primary Care, 2007. **18**(6): p. 724-735.

155. Duffield, K.E. and J.A. Spencer, *A survey of medical students' views about the purposes and fairness of assessment.* Medical Education, 2002. **36**(9): p. 879-86.

156. Finall, A., *Trainers' perceptions of the direct observation of practical skills assessment in histopathology training: a qualitative pilot study.* Journal of Clinical Pathology, 2012. **65**(6): p. 538-40.

157. Gelan, E.A., R. Essayas, and K. Gebressilase, *Perception of final year medical students about objective structured clinical examination in the Department of General Surgery.* Ethiopian Medical Journal, 2015. **53**(4): p. 183-189.

158. Gnanathasan, C.A., F.I. Achike, and J. Abdullah, *Perceptions of examinees (Students) and examiners (Faculty) of OSCE at the International Medical University (IMU), Malaysia.* Medical Education, Supplement, 2010. **44**: p. 5-6.

159. Haider, I., et al., *Perceptions of final professional MBBS students and their examiners about objective structured clinical examination (OSCE): A combined examiner and examinee survey.* Journal of Medical Sciences (Peshawar), 2016. **24**(4): p. 206-211.

160. Hays, R.B., et al., *Short and long multiple-choice question stems in a primary care oriented undergraduate medical curriculum.* Education for Primary Care, 2009. **20**(3): p. 173-177.

161. Hunter, A.R., E.J. Baird, and M.R. Reed, *Procedure-based assessments in trauma and orthopaedic training--The trainees' perspective.* Medical teacher, 2015. **37**(5): p. 444-9.

162. Ibrahim, N.K., et al., *Perceptions of clinical years' medical students and interns towards assessment methods used in King Abdulaziz University, Jeddah.* Pakistan Journal of Medical Sciences, 2015. **31**(4): p. 757-762.

163. Jafarzadeh, A., *Designing the OSCE method for evaluation of practical immunology course of medical students: In comparison to written-MCQ and oral examination.* Rawal Medical Journal, 2009. **34**(2): p. 219-222.

164. Kania, R.E., et al., *Online script concordance test for clinical reasoning assessment in otorhinolaryngology: The association between performance and clinical experience.* Archives of Otolaryngology - Head and Neck Surgery, 2011. **137**(8): p. 751-755.

165. Kasanda, C., et al., *Medical and pharmacy students' perceptions of the grading and assessment practices.* Frontiers in Psychology Vol 4 2013, ArtID 423, 2013. **4**.

166. Khairy, G.A., *Feasibility and acceptability of objective structured clinical examination (osce) for a large number of candidates: experience at a university hospital.* Journal of Family and Community Medicine, 2004. **11**(2): p. 75-8.

167. Khorashad, A.K., et al., *The assessment of undergraduate medical students' satisfaction levels with the objective structured clinical examination.* Iranian Red Crescent Medical Journal, 2014. **16**(8).

168. Korszun, A., et al., *Assessment of professional attitude and conduct in medical undergraduates.* Medical Teacher, 2005. **27**(8): p. 704-708.

169. Labaf, A., et al., *Students' concerns about the pre-internship objective structured clinical examination in medical education.* Education for health (Abingdon, England), 2014. **27**(2): p. 188-192.

170. Levine, J.C., T. Geva, and D.W. Brown, *Competency Testing for Pediatric Cardiology Fellows Learning Transthoracic Echocardiography: Implementation, Fellow Experience, and Lessons Learned.* Pediatric Cardiology, 2015. **36**(8): p. 1700-11.

171. McCourt, C., et al., *The level playing field: The impact of assessment practice on professional development.* Medical Education, 2012. **46**(8): p. 766-776.

172. McKinley, R.K., et al., *Formative assessment of the consultation performance of medical students in the setting of general practice using a modified version of the Leicester Assessment Package.* Medical Education, 2000. **34**(7): p. 573-579.

173. McLaughlin, K., et al., *Does blueprint publication affect students' perception of validity of the evaluation process?* Advances in Health Sciences Education, 2005. **10**(1): p. 15-22.

174. Nofziger, A.C., et al., *Impact of peer assessment on the professional development of medical students: a qualitative study.* Academic Medicine, 2010. **85**(1): p. 140-7.

175. Al Omari, A. and Z.M. Shawagfa, *New experience with objective structured clinical examination in Jordan.* Rawal Medical Journal, 2010. **35**(1): p. 82-84.

176. Pearce, I., et al., *The record of in-training assessments (RITAs) in urology: An evaluation of trainee perceptions.* Annals of the Royal College of Surgeons of England, 2003. **85**(5): p. 351-354.

177. Junod Perron, N., et al., *The quality of feedback during formative OSCEs depends on the tutors' profile.* BMC medical education, 2016. **16**(1): p. 293.

178. Raheel, H. and N. Naeem, *Assessing the Objective Structured Clinical Examination: Saudi family medicine undergraduate medical students' perceptions of the tool.* JPMA - Journal of the Pakistan Medical Association, 2013. **63**(10): p. 1281-4.

179. Rahman, S.A., *Promoting learning outcomes in paediatrics through formative assessment.* Medical Teacher, 2001. **23**(5): p. 467-470.

180. Smith, S., et al., *The distracted intravenous access (DIVA) test.* The clinical teacher, 2012. **9**(5): p. 320-4.

181. Suhoyo, Y., et al., *Influence of feedback characteristics on perceived learning value of feedback in clerkships: does culture matter?* BMC medical education, 2017. **17**(1): p. 69.

