THE SOURCES OF MEASUREMENT ERRORS IN THE EPA-BASED ASSESSMENT OF COMPETENCIES AND ENTRUSTMENT DECISIONS.

A SCOPING REVIEW

Submitted by: Sehar Ashraf

Assistant Professor of Psychiatry

Fellowship in Psychiatry

College of Physicians and Surgeons Pakistan

Islamabad, Pakistan

Submitted for the Degree of Master's in Health Professions Education: Assessment and Accreditation, Keele University

Date: August 22nd, 2021

Declaration

I certify

(a) That the above dissertation/project is my own account, based upon work actually carried out by me, and that all sources of material not resulting from my own experimentation, observation or specimen collecting, including observational data, have been clearly indicated.

(b) That no part of the work incorporated in the above dissertation/project is a quotation from published or unpublished sources, except where this has been clearly acknowledged as such, and that any specific direction or advice received is also properly acknowledged.

(c) That I have read, understood, and abided by the terms of Regulation VII1.5 below: CONDUCT WITH REGARD TO DISSERTATIONS, PROJECTS, ESSAYS ETC., WHICH FORM PART OF A FINAL EXAMINATION FOR ASSESSMENT PURPOSES 10 (a) Titles must be approved or specified by the Department concerned in accordance with the provisions in the Calendar. (b) The dissertation, projects or essays etc. shall be in the student's own words, except for quotations from published and unpublished sources which shall be clearly indicated as such and be accompanied by full details of the publications concerned. The source of any map, photograph, illustrations, etc. shall be similarly indicated. The student shall indicate clearly the sources, whether published or unpublished, of any material not resulting from his own experimentation, observation or specimen collecting, including observational data.

Students will be required to sign a statement to that effect. Failure to comply strictly with these requirements may be construed as cheating.

Signed… Sehar Ashraf Date: August 22nd, 2021

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Jack Boulet for his guidance, invaluable feedback and help for completing my work.

I would also like to thank Professor Janet Grant for her continuous support and encouragement during my period of study.

And of course, how can I forget Leo, for his untiring efforts to run the course and online sessions smoothly and for checking in on us from time to time, to keep us on the track.

I really appreciate the efforts of the entire team and faculty for their contributions.

Lastly, I would like to thank my parents for their prayers and for believing in me, and my son, who has always been a source of motivation for me.

Contents

List of abbreviations:

ACGME	Accreditation Council for Graduate Medical Education
CBD	Case Based Discussion
CCC	Clinical Competency Committee
CLT	Cognitive Load Theory
CU	Construct-under representation
CIV	Construct-irrelevant variance
CMBE	Competency Based Medical Education
EPA	Entrustable Professional Activities
KFEs	Key-feature Examinations
MCQs	Multiple Choice Questions
MSF	Muti-source feedback
Mini-CEX	mini Clinical Evaluation Exercise
OSCEs	Objective Structured Clinical Examinations
PD	Programme Director
WPBA	Workplace-based Assessment

Abstract

Context: Around the turn of the millennium, the increased use of Entrustable Professional Activities (EPAs) in the assessment of competencies in higher medical education makes it important for the educators to ensure the reliability and validity of these assessments by minimizing measurement errors. Due multiple tenets of Entrustable Professional Activities-based framework, from basic structure to training of faculty, feedback to the trainees, expertise in formative and summative assessments and institutional leadership, make it impossible to avoid certain pitfalls in the smooth assessment of competencies. The aim of this scoping review is to analyze the measurement errors and discuss them in the light of cognitive psychology and in the context of validity framework for entrustment decisions.

Method: I used Arksey and O'Molley's framework of scoping review to synthesise the findings in my study. Then I performed a systematic search on PubMed, Ovid and Google Scholar for articles published in English and on any date from the inception of these databases to June, 2021. Articles were screened for eligibility using the inclusion criteria. Themes were generated using the process of content analysis.

Results: Thirty articles were included in the review. Eight themes were generated: measurement errors due to factors related to trainee, rater, trainee-supervisor relationship, assessment method, task, implementation of EPA based model, entrustment scales and contextual factors.

Conclusion: Entrustable Professional Activities are essential to the assessment of competencies, however, the potential sources of measurement errors act as barriers to the successful development of EPAs and implementation of entrustment decisions. In future, high-level evidence based research is needed to target these areas to ensure validity and reliability of entrustment decisions based on EPAs.

Introduction:

Since the introduction of 'time-variable' model of medical education, the concept and practices of curriculum delivery and assessment for medical educators, trainers and programme directors have changed (Gruppen et al 2018). Time-variable training de-emphasises time and highlights the importance of 'fixed outcomes' to set the standards of health care delivery and patient safety. The individual learners enter a training programme with variable skills and levels of competency. They acquire further competencies at different rates as they progress in their training. Hence, Competency Based Medical Education (CBME) can be defined as an outcome-based method of curriculum design, implementation, and assessment using a framework characterised by fixed outcomes and variable time (Frank et al 2010).

Competency Based Medical Education requires robust assessment which is essential to effective learning (Iobst and Holmboe 2020). This systematic approach to assessment, which comprises multiple integrated encounters with assessors and supervisors longitudinally over time, is referred to as programmatic assessment (Iobst and Holmboe 2020). It helps to ensure valid and reliable evaluation of learner progression towards unsupervised practice. The role of the rater has received increasing attention as one of the strengths of performance based assessment. Multiple longitudinal clinical encounters over time can provide richer data on the trainee's abilities (van der Vleuten et al 2012).

The reliability of performance based assessments varies according to setting, training of the assessor, assessment tool being used and the level of complexity of the competency

being assessed (Pelgrim et al 2011). Performance based assessment gives a better overview of the trainee's abilities, but it also increases the potential risk of rater errors in the assessments (Feldman et al 2012). Interrater variability has raised concerns about the validity and reliability of the assessments (Gingerich et al 2011). The quality of the assessment can be improved with adequate training of the raters. However, the results of research concerning the impact of rater training on measurement errors is mixed (Lee, Brain and Martin 2017).

Assessment of complex competencies such as communicator, leadership and medical expert requires efforts in real clinical setting (Wass et al 2001). High stakes decisions are based on inferences drawn from observations of the behaviours. The efforts to translate these key concepts into objectively identifiable behaviours can fail to represent the true essence of what they actually meant to measure (Whitehead et al 2015). That is, a person's performance may not reflect his/her "true" competence as these skills and attitudes are difficult to measure.

Selection of the constructs and tasks for the assessment has major implications on the quality of the decision or the judgement of the assessor (Albanese and Case 2016). Data needs to be gathered from multiple sources on several occasions using various methods to arrive at a valid decision. The measurement tools such as forms, portfolios, multi-Source feedback (MSF), checklists, and faculty observations based on trainees' performance such as case based discussions (CBDs) and mini clinical evaluation exercise (mini-CEX) can provide data for an assessment decision (Lockyer et al 2017).

The ability of a rater to assess a clinical interaction is limited by mental capacity, time restraints, cognitive abilities and environmental factors (Gigerenzer and Selten 2002). Attention span of human mind is finite, the memory systems can become saturated, decision making can be influenced by emotions, past experiences and perceptions, and judgement is relative (Eva 2018). Raters of equal calibre, experience, and qualification, when asked to mark the same interaction of a trainee with a patient, based on a video recording produced scores with greater variations which was a source of inconsistency in pass/fail results (Cook et al 2010). Hence, the confluence of these multiple factors make the assessors more vulnerable to potential measurement errors and biases.

The purpose of this scoping review is to better understand the current practices in Entrustable Professional Activities (EPAs) based assessments and to identify the issues and factors which can be the potential sources of errors in making entrustment decisions. Secondly, my area of interest is cognitive psychology, which can help me understand the mental processes in addition to contextual factors that can lead to inter-rater variability and associated measurement errors. I will begin by giving an overview of the components of EPA-based assessments. Then I will discuss multiple theories from cognitive psychology which underpin these mental processes and rater judgement. Lastly, I will try to relate these processes to measurement errors and the validity of entrustment decisions.

Literature review

The convergence of medical education with competency based delivery of curriculum and assessment has changed the way that medical educators assess their trainees (ten Cate and Scheele 2007). Competencies of a trainee as defined by Frank et al (2010) include abilities across multiple domains of physician performance in a certain context, which are dynamic and change with time, experience, and setting. In a residency programme, the trainees are supposed to demonstrate a minimum level of competency in order to ensure readiness to act as an independent practitioner (ACGME 2012). For the assessment of these core competencies, it was argued that they are too granular for any meaningful assessment. Moreover, the individual domains of the competencies do not actually add up to the performance outcomes of the trainees (Whitcomb 2007, lobst 2010). To address this gap, a conceptual framework of Entrustable Professional Activities (EPAs) was introduced in 2005 (ten Cate 2005). This concept is important as it directly connects the competencies to clinical practice.

Entrustable Professional Activities (EPAs)

An EPA is a unit of professional work that a trainee can be entrusted with once he/she has acquired a sufficient level of competence (ten Cate and Scheele 2007). Entrustable Professional Activities are described as the skills that must be learnt at the workplace and require the learner to be able to integrate a number of competencies from several domains. Central to the idea of EPA is the concept of trust. As described by ten Cate, “Trust is a central concept for safe and effective health care. Patients must trust their physicians,

and health care providers must trust each other in a highly interdependent health care system" (ten Cate 2013, p.1).

The purpose of EPAs is to operationalise competency based medical practice linking clinical skills to progressive proficiency and autonomy of the graduates in the delivery of safe health care (ten Cate and Scheele 2007). Licensing and registration establish their readiness for unsupervised clinical practice. Description of EPAs ensure clarity amongst educators and employers as to what exactly to expect from the individuals who are trusted with this entitlement. Entrustable Professional Activities invite the trustees to legitimate practice of principles of safe medicine and health care to become members of community of professionals (ten Cate and Taylor 2020).

Entrustable Professional Activities do not tend to replace competencies, rather the idea is to integrate competencies into assessment which has a target to achieve. Competencies can be judged with the help of key constructs in medical education such as knowledge, professionalism, team work etc. which are intangible (Gruppen et al 2018). Proper mapping of EPAs against the competencies require a competency framework to identify the roles or domains; for example, Canadian Physician Competency Framework CanMEDS (CanMEDS 2015), ACGME framework of core competencies in USA (ACGME 2012) and Tomorrow's Doctors Framework in UK (GMC 2020). To run a programme successfully in a region, this mapping will yield an EPA-competencies matrix to establish the key areas or content for assessment (ten Cate et al 2015).

In 1999, Accreditation Council for Graduate Medical Education (ACGME), developed six core competencies (Patient Care, Medical Knowledge, Practice-Based Learning and Improvement, Interpersonal and Communication Skills, Professionalism, and Systems-Based Practice) (Nasca et al 2012). Next, it began restructuring the educational system based on EPAs, competencies, and milestones. CanMEDS competency framework enlists these individual roles as Medical Expert, Collaborator, Scholar, Leader, Professional, Communicator and Health Advocate (Frank, Snell and Sherbino 2015). Milestones are stages in the development of the competencies which are observable markers of physician's abilities from novice to expert (Englander et al 2017). Each EPA requires integration of multiple competencies across various domains; for example, an EPA may include competencies from domains of Professionalism, Medical knowledge and Communication and Interpersonal skills, using the ACGME framework. Each EPA can be mapped to one or more domains of competence, which may further include one or more competencies in each domain. Each speciality has specific milestones which determine the progression of the trainees to the next level of entrustment. Milestone is a smaller construct of a competency in ACGME framework and represents behaviours corresponding to the level of achievement of that competency (Englander et al 2017).

Entrustable Professional Activities based teaching and assessment have evolved as major innovations in medical education over last two decades at under-graduate as well as post-graduate levels (lobst and Holmboe 2020). Despite the wider popularity and acceptance of EPAs, they lack a common language and definition of their components. For instance, the

domain of milestones differs in Europe as compared to the one described under the ACGME framework. Milestones are conceptualised as bigger than EPAs or competencies (Englander et al 2017). Moreover, the assessment of competencies by the clinical supervisors may be incomplete or out of date (Tekian et al 2020). For example, the concept of care coordination at the time of discharge may vary from physician to physician and between two different systems of care. Providing care is a complex coordination activity and involves the patient and his/her family and usually requires an interdisciplinary team (Coleman et al 2004). Physicians may not have expertise in competencies such as managing liaison consultations and referrals, patient safety and quality measures, inter-professional team work and follow up on lab work (Tekian et al 2020). Physicians are not trained in these key areas; hence, assessment of these competencies may be deficient in entrustment decisions.

Another issue with EPA-based assessment of competencies is that faculty tend to follow their own mental models (self) while assessing the trainees. In a study in Canada, Apramian et al (2015) showed that the assessment of competencies in surgery may be shaped by the surgeon's own view of his/her preferences and principles. Principles in surgery are definite and non-negotiable whereas preferences are based on alternate approaches which may or may not be acceptable (Apramian et al 2015). Preferences may give rise to idiosyncrasies and difficulty in measurement, hence pose a threat to validity of entrustment decisions. In a fairly recent study, raters were asked to observe a station on breaking bad news in Objective Structured Clinical Examination (OSCE). Most of the faculty were not well versed

with a framework for breaking bad news. The primary frame of reference they used was self. If 'self' is the frame of reference, it leads to rater biases and causes inaccurate and ineffective entrustment judgement (Kogan et al 2010).

EPAs provide a structure for workplace-based assessment and indicate the key features in a curriculum to enable the trainees reach up to the level of entrustment (Englander 2016). Generally, there are five levels of entrustment. Level 1: Trainee is not allowed to perform the activity at all. Level 2: Trainee is allowed to perform the activity with direct supervision (supervisor present and proactive in the room). Level 3: Trainee is allowed to perform the activity with indirect supervision (supervisor not present but is immediately available if needed). Level 4: Trainee is allowed to perform the activity independently (with distant supervision not immediately available). Level 5: Trainee is allowed to provide supervision to junior learners doing the activity (Hart et al 2019).

Entrustment decisions:

Entrustment decisions may take place as a part of routine clinical interaction between the trainee and the supervisor or it can be a high-stake decision resulting in certification. The former, more formally known as ad-hoc decisions usually require fewer observations by the supervisor. On the other hand, for high-stakes summative decisions, the adequate sampling of the content, the number of observations by experts, peers or patients and consistency shown by the trainee in that specific EPA are important areas to be considered for making a grounded decision based on sufficient data (ten Cate and Taylor 2020).

Grounded trust leads to 'summative decisions' which assigns higher level of responsibility to the trainee and a graded decrease in supervision (ten Cate et al 2016). Ad-hoc as well as summative decisions enable the supervisors and programme directors make informed judgement on the level of autonomy of the trainee.

The judgement has to be holistic in order to reach an entrustment decision. Entrustment in medical education is equivalent to granting responsibility to the trainee which may include all the implicit and explicit variables in relation to the clinical responsibility (ten Cate et al 2016). Sometimes all the variables are difficult to describe. By virtue of working with the trainees, the supervisors and the clinical heads seem to 'just know' about their trainees based on their experience. Due to an abstract nature of the competencies, it is difficult to acknowledge a complete set of all the variables which come in to play as a supervisor deems a trainee worthy of entrustment. Instead of relying on checklists, when the entrustment decision is based on EPAs, it can incorporate the expert opinion of the assessors in formal assessments (ten cate et al 2016).