182. Tayem, Y.I., et al., *Medical students' perceptions of peer assessment in a problem-based learning curriculum.* Sultan Qaboos University Medical Journal, 2015. **15**(3): p. e376-e381.

183. Tsugawa, Y., et al., *Professionalism mini-evaluation exercise for medical residents in Japan: A pilot study.* Medical Education, 2009. **43**(10): p. 968-973.

184. Vishwakarma, K., et al., *Introducing objective structured practical examination as a method of learning and evaluation for undergraduate pharmacology.* Indian Journal of Pharmacology, 2016. **48**(7): p. S47-S51.

185. Wittels, K. and J.K. Takayesu, *Development of a simulation based assessment tool to measure emergency medicine resident competency.* Academic Emergency Medicine, 2013. **1**: p. S123.

186. Amr, M., D. Raddad, and Z. Afifi, *Objective Structural Clinical Examination (OSCE) during psychiatry clerkship in a Saudi university.* Arab Journal of Psychiatry, 2012. **23**(1): p. 69-73.

187. Anderson, J., N.P. Robertson, and P.E.M. Smith, *Promoting feedback in clinical neuroscience teaching.* Journal of Neurology, Neurosurgery and Psychiatry, 2012. **83**: p. A29.

188. Boehler, M.L., et al., *An investigation of medical student reactions to feedback: a randomised controlled trial.* Medical education, 2006. **40**(8): p. 746-9.

189. Chander, B., et al., *Teaching the competencies: using objective structured clinical encounters for gastroenterology fellows.* Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association, 2009. **7**(5): p. 509-14.

190. Cogbill, K.K., P.S. O'Sullivan, and J. Clardy, *Residents' Perception of Effectiveness of Twelve Evaluation Methods for Measuring Competency.* Academic Psychiatry, 2005. **29**(1): p. 76-81.

191. Hicks, P.J., et al., *To the point: medical education reviews--dealing with student difficulties in the clinical setting.* Am J Obstet Gynecol, 2005. **193**(6): p. 1915-22.

192. Jawaid, M., Z. Masood, and F. Jaleel, *Students' perception of surgical objective structured clinical examination (OSCE) at Dow University Of Health Sciences.* Journal of Postgraduate Medical Institute, 2014. **28**(1): p. 19-23.

193. Jefferies, A., et al., *Assessment of Multiple Physician Competencies in Postgraduate Training: Utility of the Structured Oral Examination.* Advances in Health Sciences Education, 2011. **16**(5): p. 569-577.

194. Kalet, A.L., et al., *Promoting professionalism through an online professional development portfolio: Successes, joys, and frustrations.* Academic Medicine, 2007. **82**(11): p. 1065-1072.

195. Kelly, S.P., et al., *Learner perception of oral and written examinations in an international medical training program.* International Journal of Emergency Medicine, 2010. **3**(1): p. 21-26.

196. Laughlin, T., A. Brennan, and C. Brailovsky, *Effect of field notes on confidence and perceived competence: survey of faculty and residents.* Canadian family physician Medecin de famille canadien, 2012. **58**(6): p. e352-6.

197. Marrero, I., et al., *Assessing professionalism and ethics knowledge and skills: preferences of psychiatry residents.* Academic Psychiatry, 2013. **37**(6): p. 392-7.

198. McLay, R., et al., *Simulating a full-length psychiatric interview with a complex patient: An OSCE for medical students.* Academic Psychiatry, 2002. **26**: p. 162-167.

199. Nagoshi, M., et al., *Using standardized patients to assess the geriatrics medicine skills of medical students, internal medicine residents, and geriatrics medicine fellows.* Academic Medicine, 2004. **79**(7): p. 698-702.

200. Nasir, A.A., et al., *Medical students' perception of objective structured clinical examination: a feedback for process improvement.* Journal of Surgical Education, 2014. **71**(5): p. 701-6.

201. Nowacki, A.S., *Making the grade in a portfolio-based system: student performance and the student perspective.* Frontiers in Psychology, 2013. **4**: p. 155.

202. Parikh, P.P., et al., *Simulation-based end-of-life care training during surgical clerkship: assessment of skills and perceptions.* Journal of Surgical Research, 2015. **196**(2): p. 258-63.

203. Rafique, S. and H. Rafique, *Students' feedback on teaching and assessment at Nishtar Medical College, Multan.* JPMA - Journal of the Pakistan Medical Association, 2013. **63**(9): p. 1205-9.

204. Sadia, S., S. Sultana, and F. Waqar, *OSCE as an assessment tool: Perceptions of undergraduate medical students.* Anaesthesia, Pain and Intensive Care, 2009. **13**(2): p. 65-67.

205. Schwaab, J., et al., *Using second life virtual simulation environment for mock oral emergency medicine examination.* Academic Emergency Medicine, 2011. **18**(5): p. 559-62.

206. Zyromski, N.J., E.D. Staren, and H.W. Merrick, *Surgery residents' perception of the Objective Structured Clinical Examination (OSCE).* Current Surgery, 2003. **60**(5): p. 533-537.

207. Olson, L.G., et al., *The effect of a Structured Question Grid on the validity and perceived fairness of a medical long case assessment.* Medical Education, 2000. **34**(1): p. 46-52.

208. Harrison, C.J., et al., *Web-based feedback after summative assessment: how do students engage?* Med Educ, 2013. **47**(7): p. 734-44.

209. Ibrahim, J., et al., *Interns' perceptions of performance feedback.* Medical education, 2014. **48**(4): p. 417-29.

210. Altahawi, F., et al., *Student perspectives on assessment: Experience in a competency-based portfolio system.* Medical Teacher, 2012. **34**(3): p. 221-225.

211. Chaffinch, C.N., et al., *Learner feedback on the pediatrics milestones assessment project appd-learn-nbme pediatrics milestones assessment group.* Academic Pediatrics, 2013. **13**(4): p. e6.

212. Linn, R., E. Baker, and S. Dunbar, *Complex, performance-based assessment: Expectations and validation criteria.* Educational Researcher, 1991. **20**(8): p. 15-21.

213. Pellegrino, J., *A learning sciences perspective on the design and use of assessment in education*, in *The Cambridge Handbook of the Learning Sciences*, R.K. Sawyer, Editor. 2015, Cambridge University Press: Chapel Hill.

214. Spencer, J. and R. Jordan, *Learner centred approaches in medical education.* BMJ, 1999. **318**: p. 1280-3.

215. Messick, S., *Standards of Validity and the Validity of Standards in Performance Asessment.* Educational Measurement: Issues and Practice, 1995. **14**(4): p. 5-8.

216. Mann, K., et al., *Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict.* Acad Med, 2011. **86**(9): p. 1120-7.

217. Thomas, A. and M. Law, *Research utilization and evidence-based practice in occupational therapy: a scoping study.* Am J Occup Ther, 2013. **67**(4): p. e55-65.

Appendices

Appendix A: Search strategy for OVID Medline

#	Search Statement	Results	Annotation
1	((evaluator* or feedback or Assessment or Assess or assessed or judgements or assignment or assignments or Exam or exams or Examination* or Questionnaire or Questions or Testing or test or tests or Tester or Evaluation or Evaluate or Rating or Scales or Appraisal or Score or Scores or Grades) adj2 (credibility or receptivity or Helpful or counterproductive or Credible or Constructive or Justifiable or defensible or Relevant or Reasonable or legitimate or Influence or productive or trust or trustworthy or trustworthiness or merit or value or acceptable or appropriate or applicable or fair or fairness)).ab,kf,ti.	71253	
2	*feedback/ or formative feedback/ or *feedback, psychological/	7389	
3	(perception* or perceived).ab,hw,kf,ti.	479825	
4	Perception/	28060	
5	3 or 4	479825	
6	2 and 5	1441	
7	1 or 6	72627	
8	education, medical/ or education, medical, graduate/ or education, medical, undergraduate/ or "internship and residency"/	128376	
9	Students, Medical/	27819	
10	Schools, Medical/	23800	
11	*Clinical Competence/	39208	
12	"clerkship*".ab,kf,ti.	4176	
13	"undergraduate medic*".ab,kf,ti.	4275	
14	"graduate medic*".ab,kf,ti.	5232	
15	((resident or residents or residency) adj3 (medicine or school or education)).ab,kf,ti.	9751	
16	((Intern or Interns or Internship) adj3 (medicine or school or education)).ab,kf,ti.	420	
17	medical students.ab,kf,ti.	30256	

18	medical schools.ab,kf,ti.	11511	
19	(Post?graduate* adj2 medic*).ab,kf,ti.	3299	
20	"house officer*".ab,kf,ti.	1771	
21	"registrar*".ab,kf,ti.	3286	
22	foundation year.ab,kf,ti.	281	
23	"junior doctor*".ab,kf,ti.	2535	
24	8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23	203551	
25	nursing home.ab,ti.	19235	
26	24 not 25	203297	
27	N.sb.	707656	
28	26 not 27	190065	
29	7 and 28	1424	
30	limit 29 to (english or french)	1365	
31	limit 30 to yr="2000 -Current"	1193	

Appendix B: Search strategy for EMBASE

#	Search Statement	Results	Annotation
1	((evaluator* or feedback or Assessment or Assess or assessed or judgements or assignment or assignments or Exam or exams or Examination* or Questionnaire or Questions or Testing or test or tests or Tester or Evaluation or Evaluate or Rating or Scales or Appraisal or Score or Scores or Grades) adj2 (credibility or receptivity or Helpful or counterproductive or Credible or Constructive or Justifiable or defensible or Relevant or Reasonable or legitimate or Influence or productive or trust or trustworthy or trustworthiness or merit or value or acceptable or appropriate or applicable or fair or fairness)).ab,hw,kw,ti.	101711	
2	constructive feedback/	329	
3	1 or 2	101711	
4	*medical education/ or *clinical education/ or *medical school/ or *residency education/ or *surgical training/ or *teaching round/	135240	
5	*medical student/	20290	
6	*resident/	4736	
7	"medical student*".ab,ti.	41166	
8	"clerkship*".ab,ti.	4713	
9	"undergraduate medic*".ab,ti.	4699	
10	"graduate medic*".ab,ti.	5472	
11	((resident or residents or residency) adj3 (medicine or school or education)).ab,ti.	12778	
12	((Intern or Interns or Internship) adj3 (medicine or school or education)).ab,ti.	627	
13	"house officer*".ab,ti.	2175	
14	"registrar*".ab,ti.	5405	
15	foundation year.ab,ti.	470	
16	"junior doctor*".ab,ti.	3744	
17	*clinical competence/	20789	
18	(Post?graduate* adj2 medicine).tw.	971	

19	4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18	204697	
20	nursing home.tw.	23934	
21	19 not 20	204328	
22	3 and 21	1651	
23	limit 22 to (english or french)	1581	
24	limit 23 to yr="2000 -Current"	1410	