To trust a trainee in an entrustment decision is more than the assessment of knowledge, skills and ability. Trusting requires careful assessment by the supervisor of the risks to the patients and benefits as to how the trainee will act in difficult, unpredictable and intense situations without anyone intervening or observing him/her (ten Cate and Hoff 2017). Mayer's Integrative Model of Organisational Trust include benevolence and integrity as essential qualities in addition to ability (Mayer, Davis and Schoorman 1995). Ability is a set of skills and competencies which are domain specific and enable the trustee to have

influence within that specific area. Benevolence is a mutual relationship between the trustor (mentor) and trustee, whereby, the mentor wants to help the trustee without any reward. Benevolence is a positive orientation of the trustee towards the trustor. This orientation would guide the trustees' motivation and intentions towards the responsibility he/she is assigned with. Integrity based on the trust between the trustor and the trustee is the trustor's perception or faith in the trustee that the trustee will adhere and conform to a set of principles acceptable by the trustor (Mayer, Davis and Schoorman 1995). Expanding on the same concept, Kennedy et al (2008) suggested that trust in a trainee depends on four attributes: conscientiousness, knowledge and skill, truthfulness and discernment of own limits. In another study, to synthesise the key trainee factors, the authors performed a content analysis of all the relevant studies in the last fifteen years (Chen and ten Cate 2020). From the descriptions of the trainee qualities, five themes were generated: reliability, humility, agency, capability and integrity. Capability is the task-specific evidence based knowledge and skill along with sound clinical judgement and decision taking ability (Kennedy et al 2008). Reliability is the readiness of the trainees to take responsibilities diligently, they are conscientious, attentive, follow through on the tasks assigned to them and want to be held accountable for their attitudes and skills (Chen and ten Cate 2020 and Kennedy 2008). Humility is willingness to accept that one is fallible in uncertain situations, be willing to seek help when needed and knowing one's limits in knowledge and skill is another quality of a trainee which is considered important (Ginsburg et al 2010, Kennedey et al 2008). Integrity in entrustment decisions is about upholding rules

and regulations, truthfulness, honesty, prioritising patient needs over one's own needs, keeping up to professional behaviours, respecting patient autonomy and being empathetic (Chen and ten Cate 2020, Ginsburg et al 2010). Lastly, Agency is demonstration of eagerness to learn by the trainees, their passion, enthusiasm for their work, offering solutions to their problems and proactive behaviour towards personal and professional growth (Ginsburg et al 2010, Kennedy et al 2008). All these personal attributes form an important intricate framework of qualities which supervisors find helpful in entrustment decisions, considering trainee features beyond knowledge and skills.

Any clinical task in medicine involves some unseen risk, therefore, entrustment decisions have this hidden component involved where this risk is estimated with respect to the readiness of the trainee to perform in uncertain situations and his/her ability to cope and adapt. In other words, assessment of entrustment includes evaluation of 'Adaptive expertise' (ten Cate et al 2021). Assessment of the prospective risk involved would indicate the right approach to estimation of adaptive expertise by the rater or supervisor; for example, "do I feel the learner is ready for less supervision on this EPA", rather than looking at it in a retrospective way, "how much supervision was provided with an activity" (ten Cate et al 2021, p. 3). In an entrustment decision, this readiness is based on EPA framework. If a trainee has been entrusted with any specific EPA, he/she is required to show the same level of expertise in all other EPAs associated with optimum level of patient care. Moreover, based on the dynamic nature of medical practice, the trainee may not be able to maintain or exhibit the same level of proficiency over a period of years, and there can be a change

in what the trainee was entrusted with and what he would be entitled to perform later on (ten Cate et al 2021). Hence, the specific requirements for entrustment decisions based on aggregate of EPAs that may include trust along with adaptive expertise are limited (Wijnen-Meijer et al 2013). The trainee attributes other than knowledge and skills, such as integrity, benevolence and patient-centeredness are not task specific. These elements may not only be considered important as a part of big decisions, instead, the educators and clinical supervisors may need to focus on these factors in relation to entrustment in their routine clinical interaction with the trainees, closely observe them whenever possible, as the learners apply their prior knowledge to new situations, exhibit help-seeking behaviours, prioritise their patients' needs over personal needs, volunteer information about any omissions in the patient care and ask for supervision in difficult situations (ten Cate et al 2021). Maybe the educators, clinical supervisors, and clinical committees need to understand, practice and gradually evolve 'entrustment thinking' which may include task-specific and task-non-specific trainee features. This sort of approach should be a part of all workplace-based assessments, clinical simulations, semi-live simulations, entrustment-based discussion and other clinical encounters (ten Cate et al 2021).

Ad-hoc entrustment decisions are made by the frontline assessors who are working with the trainees in the clinical environment (Choo et al 2014) whereas, most of the summative decisions are made by Clinical Competency Committee (CCC) and, ultimately, final recommendations are provided by the Programme Director (PD) based on supervisory role categorization, which is a type of entrustment decision. In most of the cases concerning the

final entrustment decisions, PDs agree with the CCC members, but not always. In one study based on the gap in the relationship between CCC and the PD, justifications and reasons were explored for discrepancies between the CCC members and PD. It was found that supervisors gave more justifications on the performance of the trainees as they move the trainee to a greater supervisory role; for example, his professional abilities, communication skill and trustworthiness. On the other hand, PDs gave fewer justifications as they moved the trainee to a lower supervisory role category; they based the decisions on the experience of the trainee, for instance insufficient experience, which is not consistent with the competency based medical education model (Schumacher 2019).

The supervisors may need to understand the relationship between a number of factors related to trainee, clinical environment and the skill to be assessed (Mohr, Batalden and Barach 2004). Familiarity with the trainee, resident's level of training, difficulty of the EPA, condition of the patient and availability of any other medical personnel may affect entrustment decisions. The interplay of these factors can be a source of anxiety and ambivalence for the supervisors which can lead to difficulty in making the valid and reliable entrustment decisions and can be equally harmful to the patient (Sterkenburg et al 2010).

Efforts made towards the training of the assessors have generally been unsuccessful in producing long lasting effects. It may make them more stringent or lenient but little change in inter-rater reliability has been observed (Holmboe, Hawkins and Huot 2004). Similarly, an attempt was made to make the constructs more comprehensive by increasing the items of the scale. The inter-rater reliability and feedback were compared when assessors were

asked to observe the same behaviours on a scale with two dimensions versus seven dimensions (Teverus and Eva 2013). The authors found that the quality of the assessment declined when the raters were asked to evaluate the performance more comprehensively because completing a long scale is a difficult job and the assessors relied more on their memory and on what they found more obvious and easiest to observe (Eva et al 2007). Hence, *tabula rasa* (meaning clean slate) does not exist for assessors. Applying a constructivist model, it would be wiser to help our assessors build on what they already know and expect them to stay within the limits of the human mind, following the principles of cognitive psychology. Tavares et al (2016) suggested that by decreasing the number of dimensions to be assessed and by aggregating the scores of multiple assessors, assessment designers can increase the reliability of the ratings and provide more feedback to the learner. By doing so, it will return us to the situation which competency based medical education is trying to overcome: remove the focus from an overly narrow scope of medical practice (Frank et al 2017).

Having diverse views on the same performance may not reduce the objectivity of the designed tests (ten Cate O and Regehr 2019). It is about time educators gain more insight and embrace this truth that subjectivity cannot be avoided on performance based assessments. Objectivity can be understood as a single but socially constructed shared perspective on subjectivity of the assessment. An entrustment decision is context dependent and judgement of a performance is a social event. Hence, multiple assessors mean multiple contexts and these assessors may tend to differ in their opinions of the same

performance (ten Cate and Regehr 2019). But it does not mean that it creates noise in the ratings, rather it can be perceived as variation in the perceptions by the raters in the context of a social act as there is no absolute truth about the performance. To reach a summative entrustment decision, it is mandatory for a team of experts to reach a consensus through a thorough and coherent evaluation of the readiness of a trainee to be trusted.

Mental Processes and Role of Cognitive Psychology:

The decision on the trainee's performance is a multi-attribute task and the rationality of the judgement is inevitably limited by an interplay of the psychological theories, the training of the assessor, time restraints and the extent to which all these factors can be controlled (Gigerenzer and Selten 2002). New information is acquired and judged by the assessor in the light of previously formed schemas. Schemas are mental images which have formed over a period of years, based on personal experiences and the knowledge gained (Bransford, Brown and Cocking 1999). They help us in perceiving, organising and utilising the information. Moreover, cognitive processing of the information requires a complex system to make codes, connections, networks and constructs to ensure proper storage and retrieval. Such a complex network in brain is called Memory. A rater uses his/her working memory as he/she assesses a performance, whereas the knowledge and clinical information gathered during his/her years of training is stored in long-term memory (Baddley 1986). These memory codes can be accessed and retrieved based on how schema formation has taken place. During an assessment, there is a lot of information and many

interpretations to be made, which may expose the assessor to considerable level of stress and may have implications on his/her perceptions, working memory, information processing and attention span (Byrne, Tweed and Halligan 2014). By making the assessments more objective, longer and atomised, the assessor may be overwhelmed due to divided attention and saturation of working memory systems. Hence, the whole exercise can be counter-productive and may cause incongruence between the observed and true ability of the examinee.

Heuristics and biases

The examiners and assessors are generally predisposed to errors and variability due to their personal biases and attributes when making the entrustment decisions in performance based assessments. Research on human judgement shows that there are three main mental processes that help humans in decision making: heuristics and biases, natural decision making, and social cognition theory (Berendonk, Stalmeijer and Schuwirth 2013). Heuristics are mental short cuts that help to make a decision providing general rules of thumb without empirical evidence (Tversky and Kahneman 1974). These heuristics lead to errors and biases in judgement. For example, availability heuristic is often used in situations when the people are asked to judge plausibility of an event by the ease with which occurrences can be recalled in mind (Tversky and Kahneman 1974). For instance, one may judge the average age of having a heart attack by recalling similar event in his/her acquaintances.

Natural decision making, and Social Cognition Theory

Naturalistic decision making focuses on how humans arrive at a decision in uncertain situations where actuarial methods are not feasible and a quick response is required (Berendonk, Stalmeijer and Schuwirth 2013). Social cognition theory determines the contribution of personal and social factors in decision making. Judgement cannot take place in isolation and it depends upon the motivation of the assessor, personal goals, social environment and local practices (Levy and Williams 2004).

Impact of Human Emotions on decision making

In addition to internal, external and cognitive factors, the human emotions have an important role to play in rater's judgements. Pellegrino et al (2001) suggested three facets of the learner's assessment in 'an assessment triangle': cognition (assessor's mental models), observations (task assigned to elicit learner's responses) and interpretation (reasoning from observations). 'Emotions' are considered the fourth facet (Gomez-Garibello and Young 2018). This tetrad of assessment has changed the role of an assessor from a measurement tool to a rater who supports and facilitates learning by judging a social act intertwined with cognitive and emotional components in the context of a competency based assessment.

Gestalt, role tension and idiosyncrasy

More detailed view of the cognitive and contextual factors has been discussed by Lee, Brain and Martin (2019). An assessor may struggle with the dual role he/she has to play of an

evaluator who has to assess the trainee, as well as a teacher whose duty is to identify the weaknesses of the trainee and give a constructive feedback at the same time. Gestalt and idiosyncrasy are the other factors related to the assessor (Lee, Brain and Martin 2019). Gestalt or overall impression is a strong determinant of global rating by the assessor. Few assessors have a tendency to form the overall impression within the first few minutes of interaction, whereas others like to look at the individual domains before reaching a final judgement. Idiosyncrasy is described as variable approaches to task assessment, standards and frames of reference to judge the performance, conduct of an assessor and management of the patient. Many raters use learner's peers or themselves as a frame of reference which can result in inter-rater variability. Assessors often find it difficult to do justice to their role as an assessor and feel torn, tensed and confused between these dual roles of teacher versus rater.

Now the current debate about measurement errors has switched from causes of variability to evaluation of this possibility, whether these errors can be overcome. The rater-based measurement errors can be viewed as socio-cognitive events having a definitive social and cognitive aspect. Impressions are formed based on factual knowledge, memory recall, past experiences, inferences and evaluative reactions towards the trainee (Gingerich, Regehr and Eva 2011). Researchers have suggested that preformed social images exist within the mind of the assessors. This social categorisation is automatic and happens without the voluntary participation of the assessor. Any deliberate efforts to counteract this tendency to social categorisation can actually cause adverse effects on impression formation

(Wegner 1994). A thorough understanding of the cognitive processing of information is important to help the educators make informed decision on the judgement of a performance.

Person Model

Person Model describes the characteristics of the individuals belonging to a specific category that allows to make predictions as how the person is likely to behave in a certain scenario (Park, De Kay and Kraus 1994). In a study comprising of 69 participants, the responses of raters were clustered together based on how they evaluated the interaction of a ratee with a friend in a four-minute video (Mohr and Kenny 2006). The raters described the ratee into three predominant categories. Category one describes the person as kind, considerate, expressive and friendly. Category two includes traits such as nervous, insecure, easily distractible and indecisive. Category three described the ratee as insensitive, rude and obnoxious. Person Model may act as a framework to give a plausible explanation for why a rater forms an impression of a ratee and how it causes variation in the ratings, attributable to this unique relationship between the rater and the ratee (Gingerich, Regehr and Eva 2011).

Clinical reasoning and Information processing models:

The deductive reasoning models based on the clinicians thought processes take their origin from the preliminary but seminal work done by Elstein, Shulman and Sprafka (1978) and Barrows et al (1982). Initially the researchers thought that the clinical reasoning is based

on some expert problem solving skills which are related to the expertise of clinicians rather than the knowledge. A closer look revealed that clinical problem solving skills comprise of two processes: the initial hypothesis generation which may take few seconds to minutes, followed by diagnostic accuracy, which is a stepwise systematic search for confirmation of data, a longer process. It was found that a novice would use the same process to generate a hypothesis as compared to an expert, however, an expert will do it in a better way. Secondly, the diagnostic accuracy would depend upon the content of hypothesis. If they both got it right then the diagnosis would be accurate, but if they did not, then the outcome would be a wrong diagnosis. Hence, the domain specific knowledge is the key to generate accurate diagnosis and forms the basis of clinical reasoning (Perkins and Salomon 1989, Elstein, Shulman and Sprafka 1990).

Mental workload can hamper the speed of information processing and consequently affect the ability to judge and make decisions (Wickens and Carswell 2006). Mental workload depends on capacity and saturation of mental resources. Mental resources comprise of working memory and long term memory (Lord and Maher 1990). Working memory is limited, whereas, long term memory has infinite space. The Cognitive Load Theory (CLT) (Young et al 2014) describes how the learners process the various components of the task using their working memory and long term-memory. The intrinsic load contains all the information about components of task and its interconnectivity. The extraneous load may contain the unnecessary details which are not important for the completion of task, such as controlling the simulator, worrying about patient's safety and being preoccupied with a

busy clinic. Germane load is the mental activity of understanding and rating the task, and storing it in long-term memory, which can be retrieved later. Cognitive Load Theory can be utilised to moderate the exogenous load with the endogenous demands to optimise the outcome (Mayer 2010). The mental resources need to be aligned to the demands of the task. The information processing demands of the task may compel the assessor to run multiple neural circuits simultaneously according to the dimensions of the performance and complexity of the behaviors to be assessed. In an experiment when the length of stimulus to be assessed was increased from 6 to 18 minutes, it made a difference in the rater cognition of the task (Govearts et al 2007). For less complex tasks, the performance of a novice and an expert rater was the same. However, differences were noted in the performance of an expert rater as the complexity was increased. As the complexity increases, so does the mental load, but for an expert rater, the capacity to cater for such a load also increases with expertise, hence he/she has better judgement and decision making ability. Serial processing is a cognitive strategy whereby the assessor will try to process some aspects of the performance while neglecting other dimensions, or delaying others (Tavares and Eva 2013). Another mental strategy to reduce the cognitive load is 'Degraded Concurrent Processing' in which the rater judges the dimensions of the performance at a lower level of accuracy than if he were to assess them in isolation. Similarly, use of heuristics, narrow attention span, fatigue and performance anxiety may also lead to variations in the judgements across raters.

With the dawn of new millennium, a new array of possible thinking processes associated with clinical reasoning were defined. But the debate was just the same old: whether generalizable thinking skills of physicians have a role to play in diagnostic accuracy or if content specific knowledge is the key to better clinical reasoning. The terms such as 'Critical Thinking, Metacognition and Reflection' were introduced into medical education (Monteiro et al 2020). Critical thinking is about evaluating and appraising the information which is available in the light of best possible scientific evidence prior to understanding and believing it (Aveyard, Sharp and Woolliams 2015, p.7). It requires a set of discrete skills which are more generalisable across a number of contexts. The skills may include synthesizing knowledge, evaluation, analysis and drawing conclusions (Facione 1990). Whether critical thinking is a stable trait or does it depend on certain factors? An experiment was performed by Abrami et al (2008) to find out if any educational intervention could improve the scores or otherwise. He concluded that it is an independent variable, a stable skill that is related to academic grades, whereas another study suggested that it is dependent on years of education or some intervention that can alter the scores, hence it improved with education (Scott, Markert and Dunn 1998).