Appendix C: Search strategy for PsycInfo

#	Search Statement	Results	Annotation
1	((evaluator* or feedback or Assessment or Assess or assessed or judgements or assignment or assignments or Exam or exams or Examination* or Questionnaire or Questions or Testing or test or tests or Tester or Evaluation or Evaluate or Rating or Scales or Appraisal or Score or Scores or Grades) adj2 (credibility or receptivity or Helpful or counterproductive or Credible or Constructive or Justifiable or defensible or Relevant or Reasonable or legitimate or Influence or productive or trust or trustworthy or trustworthiness or merit or value or acceptable or appropriate or applicable or fair or fairness or perception or reputation)).ab,hw,id,ti.	19854	
2	credibility/ or reputation/ or *perception/	13281	
3	*feedback/ or "knowledge of results"/	8301	
4	2 and 3	70	
5	1 or 4	19903	
6	medical education/ or medical internship/ or medical residency/	17437	
7	medical students/	10684	
8	"medical school*".ab,hw,id,ti.	6813	
9	((Intern or Interns or Internship) adj3 (medicine or school or education)).ab,hw,id,ti.	298	
10	"clerkship*".ab,hw,id,ti.	1202	
11	"undergraduate medic*".ab,hw,id,ti.	1117	
12	"graduate medic*".ab,hw,id,ti.	865	
13	((clinical or medical) adj2 residen*).ab,hw,id,ti.	4444	
14	"house officer*".ab,hw,id,ti.	258	
15	"registrar*".ab,hw,id,ti.	569	
16	foundation year.ab,hw,id,ti.	56	
17	"junior doctor*".ab,hw,id,ti.	387	
18	(Post?graduate* adj2 medicine).tw.	29	
19	6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18	25858	

20	nursing home.tw.	7707	
21	19 not 20	25755	
22	5 and 21	328	
23	limit 22 to (english or french)	319	
24	limit 23 to yr="2000 -Current"	292	

Appendix D: Search strategy for ERIC (EBSCO)

#	Query	Limiters/Expanders	Last Run Via	Results
S18	S16 AND S17	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	174
S17		Limiters - Date Published: 20000101-20171231 Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	600,901
S16	S1 AND S15	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	256
S15	S2 OR S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	14,020
S14	TI Post-graduate* N2 medic* OR AB Post-graduate* N2 medic*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	11
S13	TI ((Intern or Interns or Internship) N3 (medicine OR medical OR clinical)) OR AB	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	183

	((Intern or Interns or Internship) N3 (medicine OR medical OR clinical))			
S12	TI ((resident or residents or residency) N3 (medicine or school or education)) OR AB ((resident or residents or residency) N3 (medicine or school or education))	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	1,176
S11	TI "junior doctor*" OR AB "junior doctor*"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	22
S10	TI "foundation year" OR AB "foundation year"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	40
S9	TI registrar* OR AB registrar*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	484
S8	TI "house officer*" OR AB "house officer*"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen -	61

			Advanced Search Database - ERIC	
S7	TI "graduate medic*" OR AB "graduate medic*"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	286
S6	TI "undergraduate medic*" OR AB "undergraduate medic*"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	306
S5	TI clerkship* OR AB clerkship*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	523
S4	TI "medical student*" OR AB "medical student*"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	3,063
S3	DE "Medical Students" OR DE "Medical Schools"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	5,897
S2	DE "Medical Education" OR DE "Graduate Medical Education"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	9,974

S1	(feedback or elevator* or Assessment or Assess or assessed or judgements or assignment or assignments or Exam or exams or Examination* or Questionnaire or Questions or Testing or test or tests or Tester or Evaluation or Evaluate or Rating or Scales or Appraisal or Score or Scores or Grades) N2 (credibility or receptivity or Helpful or counterproductive or Credible or Constructive or Justifiable or defensible or Relevant or Reasonable or legitimate or Influence or productive or trust or trustworthy or trustworthiness or merit or value or acceptable or appropriate or applicable or fair or fairness or perception)	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - ERIC	24,614

Appendix E: Search strategy for Scopus

(TITLE-ABS ((evaluator* OR feedback OR assessment OR assess OR assessed OR judgements OR assignment OR assignments OR exam OR exams OR examination* OR questionnaire OR questions OR testing OR test OR tests OR tester OR evaluation OR evaluate OR rating OR scales OR appraisal OR score OR scores OR grades) W/2 (credibility OR receptivity OR perception OR helpful OR counterproductive OR credible OR constructive OR justifiable OR defensible OR relevant OR reasonable OR legitimate OR influence OR productive OR trust OR trustworthy OR trustworthiness OR merit OR value OR acceptable OR appropriate OR applicable OR fair OR fairness))) AND (KEY (((student* OR undergraduate OR graduate OR post-graduate OR clerkship* OR resident* OR intern* OR school*) W/2 (medicine OR medical OR clinical))) AND PUBYEAR > 1999) AND (LIMIT-TO (SUBJAREA , "MEDI ")) AND (LIMIT-TO (LANGUAGE , "English "))

Appendix F: Characteristics of all included articles		
#	**First author, journal, year**	**Objectives**
1.	Al-Kadri, HM; Medical Education Online; 2012	To explore the relationship between students' perceptions and practices of self-assessment and their study strategies within a community of clinical practice. To assess the impact of student and supervisor self-assessment and feedback training on students' perceptions and practices of self-assessment.
2.	Omari, AAL; Rawal Medical Journal; 2010	To see our students' evaluation of OSCE, so as to avoid pitfall in case of application at postgraduate evaluation of medical residents and national board exam.
3.	Amr, M; The Arab Journal Of Psychiatry; 2012	To investigate the validity of the OSCE by comparing student performance on the OSCE with traditional forms of evaluation and through a student opinion survey at the end of examination. To examine the effect of gender on performance and acceptability of OSCE.
4.	Anderson, J; Journal of Neurology, Neurosurgery and Psychiatry; 2012	"To assess student satisfaction with newly implemented personal tutor system and formative assessment."
5.	Arnold, L; Journal of General Internal Medicine; 2012	To identify factors that, according to students themselves, would encourage or discourage their participation in peer assessment.
6.	Atahawai, F; Medical Teacher; 2012	To capture [learners'] perceptions in order to gain insights into the strengths and weaknesses of a competency-based assessment system.