Metacognition is thinking about one's own thinking (Monteiro et al 2020). It includes the processes that monitor and assess one's own knowledge, awareness and regulation of thinking patterns and actions. In the context of medical education, the initial literature suggested that like most of the physician specific activities, metacognition comprises of more general self-regulated skills which help the physicians reflect upon themselves and

improve the practice of medicine (Flavell 1976). Reflection in the context of metacognition is awareness of one's own thinking that allows critical appraisal of actions and thoughts from recent or distant past to improve performance (Mann, Gordon and MacLeod 2009). The physicians can build on their clinical skills by improving their reasoning and self-monitoring. These strategies may improve the behavioral outcomes in terms of their attitudes, however, it has no effects on diagnostic accuracy (Ackerman and Thompson 2017).

To further the understanding of cognitive processes, in 2011 psychologist Kahneman put forward his nobel prize winning Dual-process theory of thinking and clinical reasoning (Kahneman 2011). He suggested that thinking can be sub-divided in fast and slow components. The fast component may be called System 1 whereas slow component as System 2. System 1 is more automatic, based on instincts and intuitions, and comes as a reflex without the conscious awareness. For example, recognising the different colors and identifying facial expressions. System 2 on the other hand involves more problem solving approach, delving into deep conscious analysis, comprises of secondary-process thinking, which is deliberate, depends on multiple cognitive resources and is more systematic. For instance, performing complex mathematical calculations. System 1 thinking may also include the use of heuristics to complete certain patterns to achieve coherent and complete understanding of situations without indulging into secondary process thinking. This may lead to cognitive biases and causes inter-rater variability. Hence, physicians need

a comprehensive understanding of this complex dual process theory and how it facilitates their clinical decision making and personal judgements.

Health care professionals are different in their values, beliefs and attitudes towards the delivery of patient care. It leaves room for individualized way of thinking and decision making in different situations instead of rigidly and blindly following the clinical guidelines. No two situations are entirely similar; hence, it calls for essential inquiry into checking out their assumptions and thinking logically based on the evidence available to help them make the informed decisions. Self-regulatory judgement requires an extensive training in to analytical mode of thinking (Croskerry 2013). Being more vigilant about one's own thinking patterns and deliberate efforts to decouple intuitive thinking from analytical thinking requires mindfulness of one's own cognitive predispositions and vulnerabilities. The software of the mind, which is usually referred to as 'Mindware' comprises of complex structures of rules, knowledge-rich networks called schemas and knowledge-based problem solving procedures and strategies that the person can use and retrieve from long term memory (Stanovich 2010). Decoupling will require de-biasing techniques. These techniques may include corrective mechanisms involving use of critical thinking, metacognitive approach and self-reflection. When the educators and assessors are mindful of their thinking processes, they can easily incorporate their personal judgements, values and cultural beliefs into more formal assessment decisions without creating noise and variability.

Method

A scoping review was performed on the literature available on measurement errors and biases in relation to entrustment decisions based on EPAs.

The following steps were included in data gathering, based on Arksey and O'Malley (2005) framework of conducting a scoping review:

I. Identify the research question: My research question was; 'What are the sources of measurement errors in competency-based assessment where EPAs are employed and decisions are made regarding entrustability?'

II. Identify relevant studies: An initial literature search was performed on the following databases: PubMed, Ovid and Google scholar. As expected, the search turned out only few articles referring to entrustment decisions. Key terms used were: 'Competency-based Medical Education' OR 'EPA-based assessments' OR 'Competency based assessments' OR 'Entrustment decisions' OR 'Entrustable professional activities' OR 'Assessing professional competency' OR 'Programmatic Assessment' AND 'Validity of decisions' OR 'Measurement errors' OR 'Inter-rater variability' OR 'Rater cognition' OR 'Rater judgement'

III. Study selection: Inclusion criteria: The papers with a focus on measurement errors and rater biases were included, in relation to competency, performance based assessments and entrustment decisions based on EPAs. Not too many papers focused on measurement errors on EPA based-assessments were found.

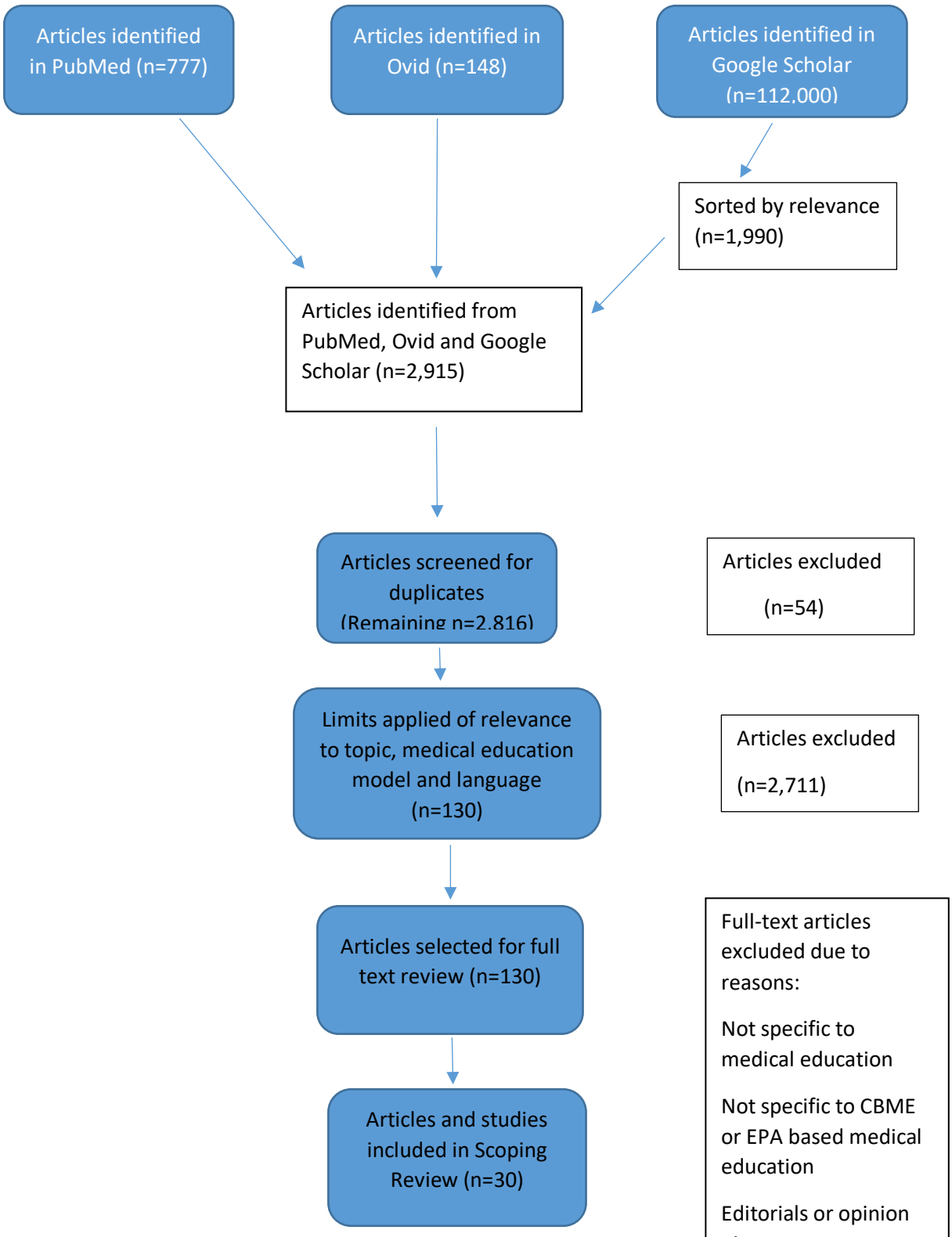
The papers in only English language and published on any date till June, 2021 were included.

Exclusion criteria: Papers in languages other than English. Articles addressing an assessment model other than performance, EPA or competency based were excluded.

IV. Next, I reviewed the abstracts to find out which papers were eligible for full-text review. A total of 130 relevant articles were identified using the search terms or the key words, followed by searching the electronic data bases and hand searching of articles from journals and conferences.

Results:

The initial search by the principal investigator resulted in a total number of 2,915 articles. After removing the duplicates and applying limits of language, followed by screening by relevance to topic and model of medical education, 130 studies were included in full text review. Finally, a set of 30 studies were selected for synthesis of data, content analysis and to be included in the scoping review.


Articles identified in PubMed (n=777)
Articles identified in Ovid (n=148)
Articles identified in Google Scholar (n=112,000)
Sorted by relevance (n=1,990)
Articles identified from PubMed, Ovid and Google Scholar (n=2,915)
Articles screened for duplicates (Remaining n=2,816)
Articles excluded (n=54)
Limits applied of relevance to topic, medical education model and language (n=130)
Articles excluded (n=2,711)
Articles selected for full text review (n=130)
Full-text articles excluded due to reasons:
Not specific to medical education
Not specific to CBME or EPA based medical education
Editorials or opinion pieces
Articles and studies included in Scoping Review (n=30)

Description of the themes generated from the included studies

The studies comprised of a mix of systematic reviews, mix method designs, faculty surveys, trainee evaluations and narratives of data base searches. Details of the articles are given in table 1.

Eight themes were generated: errors due to factors related to trainee, rater, trainee-supervisor relationship, assessment method, task, implementation of EPA based model, entrustment scales and contextual factors.

Three studies (Hauer et al 2014, ten Cate et al 2016, Caro Monroig et al 2021) identified a number of factors which may affect entrustment decisions in relation to supervisor, trainee, supervisor-trainee relationship, task and contextual factors. However, the specific measurement errors in relation to entrustment were highlighted sporadically in a few studies.

Table 1: Studies included in Scoping Review

	Type of Measurement error	**Author and year**	**Study type**	**Data collection method**	**Outcome**
1.	Errors due to Rating scales	Schott et al 2015	Prospective observational study	Checklist and milestones rating scales were used	The checklist format and milestone rating instrument both had very low inter-rater reliability. It could be because of rater error or instrument error.
2.	Inter-rater variability	McGill, van der	Evaluation and	End-of-rotation, workplace based	Increased inter-rater variability was observed due to poor sampling

		Vleuten and Clarke 2011	psychometrics	trainee assessment	technique, inconsistent assessment of competencies and inadequate number of raters.
3.	Rater errors	Dudek et al 2005	Faculty survey	Qualitative analysis of semi-structured interviews	Four major areas were identified: supervisors do not know what to document, they fail to document, fear of appeal process and no system in place for remediation.
4.	Rater errors	Charlin et al 2006	Evaluation and psychometrics	Member of study panel were asked to fill out a test based on Script Concordance Test	Items with variability have the power to discriminate details towards solutions to clinical problems. However, items with large margins would not indicate clinical experience, rather it shows disagreement among raters.
5.	Rater errors	St-Onge et al 2016	Survey and psychometrics	semi-structured interviews	Raters make judgements by balancing two sources of information: external and internal involving specific cognitive processes.
6.	Rater errors	Thomas et al 2011	Longitudinal prospective study	Faculty-on-resident and group faculty-on-resident scores were compared	Overall mean scores were higher for group versus individual assessments. Halo effect was greatly reduced for resident assessments across a number of performances by adding group assessment to individual faculty assessment.
7.	Rater judgement	Cox 2000	Narrative review	Narrative	Clinical performance can only be assessed in the context of daily clinical work where rater has open-ended time to assess the trainees for procedural skill, case management, personal management and knowledge, rather than having a reductionist attempt at measurement which leads to errors in judgement.
8.	Rater errors	Kreiter et al 2016	Prospective longitudinal study	Students' end-of-clinical shift evaluations by the same rater	If the same rater assesses the student repeatedly, it is less likely that the ratings will give a reliable score. It can contribute to construct irrelevant

					variance, which can be attributed to extraneous personality factors due to the rater.
9.	Errors due to assessment method	Tavares et al 2020	Faculty survey	Qualitative in-depth semi-structured interviews, content analysis	The data on which formative and summative decisions are made in CBME is incomplete. In terms of validity of the decisions, assessment of competencies based on an outcome framework may not translate into all the essential parameters of assessment framework, due to quality of tools being used, untrained faculty and contextual social and practical factors.
10.	Errors due to Assessment method	Brown et al 2021	5-year pilot study, evaluation for entrustment on the basis of workplace based assessments data from multiple schools	Trained entrustment groups examined learners' readiness for unsupervised practice and availability of data for entrustment decisions	For majority of EPA specific data entrustment decisions were made, but for 28.4 % of data no specific decision could be made because of insufficient or limited data quality and quantity.
11.	Errors due to lack of shared mental model for faculty development	Favreau et al 2017	Critical synthesis of EPA literature to discuss a shared mental model for faculty developmen t for entrustment decisions	Literature review	For core EPA faculty development, four-dimensional shared mental model was proposed which included adequate observational and workplace based assessment skills for teaching and assessment of core EPAs, feedback and coaching skills, self-reflective skills as they act as role models to help the trainees inculcate trustworthiness, and lastly, to establish communities of practice based on core EPAs.

12.	Errors due to ineffective implementation of EPA-based framework	Holomboe et al 2017	Narrative review	Narrative	Criticism of CBME includes reductionist approach, lack of evidence base, lacks universal definition of its components, workplace based assessments are not up to the mark, faculty is untrained, institutional resources are lacking, lacks psychometric evidence and current training systems cannot ready to change to time-variable model.
13.	Errors due to how entrustment decisions are perceived and enacted across various specialties	Melvin et al 2020	Faculty survey	Semi-structured interviews, content analysis	Several themes were identified which could be the reasons for tensions in entrustment decisions in Internal Medicine: faculty seemed to be confused about the language and concept of trust, entrustment and competency, entrustment decisions are not made on the basis of attendings but rather decided on the basis of training programme and the level of training. Thirdly, entrustment is not a single point-in-time discrete assessment due to longitudinality of tasks and previous relationship with the supervisor. The language of entrustment scale does not coincide with the decision of the raters, and lastly, the entrustment decisions affect more the attendings and faculty rather than the trainees.
14.	Errors due to lack of validity of entrustment decisions	Touchie et al 2021	Narrative review	Narrative	Validity evidence should be gathered as an on-going process based on Kane and Messick frameworks to ensure defensibility of summative entrustment decisions.
15.	Errors due to lack of progression of performance at the end of training in	Schumacher et al 2020	Prospective longitudinal assessment of trainees over time using EPA-	17 general Pediatric EPAs were assessed using entrustable	90% of the trainees managed to achieve the level of unsupervised practice for only 8 out of 17 EPAs at the end of training period. This study indicates gaps in the observed readiness of the trainees for unsupervised practice and the actual

	EPA-based summative decisions		based assessment data	supervision scales	standards set for meeting the optimum needs of the health care system and patient safety.
16.	Errors due to context complexity and use of entrustability scales	Rekman et al 2016	Narrative review	Narrative	Entrustment-aligned tools lack this quality to completely cater for all aspects of observation of the trainee in complex performance covered by a specific EPA.
17.	Errors due to lack of validity evidence in EPAs development	Taylor et al 2021	Narrative review	Narrative	Construct validity evidence for EPA development will ensure accurate prediction of graduates' assessment for readiness for practice without supervision. It comprises of five steps: selection of experts, identification of EPAs, revisions of EPAs, evaluation process of the selected EPAs and adoption of EPAs into the curriculum. Avoiding biases in expert selection will ensure accurate reflection of clinical practice and that the right people are involved in entrustment decisions and identifying EPAs. It will avoid pitfalls in deployment of EPAs in the assessment of training programmes.
18.	Rater cognitive biases	Marchegiani, Reggiani and Rizzolli 2013	Laboratory experiment	Study behavioral implications of a model in a laboratory experiment	If a rater has to make an error, it is better if he/she tends to be lenient than stringent. If assessors fail to reward deserving students, it is more detrimental than if they reward undeserving students.
19.	Errors due to ineffective assessment of 'trust' in entrustment decisions	Ten Cate 2021	Narrative review	Literature search	The concept of entrustment holds in it an inherent responsibility of the educators to ensure that the learner is ready to practice without supervision along with the readiness of supervisor to accept the risk involved which requires observation of more than 'doing'. Assessment methods need to include

					the future prediction of performance and risk assessment in health care tasks and patient safety.
20.	Rater Cognitive biases	Iramaneerat and Yudkowsky 2007	Students evaluation	Students skills assessment with standardized patients	SPs showed differences in their ratings, exhibiting rater leniency. Similarly, four SPs showed inconsistency, one showed halo effect and four had restriction the range in their ratings.
21.	Rater cognitive bias	Kogan et al 2010	Faculty evaluation and psychometrics	Faculty completed a demographic questionnaire, then performed 8 standardized patient encounters and rated standardized residents' videos using mini-CEX. Faculty interaction with SP and process performance was compared with their mini-CEX ratings of residents.	Faculty higher performance scores in history taking were associated with rater stringency in interviewing and organizational skills. Higher faculty process scores were associated with stringency in interviewing, physical examination and organizational skills.
22.	Errors due to entrustability scales	Hatala et al 2019	Narrative	Review of literature	In Internal Medicine, certain non-procedural tasks are not measureable and not observable, and EPAs are too broad to capture the activities of daily encounters such as admitting the patients. Moreover, rating scales have limitations, being more quantitative and psychometric oriented. In Medicine, supervisors do not speak language of trust and entrustment. Rather they are more comfortable describing their

					feelings and are more descriptive. Hence, entrustability ratings do not provide committees with robust data to make defensible decisions.
23.	Errors due to supervisor-trainee relationship	Caro Monroig et al 2021, Huaer et al 2014, ten Cate et al 2016 and 2021.	Narrative reviews and	Narrative reviews and	A longitudinal relationship with supervisor may actually help the trainees in learning but it may cause interference and biases in assessments and may jeopardize the role of a coach as an assessor by causing role conflict.
24.	Errors due to use of comments in assessments and feedback on entrustment decisions	Ginsburg et al 2021	Literature review	Review	In programmatic assessments, words serve a double purpose: to provide feedback and to be used in summative decisions. Supervisors use the same words regardless of the purpose. This may result in a low quality feedback to the trainees and difficulty aligning numbers to the written comments in the entrustment decisions.
25.	Errors due to poor description of the anchors of the entrustment scale and rater variability	Krupat et al 2017	Narrative	Review of literature	Core EPAs comprises of hospital based activities. Instead of focusing on professionalism and medical knowledge, the undergraduate students are compelled to learn discrete tasks, which sound more like educational objectives than goals. Moreover, the assessment of the core EPAs is all about subjective process of the judgement of the assessors. There are many layers of inference between observation and assessment which may cause variability in scores. The layers may include observation of performance, a rating system, well described anchors, context specificity, entrustment judgement based on trainee characteristics such as trustworthiness in addition to skill and

					level of supervision according to the agreement of raters.
26.	Diagnostic and management errors due to Cognitive biases and personality traits of raters	Saposnic et al 2016	Systematic review	Review of literature	Cognitive biases and personality traits affect the clinical reasoning of raters and may lead to diagnostic errors and have implications on patients' outcomes.
27.	Errors due to contextual factors	Lucey et al 2020	Review of literature	Review of literature	Contextual influences due to diversity in culture, medical training programmes and systems may put learners from minority groups at a disadvantage. It induces inequity in assessments due intrinsic design of assessment, tool being used and context in which assessment is taking place. It leads to structural and interpersonal biases.
28.	Rater biases	Schwartz et al 2020	Trainee evaluation and psychometrics	Two cohorts of trainees from pediatric residency programme were assessed by actual CCC and virtual CCC	The trainees received higher ratings by their own CCC as compared to virtual CCC. It could be because their greater comfort with entrustment decisions and contextual knowledge of their own system. And partially due to knowledge of their own trainees and personal relationship with them. Moreover, trainees who graduated from schools following traditional model of medical education were given similar ratings to those graduating from school following time-variable competency based model.

Discussion:

Assessment of performance is a critical and mandatory part of health professional education. It comes with challenges and opportunities for improving patient safety and health care quality (Eva 2018). One of the challenges of assessment of performance

derives from the fact that there are multiple domains of competencies which are abstract enough to make them difficult to measure and hence, subject to significant errors. Various studies have put forward a number of explanations for these errors, however, the fact remains that it is difficult to objectify all the components of the assessments into measurable numerical scores. It is therefore imperative to come up with a framework for assessment of competencies which is relatively simpler, conforms to the universal standards of assessment, easy to adopt by the educators and helps them to identify the measurement errors and minimize them

The central tenet of CBME is holism and integration as in CBME the whole is considered bigger than the sum of its parts (Schuwirth and Ash, 2013). If this new competency-based medical education system has to survive, the educators need to adopt an integrated approach to measure 'competence' which is not based on the concept of testing of discrete granular parts, rather it needs to be done as a whole. In this study, I have reviewed the common sources of measurement errors in EPA based assessments.

It is important to talk about the factors leading to measurement errors but before I do that I would like to establish some grounds for the reliability and validity arguments for entrustment decisions.

Validity and its relation to Measurement errors:

All the factors which interfere with the accurate measurement of scores will reduce the validity of an assessment. Validity is the extent to which a test or an assessment

accurately measures what it is supposed to measure. In medical education, validity refers to the extent of meaningful interpretation of the test scores (Downing and Haladyna 2004). Messick (1989) identified two sources of measurement errors: Contruct-under representation (CU) and Construct-irrelevant variance (CIV). Construct under representation means inadequate sampling of the content by the instrument being utilised. Construct irrelevant variance refers to the errors due to other variables or the systematic errors resulting into errors in the assessment not due the construct being measured (Downing and Haladyna 2004). In CBME one of the major reasons for CU is too few cases representative of the domain to generalise the scores which can threaten the validity of the assessment. Construct irrelevant variance in performance assessments can be introduced because of the rating scales, rating methods and rater biases. If the ratings are unstandardised or judged by untrained raters, it may result in CIV.

Content, development and validity of EPAs:

A framework of delivery of medical education based on EPAs may not comprise of a single approach towards its development. In a systematic review of eight years, no single method for the content and development of EPA was reported (O'Dowd et al 2019). These methods may include Delphi groups, focus groups and literature review to develop a census among the subject experts. To defend validity arguments for the development of EPA framework, a five-steps method has been described by Taylor et al (2020) based on Messick framework. These steps include: selection of experts for the

process, identification of EPAs, iterative revisions of EPAs, evaluation of the consensus EPAs and its adoption and implementation.

Careful selection of experts in a specific specialty may ensure content validity and accurate judgements in the evaluation process. The selection may be made on the basis of their expertise and representativeness amongst the group of individuals with relevant experience and skills. While working towards the construction of EPAs, it is important that these experts should have a clear understanding of the goals and method of development. EPAs can be identified by keeping a log of prospective clinical work, prior practices and peer-reviewed literature (Taylor et al 2018). This approach can subject to availability bias. If EPAs are based on clinical practices from some other jurisdiction, it may result in missing out on very important details and differences in organization of training and residency which can lead to errors in assessment. Moreover, certain other biases may appear as a result of panel discussions, such as group think and discussion biases. Here the discussion process can be led and influenced by the most senior or charismatic member. Consequently, such a group may fail to cater for the contradictory opinions of other members and lead to conflicting results (Wittenbaum 1998). Similarly, programme directors and supervisors naturally tend to gravitate towards their peers with whom they have some prior work experience. It can create homogeneity in the group which can lead to systematic errors (Downing 2009). Hence, selecting the experts is the most important part in the development of EPAs.

The next steps towards the validity are as follows: Identification of candidate EPAs for high stake assessment need to be based on EPAs which represent important professional tasks. It can be done through a face-to-face consensus or based on Delphi approach. It is an iterative process which generate robust evidence in favor of more valid EPAs. Revisions may include multiple rounds of getting feedback from the experts and interpreting the tasks and titles of EPAs. Wide variability in the experts' opinions may reduce the interrater reliability. By introducing a number of checks such as through triangulation (review by a third party), the authenticity of the arguments can be increased further. Evaluation of the EPAs may include looking for evidence to ensure alignment with constructs, in-line with standard practices and adequate utilization in workplace-based assessments. Statistical methods such as coefficient correlation and factor analysis of the survey data can be used to evaluate the agreement between the raters (Taylor et al 2018, 2021). The content can be evaluated by exact mapping of the EPA to the construct or the professional domain to be assessed. Finally, the formal implementation and adoption of EPAs by an institution set the standards for the summative entrustment decisions. Consequently, it may affect the patient care and protects community interests at large. Hence, the validity and reliability data of entrustment decisions may include health care and patient-safety outcomes, which will eventually bring all the stakeholders in the picture such as learners, policy makers, health care providers, patients and society (Taylor et al 2021).

Validity of Entrustment decisions based on Kane's Validity Framework:

Validity is the defensibility of the scores and data produced as a result of an assessment (Messick 1989). Validity is the evidence gathered as a result of the process of assessment to support its purpose, it is not a feature of an instrument being used in an assessment (Messick 1989). Validation is the investigative process to support the interpretation of test scores through theoretical basis or plausible evidence to prove the desired scores (Downing and Haladyna 2004). The initial and a fundamental step in the validation of an assessment is to define the construct. Any achievement or performance test is based on assessment of knowledge, cognitive abilities and attitude. MCQs format of assessment method can be utilized to assess types of knowledge. The other form of construct comprises of cognitive abilities (Downing and Haladyna 2004). Each cognitive ability further constitutes of contextualised schemas, mental models and complex performance component. The most appropriate method to assess a complex ability is performance based assessment. Validity of a performance based assessment may depend on the test content, response processes, internal structure, evidence based on relationship with other variables and consequences of assessment (Messick 1989).

Validity of summative entrustment decisions need to be based on robust evidence as it directly involves patient care and safety (Touchie et al 2021). Eight steps were proposed to organize the evidence for entrustment decisions (Cook and Hatala 2016). First of all, the construct has to be defined which is most often a competency, in the context of EPA based assessment. At step two, the purpose of the ad-hoc or summative decision

must be made very clear, as regard to the level of training and the supervision. In steps three through six, the evidence is collected from the observation to the decision making. The evidence in relation to what the teachers have observed, on how many occasions, observed by how many teachers, is the cumulative data sufficient for the summative decision, is the data based on encounter with different types of patients in different settings, what are the sources of information, is the data enough to proceed with an entrustment decision, can the entrustment decision can be generalized to new situations and does the decision relate to any other sources of information such as credentialing or licensure examinations and remediation decisions. Finally, how did the clinical committee make the decision, is it based on clear guidelines, can the results predict the future expectations from the trainee in terms of unsupervised safe clinical practice, and lastly, could there be any unintended implications of this decision? An entrustment decision based on the above mentioned information would be considered valid by the committee members.

Similarly, validity of an entrustment decision can also be proved using Kane validity framework (Kane 1992). Kane emphasizes the chain of evidence from the point of observation to final interpretation of the scores and entrustment decision making. Kane framework is based on four inferences: scoring, generalization, extrapolation and implication. Scoring for summative entrustment decisions starts when the clinical supervisors and committee members review the data to determine the scores of entrustment decision. The inference may include data gathering as to how these

decisions are made. How the observations are converted to scores, the construction and interpretation of items, management of group processes, administration and execution of simulation and rater training. For scoring inference, entrustment provides construct alignment of the cognitive judgement processes of the members of the clinical committee to the tasks being judged (Touchie et al 2021). The data may hold different weightage for each of the member of the committee. Different cognitive views of the data may lead to divergence in the decision making. However, a coherent constructive discussion may still keep the unanimity of the group intact. Inter-rater variability does not necessarily mean reduced validity (ten Cate and Regher 2019). A shared understanding or mental model of the entrustment decisions based on the subjective impression of the assessors combined with adequate training of raters will result into mitigation of cognitive errors (Touchie et al 2021).

Generalisation in the context of EPA based assessments would mean how exactly the scores assigned to the trainees represent the all the possible observed performances in a number of situations (ten Cate 2020). The data collection process, adequate sampling of clinical cases, differentiating evidence to discriminate different levels of supervision and reliability of the CCC to be able to reach the same decision if the whole process is repeated, are the important areas with respect to generalization of the results. Data collection from direct sources of observed performances such as mini-CEX, CBDs, MSF, simulation-based assessments and end-of-rotation progress assessments are important sampling strategies (Duijn et al 2019). The psychometric properties of

complex subjective entrustment ratings are difficult to elicit (Touchie et al 2021). To date there are no studies to provide any evidence on blueprinting of entrustment decisions and psychometric evaluation of programmatic assessments.

Extrapolation of entrustment decisions would not only mean that the current summative decisions can defend the current proficiency of the trainees but it should predict all the possible unseen scenarios of the performance, based on EPAs in future (Touchie et al 2021). It essentially takes into consideration the quality care measures as a part of evaluation of the processes and outcomes of a training programme (Schumacher et al 2020). That would mean extrapolation of trust in performance of trainees beyond graduation.

Implication of entrustment decisions is about providing the evidence for different entrustment levels, frames of reference and predefined shared mental models, theories that form the basis of informed decisions, how thresholds are set for remediation and pass/fail decisions (entrusted/ not entrusted) and the consequences of decisions (Touchie et al 2021). Evidence for consequences of the decisions may include the effects on learning, the patient care quality and safety measures and professional identity formation of the trainees and their well-being. The consequences of entrustment may also be evident in the form of variation in time of duration of training and a need for a more flexible training programme to accommodate such changes in the form of extension or remediation.

Table 2: Sources Of Measurement Errors							
Errors due to Assessment method	**Errors due to rater**	**Errors due to trainee**	**Errors due to Supervisor-trainee relationship**	**Errors due to failure to implement EPA based model of education**	**Errors due to Task**	**Errors due to Contextual Factors**	**Errors due to Entrustment scale**
Inadequate sampling	Experience as an assessor, trainer and supervisor	Belonging to a minority, marginalised or different ethnic group	Length and intensity of relationship [a]	Incomplete translation of outcomes or competencies into effective assessment strategies	Level of complexity of competency	Time pressures	Reliability of scoring scale
Poor alignment of construct to rating scale	Untrained faculty	Burnout or excessive workload	Interference with assessment [b]	Assessment inequity	Patient risk involvement [a]	Environment and support of the staff where assessment is being carried out	Poor selection of rating instrument
Incomplete data and information of trainee performance scores	Cognitive biases	Level of supervision	Role conflict: assessor versus coach	Lack of shared mental model for the faculty development	Level of urgency [a]	Political, legal or organisational constraints [a]	
No shared mental model for processin	Poor attention and concentration levels	trustworthiness	Previous knowledge of the learner	A reductionist approach		Resources and workplace culture [a]	

g of information by CCC in entrustment decisions							
Dominance of a leader in decision making process	Performance anxiety and compromised psychological health		Quality and quantity of feedback [b]	lack of evidence base		Situational hectic circumstances [a]	
Too few faculty members for observation of assessment	Sense of responsibility, autonomy towards the patients, trainees and institutes [b]		Defining and communicating shared expectations from the task [b]	lacks universal definition of its components			
	Personality traits			workplace based assessments are not up to the mark			
	Effect of emotions			faculty is untrained			
				institutional resources are lacking			
				lacks psychometric evidence			

				current training systems are not ready to change to time-variable model			

a Terminologies borrowed from ten Cate et al 2016.

b Terminologies borrowed from Caro Monroig et al 2021

SOURCES OF MEASUREMENT ERRORS

Errors due to Assessment methods:

The total number of studies included in the scoping review in relation to assessment methods were eight. Generally, the characteristic features of workplace assessments have been discussed in relation to assessment of competencies and observation of performance. But any specific article highlighting the dos and don'ts of utilization of assessment methods with reference to entrustment decision was not found. However, several studies compared the validity of workplace based performance in relation to CBME (Dijksterhuis et al 2009, Duijn et al 2019, Daniel et al 2019, Durning et al 2019 and ten Cate et al 2021). It is beyond the scope of this study to evaluate the feasibility, validity and reliability of each of performance based assessment methods utilised in entrustment decisions. However, I have tried to give an overview of few of the characteristic features of WPBA and to analyse them in the context of EPA based assessments, and the measurement errors caused by them.