7.	Baerheim, A; Medical Education; 2003	"To evaluate how sixth year medical students experienced the project, and to what extent their performance in the examination was influenced."
8.	Bleasel, J; BMC Medical Education; 2016	To describe students' experience of using the ePortfolio, and receiving feedback on written long cases. To explore the relationship between quantity and quality of feedback.
9.	Boehler, ML; Medical Education; 2006	"To evaluate learning outcomes and perceptions in students who received feedback compared to those who received general compliments."
10.	Brown, JM; Medical Teacher; 2014	"To explore the perception of STs and their assessors on MSF as a work based assessment tool."
11.	Burford, B; Medical Education; 2010	To compare perceptions of two tools for giving MSF to UK junior doctors, of which one provides mainly textual feedback and one provides mainly numerical feedback and also compared to: raters giving feedback, and supervisors delivering feedback.
12.	Burgess, A; The Clinical Teacher; 2015	To investigate students' views on receiving verbal feedback from their peers during their formative long case examination.
13.	Carter, H; Medical Education; 2010	This study was to find out how helpful Child Health feedback was.
14.	Castanelli, DJ; Canadian Journal of Anaesthesia; 2016	To explore how Australian and New Zealand College of Anaesthetists (ANZCA) trainees and supervisors of training considered their experience with the mini-CEX 18 months after their compulsory introduction into anesthesia training.

15.	Chander, B; Clinical Gastroenterology and Hepatology; 2008	To describe the process of developing and implementing a 4-station OSCE to assess the interpersonal and professionalism competencies of gastroenterology fellows. To provide pilot data on fellows' levels of competence in these areas as assessed through OSCE performance. To share data and insights on the feasibility, acceptability, and usefulness of OSCEs for assessing competence, evaluating training, and improving faculty feedback.
16.	Cho, SP; Clinical Medicine; 2014	To evaluate: feasibility, validity; educational impact; and the role of SLEs in the ARCP. To report the perception of trainees and trainers on the educational value of SLEs. To explore whether SLEs were able to improve the trainees' and trainers' perceptions of WPBAs.
17.	Cogbill, KK; Academic Psychiatry; 2005	To solicit residents' perceptions of how effectively different evaluation methods assessed their competency for each of the 25 required skills defined by ACGME.
18.	Cottrell, E; Education for Primary Care; 2015	To assess the utility of the learning needs analysis (LNA), academic supervisor report (ASR), and current WPBAs in the academic setting.
19.	Craig, S; Emergency Medicine Australasia; 2010	To present ACEM trainees' perceptions and experiences of assessment, supervision and feedback.
20.	Curran, VR; Medical Teacher; 2009	To evaluate the usefulness and merit of the stimulated clinical examination (SCE) as a means of assessing the clinical-skill competencies of entering PGY1 family-medicine residents.
21.	Dadgar, SR; Journal of the Pakistan Medical Association; 2008	“To compare medical students' perceptions regarding Objectively Structured Clinical Examination (OSCE) with Multiple Choice Questions and Oral exam in their semiology course.”

22.	De Almeida Troncon, LE; Sao Paulo Medical Journal; 2004	To report on the student and faculty member responses to this attempt, which highlighted some of the difficulties that may be found in the management of educational change.
23.	Dijksterhuis, MGK; Medical Teacher; 2014	"To qualitatively explore trainees' and supervisors' perceptions on what factors determine active engagement in formative assessment."
24.	Dowling, S; Education for Primary Care; 2007	To evaluate this tool with particular reference to its acceptability, feasibility, and educational impact.
25.	Duffield, KE; Medical Education; 2002	To begin the process of gathering student opinion about assessment at Newcastle Medical School.
26.	Dujin, CCMA; Perspectives of Medical Education; 2017	To determine students' perceptions of meaningful feedback required to prepare for performing an EPA at a designated level of supervision
27.	Finall, A; Journal of Clinical Pathology; 2012	To determine the perceptions and experiences of trainers carrying out direction observation of practical skills (DOPS) assessments in a histopathology setting?"
28.	Foucault, A; Medical Teacher; 2014	"To gain insight into the Concordance Judgment Learning Tool (CJLT) experience."
29.	Gelan, EA; Ethiopian Medical Journal; 2015	"To assess perceptions of final year medical students about the Organized Structured Clinical Exam (OSCE)."

30.	Gnanathasan, CA; Medical Education; 2010	To evaluate IMU students' and examining faculty perceptions of OSCE in the Phase-I medical curriculum.
31.	Green, AR; Ethnicity and Disease; 2007	To better understand and improve an OSCE station emphasizing cross-cultural communication skills (ccOSCE) and interviewed students
32.	Gulbas, L; BMC Medical Education; 2016	"To demonstrate that medical students share an understanding of qualities inherent to high-quality and low-quality written comments and to determine features identifying high and low quality comments to clinical medical students."
33.	Haider, I; Journal of Medical Science; 2016	"To explore the perceptions of MBBS students and their examiners regarding OSCE who appeared in the final professional MBBS examination in 2015."
34.	Harrison, CJ; Perspectives of Medical Education; 2016	To determine the factors within medical schools' assessment systems which aid or hinder student receptivity to feedback.
35.	Harrison, CJ; Medical Teacher; 2015	To investigate the feasibility of electronic audio feedback in OSCEs.
36.	Hays, RB; Education for Primary Care; 2009	To report a comparison of questions in two versions - with and without a brief, generalist clinical scenario - in Year 1 of a new curriculum in one UK undergraduate medical school.
37.	Heeneman, S; Medical Education; 2015	To gain more insight into the following research questions: (i) which elements of the comprehensive programme of assessment do students perceive as supporting or as inhibiting their learning? (ii) what are the factors that students consider important for the active construction of their learning in an assessment for learning environment?