A complete assessment of a training programme may include assessment of the knowledge, skills and attitudes of the training physicians or the graduating students (Dijkstra et al 2010). The purpose of assessment as described by Kane is to assess the ability of the learner 'to use the appropriate knowledge, skills, and judgment to provide effective professional services over the domain of encounters defining the area of practice' (Kane 1992, p.167). A number of assessment methods have been developed such as MCQs, OSCEs and workplace based assessments (Norcini and Burch 2007). The goal of assessment may be to allow entry to a student to a medical school or to determine proficiency of a trainee to promote him/her to the next level of training (McKinley and Norcini 2014). The assessment goal determines whether to compare the scores of the students to other students who are also taking the same test or to evaluate the learners against a predefined passing score. In other words, the standards can be set as a relative parameter or absolute, respectively. The former is termed as norm-referenced and latter as criterion-referenced (Livingstone and Zieky 1982). In high stakes assessment, absolute standard setting approaches are more appropriate as the proficiency level and the difficulty of the exams may act as confounding factors for the appropriateness of the passing scores (Mckinley and Norcini 2014). In CBME, the passing scores are determined and set by a group of panelists which may comprise of faculty members, programme directors and clinical supervisors for the assessment of attitudes, behaviors and knowledge.

The top of Miller's pyramid 'Does' may require the educators to use methods involving multiple data points, such as 360-degree evaluation and multisource feedback (Norcini and Burch 2007). Assessment methods need to include the future prediction of performance and risk assessment in health care and patient safety tasks. Tools which are utilized in the assessment of 'Does' may be used to capture all the dimensions of trust, such as DOPS, MSF and mini-CEX. However, tools which can adequate cover all facets of entrustment decisions are limited (Wijnen-Meijer et al 2013). At the highest point of Miller pyramid, a critical yet well informed decision based on evidence and cumulative scores may guide the committee in to entrusting/not entrusting a trainee.

In a systematic review by Duijn et al (2019), around 80 studies described 67 assessment tools. The most commonly utilised workplace assessment tools were MiniCEX, Direct Observation of Professional Skills (DOPS) and OSCEs. Case Based Discussions (CBDs) and Multi-Source Feedback are worthwhile addition to the list, however, not many studies have explored the tools with reference to entrustment decisions. The validity evidence of these tools has been established but further studies are needed in future to explore factors which may lead to successful implementation of these tools in the context of summative entrustment decisions. Nevertheless, many of the tools can be used for providing feedback to the students and to establish the level of supervision and autonomy (Duijn et al 2019).

Multiple clinical direct observations usually in the form of WPAs are used in formative assessments whereas global assessments and data from multisource feedback may

count towards summative assessments (Daniel et al 2019). Written notes in the form of a post-encounter report may be another structured way of communicating a detailed explanation of clinical reasoning and may influence rater judgement (Durning et al 2012). WPAs are known as great tools for assessment of clinical reasoning, such as hypothesis generation, information gathering, problem representation, generation list of differential diagnosis and discussing management plan (Daniel et al 2019). In relation to adequate content coverage, WPAs may cause problems in over or under-representation of appropriate cases. Moreover, certain psychometric properties based on the internal structure item analysis, standard errors of measurement and reliability of the scoring scale may also pose some problems as these characteristics are based on the observation of the assessor. Hence, biases and cognitive errors are an inherent part of summative entrustment decisions (Kogan et al 2012, Daniel et al 2019, Gingerich et al 2014). This threat to validity can be minimised by utilizing a number of WPAs with adequate sampling of relevant clinical cases involving multiple raters over time. For instance, at least 12 to 14 mini-CEX stations are needed for adequate sampling to bring reliability up to acceptable levels (Daniel et al 2019). Implementation of WPAs is resource intensive. WPAs require vigorous faculty development programmes, time and identifying those trainers who would be interested in conducting performance based assessments (Daniel et al 2019). Competency based assessments cannot rely heavily on non-WPAs as these partly assess clinical reasoning. However, Multiple Choice Questions (MCQs) and Key-feature Examinations (KFEs) are used in progress tests,

credentialing decisions and licensing exams (Daniel et al 2019). MCQs comprise of a main stem followed by multiple options or alternatives to choose from and requires the students to select one best answer. KFEs are based on clinical vignettes and may require the students to answer to two or three questions in relation to clinical decision making. A combination of MCQs and KFEs with global judgements, direct observations and Objective Structured Clinical Examinations (OSCEs) may render an outcome which will possibly cover multiple aspects of attitude, skills and knowledge in relation to programmatic competency-based assessment (Daniel et al 2019).

Errors due to ineffective implementation of EPA based model

Inadequate translation of EPA framework, which is based on tenets of competencies or outcomes, into strategies to assess these components may include a number of in-articulated factors (Tavares et al 2020). These factors may include quality of the tool, unprepared faculty, poor alignment of constructs to the scales and social, practical and political contexts which may render the process of translation of outcomes into entrustment decisions ineffective and incomplete.

Moreover, Lucey et al (2020) talks about inequity in assessments. Equity in assessment means that all the learners are provided with equal learning, coaching and grading opportunities irrespective of race, geographical background or ethnicity. Neither learning nor assessment should be influenced by personal or structural biases of the assessor or the learner in a social and interpersonal context (Banks and Banks 1995).

Intrinsic equity means the assessment design and tools equally favor the students who are in minority or marginalized. Contextual equity in assessment refers to the environment, social and structural factors making it a conducive experience for both the learners and the assessors. Instrumental equity is in relation to use of the assessment data to provide equal opportunities to the learner subsequently. In entrustment decisions, equity would refer to the sufficient training of the assessor to give similar ratings to different levels of performances due to cultural and system differences amongst trainees.

Favreau et al (2017) proposed a shared mental model for the faculty development for entrustment decisions. It included adequate observational and workplace based assessment skills for teaching and assessment of core EPAs, feedback and coaching skills, self-reflective skills as they act as role models to help the trainees inculcate trustworthiness, and lastly, to establish communities of practice based on core EPAs. A lack of similar model may hamper the process of effective generation of trainee scores for summative entrustment decisions. Similarly, Holomboe et al (2017) highlighted the deficiencies as follows: a reductionist approach, lack of evidence base, lacks universal definition of its components, workplace based assessments are not up to the mark, faculty is untrained, institutional resources are lacking, lacks psychometric evidence and current training systems cannot ready to change to time-variable model. In relation to EPA based practices across various specialties, Melvin et al (2020) identified several reasons, for dissimilarity in how entrustment decisions are perceived and enacted

across various specialties. He deduced his findings on the basis of content analysis of a faculty survey in Internal Medicine. The reasons for tensions in entrustment decisions were: faculty seemed to be confused about the language and concept of trust and competency, entrustment decisions are not made on the basis of attendings but rather decided on the basis of training programme and the level of training. Thirdly, entrustment is not a single point-in-time discrete assessment due to longitudinality of tasks and previous relationship with the supervisor also influenced the decision making process. The language of entrustment scale also did not coincide with the decision of the raters, and lastly, the entrustment decisions seemed to affect more the attendings and faculty rather than the trainees.

Role of Direct Observation use of Entrustment scales in Entrustment decisions:

Direct observation is assessed by a faculty member in a clinical setting using an assessment tool such as MiniCEX. Personal judgement or clinical reasoning is usually included as a part of direct observation in addition to a scoring checklist (Daniel et al 2019). A learner's performance is evaluated in a real clinical setting based on an interaction with a real patient. Mostly in training programmes, the busy faculty rely on retrospective entrustment ratings for summative decisions that they had not observed (Sibbald et al 2021). Such an approach involves decision making based on the ability of the supervisor to recall and summarize a number of multiple informal contacts with the resident. These summative decisions may have a component of rater biases such as primacy, recall bias and recency effects. In one study, real-time observations were

compared to retrospective ratings (Sibbald et al 2021). It was found that one in three of the retrospective entrustment ratings was reclassified based on observation of the same learner in real-time standardised settings. Around 40 % of the raters changed their ratings based on direct observation. It provides validity evidence for entrustment decisions based on real-time ratings, even for cognitive tasks. It may have implications on the quality of training and assigning independent practice to trainees without supervision based on integrating prior ratings in high-stakes entrustment decisions.

Time for assessment:

The time allocated by the faculty and clinical supervisors to the assessment may greatly determine the quality of assessment. Involvement of faculty in planning and development of the material and implementation may affect the overall feedback and the delivery of educational programme (Williams et al 2014, Hauer et al 2016).

Measurement errors due to raters:

A rater's judgement may involve a number of cognitive processes that come into play as he/she uses these knowledge rich networks to draw upon the observation of the performance. It has been described as a three-phase process by Gauthier, St-Onge and Tavares (2016): The observation phase involves generation of the impressions at the sub-conscious level through automatic as well as controlled mechanisms. The rater may rely on active selection of the information in relation to relevant behaviors to be observed and judged. At this point, attention and adequate concentration levels are

mandatory, as it can impact attending to important aspects of the performance to be observed. Sometime raters form high-level inference instead of observing the behaviors. Moreover, extreme tiredness, burnout and intense emotional state at the end of the rater may also have deleterious effects on the alertness of the rater. Processing is the second phase which requires active management of the information to categorise it further on the basis of previously formed schemas, to make some sense out of it. Working as well as long term memory play active roles in these complex processes. The current processing takes place in working-memory compartment, whereas, the previously learned information retrieved from long-term memory is incorporated in it. The situational and contextual factors influence these mental processes which can consequently affect the implicit synthesis of information in the next phase. In processing phase, the rater analyses the components of the competency in the performance and the level of expertise exhibited by the trainee. The rater's use of the exemplars, their own conception of competency and ability to form their judgements according to the task may actually determine the overall quality of assessment. In the integration phase, the raters synthesize information in to final global judgements, or may convert it into scores or narratives.

Moreover, the variability in rater responses could be a hallmark of the clinical experience which enhances the ability to deal with uncertain situation and to defend the personal judgements on the grounds of clinical reasoning (Charlin et al 2006). In the same study, noise and variability in rater judgements was assessed. It was concluded

that in case of ill-defined clinical problems, clinical experience is associated with better clinical reasoning. Secondly, clinical reasoning in case of ill-defined problems is a source of variability which is due to a different cognitive pathway that a rater may take while reaching a decision. An assessment tool which is designed to evaluate the variability in judgements can tell about the acceptable levels of disagreement without creating noise in assessments. Lastly, with greater variability in test items or rater judgement shows better discrimination between extreme groups of students. However, extreme variability is considered undesired and is a measurement error which results in noise in assessments. Hence, if the role of the examiners is to differentiate good students from weak students, then certain tools which induce variability in the rater judgement will prove to be effective strategies to add to the clinical experience of the raters to help them discriminate the performances and make sound judgements in ill-defined situations.

The major part in rater errors is due to cognitive biases such as rater inconsistency, leniency, severity and halo effect. Rater inconsistency occurs when he/she does not use the rating scale as other raters which may result into high rater variability or inconsistency in results (Iramaneerat and Yudkowsky 2007). Hence, it may not help them differentiate the good trainee from a weak trainee. Leniency bias is when a rater gives a high score to a performance when he/she does not deserve it and severity bias occurs when a rater gives a low score to a student who deserves a better score (Marchegiani L, Reggiani T, Rizzolli 2013). These biases are also termed as dove or hawk

effects, respectively. Sometimes, faculty members use a different frame of reference than what they are supposed to follow, for instance, they may rely on their own practice style as a reference instead of a standard (kogan et al 2010). Similarly, if a rater tends to identify with a student or knows the learner, he/she is likely to assess him more positively which is called 'halo' effect (Iramaneerat and Yudkowsky 2007).

The decision of trusting a trainee may have dimensions other than knowledge and skills. In a study based on content analysis of the process of entrustment judgements of trainees at the end of first year training by programme directors, five underlying themes were identified which were based on interpersonal traits and personal characteristics rather than the performance skills and knowledge (Yoon et al 2020). The trust was considered as a stand-alone entity, not a part of core competencies defined by ACGME, and competencies such as professionalism, leadership qualities, curiosity, empathy, collaboration, team work and personal traits were considered important in trust decisions. Moreover, student engagement in medical honor societies, such as Alpha Omega Alpha in medical college received 5 % higher positive trust ratings. This could be because student involvement in similar trusts may signify the presence of similar positive personality attributes which make a good physician. Unlike involvement in a trust, referral of a trainee to a student promotion committees during the training period as a result of remediation decisions, did not result in an increase in negative trust ratings. In fact, it had no effects. Hence, the decision of trust is complex and it may

slowly and gradually evolve from undergraduate to graduate years as the trainees develop cognitively, personally and intellectually.

I think variability is just a dimension in rater thinking process. In entrustment decision, the beauty of how the rater judgement unfolds speaks of the credibility of a trainee or otherwise. To give due weightage to the implicit cognitive processes, we need a tangible framework to include rater judgement in the form of viable formal decisions which are defensible and may be subjected to some sort of psychometric analysis, for which further studies are needed to ensure its validity.

Measurement errors due to trainee

Trainee's abilities, confidence, awareness of his limitations and his efforts for participation may help him to earn the 'trust' (Hauer et al 2014). The trainee qualities such as conscientiousness, humility, empathy, agency are considered an essential part of entrustment (Chen and ten Cate et al 2020). Moreover, insight into the personal strengths and weaknesses, self-reflection, goal-directedness and focus on learning are positive attributes which count towards future success and safe patient care (Hauer et al 2014). On the other hand, long duty hours, burnout and work over load are considered as negative attributes which may affect performance of trainees in assessments.

Contextual factors

Contextual factors include workplace culture, environment, staffing, resources, work-shifts duty hours, workload, urgency to make the decision, availability of other trainees and local health care system (Hauer et al 2014). Supportive environment and opportunities for the trainees to practice autonomy may greatly increase the time spent in accomplishing challenging tasks. Consequently, it would enhance his/her ability to apply knowledge and generalize the practical skills to a number of settings, which is an essential part of the trustworthiness.

Factors related to task

The task complexity, the risk involved and sequencing can affect entrustment by the supervisor (Hauer et al 2014). A simple task assigned to a trainee and executed correctly may readily lead to an ad-hoc entrustment decision. However, for a complex task, multiple observations and multiple raters may be required to ensure validity of the decision. As the trainee progresses through multiple tasks, supervisor may want to increase the degree of complexity to test trainee's interpersonal, communication, case management and psychomotor skills (Sterkenburg et al 2006). As trainee goes up in his/her experience of performing complex tasks, supervisor can allow him/her to proceed further and grant higher level of trust and independence with lesser degree of supervision. Sequencing is graded exposure of the learner according to learning needs and corresponding progressive advancement in clinical responsibilities towards the summative decisions. This information needs to be integrated in an informed way along with narratives by the supervisor and perspectives of other team members and patients

to form a holistic picture of trainee's proficiency (ten Cate et al 2021). Any component found missing or EPAs implemented without the basic principle of progressive entrustment may lead to unintended consequences and measurement errors (ten Cate et al 2021).

Measurement errors due to Supervisor-trainee relationship

Depending upon the duration of the relationship longitudinally, the supervisor who seem to know more about their trainees are better equipped to make the right entrustment decisions (Huaer et al 2014). On the other hand, shorter duration of relationship may actually count as a barrier. Reciprocity from the trainee may actually help him/her in self-directed learning (ten Cate et al 2021). At the same time, knowledge of the trainees in an assessment may lead to personal biases as well (Schwartz et al 2020). Moreover, it may jeopardize the role of a trainer as an assessor by causing role conflict (Caro Monroig et al 2021, ten Cate 2016).

What have I learned? (Suggestions and recommendations):

Researchers in medical education need to systematically evaluate the measurement errors based on the knowledge of cognitive psychology, and form guidelines to reduce the thinking errors in clinical reasoning, diagnostics and personal judgements in high or low-stakes examinations.

If a cognitive bias has been identified, it needs to be addressed by training of the assessor, helping him/her engage in to analytical thinking rather than intuitive patterns

of cognitive processing of information. Critical thinking and reflective strategies may also be introduced as a regular part of assessors training workshops.

Absolute objective assessment of all the aspects of clinical performance in terms of numerical scoring and psychometric testing is a myth, the earlier the educators try to bust the myths in medical education and assessment of competencies, the more efforts they will try to put into working towards devising methods to incorporate subjective judgement of the raters in summative assessments.

Data on performance of trainees need to be gathered from multiple sources, on multiple occasions by clinical advisors and trainers. The result may be maintained in the form of electronic portfolio which should be easily accessible and available to the faculty and the trainee.

The learners be considered as equal stakeholders in CBME and may allow to advocate their learning plans and training trajectory. Learning plans need to be reviewed and discussed with the trainees in detail, including the important benchmarks and the consequences of not achieving those benchmarks and the impact of remediation on the period of training. Flexible time-variable training plans should be introduced for exceptionally bright and weak students to help them achieve their goals.

I think the close supervision model in CBME is not much different from the apprenticeship model in the old traditional training model. However, in CBME, it is more structured and define the levels of autonomy and supervision needed at each

level. Hence, it makes the supervisor more accountable in terms of trusting the trainee with the responsibility assigned to him/her.

The undocumented data has been considered as problematic evidence for so long. It needs to be incorporated into documented data in ad-hoc and summative assessments to get a holistic picture of the trainee performance.

The context is very important in EPA based assessments. All the variables need to be defined and explained to the trainee and the rater in detail. The result of ad-hoc decisions has to be more than just a tick box exercise. The scores along with narrative decisions should be well documented, which can be utilised in summative decisions later on.

The retrospective and prospective scales, when used individually, need to be made explicit, describing the anchors of scoring and judgement in detail. Adequate training of raters need to be ascertained before they are allowed to evaluate the trainees. Faculty training workshops should be made a regular feature of all the under-graduate and graduate EPA-based training programmes. Supervisory grades may be assigned to the clinical supervisors and advisors according to their experience as a trainer and assessor. Only senior raters may be allowed to judge the complex performances, the relevant milestones and the advanced EPAs.

Patient safety markers, rate of relapses and hospitalisation, number of follow up visits, patient satisfaction and feedback from family and care givers may be included in

evaluation of the training programmes. Participation of the physicians in sub-specialities and research activities is another marker of the quality of training.

The role of each member of CCC needs to be redefined and reinforced. Proper mandate, structure and functioning has to be devised, closely monitored and documented. An internal audit team can be helpful in this regards. The audit team can reflect on and critically evaluate the actions of each team member and overall functioning of the committee. They can give a feedback to the chair of the committee on quarterly basis. The chair can advise for further training of any of the committee member or can request for removal or replacement.

Last but not the least, I think as local health corporate sector along with national government bodies is seeking to define medical practice in terms which makes it defensible, the competence based medical model will eventually incorporate in it an 'occupational competency' which will be based on job performance and occupational outcomes, assessed by colleagues, patients and other stakeholders.

References:

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R. and Zhang, D. (2008). Instructional Interventions Affecting Critical Thinking Skills and Dispositions: A Stage 1 Meta-Analysis. *Review of Educational Research,* 78(4), 1102–1134.

Accreditation Council for Graduate Medical Education 2012 Milestones. http://www.acgme-nas.org/milestones.html.

Ackerman, R. and Thompson, V. A. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in cognitive sciences*, *21*, 8, 607–617.

Albanese, M. and Case, S. M. (2016). Progress testing: Critical analysis and suggested practices. *Advances in Health Science Education Theory and Practice,* 21, 221–234.

Apramian, T., Cristancho, S., Watling, C., Ott, M. and Lingard, L. (2015). Thresholds of principle and preference: exploring procedural variation in postgraduate surgical education. *Academic Medicine*, 90, 11, 70–76.

Arksey, H. and O'Malley, L. (2005) Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8, 19-31.

Aveyard, H., Sharp, P. and Woolliams, M. (2015). *A Beginner's Guide to Critical Thinking and Writing in Health and Social Care.* 2nd edition. Maidenhead: Open University Press.

Baddeley, A. D. (1986). *Working Memory*. Oxford: Oxford University Press

Bandura, A. and Walters, R. H. (1977). *Social learning theory*: Prentice hall Englewood Cliffs, NJ.

Banks, C. A. M. and Banks, J. A. (1995). Equity pedagogy: An essential component of multicultural education. *Theory Practice*, 34, 152–158.

Barrows, H. S., Norman, G. R., Neufeld, V. R. and Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medical practice. *Clinical and investigative medicine. 5*, 1, 49–55.

Berendonk. C., Stalmeijer, R.E. and Schuwirth, L.W. (2013). Expertise in performance assessment: assessors' perspectives. *Advances in Health Sciences Education Theory Practical*,18, 559–71.

Brown, D. R., Moeller, J. J., Grbic, D., Biskobing, D. M., Crowe, R., Cutrer, W. B., Green, M. L., Obeso, V. T., Wagner, D. P., Warren, J. B., Yingling, S. L., Andriole, D. A., & Core Entrustable Professional Activities for Entering Residency Pilot (2021). Entrustment Decision-Making in the Core EPAs: Results of a Multi-Institutional Study. *Academic medicine: journal of the Association of American Medical Colleges*, 10.1097/ACM.0000000000004242. Advance online publication.

Bransford, J.D., Brown. A.L. and Cocking, R.R. (1999) *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press**.**

Byrne, A., Tweed, N. and Halligan, C. (2014) A pilot study of the mental workload of objective structured clinical examination examiners. *Medical Education*, 48, 262–267.

CanMEDS Framework (2015). https://www.royalcollege.ca/rcsite/canmeds/canmeds-framework-e

Caro Monroig, A. M., Chen, H. C., Carraccio, C., Richards, B. F., ten Cate, O. and Balmer, D. F. (2021). Medical Students' Perspectives on Entrustment Decision Making in an Entrustable Professional Activity Assessment Framework: A Secondary Data Analysis. *Academic Medicine, 96,* 8, 1175-1181.

Charlin, B., Gagnon, R., Pelletier, J., Coletti, M., Abi-Rizk, G., Nasr, C., Sauvé, E. and van der Vleuten, C. (2006). Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Medical education*, *40*, 9, 848–854.

Chen, H. C. and ten Cate, O. (2020). The ingredients of a rich entrustment decision. Medical Teacher, 42, 12.

Choo, K. J., Arora, V. M., Barach, P., Johnson, J. K. and Farnan, J. M. (2014). How do supervising physicians decide to entrust residents with unsupervised tasks? A qualitative analysis. *Journal of Hospital Medicine*, 9, 169–175.

Coleman, E. A., Min, S. J., Chomiak, A. and Kramer, A. M. (2004) Posthospital care transitions: patterns, complications, and risk identification. *Health Services Research*, 39, 5, 1449–1465.

Cook, D., A. and Hatala, R. (2016). Validation of educational assessments: a primer for simulation and beyond. *Advanced Simulation* (Lond). 1:31.

Cook, D. A., Beckman, T. J., Mandrekar, J. N. and Pankratz, V. S. (2010). Internal structure of mini-CEX scores for internal medicine residents: Factor analysis and generalizability. *Advances in Health Sciences Education,* 15, 5, 633–645.

Cox K. (2000). Examining and recording clinical performance: a critique and some recommendations. *Education for health (Abingdon, England)*, *13*(1), 45–52.

Croskerry, P. (2013). From Mindless to Mindful Practice — Cognitive Bias and Clinical Decision Making. *The New England Journal of Medicine, 368* (26), 2445-2448.

Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., Ratcliffe, T., Gordon, D., Heist, B., Lubarsky, S., Estrada, C. A., Ballard, T., Artino, A. R., Jr, Sergio Da Silva, A., Cleary, T., Stojan, J. and Gruppen, L. D. (2019). Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Academic medicine: journal of the Association of American Medical Colleges*, *94*(6), 902–912.

Dijksterhuis, M. G., Voorhuis, M., Teunissen, P. W., Schuwirth, L. W., ten Cate, O. T., Braat, D. D. and Scheele, F. (2009). Assessment of competence and progressive independence in postgraduate clinical training. *Medical education*, *43*, 12, 1156–1165.

Dijkstra, J., Van der Vleuten, C. P. and Schuwirth, L. W. (2010). A new framework for designing programmes of assessment. *Advances in health sciences education: theory and practice*, *15*(3), 379–393.

Downing, S. M. and Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38: 327–333.

Downing, S. M. and Yudkowsky, R. (2009). *Assessment in Health Professions Education*. New York: Taylor and Francis.

Dudek, N. L., Marks, M. B. and Regehr, G. (2005). Failure to fail: the perspectives of clinical supervisors. *Academic medicine: journal of the Association of American Medical Colleges*, 80, 10, 84–87.

Duijn, C., Dijk, E., Mandoki, M., Bok, H. and Cate, O. (2019). Assessment Tools for Feedback and Entrustment Decisions in the Clinical Workplace: A Systematic Review. *Journal of veterinary medical education*, *46*(3), 340–352.

Durning, S. J., Artino, A., Boulet, J., La Rochelle, J., van der Vleuten, C., Arze, B. and Schuwirth, L. (2012). The feasibility, reliability, and validity of a post-encounter form for evaluating clinical reasoning. *Medical teacher*, *34*(1), 30–37.

Elstein, A. S., Shulman, L. S. and Sprafka, S. A. (1978). *Medical Problem Solving an Analysis of Clinical Reasoning.* Cambridge, MA: Harvard University Press.

Elstein, A. S., Shulman, L. S. and Sprafka, S. A. (1990) Medical problem solving: a ten-year retrospective. *Evaluations in Health Profession*. 13, 1, 5-36.

Englander, R., Flynn, T., Call, S., Carraccio, C., Cleary, L., Fulton, T.B., Garrity, M.J., Lieberman, S.A., Lindeman, B., Lypson, M. L., Rebecca M. Minter, R.M., Rosenfield, J., Thomas, J., Wilson, M.C. and Aschenbrener. C.A. (2016) Toward Defining the Foundation of the MD Degree: Core Entrustable Professional Activities for Entering Residency. *Academic Medicine*, 91,10.

Englander, R., Frank, J. R., Carraccio, C., Sherbino, J., Ross, S., Snell, L. and ICBME Collaborators. (2017) Toward a shared language for competency-based medical education, *Medical Teacher*, 39, 6, 582-587.

Eva., K. W. (2018) Cognitive Influences on Complex Performance Assessment: Lessons from the Interplay between Medicine and Psychology. *Journal of Applied Research in Memory and Cognition*, 7, 177-188.

Eva, K. W., Solomon, P., Neville, A. J., Ladouceur, M., Kaufman, K., Walsh, A. and Norman, G. R. (2007). Using a sampling strategy to address psychometric challenges in tutorial-based assessments. *Advances in Health Sciences Education,* 12, 19–33.

Facione PA. (1990). *Critical Thinking*: *A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*. Millbrae, CA: The California Academic Press.

Favreau, M. A., Tewksbury, L., Lupi, C., Cutrer, W. B., Jokela, J. A., Yarris, L. M. and AAMC Core Entrustable Professional Activities for Entering Residency Faculty Development Concept Group (2017). Constructing a Shared Mental Model for Faculty Development for the Core Entrustable Professional Activities for Entering Residency. *Academic medicine: journal of the Association of American Medical Colleges*, *92*(6), 759–764.

Flavell, J. H. (1976) Metacognitive aspects of problem solving. In: Resnick LB, ed. *The Nature of Intelligence*. Hillsdale, NJ: Lawrence Erlbaum Associates, 231-235.

Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R., Harris, P., Glasgow, N. J., Campbell, C., Dath, D., Harden, R. M., Iobst, W., Long, D. M., Mungroo, R., Richardson, D. L., Sherbino, J., Silver, I., Taber, S., Talbot, M. and Harris, K. A. (2010). Competency-based medical education: theory to practice. *Medical teacher*, *32*(8), 638–645.

Frank, J. R., Snell, L., Englander, R., Holmboe, E. S. and ICBME Collaborators (2017). Implementing competency-based medical education: Moving forward. *Medical teacher*, *39*(6), 568–573.

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

Gauthier, G., St-Onge, C. and Tavares, W. (2016). Rater cognition: review and integration of research findings. *Medical education*, *50*(5), 511–522.

Gigerenzer, G. and Selten, R. (2002). *Bounded rationality: The adaptive toolbox.* Cambridge, MA: MIT Press.

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M. and Holmboe, E. (2014). Seeing the 'black box' differently: assessor cognition from three research perspectives. *Medical education*, *48*(11), 1055–1068.

Gingerich, A., Regehr, G. and Eva, K.W. (2011) Rater-based assessments as social judgments: rethinking the aetiology of rater errors. *Academic Medicine*, 86,10, 1–7.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K. and Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine*, 85, 780–786.

Ginsburg, S., Watling, C. J., Schumacher, D. J., Gingerich, A. and Hatala, R. (2021). Numbers Encapsulate, Words Elaborate: Toward the Best Use of Comments for Assessment and Feedback on Entrustment Ratings. *Academic medicine: journal of the Association of American Medical Colleges*, *96*, 7, 81–86.

Gomez-Garibello, C. and Young, M. (2018). Emotions and assessment: considerations for rater-based judgements of entrustment. *Medical education*, *52*(3), 254–262.

Govaerts, M., van der Vleuten, C., Schuwirth, L. and Muijtjens, A. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education*, 12, 2, 239–260.

Gruppen, L. D., ten Cate, O., Lingard, L. A., Teunissen, P.W. and Kogan, J. R. (2018). Enhanced Requirements for Assessment in a Competency-Based, Time-Variable Medical Education System. *Academic Medicine*, 93, 3, 17-21.

Hart, D., Franzen, D., Beeson, M., Bhat, R., Kulkarni, M., Thibodeau, L., Weizberg, M. and Promes, S. (2019). Integration of Entrustable Professional Activities with the Milestones for Emergency Medicine Residents. *The western journal of emergency medicine*, 20, 1, 35–42.

Hatala, R., Ginsburg, S., Hauer, K. E. and Gingerich, A. (2019). Entrustment Ratings in Internal Medicine Training: Capturing Meaningful Supervision Decisions or Just Another Rating? *Journal of general internal medicine*, *34*, 5, 740–743.

Hauer, K. E., ten Cate, O., Boscardin, C. K., Iobst, W., Holmboe, E. S., Chesluk, B., Baron, R. B. and O'Sullivan, P. S. (2016). Ensuring Resident Competence: A Narrative Review of the Literature on Group Decision Making to Inform the Work of Clinical Competency Committees. *Journal of graduate medical education*, *8*, 2, 156–164.

Hauer, K. E., ten Cate, O., Boscardin, C., Irby, D. M., Iobst, W. and O'Sullivan, P. S. (2014). Understanding trust as an essential element of trainee supervision and learning in the workplace. *Advances in health sciences education: theory and practice*, *19*, 3, 435–456.

Holmboe, E. S., Hawkins, R. E. and Huot, S. J. (2004). Effects of training in direct observation of medical residents' clinical competence. *Annals of Internal Medicine,* 140, 874–881.

Holmboe, E. S., Sherbino, J., Englander, R., Snell, L., Frank, J. R. and ICBME Collaborators (2017). A call to action: The controversy of and rationale for competency-based medical education. *Medical teacher*, *39*, 6, 574–581.

Iramaneerat, C. and Yudkowsky, R. (2007). Rater Errors in a Clinical Skills Assessment of Medical Students. *Evaluation and the Health Professions*. 30, 3, 266-83.