38.	Hicks, PJ; Academic Medicine; 2016	To examine the utility of the assessment procedures developed, the resources required for the assessments, and responses to the project by participating learners and program directors.
39.	Hunter, AR; Medical Teacher; 2015	To gain an understanding of the attitudes of trauma and orthopaedic trainees across the UK regarding their use of PBAs and identify factors influencing any perceived educational benefit.
40.	Ibrahim, NK; Pakistani Journal of Medicine; 2015	To determine the perception of clinical years' medical students and interns about assessment methods used in Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia.
41.	Ingram, JR; Medical Teacher; 2013	"To explore medical specialty trainees' perceptions of MSF to discover whether it had achieved its formative educational purpose… to evaluate to what extent this was affected by trainees' acceptance of the assessment tool and by their experiences of feedback, both as a recipient and as a rater of other trainees."
42.	Jafarzedeh, A; Rawal Medical Journal; 2009	"To design an objective structured clinical examination (OSCE) method for evaluation of medical students in practical immunology course and to compare their perceptions regarding OSCE with multiple choice questions (MCQ) and oral examination."
43.	Jawaid, M; Journal of Postgraduate Medical Institute; 2013	"To evaluate overall perception of students at the end of surgical OSCE examination with an aim to determine their acceptability of this assessment process."
44.	Jefferies, A; Advances in Health Sciences Education; 2011	To describe the utility of a structured oral exam that was designed to assess the 7 CanMEDS roles in a postgraduate subspecialty training program.
45.	Johnson, G; Clinical Medicine; 2008	To describe the methods of using feedback from trainees and supervisors to evaluate the effectiveness of the core medical training (CMT) package (curricula, appraisal, assessment, and the e-portfolio). To report the results and conclusions that have informed the nationwide launch of the CMT programme.

46.	Kalet, AL; Academic Medicine; 2007	To describe the development of the Professional development portfolio (PDP) and share four years of experience with its implementation. To describe the experiences and attitudes of the first students to participate in this program as reported in an annual student survey.
47.	Kania, RE; Archives of Otolaryngology Head and Neck Surgery; 2011	“To report on the creation and administration of an online Script Concordance test (SCT) for ear, nose and throat (ENT), the ENT-SCT.”
48.	Kasanda, CD; Frontiers in Psychology; 2013	To obtain some empirical evidence that would ascertain undergraduate students' perceptions of the University of Namibia's grading and assessment. To find out from the students their perceived evaluation of the quality of the assessment regimes applied by lecturers in the Schools of Medicine and Pharmacy at the University of Namibia. Specifically, the study sought students' perceptions on: general assessment practices used by lecturers in the Schools of Medicine and Pharmacy; provision of feedback to students on assignment and test results; administration and conduct of examinations; fairness of assessment practices; impact of assessment regimes on student morale, study progression, graduation and study costs; ethical conduct of assessments.
49.	Kelly, SP; International Journal of Emergency Medicine; 2010	To evaluate learner perception of four common examination methods in an international educational curriculum in emergency medicine: structured oral case simulations, MCQs; semi-structured oral tests, and essay tests.
50.	Khairy, GM; Journal of Family and Community Medicine; 2005	To examine the feasibility and acceptability of an objective structured clinical examination (OSCE) which was used to examine a large number of medical students for the first time at our institution.
51.	Khorashad, AK; Iranian Red	"To determine the level of satisfaction of the undergraduate medical students of internal medicine at Ghaem Hospital, Mashhad, Iran, in order to detect such problems and contribute to the improvement of OSCE."

	Crescent Medical Journal; 2014	
52.	Kim, H; Practical Radiation Oncology; 2016	To assess the national perception of the American College of Radiology (ACR) in-training examination (ITE) and collect specific feedback regarding the test."
53.	Korszun, A; Medical Teacher; 2009	To evaluate the effectiveness of this longitudinal assessment of professional attitude and conduct (AC) by both reviewing results and obtaining feedback of the perceptions of participating teachers and students after one year of implementation.
54.	Labaf, A; Education for Health; 2014	To assess acceptance of the recently introduced tool by the students.
55.	Laughlin, T; Canadian Family Physician; 2012	"To evaluate the effectiveness of field notes in assessing teachers' confidence and perceived competence, and the effect of field notes on residents' perceptions of their developmental competence."
56.	Lefroy, J; Medical Education; 2015	To understand first the meaning which medical students construct from WBA and feedback with and without grades and, second, how this is influenced by the students' internal and external environments. To use this information to develop more effective, individually tailored feedback processes.
57.	Levine, JC; Pediatric Cardiology; 2015	To describe this system [formal feedback tool to assess noninvasive imaging skills to perform transthoracic echocardiography] and the results of a survey designed to assess fellow's experience with it.
58.	Malhotra, S; Medical Teacher; 2009	To investigate internal medicine residents'' perceptions of the mini-CEX as an educational tool when implemented as a method of formative in-training assessment used on a regular basis during their residency training.
59.	Marrero, I; Academic Psychiatry; 2013	To examine the perspectives of psychiatry trainees regarding methods of evaluating professionalism, utilizing a subset of data from a larger survey of psychiatry residents' perspectives on ethics and professionalism in training.