Kahneman D. (2011). *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Kane, M. T. (1992). An argument-based approach to validity. Psychology Bulletin. 112, 3, 527–535.

Kennedy, T. J. T., Regehr, G., Baker, G. R. and Lingard, L. (2008). Point-of-care assessment of medical trainee competence for independent clinical work. *Academic Medicine*. 83, 10, 89–92.

Kogan, J. R., Conforti, L. N., Bernabeo, E. C., Durning, S. J., Hauer, K. E., & Holmboe, E. S. (2012). Faculty staff perceptions of feedback to residents after direct observation of clinical skills. *Medical education*, *46*, 2, 201–215.

Kogan, J. R., Hess, B. J., Conforti, L. N. and Holmboe, E. S. (2010) What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Academic Medicine*, 85, 10, 25–28.

Kreiter, C. D., Wilson, A. B., Humbert, A. J. and Wade, P. A. (2016). Examining rater and occasion influences in observational assessments obtained from within the clinical environment. *Medical education online*, *21*, 29279.

Krupat E. (2017). Critical thoughts about the core entrustable professional activities in undergraduate medical education. Academic Medicine. 93, 3, 1.

Lee, V., Brain. K. and Martin, J. (2017) Factors influencing Mini-CEX rater judgments and their practical implications: A systematic literature review. *Academic Medicine*, 92, 6, 880-887.

Lee, V., Brain, K. and Martin, J. (2019) From opening the 'black box' to looking behind the curtain: cognition and context in assessor-based judgements. *Advances in Health Sciences Eduction*, 24**,** 85–102.

Levy, P. and Williams, J. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, 30, 6, 881–905.

Livingston, S. and Zieky, M. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Lobst, W. F. and Holmboe, E. S. (2020) Programmatic Assessment: The Secret Sauce of Effective CBME Implementation. *Journal of Graduate Medical Education*,12, 4, 518-521.

Lobst, W.F., Sherbino, J., ten Cate, O., Richardson, D. L., Dath, D., Swing, S. R., Harris, P., Mungroo, R., Holmboe, E.S. and Frank. J. R. (2010) Competency-based medical education in postgraduate medical education. *Medical Teacher*, 32,651–656.

Lockyer, J., Carraccio, C., Chan, M. K., Hart, D., Smee, S., Touchie, C., Holmboe, E. S., Frank, J. R. and ICBME Collaborators (2017). Core principles of assessment in competency-based medical education. *Medical teacher*, *39*, 6, 609–616.

Lord, R. G. and Maher, K. J. (1990). Alternative information-processing models and their implications for theory, research, and practice. The Academy of Management Review, 15, 1, 9–28.

Lucey, C. R., Hauer, K. E., Boatright, D. and Fernandez, A. (2020). Medical Education's Wicked Problem: Achieving Equity in Assessment for Medical Learners. *Academic medicine: journal of the Association of American Medical Colleges*, *95*, 12, Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments), 98–108.

Mann, K., Gordon, J. and MacLeod, A. (2009). Reflection and reflective practice in health professions education: a systematic review. *Advances in health sciences education: theory and practice*, *14*, 4, 595–621.

Marchegiani L, Reggiani T, Rizzolli M. (2013). Severity vs. *Leniency Bias in Performance Appraisal: Experimental Evidence*. Bozen Economics and Management Paper Series.

Mayer, R. E. (2010). Applying the science of learning to medical education. *Medical Education*, 44, 6, 543–549.

Mayer, R., Davis, J. and Schoorman, F. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review, 20*, 3, 709-734.

McGill, D. A., van der Vleuten, C. P. and Clarke, M. J. (2011). Supervisor assessment of clinical and professional competence of medical trainees: a reliability study using workplace data and a focused analytical literature review. *Advances in health sciences education: theory and practice*, *16*, 3, 405–425.

Mckinley, D. and Norcini, J. (2014). How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher, 36*, 110 - 97.

McMillan, J.H. (2017). Classroom assessment: *Principles and practice that enhance student learning and motivation*. Pearson.

Melvin, L., Rassos, J., Stroud, L. and Ginsburg, S. (2020). Tensions in Assessment: The Realities of Entrustment in Internal Medicine. *Academic Medicine, 95,* 4, 609-615.

Messick S. Validity. (1989). In: Linn RL, ed. *Educational Measurement*. 3rd edn. New York: American Council on Education, Macmillan,13–104.

Mohr, C. D. and Kenny, D. A. (2006). The how and why of disagreement among perceivers: An exploration of person models. *Journal of Experimental Social Psychology, 42*(3), 337–349.

Mohr, J., Batalden, P. and Barach, P. (2004) Integrating patient safety into the clinical microsystem. *Quality Safe Health Care*,13, 2, 34-38.

Monteiro, S., Sherbino, J., Sibbald, M. and Norman, G. (2020). Critical thinking, biases and dual processing: The enduring myth of generalisable skills. *Medical education*, *54*, 1, 66–73.

Nasca, T. J., Philibert, I., Brigham, T. and Flynn, T. C. (2012). The next GME accreditation system – Rationale and benefits. *New England Journal of Medicine*, 366,1051–1056.

Norcini, J. and Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher*, 29:855–871.

O'Dowd, E., Lydon, S., O'Connor, P., Madden, C. and Byrne, D. (2019). A systematic review of 7 years of research on entrustable professional activities in graduate medical education, 2011-2018. *Medical education*, *53*(3), 234–249.

Park, B., DeKay, M. L. and Kraus, S. (1994). Aggregating social behavior into person models: Perceiver-induced consistency. *Journal of Personality and Social Psychology, 66*(3), 437–459.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington: National Academy Press.

Perkins, D. N. and Salomon, G. (1989). Are cognitive skills context-bound? Ed Res, 18, 1, 16-25.

Pelgrim, E.M., Kramer, A.M., Mokkink, H.A., van den Elsen, L., Grol, R.M. and van der Vleuten, C.M. (2011) In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in Health Sciences Education*, 16,131–142.

Rekman, J., Gofton, W., Dudek, N., Gofton, T. and Hamstra, S. (2016). Entrustability Scales. *Academic Medicine, 91* (2), 186-190.

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC medical informatics and decision making*, *16*(1), 138.

Schott, M., Kedia, R., Promes, S. B., Swoboda, T., O'Rourke, K., Green, W., Liu, R., Stansfield, B., & Santen, S. A. (2015). Direct Observation Assessment of Milestones: Problems with Reliability. *The western journal of emergency medicine*, *16*(6), 871–876.

Schumacher, D. J., Poynter, S., Burman, N., Elliott, S. P., Barnes, M., Gellin, C., del Rey, J. G., Sklansky, D., Thoreson, L., King, B. and Schwartz, A. (2019). Justifications for Discrepancies Between Competency Committee and Program Director Recommended Resident Supervisory Roles. *Academic Pediatrics*, 19, 5, 561-565.

Schumacher, D. J., West, D. C., Schwartz, A., Li, S. T., Millstein, L., Griego, E. C., Turner, T., Herman, B. E., Englander, R., Hemond, J., Hudson, V., Newhall, L., McNeal Trice, K.,

Baughn, J., Giudice, E., Famiglietti, H., Tolentino, J., Gifford, K., Carraccio, C. and Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network General Pediatrics Entrustable Professional Activities Study Group (2020). Longitudinal Assessment of Resident Performance Using Entrustable Professional Activities. *JAMA network open*, *3*, 1, e1919316.

Schuwirth, L. and Ash, J. (2013). Assessing tomorrow's learners: In competency-based education only a radically different holistic method of assessment will work. Six things we could forget, *Medical Teacher*, 35:7, 555-559.

Schuwirth, L., van der Vleuten, C. and Durning, S. J. (2017). What programmatic assessment in medical education can learn from health care. *Perspectives on medical education*, *6*(4), 211–215.

Schwartz, A., Balmer, D. F., Borman-Shoap, E., Chin, A., Henry, D., Herman, B. E., Hobday, P., Lee, J. H., Multerer, S., Myers, R. E., Ponitz, K., Rosenberg, A., Soep, J. B., West, D. C. and Englander, R. (2020). Shared Mental Models Among Clinical Competency Committees in the Context of Time-Variable, Competency-Based Advancement to Residency. *Academic medicine: journal of the Association of American Medical Colleges*, *95*(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 59th Annual Research in Medical Education Presentations), 95–102.

Scott, J. N., Markert, R. J. and Dunn, M. M. (1998) Critical thinking: change during medical school and relationship to performance in clinical clerkships. Medical Eduction. 32, 1, 14-18.

Sibbald, M., Mansoor, M., Tsang, M., Blissett, S. and Norman, G. (2021). The critical role of direct observation in entrustment decisions. *Canadian Medical Education Journal*. (In press).

Stanovich, K. E. (2010). Rationality and the reflective mind. New York, NY: Oxford University Press.

Sterkenburg, A., Barach, P., Kalkman, C., Gielen, M. and ten Cate, O. (2010) When do supervising physicians decide to entrust residents with unsupervised tasks? *Academic Medicine*, 85, 9, 1408-17.

St-Onge, C., Chamberland, M., Lévesque, A. and Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in health sciences education: theory and practice*, *21*(3), 627–642.

Tavares, W. and Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education*, 18, 291–303.

Tavares, W., Ginsburg, S. and Eva, K. W. (2016). Selecting and simplifying: Rater performance and behaviour when considering multiple competencies. *Teaching and Learning in Medicine*, 28, 41–51.

Tavares, W., Rowland, P., Dagnone, D., McEwen, L. A., Billett, S. and Sibbald, M. (2020). Translating outcome frameworks to assessment programmes: Implications for validity. *Medical education*, *54*(10), 932–942.

Taylor, D., Park, Y. S., Smith, C., Cate, O. T., and Tekian, A. (2021). Constructing Approaches to Entrustable Professional Activity Development that Deliver Valid Descriptions of Professional Practice. *Teaching and learning in medicine*, *33*, 1, 89–97.

Taylor, D. R., Park, Y. S., Smith, C. A., Karpinski, J., Coke, W. and Tekian, A. (2018). Creating Entrustable Professional Activities to Assess Internal Medicine Residents in Training: A Mixed-Methods Approach. *Annals of internal medicine*, *168*, 10, 724–729.

Tekian, A., ten Cate, O., Holmboe, E., Roberts, T. and Norcini, J. (2020) Entrustment decisions: Implications for curriculum development and assessment, *Medical Teacher*, 42,6, 698-704.

ten Cate, O. (2005) Entrustability of professional activities and competency-based training. *Medical Education*. 39,1176–1177.

ten Cate O. (2013) Nuts and bolts of Entrustable Professional Activities. *Journal of Graduate Medical Education*, 5, 157–158.

ten Cate, O. (2016). Entrustment as assessment: recognizing the ability, the right, and the duty to act. *Journal of Graduate Medical Education*, 8:261–2.

ten Cate, O., Carraccio, C., Damodaran, A., Gofton, W., Hamstra, S. J., Hart, D. E., Richardson, D., Ross, S., Schultz, K., Warm, E. J. , Whelan, A. J. an Schumacher, D. J. (2021). Entrustment Decision Making: Extending Miller's Pyramid. *Academic Medicine, 96* (2), 199-204.

ten Cate, O. and Scheele. F. (2007) Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice? *Academic Medicine*, 82, 6, 542–547.

ten Cate, O., Balmer, D. F., Caretta-Weyer, H., Hatala, R., Hennus, M. P. and West, D. C. (2021). Entrustable Professional Activities and Entrustment Decision Making: A Development and Research Agenda for the Next Decade. *Academic medicine: journal of the Association of American Medical Colleges*, *96*, 7, 96–104.

ten Cate, O., Chen, H. C., Hoff, R. G., Peters, H., Bok, H. and van der Schaaf, M. (2015). Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. *Medical teacher*, *37*, 11, 983–1002.

ten Cate, O., Hart, D., Ankel, F., Busari, J., Englander, R., Glasgow, N., Holmboe, E., Iobst, W., Lovell, E., Snell, L. S., Touchie, C., Van Melle, E., Wycliffe-Jones, K. and International Competency-Based Medical Education Collaborators (2016). Entrustment Decision Making in Clinical Training. *Academic medicine,* 91, 2, 191–198.

ten, Cate, O. and Hoff, R. G. (2017). From case-based to entrustment-based discussions. *Clinical Teacher*. 14:385–389.

ten Cate O and Regehr G. (2019) The Power of Subjectivity in the Assessment of Medical Trainees. *Academic Medicine*, 94, 3, 333-337.

ten Cate, O. and Taylor, D. R. (2020) The recommended description of an entrustable professional activity: AMEE Guide No. 140, *Medical Teacher*.

Thomas, M. R., Beckman, T. J., Mauck, K. F., Cha, S. S., & Thomas, K. G. (2011). Group assessments of resident physicians improve reliability and decrease halo error. *Journal of general internal medicine*, *26*(7), 759–764.

Tomorrow's Doctors GMC (2020). https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/outcomes-for-graduates

Touchie, C., Kinnear, B., Schumacher, D., Caretta-Weyer, H., Hamstra, S. J., Hart, D., Gruppen, L., Ross, S., Warm, E., Ten Cate, O. and ICBME Collaborators (2021). On the validity of summative entrustment decisions. *Medical teacher*, 1–8. Advance online publication.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

van der Vleuten, C.P., Schuwirth, L.W., Driessen, E.W., Dijkstra, J., Tigelaar, D., L K J Baartman, L. K. J. and Tartwijk, J. (2012) A model for programmatic assessment fit for purpose. *Medical Teacher*,34, 3, 205–214.

Wass, V., van der Vleuten, C. P., Shatzer, J. and Jones, R. (2001) Assessment of clinical competence. *Lancet,* 357(9260):945-9.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychology Review*, 101, 34 –52.

Whitcomb, M. E. (2007) Redirecting the assessment of clinical competence. *Academic Medicine*, 82, 527–528.

Whitehead, C. R., Kuper, A., Hodges, B. and Ellaway, R. (2015) Conceptual and practical challenges in the assessment of physician competencies. *Medical Teacher*, 37, 245–251.

Wickens, C. D. and Carswell, C. (2006). *Handbook of human factors and ergonomics* (3rd ed.). Hoboken, N J: Wiley.

Wijnen-Meijer, M., Van der Schaaf, M., Booij, E., Harendza, S., Boscardin, C., Van Wijngaarden, J. and Ten Cate, T. J. (2013). An argument-based approach to the validation of UHTRUST: can we measure how recent graduates can be trusted with

unfamiliar tasks? *Advances in health sciences education: theory and practice*, *18*(5), 1009–1027.

Williams, R. G., Chen, X. P., Sanfey, H., Markwell, S. J., Mellinger, J. D. and Dunnington, G. L. (2014). The measured effect of delay in completing operative performance ratings on clarity and detail of ratings assigned. *Journal of surgical education*, *71*, 6, 132–138.

Wittenbaum GM. (1998). Information sampling in decision making groups: the impact of members' task-relevant status. *Small Group Research*. 29, 1, 57–84.

Yoon, M. H., Kurzweil, D. M., Durning, S. J., Schreiber-Gregory, D. N., Hemmer, P. A., Gilliland, W. R. and Dong, T. (2020). It's a matter of trust: exploring the basis of program directors' decisions about whether to trust a resident to care for a loved one. *Advances in health sciences education: theory and practice*, *25*, 3, 691–709.

Young, J. Q., Van Merrienboer, J., Durning, S. and ten Cate, O. (2014). Cognitive load theory: implications for medical education: AMEE Guide No. 86. *Medical Teacher*, 36, 371–84.