60.	McCoubrie, P; Medical Teacher; 2004	To develop an evidence-based strategy to use MCQs more fairly so that MCQs can continue to have an important role in assessment and a positive effect on learning.
61.	McCourt, C; Medical Education; 2012	To use a naturalistic approach to explore why students partcipate in passing examination information and how assessment practices may affect students' professional attitudes and behaviour. A secondary aim was to evaluate the acceptability of the corralling procedure from the student perspective.
62.	McKavanagh, P; Postgraduate Medical Journal; 2012	To investigate how foundation doctors in their second postgraduate year perceive WPBAs.
63.	McKinley, RK; Medical Education; 2000	To describe an evaluation of the use of a modified LAP in the formative assessment of the performance of medical students consulting with real patients with particular reference to validity, inter-assessor reliability, acceptability, feasibility, and educational impact.
64.	McLaughlin, K; Advances in Health Sciences Education; 2005	"To examine the effect of blueprint publication on students' perceptions of the validity of the evaluation process."
65.	McLay, RN: Academic Psychiatry; 2002	"To test the value of such an OSCE within the third-year psychiatry clerkship at Tulane University, we established an exam using a single simulated case."
66.	Miller, A; British Medical Journal; 2010	"To investigate the literature for evidence that workplace based assessment affects doctors' education and performance."
67.	Murdoch-Eaton, D; Medical Education; 2012	“To investigate how undergraduate medical students recognize, respond to and utilize feedback” “To determine if there are maturational differences in understandings of the role of feedback across different year cohorts in a medical school”

68.	Nagoshi, M; Medical Education; 2004	To describe our experience with a multistation, Geriatrics Standardized Patient examination (GSPX) developed to evaluate the clinical skills of trainees at three levels: medical students, residents, and geriatrics medicine fellows.
69.	Nasir, AA; Journal of Surgical Education; 2014	To explore students' perceptions about the acceptability of OSCE process and to provide feedback to be used to improve the assessment technique.
70.	Nesbitt, A; The Clinical Teacher; 2013	To assess student perception of WPBA at UCLMS, and to determine whether re-designing the form had altered this perception.
71.	Nestel, D; Assessment & Evaluation in Higher Education; 2011	To explore students' responses to OSCEs and the IPPI when used as formative assessment.
72.	Nikendei, C; Medical Teacher; 2007	To determine the main benefits and impressions of [ward round] approach from the perspective of (a) final year students and (b) standardized patients
73.	Nofziger, AC; Academic Medicine; 2010	To determine what types of peer feedback do medical students remember months to years later, what kinds of immediate and delayed reactions do students have to peer feedback, and what transformations in attitude and/or behaviours do students make in response to peer feedback.
74.	Nowacki, AS; Frontiers in Psychology; 2013	To investigate the impact of assessment method, portfolio only (P) vs. portfolio and grade (PG), on student performance and student perception of their learning experience."
75.	Olson, LG; Medical Education; 2000	"To determine whether the structured question grid achieved its purpose of improving the perceived reliability and fairness of the assessment…"

76.	Papinczak, T; Advances in Health Sciences Education; 2007	To explore student attitudes to, and perceptions of, peer assessment concerning student fulfilment of roles and responsibilities within their PBL tutorials.
77.	Parikh, PP; Journal of Surgical Research; 2015	To report on the relationship between simulation-based palliative and end-of-life care OSCE ratings and the key psychosocial competencies of communication skills, trust, and self-assessed empathy as measured by standardized instruments. To examine the perceptions and experiences of students regarding their palliative care and end-of-life OSCE training.
78.	Parslow, GR; Biochemistry and Molecular Biology Education; 2010	N/A
79.	Pearce, I; Annals of the Royal College of Surgeons of England; 2003	To evaluate the effect of the Calman reforms on specialist registrars (SpRs) in urology with respect to their educational goals, their experience of the RITA process and its value in preparing them for their chosen consultant careers.
80.	Perera, J; Medial Teacher; 2008	To investigate the relationship and degree of correspondence between perceptions of teachers with student perceptions and expectations with regard to feedback received during learning.
81.	Perron, NJ; BMC Medical Education; 2016	To evaluate whether the content and process of feedback varied according to the tutors' profile.
82.	Pierre, RB; BMC Medical Education; 2004	To evaluate student overall perception of the end-of-clerkship OSCE, determine student acceptability of the process and provide feedback to enhance further development of the assessment.

83.	Plant, JL; Advances in Health Sciences; 2013	To examine the process of informed self-assessment in action in a specific educational context, with a goal to better understand how, why, and to what extent resident physicians adjust their self-assessment based on external information.
84.	Rafique, S; Journal of the Pakistan Medical Association; 2013	"To obtain student feedback on teaching and assessment at Nishtar Medical College, Multan, Pakistan, which is a well-reputed public sector medical institution of the country."
85.	Raheel, H; Journal of the Pakistan Medical Association 2013	"To explore the perceptions of undergraduate medical students about the OSCE"
86.	Rahman, SA; Medical Teacher; 2001	To develop appropriate formative assessment strategies as a means of promoting relevant learning outcomes in paediatrics.
87.	Rees, C; Medical Education; 2002	To determine students' views on how communication skills are assessed, specifically what students like and dislike about communication skills assessment.
88.	Ringsted, C; Medical Education; 2004	To investigate the experiences and thoughts of programme directors, assessors and trainees about a recently introduced ITA programme.
89.	Rudland, JR; Medical Teacher; 2011	To gather feedback from students on their perceptions of the computerised test, focusing on: the acceptability (ease of use and accessibility) of the computer-based format; whether resources were used to complete the test; the value of the immediate feedback (a score); perceived positive aspects of the computer format and areas for improvement.
90.	Sabey, A; Education for Primary Care; 2011	To establish how the new system of WPBA is working in day-to-day practice for a cohort of GPSTs in hospital posts, seeking their views of the process and experience of assessments, their perceptions of assessors' understanding and skills, and their suggestions for improvement.