Appendices

Appendices

Appendix 1: Table of search terms, databases and number of studies

Name of Database	Pubmed	OVID (all resources)	Google Scholar
Query # 1	measurement errors AND assessment of competencies AND medical education	Entrustment decisions AND validity	Measurement errors AND competency based medical education
Query # 2	Entrustable Professional Activities AND summative decisions	Rater cognition AND programmatic assessment	Measurement errors AND entrustment decisions
Query # 3	Validity AND entrustment decisions	Rater biases AND entrustable professional activities	Rater biases AND EPA-based assessments
Query # 4	Inter-rater variability AND entrustment decisions		
Query # 5	Rater cognition OR rater judgement AND entrustment decisions		
Number of results	777	148	112,000 (sorted by relevance) 1,990

Appendix 2: Sources of Measurement Errors on the basis of Kane's validity framework

Table 2: Sources of Measurement Errors on the basis of Kane's validity framework			
Sources of Errors or Inference			
Scoring	**Generalization**	**Extrapolation**	**Implications**

Direct observation	Too few raters	No evidence of patient related outcomes	How criterion for assessment is set for summative assessments
Entrustability scales	Inadequate sampling of cases	No evidence of quality health care measures	Impact on quality of training programme
Construct alignment	Inability of raters to discriminate levels of entrustment or supervision	Paucity of evidence of accountability of our physicians	Impact on graduation and CPD outcomes
Scoring and data gathering process	Inter-rater variability in the members of CCC	No records of Continuous Professional Development (CPD) activities or post-graduation employment data	Standard setting for entrustment decisions and remediation
Context specificity	Generalizability analysis	Authentic assessment method	Impact on learners
Rater	Factor analysis		Impact on all stake holders
Trainee			
Supervisor-trainee relationship			
Shared mental model			
Functioning of CCC			

Appendix 3: Sources of Measurement Errors on the basis of Messick Framework

Table 3: Sources of Measurement Errors on the basis of Messick Framework
Sources of Errors

Content	**Response process**	**Internal structure**	**Relationship to other variables**	**Consequences**
Selection of expert	Clear understanding of assessment goals	Generalizability analysis	Quality care measures	Impact on learners
Identification of EPAs based on professional tasks	Untrained raters	Factor analysis	Patient related outcomes	Quality of training programme
Mapping of EPAs to the constructs to be assessed	Use of entrustability scales		Accountability of our physicians	Standards setting for entrustment decisions and remediation
	Direct observation of performance			
	Data gathering process			
	Working of CCC			

Appendix 4: ISSUES AND MAJOR CHALLENGES TO IMPLEMENTATION OF EPA-BASED ASSESSMENTS:

As the interest rises in CBME around the turn of the millennium, many reports have concluded that the results are not as satisfying as expected (Holomboe et al 2017). In 1978, Mc Gaghie et al (p. 18) described the goal of CBME as “The intended output of a

competency-based program is a health professional who can practice medicine at a defined level of proficiency, in accord with local conditions, to meet local needs". With a philosophy revolving around patient-centeredness, the interests of public and the general concerns in the community and policy makers about the need to justify the investment of resources, the primary aim of competency based medical education to produce competent physicians is methodologically daunting and controversial (Christine et al 2018). The bare minimum requirement for the structure of CBME may include the educational support in the form of technology, cost and training of faculty members, availability and easy access to electronic data, regular and constructive feedback to the trainees, expertise in formative and summative assessments and institutional leadership to run the programme successfully (Andrew et al 2020). The programmatic approach to assessment is more than getting the system functional. Assessment of individual EPAs using an entrustment scale, giving judgement on achievement of milestones along with a narrative feedback is considered overwhelming by many faculty members (Andrew et al 2020). Moreover, assessments based on individual milestones may provide little information to base the formal progression decisions by the clinical committees. Consequently, many faculty members like to disengage from this responsibility of assessing milestones if given an option (Andrew et al 2020). Similarly, they also find giving constructive feedback a cumbersome job. One faculty from Emergency Medicine describes the difficulties as "Even though I'm very motivated to provide targeted feedback, your brain just naturally thinks globally, you

have to actually make that cognitive step from your overall gestalt about somebody down to the task, and then really think about the very specifics of the task in order to give good feedback. That's not easy, and I still fail sometimes (Andrew et al 2020, p.5)." Lastly, clinical tutors are confused about their role of coaching the trainees, mentoring and evaluating them. Hence, if educators like to keep the CBME programmes up and running, they need faculty members who are honest, motivated to learn, innovate and are ready to give constructive feedback to help and guide the clinical committees to make informed decisions on trainees' progress.

Deconstruction and reductionist approach:

Certain operational challenges threaten the successful implementation of CBME. To assess the knowledge and attitudes associated with a competency, it is mandatory to split it in to sub components, which can be observed and assessed (ten Cate et al 2016). It is also essential to give effective feedback to the trainees based on these components to help them progress in their trajectory and contribute towards cumulative scores in the form of summative assessments. The feedback needs to be constructive and highly focused, even though reductionist (granular). During the training, a number of workplace based assessments, as a part of formative assessments may target individual EPAs, the respective milestones and integrates a number of competencies. Competencies count towards the entrustment decisions. Isolated assessment of individual EPAs is useful in providing useful pieces of information on certain aspects of performance rather than the bigger picture (Andrew et al 2020). A trainee who is expert

in one of the competencies may not be so proficient in all the aspects of other competencies. When clinical supervisors are asked, how are the things going overall, they may not be able to give a satisfactory response regarding the general progress of the trainees (Andrew et al 2020). The exhibited benefit of CBME in terms of patient-centeredness and quality reforms may still need vigorous faculty development programmes, targeted feedback to trainees, evidence-based informed decisions by clinical committees and more frequent assessments to monitor the on-going learning of trainees to generate real-time data to sustain this change and implement CBME successfully (Andrew et al 2020).

No updated universal practical guidelines for implementation of EPA based assessments and working of Clinical Competency Committees

The key components of EPA based competency framework such as ad-hoc assessments, entrustment decisions, subjective judgements and supervisions scales lack a universal language and best practice guidelines which has created confusion in the minds of educators. As CBME promises to address the physicians' accountability towards public and stake holders, and to ensure utmost patients safety and health care standards, it may take this risk on the shoulders of the raters who claim to know the nuts and bolts of assessments in terms of fairness and explicitness. But if the assessors cannot document their decisions according to the published standards, cannot describe all the aspects of observed performances in black and white, and fail to capture in words what they see in their routine clinical hours spent with their trainees in emergency and

operating rooms, then perhaps it does not count (van Enk and ten Cate 2020). Sometimes the clinical advisors 'just know' about their learners by virtue of their practical experience and occurs as an implicit background knowledge which cannot be expressed due to difficulty in documenting it the form of a scores or completing the standard forms. The concept of entrustment comes handy in this regard as it is based on informed judgement by the assessors involving all the implicit and explicit variables that may count towards assigning responsibility to the trainee in the form of a formal decision. It will take several years to actually find out to what extent the 'tacit expert knowledge' counts towards the fairness of assessment through published data, as current CBME based iterations are still relatively new (van Enk and ten Cate 2020).

Similarly, there is disparity in language used for ad-hoc and summative assessments. Ad-hoc assessment can take place as a part of the routine clinical work on the basis of expertise of the trainee in any specific EPA, the supervision required and urgency of the task (van Enk and ten Cate 2020). There can be variation in the criteria as how these ad-hoc assessments are employed towards the formal summative decisions. In North America for instance, only ad-hoc form of entrustment count towards formal decisions. On the other hand, in Netherlands, summative entrustment decisions based on EPAs are the main focus of assessment (van Enk and ten Cate 2020). The role of Clinical Competency Committees becomes important in summative entrustment decisions. The ad-hoc and summative decisions are made at different points during the training period and the requirement for formal documentation by the assessors based on their

judgement can be an area of tension that needs more evidence-based collection of data from multiple sources or individuals who have worked with the trainee (van Enk and ten Cate 2020). Ad-hoc decisions utilize scales that look back at the performance of the trainee in time, as regard to how much supervision was provided to the trainee and in which specific situation. It involves assessor's subjective judgement along with a tick box and unelaborated score, which is based on workplace-based task, documented into trainee's portfolio or electronic records (van Enka and ten Cate 2020). The idea is to keep a record for formal decision making in future and the level of autonomy allowed. If ad-hoc retrospective scales are the only means of evaluating the trainees by the clinical committees, then conclusions drawn on undocumented information about the trainees may create biases (Tam et al 2020). The undocumented data may include the 'gut feelings' or the personal experiences with the trainees, conversations and interactions, personal anecdotes and concerns regarding the use of documented data (Tam et al 2020). On one hand, the members of the committee felt the need to put together the informal data with more formal documented evidence to create a holistic approach. But controversially, it also thought to interrupt the structured discussion by bringing in new issues. However, the new issues may help the key supervisor to perform an accuracy check as well before reaching the final decision of the committee. For instance, he/ she may inquire other members to give their opinions in agreement or disagreement regarding any aspect of the trainee performance. Hence, this discussion may prove to be fruitful in enhancing, correcting or countering any specific impression

about the trainee. May be the members can learn how to utilize this data in a more constructivist manner, building on what is already known to them, in a well regulated structured way, without compromising the confidentiality of the trainee, working through the whole process of analyzing it and deciding the trajectory. The undocumented data by the key supervisor can also be documented formally at the time of the observation along with the contextual factors that led to such an observation, so that it can be used later without creating any controversies for the committee members. The potential negative cognitive biases such as relying on incomplete or non-representative data (selection bias), easy access to the most recent data (availability bias) and over acceptance or relying on the group opinion (group-think bias) make create unnecessary noise in the reliability of the results (Tam et al 2020). In addition to this, too few faculty members, increased workload, lack of adequate trainee data collection and easy access, and lack of proper mandate, decision making process and structure of working of clinical competency committees may add to the list of problems. Starting with curriculum development, formative assessments and final entrustment decisions, it has to be a smooth journey avoiding frequent hiccups by devising a shared mental model and universal operating procedures for ad-hoc and summative decisions.

Paucity of evidence that relates trainee care quality to health care outcomes

The practical implications to implementation of CBME may include the measurement of clinical and educational outcomes. The short-term trainee outcomes which can help educators assess quality of CBME, which are easily measurable may include number of

EPAs covered during the early years of training, the quality of feedback provided to the trainees and the remediation required later on, to meet the deficiencies to complete the training. These can also prove to be of greater help in evaluation of the ongoing programme in terms of its improvement (Chan et al 2015). The trainee focused factors such as quality of care provided by physicians, the degree of task-based simulation used in assessment and training, the time that the faculty spend in direct observation of the performance of trainees in real clinical settings, and health and psychological wellbeing of trainees are short-term outcomes. On the other hand, ensuring accountability of our physicians by keeping the system of education transparent, patient safety, evidence based practice, efficiency of the health care system and impact on hospital flow of patients and maintaining the desired impact of the training are short as well as long-term goals of CBME. The outcomes of Graduate Medical Education are broad where health care system budget, resource management, post-residency trainee disposition and impact on patient safety and stake holders' satisfaction are key features. The educators may deem it important to compare the costs of time-variable training versus the time-fixed traditional system where funding is needed by national bodies. Multi-source data based on assessments in clinical settings and the individual educational EPA metrics may provide some link between the clinical and educational outcomes. Close collaboration between educators and quality assurance researchers is needed to link up the educational outcomes with other quality measures to determine the impact of CBME.

Moreover, CBME operates in an environment which comprises of multiple inter-professional health care providers. In such a scenario, the health care team is the frame of reference for high stakes examinations, which may involve assessments based on technology such as multi-source electronic data from workplace based assessments, simulation based problem solving skills, content- specific knowledge and attitudes of trainees, feedback from clinical advisors and multi-level cognition of clinical competency committees. The psychometric paradigm that has served the traditional system of assessment for so long may not be able to cater for uncertainty and complexity, which are the hallmarks of CBME (Eric et al 2017). CBME needs high stakes decisions as a part of system of assessments, which are statistically justified and take in to account complexity, to equally satisfy policy makers and health care providers.

Inadequate training of faculty in using prospective entrustment-supervision scales:

Entrustment decisions in EPA based framework rely on the use of supervision scales. In these scales, the raters estimate the readiness of a trainee for a specific EPA and the level of supervision needed. During the assessments, the expertise of the raters to handle the supervision scale varies (Postmes et al 2021). There are two types of entrustment supervision scales: retrospective (with a focus on the past) and prospective (with a focus on the future) (ten Cate et al 2020). The concept of prospective assessment is new and different than traditional retrospective assessment. The clinical teachers may have to draw additional conclusions and inferences not only on the quality of performance, but on trustworthiness, integrity and conscientiousness

of the trainee (Krupat et al 2017). It was observed that the raters had to rely on the descriptions of the scales during the assessment. They found it difficult to draw a line between the sub-scales of each level, for example, they could not differentiate a, b and c sub-scales of Indirect Supervision (Postmes et al 2021). Few of the raters found the scale as a continuum. Additional inference may put an extra burden on the raters in terms of cognitive load, hence it may affect the reliability of the assessment and consequently, validity of the scores. Moreover, other factors which may affect the clinician judgement include the students' self-assessment, how they put forward their clinical reasoning during the evaluation and the targeted level of supervision. The students' self-assessment and how they reason to justify how they look at themselves competency wise may influence the judgement of the rater. "I should declare them competent for 3a, also because the students referred to it often", is how one of the raters described it in an interview (Postmes et al 2021, p.5). In addition to giving weightage to the performance, the raters may get biased into thinking as to what level of supervision the student might need to complete the clerkship. Similarly, Young et al (2020) and Weijnen Meijer et al (2013) reported low reliability of prospective entrustment-scales when the raters were asked to judge the performance in addition to trainee features. On the other hand, if retrospective and prospective scales were used concurrently, the raters did not experience this problem (Cutrer et al 2020, Weijnen Meijer et al 2013). Hence, interpretation of supervision scales should be made as explicit as possible in high stakes EPA-based summative assessments.

Concept of Collective Competency and implications for its assessment:

It is a common phenomenon to find coupling of a trainee with a supervisor or clinical advisor during training periods (Sebok-Syer et al 2018). It can pose certain challenges to the basic principles of CBME, where the idea is to prepare and assess the trainee for independent clinical practice without any supervision (Gingerich 2015). Coupling in a clinical scenario may be described as structure of care delivery where a supervisor may work, guide or collaborate closely with a trainee (Sebok-Syer et al 2018). The nature and the extent to which the trainee is dependent on his supervisor may vary according to the level of supervision needed, clinical context and difficulty of the case under consideration. For example, in operating rooms, this coupling can be so strong that the residents may question their independent existence in terms of their clinical performance prior to graduation (Sebok-Syer et al 2018). This may give an insight into their 'interdependence'. The trainees' interdependence may positively or negatively, enhance or inhibit their independence (Goldszmidt et al 2015). This complex relationship may secondarily depend on a number of institutional factors, such as admission or discharge pressures and patient census (Goldszmidt et al 2015). Such interdependence may have practical implications for workplace- based assessment of trainee performance. In order to assess a trainee whose training was overshadowed by coupling with another senior trainee or a supervisor, needs a system that can evaluate multiple individuals, can cater for different task dimensions and the degree of interdependence or independence. Andrews et al (2017) proposed a multivariate

model based on Item Response Theory (IRT) to assess impact of interaction between two members of a couple on performance outcomes. IRT model addresses tendencies to behave in a close collaboration, in a team and translate it into nominal categories. In the study, a Multivariate Andersen/Rasch IRT Model was used to observe the tendency to interact and exhibit any specific behaviour pattern as one member of the team interacts with the other. These behaviour tendencies are likely to be pervasive, stable and recur in similar situations (Endler and Parker 1992). In this study, patterns of behaviour were evoked by simulation-based problem solving task and studied in dyads (combination of two vectors). Four interaction/ behaviour patterns were identified: collaborative, fake collaboration, dominant/ dominant and cooperative. Out of all the four interactions, cooperative pattern was most common to be witnessed, whereas, collaborative and cooperative behaviors yielded highest scores in terms of performance among the dyads. Dominant/dominant pattern was associated with lower scores. In another similar study, Storch (2002) identified four patterns of behavior amongst members in a group: expert/novice, collaborative, dominant/dominant and dominant/ passive. These coupling configurations suggest patterns of bahavior in a group which determine the mutual learning and can affect performance. All these approaches guide the raters towards assessment of collective competency where they can assess the independent as well as the interdependent components of performance. If behavior is an integral component of sociocultural and environmental factors, then assessment has

to be built around intersecting behaviors of trainee and supervisor (Sebok-Syer et al 2018).