91.	Sadia, S; Anaesthesia Pain & Intensive Care; 2009	"To describe the perceptions of undergraduate medical students regarding OSCE and its comparison to MCQ, essay questions and viva voce."
92.	Sargeant, J; Academic Medicine; 2011	To understand how learners and physicians familiar with structured self-assessment interventions perceived and used self-assessment in clinical learning and practice. To determine the components and processes that comprise self-assessment and the factors that influence them.
93.	Schwaab, J; Academic Emergency Medicine; 2011	To explore the use of SL virtual simulation technology to administer mock oral examinations to EM residents.
94.	Shafi, R; Medical Teacher; 2010	To share our experience of bringing relevance to basic science laboratory practice examinations by conducting competency-based IPEs, and to analyze its efficacy for the students.
95.	Sharma, N; British Journal of Hospital Medicine; 2015	To assess medical students' perceptions of the situational judgment test. To quantitatively assess whether students, having taken the situational judgment test, felt that it in fact was a worthwhile measure of these five attributes. To qualitatively gauge medical students' opinions of the situational judgment test following completion.
96.	Sharma. N; Medical Teacher; 2012	To develop and evaluate this method of assessment in a clinical clerkship over a single academic year, focusing on the feasibility and acceptability of the method to students and assessors.
97.	Sharma, S; International Journal of Applied and Basic Medical Research; 2015	"To explore perceptions of PGs and teachers about factors that determines active engagement in formative assessment."

98.	Shenwai, MR; Journal of Clinical and Diagnostic Research; 2013	To introduce structured oral examination (SOE) as a novel assessment tool to first year MBBS students in physiology and evaluating the process by taking feedback from the students and faculty.
99.	Smith, S; The Clinical Teacher; 2012	To develop and evaluate a test of automaticity of peripheral venous cannulation skill, appropriate to the level of a medical student.
100.	Spanager,L; International Journal of Medical Education; 2015)	To determine what characterizes the content of feedback conversations regarding trainee surgeons' non-technical skills when stimulated by a tool, what characterizes feedback conversations regarding trainee surgeons' non-technical skills in terms of feedback style used, and how trainee surgeons and their supervisors perceive the usefulness of the feedback stimulated by a tool.
101.	Suhoyo, Y; BMC Medical Education; 2017	To validate the influence of five feedback characteristics on students' perceived learning value of feedback in an Indonesian clerkship context.
102.	Tayem, YI; Sultan Qaboos University Medical Journal; 2015	To examine medical students' perceptions of intragroup peer assessment in a problem-based learning setting.
103.	Tsugawa, Y; Medical Education; 2009	To evaluate the reliability and validity of a Japanese version of the P-MEX by assessing the professionalism of senior residents in internal medicine at a Japanese teaching hospital.
104.	Tweed, M; Medical Education; 2001	To gain experience in the development and use of tools for determining face validity, and by these means to evaluate a new method of assessment for final-year students.
105.	Vanlint, A; Australasian	To examine the feasibility and acceptability of an OSCE.

	Journal of Ageing; 2016	
106.	Vishwakarma, K; Indian Journal of Pharmacology; 2016	To evaluate OSPE for the assessment of practical skills in pharmacology examination for undergraduate medical students and compared it with conventional practical examination (CPE).
107..	Wade, L; Advances in Health Sciences Education; 2011	"To compare students' perceptions of and preparations for the Progress Test (PT) at two medical schools."
108.	Watling, C; Academic Medicine; 2008	“To describe residents’ experiences with, perceptions of, and reactions to the ITER process at a large medical school in Canada”
109.	Weller, JM; British Journal of Anaesthesia; 2009	To explore the attitudes of trainees and specialists towards the mini-CEX and develop recommendations for assessor training and implementation of mini-CEX in anaesthesia.
110.	Wiener-Ogilvie, S; Education for Primary Care; 2012	To examine whether trainees identified during the fCSA as 'at risk of failing the MRCGP CSA exam' are more likely to fail the MRCGP CSA exam later on in the year than those trainees that were not identified as 'at risk'. To assess the acceptability and value of fCSA to trainees and trainers.
111.	Winkel, AF; Evaluation and the Health Professions; 2016	To determine learners' perception of this type of assessment.
112.	Wittels, K; Academic Emergency Medicine; 2013	To assess the inter-observer reliability of a checklist tool compared to Dreyfus five-level assessment of performance using both direct faculty observation and delayed video observation of simulated cases in sepsis and cardiogenic shock resuscitation.

113.	Wu, V; Journal of Population Therapeutics and Pharmacology; 2015	"To develop and validate the McMaster Prescribing Competency Assessment (MacPCA), on online tool suitable for evaluating clinical pharmacology knowledge and prescribing skills of medical trainees in Canada."
114.	Zyromski, NJ; Current Surgery; 2003	To specifically evaluate resident perception of the OSCE.